Open Research Data and Materials

University POLITEHNICA of Bucharest

Based on: https://open-science-training-handbook.github.io/

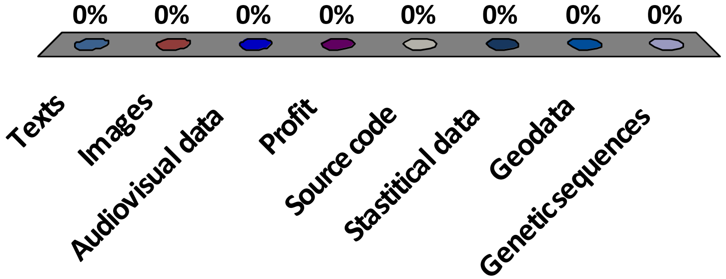
Contents

- Research Data
    - Types of research data
    - Definition
    - Research data types
    - Research data COVID-19 challenge
- Open Research Data
    - What is it? - Understanding Open Data
    - Rationale
    - Learning objectives
    - Key components
    - Questions, obstacles, and common misconceptions
    - Further reading

Research data

What would you associate research data with specifically?

- Texts
- Images
- Audiovisual data
- Profit
- Source code
- Stastitical data
- Geodata
- Genetic sequences



Definition

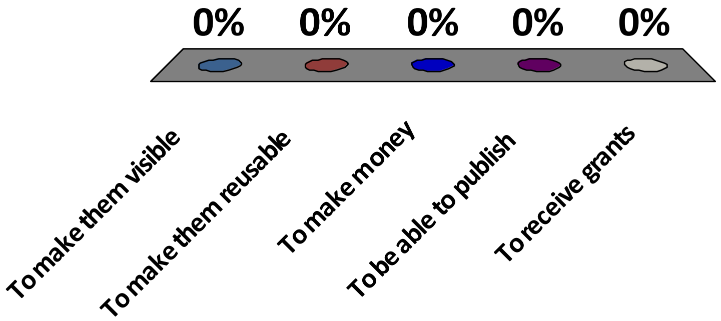
- Several definitions are possible based on specific fields, institutions and organizations.

- Research data are defined as **factual records** (numbers, texts, images and sounds), which are used as **principal sources for scientific research** and which are often recognized by the scientific community as being **necessary to validate research results**.

Organization for Economic Cooperation and Development (OECD)

Why would you make your research data available in a repository?

A. To make them visible
B. To make them reusable
C. To make money
D. To be able to publish
E. To receive grants



Limited incentives to give evidence against yourself

- We know that no one wants to incriminate themselves, and also that no one is infallible.
    - The Fifth Amendment to the United States Constitution includes a clause that no one “shall be compelled in any criminal case to be a witness against [them]sel[ves]”. (Edited to gender-neutral language.)
    - To “plead the fifth” means that someone chooses not to give evidence that there might have been something wrong in their past behaviour. They have the right to remain silent.
- Putting your code and data online can be very revealing and intimidating, and it is part of the human condition to be nervous of being judged by others.
- Although there is no law governing the communication of reproducible research - unless you commit explicit fraud in your work - sharing errors that you find in your work is heavily disincentivised.

IF I SHARE MY DATA.... PEOPLE MIGHT FIND MISTAKES
NEED TO MAKE IT OK TO BE HUMAN
SPRINGER NATURE

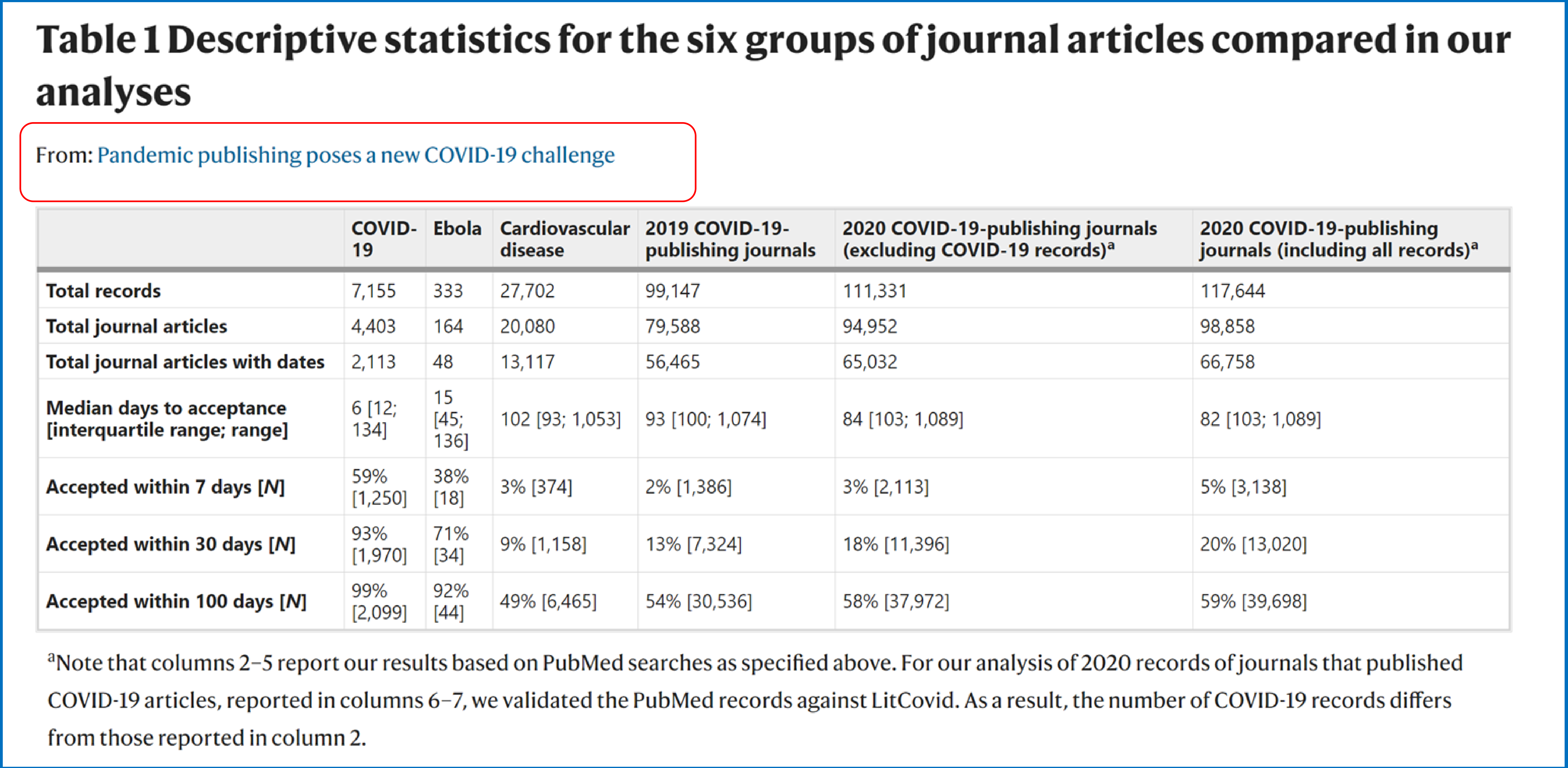


Table 1 Descriptive statistics for the six groups of journal articles compared in our analyses

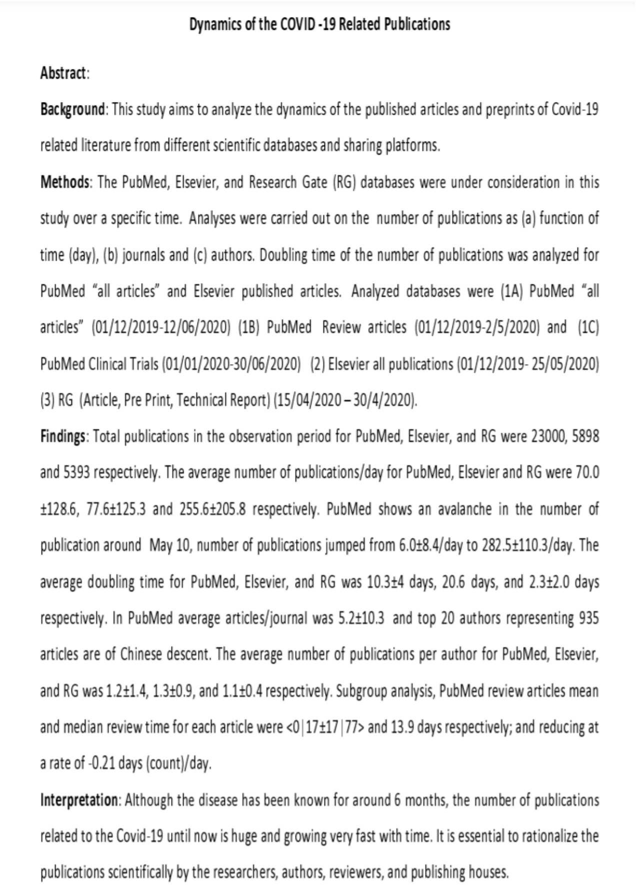
From: Pandemic publishing poses a new COVID-19 challenge

	COVID-19	Ebola	Cardiovascular disease	2019 COVID-19-publishing journals	2020 COVID-19-publishing journals (excluding COVID-19 records)[a]	2020 COVID-19-publishing journals (including all records)[a]
Total records	7,155	333	27,702	99,147	111,331	117,644
Total journal articles	4,403	164	20,080	79,588	94,952	98,858
Total journal articles with dates	2,113	48	13,117	56,465	65,032	66,758
Median days to acceptance [interquartile range; range]	6 [12; 134]	15 [45; 136]	102 [93; 1,053]	93 [100; 1,074]	84 [103; 1,089]	82 [103; 1,089]
Accepted within 7 days [*N*]	59% [1,250]	38% [18]	3% [374]	2% [1,386]	3% [2,113]	5% [3,138]
Accepted within 30 days [*N*]	93% [1,970]	71% [34]	9% [1,158]	13% [7,324]	18% [11,396]	20% [13,020]
Accepted within 100 days [*N*]	99% [2,099]	92% [44]	49% [6,465]	54% [30,536]	58% [37,972]	59% [39,698]

[a] Note that columns 2–5 report our results based on PubMed searches as specified above. For our analysis of 2020 records of journals that published COVID-19 articles, reported in columns 6–7, we validated the PubMed records against LitCovid. As a result, the number of COVID-19 records differs from those reported in column 2.

Nature Human Behaviour, vol. 4, pp. 666–669, (2020)

Dynamics of the COVID -19 Related Publications



Dynamics of the COVID -19 Related Publications

Abstract:

Background: This study aims to analyze the dynamics of the published articles and preprints of Covid-19 related literature from different scientific databases and sharing platforms.

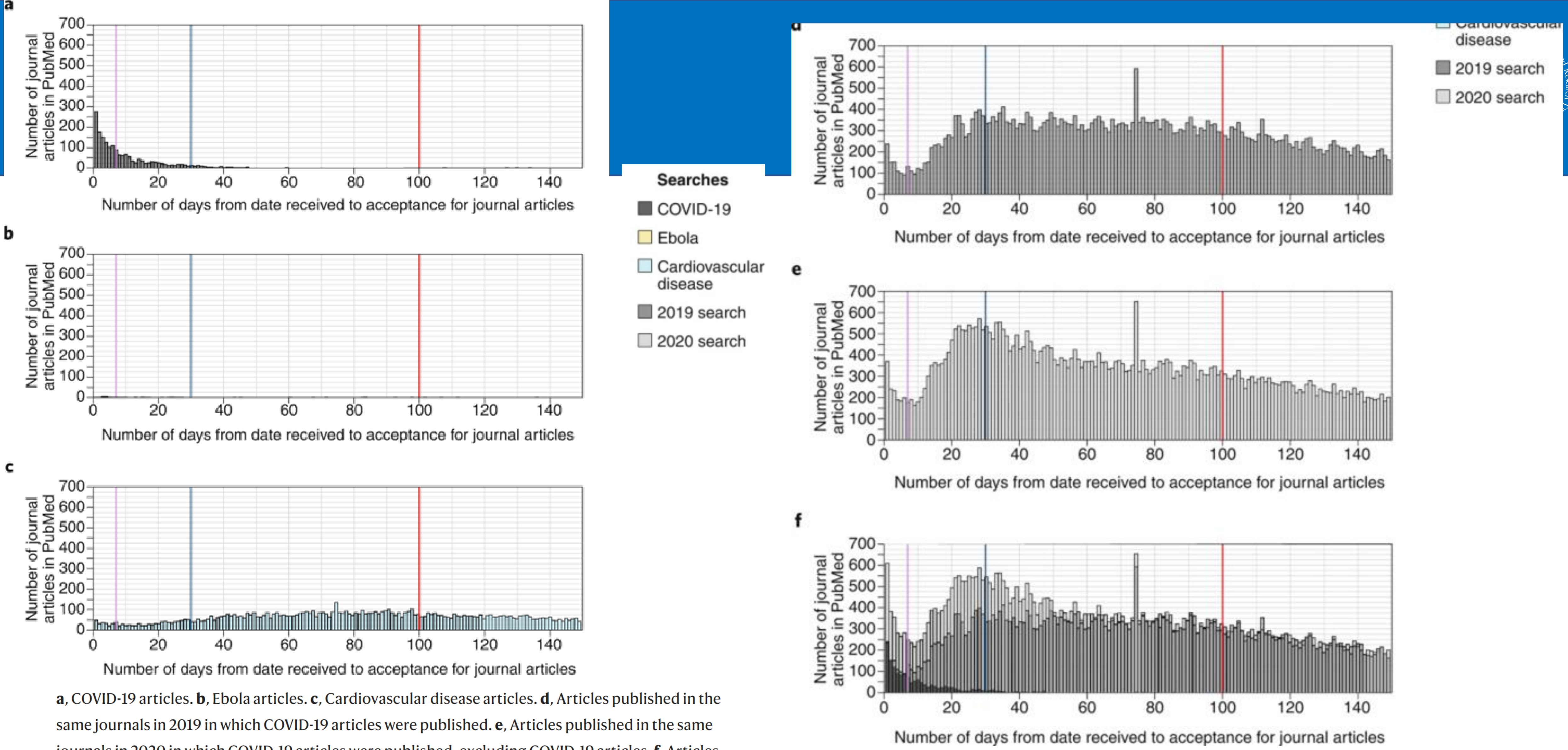
Methods: The PubMed, Elsevier, and Research Gate (RG) databases were under consideration in this study over a specific time. Analyses were carried out on the number of publications as (a) function of time (day), (b) journals and (c) authors. Doubling time of the number of publications was analyzed for PubMed “all articles” and Elsevier published articles. Analyzed databases were (1A) PubMed “all articles” (01/12/2019-12/06/2020) (1B) PubMed Review articles (01/12/2019-2/5/2020) and (1C) PubMed Clinical Trials (01/01/2020-30/06/2020) (2) Elsevier all publications (01/12/2019- 25/05/2020) (3) RG (Article, Pre Print, Technical Report) (15/04/2020 – 30/4/2020).

Findings: Total publications in the observation period for PubMed, Elsevier, and RG were 23000, 5898 and 5393 respectively. The average number of publications/day for PubMed, Elsevier and RG were 70.0 ±128.6, 77.6±125.3 and 255.6±205.8 respectively. PubMed shows an avalanche in the number of publication around May 10, number of publications jumped from 6.0±8.4/day to 282.5±110.3/day. The average doubling time for PubMed, Elsevier, and RG was 10.3±4 days, 20.6 days, and 2.3±2.0 days respectively. In PubMed average articles/journal was 5.2±10.3 and top 20 authors representing 935 articles are of Chinese descent. The average number of publications per author for PubMed, Elsevier, and RG was 1.2±1.4, 1.3±0.9, and 1.1±0.4 respectively. Subgroup analysis, PubMed review articles mean and median review time for each article were <0|17±17|77> and 13.9 days respectively; and reducing at a rate of -0.21 days (count)/day.

Interpretation: Although the disease has been known for around 6 months, the number of publications related to the Covid-19 until now is huge and growing very fast with time. It is essential to rationalize the publications scientifically by the researchers, authors, reviewers, and publishing houses.

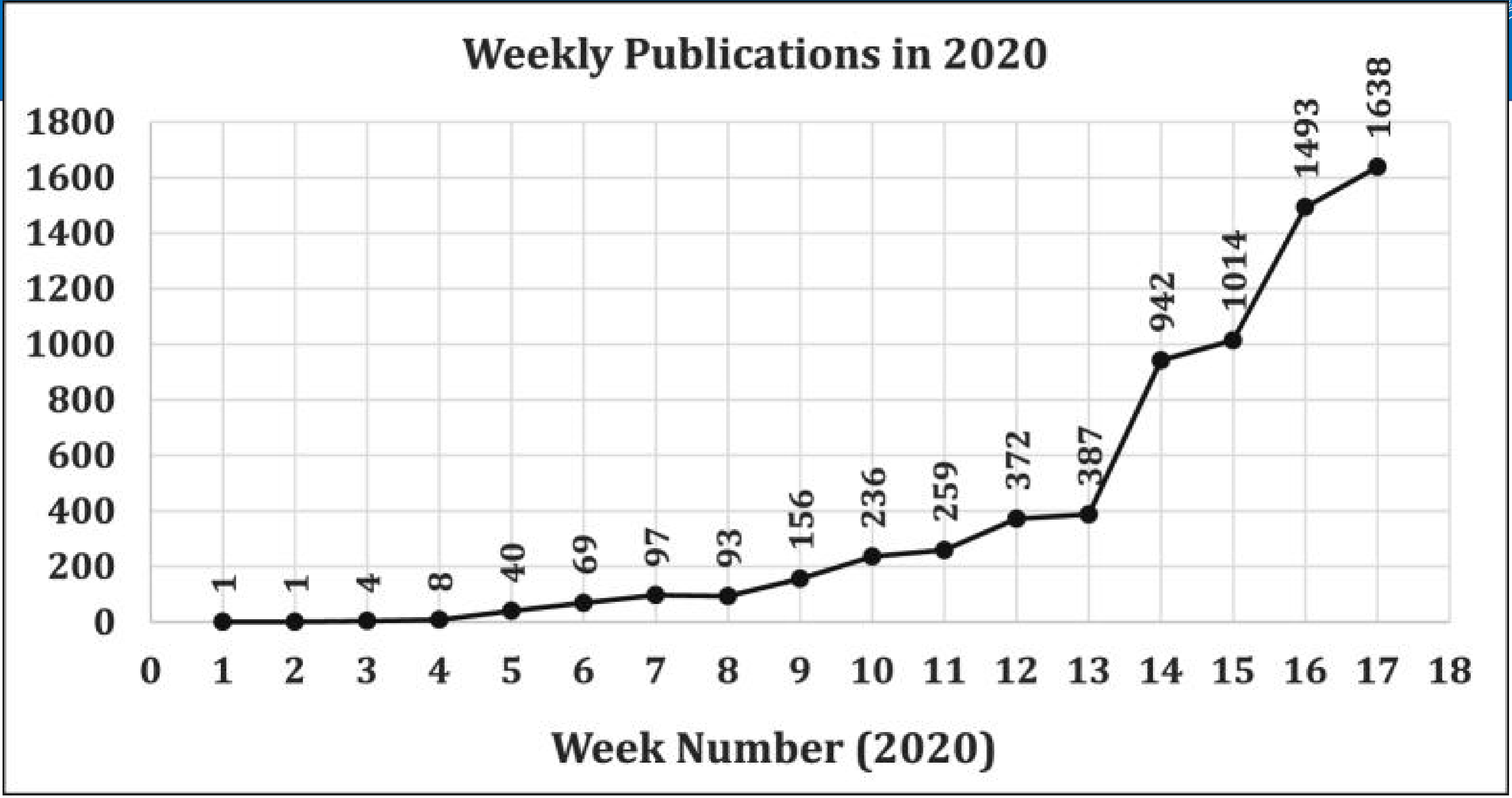
- **Findings**
  - Total publications in the observation period for PubMed, Elsevier, and RG were 23000, 5898 and 5393 respectively.
  - The average number of publications/day for PubMed, Elsevier and RG were 70.0 ±128.6, 77.6±125.3 and 255.6±205.8 respectively.
  - PubMed shows an avalanche in the number of publication around May 10, number of publications jumped from 6.0±8.4/day to 282.5±110.3/day.
  - The average doubling time for PubMed, Elsevier, and RG was 10.3±4 days, 20.6 days, and 2.3±2.0 days respectively.
  - **Publication pipeline is at extraordinary speed – 6 days for COVID-19 article**
  - **Measures are required to safeguard the integrity of scientific evidence**

doi: https://doi.org/10.1101/2020.08.05.237313



a, COVID-19 articles. **b**, Ebola articles. **c**, Cardiovascular disease articles. **d**, Articles published in the same journals in 2019 in which COVID-19 articles were published. **e**, Articles published in the same journals in 2020 in which COVID-19 articles were published, excluding COVID-19 articles. **f**, Articles published in the same journals in 2020 in which COVID-19 articles were published, including COVID-19 articles. In all panels, the purple line represents 7 days, the blue line 30 days and the red line 100 days from time of article receipt.

Nature Human Behaviour, **volume 4**, pp. 666–669(2020)



Publications related to Covid-19 in 2020 until the 17th week

J Clin Orthop Trauma. 2020 May; 11(Suppl 3): S304–S306. doi: 10.1016/j.jcot.2020.(

Open Research Data

Understanding Open Research Data

What is open data

- Definition by Open Definition : “***Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike***”.
- The full definition here
- Most important characteristics:
  - **Availability and Access**: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
  - **Re-use and Redistribution**: the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
  - **Universal Participation**: everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

What is it? - Understanding Open Data

- Open data, especially open government data, is a tremendous resource that is as yet largely untapped.
- Many individuals and organisations collect a broad range of different types of data in order to perform their tasks.
- Government is particularly significant in this respect, both because of the quantity and centrality of the data it collects, but also because most of that government data is public data by law, and therefore could be made open and made available for others to use.
- Why is that of interest?
  - **The key point is that when opening up data, the focus is on non-personal data, that is, data which does not contain information about specific individuals.**
  - Similarly, for some kinds of government data, national security restrictions may apply.

Rationale

- Research data are:
  - often the most valuable output of many research projects,
  - used as primary sources that underpin scientific research and enable derivation of theoretical or applied findings.
  - used to make findings/studies replicable, or at least reproducible or reusable (reference to Reproducible Research and Data Analysis) in any other way,

The best practice recommendation for research data is to be as open and FAIR as possible, while accounting for ethical, commercial and privacy constraints with sensitive data or proprietary data.

Learning objectives

- **Gain an understanding** of the basic characteristics and principles of open and FAIR research data, including appropriate packaging and documentation, to enable others to understand, reproduce, and re-use in alternative ways.
- **Familiarity** with the sorts of data that might be considered sensitive, and the restrictions or constraints on openly sharing them.
- **Be able to:**
  - **convert** a ‘closed’ dataset into one which is ‘open’ by implementing the necessary measures in a data management plan, with appropriate data stewardship and metadata.
  - **to use research data management plan** and to make your research results findable and accessible, even if it contains sensitive data.
- **Understand**:
  - the pros and cons of openly sharing different types of data (e.g., privacy, sensitivity, de-identification, mediated access).
  - the importance of appropriate metadata for sustainable archiving of research data.
  - the basic workflows and tools for sharing research data.

Key components

- Knowledge & Skills
- Questions, obstacles, and common misconceptions
- Learning outcomes
- Further reading

Knowledge & Skills

- FAIR principles
- Data publishing
- Data citation
- Data packaging
- Sharing sensitive and proprietary data
- Data brokers
- Analysis portals
- De-identified and synthetic data
- DataTags
- Open Materials
- Reagents
- Protocols
- Notebooks, containers, software, and hardware

FAIR principles

- A core set of principles to optimize the reusability of research data or any digital object:
  - **F**indable - It should be easy to find the data and the metadata for both humans and computers. Depends on machine-readable persistent identifiers (PIDs) and metadata.
  - **A**ccessible - The (meta)data should be retrievable by their identifier using a standardized and open communications protocol, possibly including authentication and authorisation. Also, metadata should be available even when the data are no longer available.
  - **I**nteroperable - The data should be able to be combined with and used with other data or tools. The format of the data should therefore be open and interpretable for various tools, including other data records.
  - **R**e-usable - To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. Also, the reuse of the (meta)data should be stated with (a) clear and accessible license(s).

Data publishing

- There are several distinct ways to make research data accessible:
  - **Publishing data as supplemental material** associated with a research article, typically with the data files hosted by the publisher of the article.
  - **Hosting data on a publicly-available website**, with files available for download.
  - **Depositing data in a repository** that has been developed to support data publication, e.g., Dataverse, [Dryad] (https://en.wikipedia.org/wiki/Dryad_(repository)), figshare, Zenodo.
    - A large number of general and domain or subject specific data repositories exist which can provide additional support to researchers when depositing their data.
  - **Publishing a data paper about the dataset**, which may be published as a preprint, in a journal, or in a data journal that is dedicated to supporting data papers. The data may be hosted by the journal or hosted separately in a data repository.

Data citation

- Data citation services help research communities discover, identify, and cite research data (and often other research objects) with confidence.
- This typically involves the creation and allocation of **Digital Object Identifiers (DOIs)** and accompanying metadata through services such as
  - DataCite (https://www.datacite.org), and can be integrated with research workflows and standards.
- This is an emerging field, and involves aspects such as conveying to journal publishers the importance of appropriate data citation in articles, as well as enabling research articles themselves to be linked to any underlying data.
- Through this, citable data become legitimate contributions to the process of scholarly communication, and can help pave the way for new metrics and publication models that recognize and reward data sharing.

Data packaging

- Data packages are **containers** for describing and sharing accompanying data files, and typically comprise a metadata file describing the features and context of a dataset.
- Include aspects such as: creation information, provenance, size, format type, field definitions, as well as any relevant contextual files, such as data creation scripts or textual documentation.
  - **Data are forever**: Datasets outlive their original purpose. Limitations of data may be obvious within their original context, such as a library catalog, but may not be evident once data is divorced from the application it was created for.
  - **Data cannot stand alone**: Information about the context and provenance of the data--how and why it was created, what real-world objects and concepts it represents, the constraints on values--is necessary to helping consumers interpret it responsibly.
  - Structuring metadata about datasets in a standard, machine-readable way encourages the promotion, shareability, and reuse of data.

Sharing sensitive and proprietary data

- With appropriate data management planning much sensitive and proprietary data can be shared, reused, and FAIR.
- The metadata can almost always be shared.
- Guidance and best practices for sharing sensitive data are necessarily region-specific because of differing regulations

Data brokers

- Data brokers are knowledgeable, independent parties who act as data stewards for sensitive data.
- Researchers can transfer their sensitive data and jurisdiction over access to that data to the broker.
- This is especially common with patient-level data from clinical studies.
- Brokers provide a level of independence in the evaluation of whose data requests are scientifically valid and will not violate the privacy of research participants.

Analysis portals

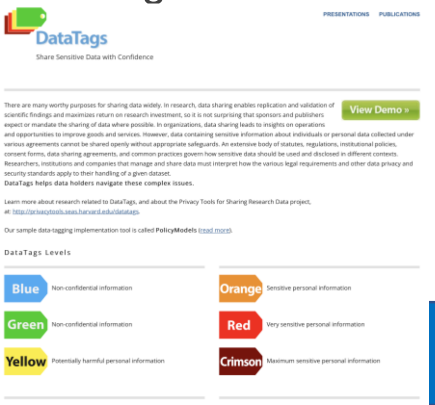
- Analysis portals are platforms that allow approved analysis of data without allowing full access (viewing or downloading) or controlling where and who gets access.
- Some data brokers also use analysis portals. Analysis portals control what additional datasets can be pooled with the sensitive data as well as what analyses can be run to ensure that personal information is not revealed during reanalysis.
- Examples of virtual analysis portals include Project Data Sphere, Vivli, RAIRD, Corpuscle, and INESS.
- Social science and other researchers with sensitive data use a single-site analysis portal that can be accessed only under controlled regime.
- Approved researchers can access the data on-site, in a safe room, for scientific purposes. However, the metadata describing the data should be openly available and adhering to the FAIR principles.

De-identified and synthetic data

- Many datasets containing participant-level private information can be shared once the dataset has been de-identified (**Safe Harbor method**) or a expert has determined that the dataset is not individually identifiable (**Expert Determination method**).
- Consult with your Research Ethics Board / Institutional Review Board to learn how to do this with your data. We also recommend the CESSDA Expert Tour Guide on Data Management, which provides information and practical examples on how to share personal data. However, **some datasets cannot be safely de-identified and shared**.
- Researchers can still improve the **openness** of research on such data by **creating and sharing synthetic data**.
  - Synthetic data is similar in structure, content, and distribution to the real data and aims to attain "analytic validity": statistical analysis will return the same results for the synthetic data as the real data.
- The United States Census Bureau, for example, uses synthetic data and analysis portals in combination to allow reuse of highly sensitive data.

DataTags

- DataTags is a framework designed to enable computer-assisted assessments of the legal, contractual, and policy restrictions that govern data sharing decisions.
    - The system asks a user a series of questions to elicit the key properties of a given dataset and applies inference rules to determine which laws, contracts, and best practices are applicable.
    - The output is a set of recommended DataTags, or simple, iconic labels that represent a human-readable and machine-actionable data policy, and a license agreement that is tailored to the individual dataset.
    - It is being designed to integrate with data repository software, and it will also operate as a standalone tool.
    - Developed at Harvard University. In Europe, DANS is working on adjusting DataTags to European legislation / General Data Protection Regulation (GDPR) (cf. DANS GDPR DataTags).



PRESENTATIONS PUBLICATIONS

DataTags

Share Sensitive Data with Confidence

There are many worthy purposes for sharing data widely. In research, data sharing enables replication and validation of scientific findings and maximizes return on research investment, so it is not surprising that sponsors and publishers expect or mandate the sharing of data where possible. In organizations, data sharing leads to insights on operations and opportunities to improve goods and services. However, data containing sensitive information about individuals or personal data collected under various agreements cannot be shared openly without appropriate safeguards. An extensive body of statutes, regulations, institutional policies, consent forms, data sharing agreements, and common practices govern how sensitive data should be used and disclosed in different contexts. Researchers, institutions and companies that manage and share data must interpret how the various legal requirements and other data privacy and security standards apply to their handling of a given dataset.
DataTags helps data holders navigate these complex issues.

Learn more about research related to DataTags, and about the Privacy Tools for Sharing Research Data project, at: http://privacytools.seas.harvard.edu/datatags.

Our sample data-tagging implementation tool is called **PolicyModels** (read more).

DataTags Levels

Blue	Non-confidential information	Orange	Sensitive personal information
Green	Non-confidential information	Red	Very sensitive personal information
Yellow	Potentially harmful personal information	Crimson	Maximum sensitive personal information

Open Materials

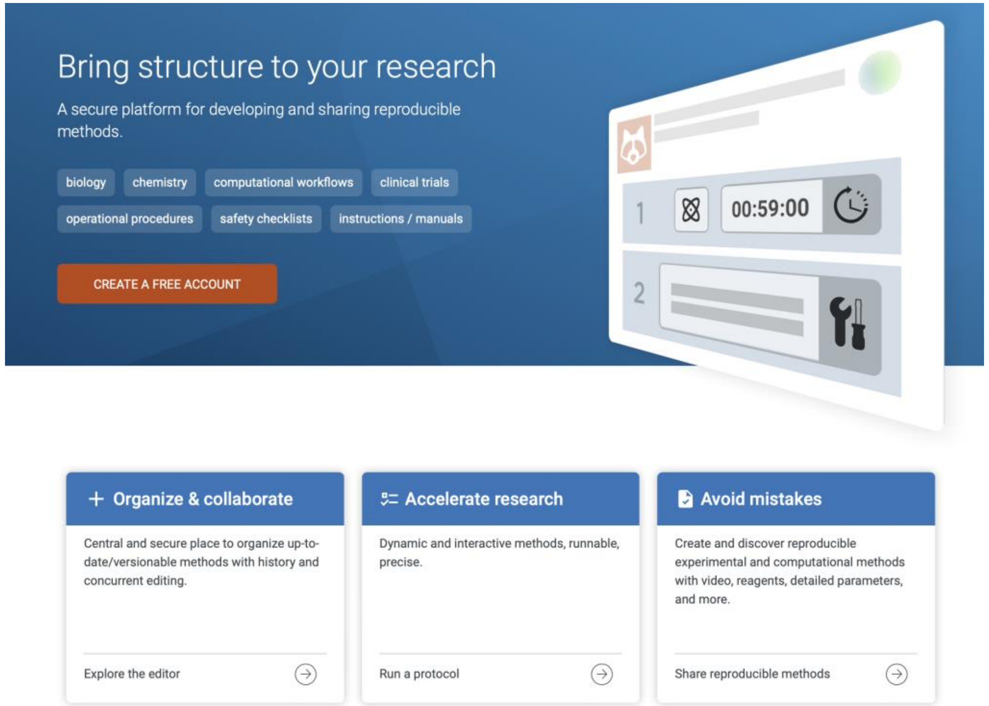
- In addition to data sharing, the openness of research relies on sharing of materials. What materials researchers use is discipline-specific and sometimes unique to a lab.
- Below are examples of materials you can share, although always confer with peers in your discipline to identify which repositories are used:
  - Reagents
  - Protocols
  - Notebooks, containers, software, and hardware
- When you have materials, data, and publications from the same research project shared in different repositories, cross-reference them with a link and a unique identifier so they can be easily located.

Reagents

- A reagents is a substance, compound or mixture that can be added to a system in order to create a chemical or other reaction.
- Reagents can be deposited with repositories like Addgene, The Bloomington Drosophila Stock Center, and ATCC to make them easily accessible to other researchers.
- License your materials so they can be reused by other researchers.

Protocols

- A protocol describes a formal or official record of scientific experimental observations in a structured format. Deposit virtual protocols for citation, adaptation, and reuse using Protocols.io.
- Read-only protocols should be deposited in your disciplines registry such as ClinicalTrials.gov and SocialScienceRegistry or a general registry like Open Science Framework.
- Many journals, such as Trials, JMIR Research Protocols, or Bio-Protocol, will publish your protocol.


Bring structure to your research

A secure platform for developing and sharing reproducible methods.

+ Organize & collaborate

Central and secure place to organize up-to-date/versionable methods with history and concurrent editing.

Explore the editor

Accelerate research

Dynamic and interactive methods, runnable, precise.

Run a protocol

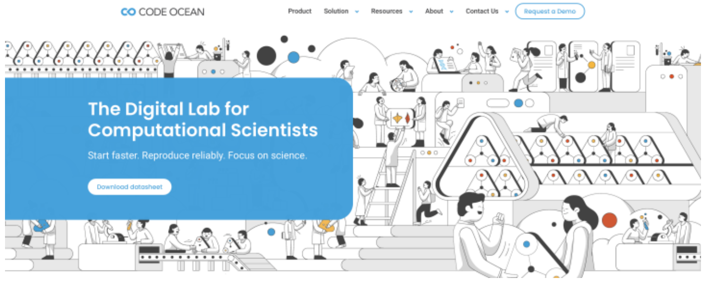
Avoid mistakes

Create and discover reproducible experimental and computational methods with video, reagents, detailed parameters, and more.

Share reproducible methods

Notebooks, containers, software, and hardware

- Reproducible analysis is aided by the use of literate programming, container technology, and virtualization.
- Besides code and data, you can also share
  - Jupyter notebooks, Docker images, or other analysis materials or software dependencies.
- Share notebooks with Open services such as mybinder that allow for public viewing and execution of the entire notebook on shared resources.
- Containers and notebooks can be shared with Rocker or Code Ocean.
  - Software and hardware used in your research should be shared following best practices for documentation as outlined in Section 3.



The Digital Lab for Computational Scientists

Start faster. Reproduce reliably. Focus on science.

Questions, obstacles, and common misconceptions

- Is it sufficient to make my data openly available?
- What do the FAIR principles mean/imply for different stakeholders / audiences?
- Is making my data FAIR a lot of extra work?
- I want to share my data. How should I license them?
- I cannot make my data directly available—they are too large to share conveniently / have restrictions related to privacy issues. What should I do?

Question

- Would you consider publishing your research data?
  - after completion of the project for which the data were collected
  - After „traditional“ publication of the results
  - at a later date
  - never

Learning outcomes

- Understand the characteristics of open data, and in particular the FAIR principles.
- Be familiar with some of the arguments for and against open data.
- Be able to differentiate and address sensitive data and open FAIR data; these two categories are not necessarily incompatible.
- Be able to transform a dataset into one that is sufficient for open sharing (non-proprietary format), meets the standards of the FAIR principles, and is designed for maximized accessibility, transparency and re-use by providing sufficient metadata.
- Know the difference between raw and processed (or cleaned) data, and the importance of version labels.
- Know commonly used file formats and community standards for maximum re-usability.
- Be able to write a data management plan.

Further reading

- Averkamp et al. (2018). Data packaging guide. github.com/saverkamp/beyond-open-data/blob/master/DataGuide.md.
- Barend et al. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. doi.org/10.3233/ISU-170824
- Brase et al. (2009). Approach for a joint global registration agency for research data. doi.org/10.3233/ISU-2009-0595
- Candela et al. (2015). Data journals: A survey. doi.org/10.1002/asi.23358
- CESSDA Training Working Group (2017-2018a). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. cessda.eu/DMGuide
- CESSDA Training Working Group (2017-2018b). CESSDA Data Management Expert Guide: Citing your data. Bergen, Norway: CESSDA ERIC.cessda.eu/DMGuide/citingdata
- FAIRsharing.org (2016). FAIR. The FAIR Principles. doi.org/10.25504/FAIRsharing.WWI10U
- Force 11 (n.y.). Guiding principles for Findable, Accessible, Interoperable, and Re-usable data publishing Version B1.0. force11.org/fairprinciples
- Gorgolewski et al. (2013). Making data sharing count: a publication-based solution. doi.org/10.3389/fnins.2013.00009
- Kratz and Strasser (2015). Making Data Count. doi.org/10.1038/sdata.2015.39
- Piwowar and Vision (2013). Data reuse and the open data citation advantage. doi.org/10.7717/peerj.175
- Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi.org/10.1038/sdata.2016.18
- Wilkinson et al. (2918). A design framework and exemplar metrics for FAIRness. doi.org/10.1038/sdata.2018.118

Open Science and Research Data Management Innovative and Distributed Training Programme

THANK YOU!

Follow us