Faculdade de Medicina

Preparing data for Open Access and Sharing

Mariana M S Santos, PhD
Investigator Assistant
ERA Chair iSTARS

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

The Open Science Revolution

Why Share? Beyond Publication

- **Transparency:** Allows others to verify your results (Scientific Integrity).
- **Efficiency:** Prevents duplication of effort; other researchers can build upon your work.
- **Funding Requirements:** Most grants (Horizon Europe, FCT) now **mandate** Open Access.
- **Impact:** Articles with shared data get more citations and visibility.

The FAIR Principles

Making Data Useful, not just Available

For data to be truly "Open," it must be FAIR:

Findable: It must have a unique persistent identifier (DOI) and rich metadata.

Accessible: It must be stored in a trusted repository with clear, well-defined access terms.

Interoperable: It must use open formats (e.g., CSV instead of proprietary Excel) and standard vocabularies (e.g., SNOMED, LOINC).

Reusable: It must include a clear data license (e.g., Creative Commons) and sufficient documentation.

The "Open Data" vs. "Privacy" Dilemma

As Open as Possible, as Closed as Necessary

The Conflict: Health data is sensitive (GDPR). How to share?

- **Tiered Access:** You don't have to put everything on the public internet.
  - *Public:* Metadata and Code.
  - *Controlled:* Sensitive raw data (requires a request and DTA).

- **Anonymization is mandatory:** Never share pseudonymized data in Open Access.

Technical Preparation (The "To-Do" List)

Prepping your Dataset for the World

1. Remove "Dead Weight": Delete intermediate files, temporary variables, and unused columns.

2. File Formats: Save data as .csv or .json. Save code as .py or .R. Avoid .xlsx or .sas7bdat.

3. The README file: The most important document. It must explain:

- What the dataset is about.
- What each variable means.
- Which version of Python/R was used.

4. Code Cleaning: Remove hardcoded paths (e.g., C:/Users/Joao/Desktop/data) and API keys/passwords

Where to Publish? (Repositories)

Reliable Homes for your Data

General Purpose:

- **Zenodo:** (CERN) Excellent, gives a DOI, very common in the EU.
- **Figshare / Dryad:** Widely used in academic publishing.

•**Institutional:** Hospital / University Repository

Specific for Bio/Medical:

- **PhysioNet:** (Complex clinical data).
- **dbGaP / EGA:** (Genomic data - usually requires controlled access).

Licenses – Telling people what they can do

Don't forget the License!

Without a license, your data is in a "legal limbo," and people will be afraid to use it.

- **CC0 (Public Domain):** "Do whatever you want.”
- **CC-BY (Attribution):** "You can use it, but you must cite me." (The most common in science).
- **CC-BY-NC (Non-Commercial):** "You can use it, but not for commercial purposes."

Closing the Circle – The Ethical Data Journey

From Patient Consent to Global Knowledge

1. The Foundation (Ethics & Law)

- It starts with the patient. We respect their rights (GDPR/Ethics Committees) because without their trust, there is no data.

2. The Architecture (Collection & Curation)

- We build secure "homes" for data (REDCap/Limesurvey). We curate and document so that our data has meaning and integrity.

Faculdade de Medicina

Closing the Circle – The Ethical Data Journey

From Patient Consent to Global Knowledge

3. The Insight (Statistics & Analysis)

- We apply rigorous methods to separate signal from noise. We look for clinical relevance, not just p-values, and we audit our models for bias.

4. The Legacy (Open Science & Sharing)

- We don't keep the knowledge to ourselves. We share our code and data (FAIR principles) so that science can move faster and save more lives.

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Faculdade de Medicina

LISBOA
UNIVERSIDADE DE LISBOA

"Data Science in Healthcare is not just about algorithms; it's about the responsible transformation of human experience into medical progress."

Feitos de outra matéria

Faculdade de Medicina

Comments & Questions

Mariana M S Santos
marianamsantos@medicina.ulisboa.pt

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria