Faculdade de Medicina

Data Collection & Curation

The Foundation of High-Quality Data Science

Mariana M S Santos, PhD
Investigator Assistant
ERA Chair iSTARS

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Faculdade de Medicina

Data Collection & Curation

The Foundation of High-Quality Data Science

The Golden Rule ➡ Garbage In → Garbage Out

Vocabulary

Collection: The process of gathering raw information.

Curation: The "art" of organizing, cleaning, and enriching data so it remains useful over time.

Moving from chaotic Excel sheets to structured, auditable databases.

LISBOA
UNIVERSIDADE DE LISBOA

Feitos de outra matéria

Why move away from Excel?

- **Security:** Audit trails (who changed what and when?).
- **Validation:** Constraints at the moment of entry (e.g., age cannot be 200).
- **Centralization:** Multiple researchers entering data simultaneously.
- **Compliance:** Role-based access control (RBAC) aligned with GDPR.

REDCap (Research Electronic Data Capture)

A usefull tool for academic clinical research

What it is: A secure web application for building and managing online databases and surveys.

Key Features for Data Science:

- **Data Dictionary:** Exportable metadata (standardized variable names).
- **Branching Logic:** Dynamic forms that adapt to previous answers.
- **Longitudinal Tracking:** Perfect for follow-ups and repeat events.
- **API Integration:** Direct connection to R/Python for automated data pulling.

Best for: Clinical trials, munticentered studies, longitudinal studies, and complex medical registries.

LimeSurvey

A Powerful Alternative for Complex Survey Logic

What it is: An open-source tool focused on advanced survey structures.

Key Features:

- **Advanced Logic:** Complex "if-then" conditions for questionnaire flow.
- **Anonymous Mode:** Stronger "out-of-the-box" settings for privacy-first surveys.
- **Multilingual:** Excellent support for studies across different countries.

Best for: Epidemiological surveys, Patient-Reported Outcome Measures (PROMs), and community-based research.

Hidden Cost of Data Wrangling

Why Curate? 80% of Science happens before the Analysis

The Reality: Data Scientists spend most of their time cleaning "messy" data rather than building models.

The Goal of Curation: Shift the effort from "fixing errors" to "extracting insights."

Quality at the Source: Using REDCap/LimeSurvey to prevent errors *during* entry (Range checks, required fields).

Result: A dataset that is "Analysis-Ready" from day one.

Metadata Dictionary

Data without Description is just Noise

What is Metadata? It’s the "Manual" for your dataset (Variable names, labels, and types).

Why it Matters:

Context: Does "99" mean 99 years old or "Missing Data"?

Units: Are we measuring in mg or µg? In Celsius or Fahrenheit?

Scales: Is a score of 1 "Best" or "Worst"?

The Golden Rule: If you can’t share your data dictionary, you can’t share your science.

Semantic Interoperability

Speaking the Same Language (Standards)

The Problem: Every hospital has its own way of naming a "Blood Pressure" variable.

The Solution: Using International Standards within your collection tools.

- ✓ **ICD-10/SNOMED:** For Diagnoses.
- ✓ **LOINC:** For Laboratory tests.
- ✓ **CDISC/OMOP:** For standardizing the entire database structure.

Benefit: Allows you to merge your data with international registries seamlessly.

Data Provenance & Audit Trails

The "Biography" of your Data

Reproducibility: Can we trace a value back to its origin?

Audit Trail: Who entered the data? When? Was it edited later? Why? *Essential for Clinical Trials and High-Impact Research.*

Source Verification: Knowing if a data point came from a patient's wearable, a manual entry, or an automated EHR export.

Integrity: Protection against accidental (or intentional) data manipulation.

Faculdade de Medicina

Data Curation for Data Scientists

Thinking Ahead for Analysis

- **Standardized Vocabularies:** Using ICD-10, SNOMED, or LOINC within your collection tools.
- **Missing Data Management:** Distinguishing between "Not Applicable," "Unknown," and "Refused.”
- **Audit Trails:** Ensuring the provenance of every data point (Crucial for Reproducibility).
- **Data Export:** How to choose between CSV, JSON, or direct Database connection (API).

LISBOA
UNIVERSIDADE DE LISBOA

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