Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Data Science Seminars

Real World Data – A Gamechanger for Life Sciences. Challenges and insights in building a Global Trial Optimisation Network

Authors:

José Antonio Ruiz Heredia

Teacher:

Mr. Brecht Claerhout

Date:
April 7, 2025

1 Assessing Completeness in *Real-World Data (RWD)* Datasets

Completeness refers to whether a dataset contains all the necessary information for its intended purpose. In the context of *Real-World Data (RWD)*, which includes sources like *Electronic Health Records (EHRs)*, insurance claims, and patient registries, ensuring completeness is vital for producing accurate and reliable analyses.

For this reason, ensuring the completeness of the data is essential to generate accurate insights, as missing information can result in incomplete analyses and unreliable decision making. Inadequate data quality poses the risk of biased conclusions, which can compromise the validity of research findings or the effectiveness of business strategies [1].

2 Questions to Ask

Evaluating the quality of *RWD* begins with a set of key questions to understand its **reliability**, **completeness**, and **suitability** [2].

- **Missingness:** What percentage of data is missing for critical variables? Are key fields (e.g., demographics, outcomes) complete? Are there known limitations or systemic missingness patterns? How is missing data represented? (e.g., *NULL*, *NA*)
- **Relevance:** Does the dataset include all variables necessary to address the research question? Are any essential data elements absent?
- **Timeliness:** Is the data up-to-date? What time range does the dataset cover? Are there any known gaps?
- **Provenance:** What is the source of the data? (e.g., *EHR*, claims, pharmacy systems) How has it been transformed or normalized? Are there documentation of these transformations?
- **Representativeness:** Does the dataset reflect the target population in terms of demographics, clinical characteristics, and geographic coverage?

3 Data Analysis Techniques

To analyze and build trust in our dataset we need to perform different analysis [3] [4].

- **Missing Data Analysis:** Calculate missingness rates for each variable and identify patterns of missingness (e.g., random vs. systematic).
- **Validity Checks:** Assess whether data values are within expected ranges (e.g., physiological limits for lab results).
- **Completeness Thresholds:** Compare completeness against predefined thresholds based on clinical expectations or regulatory standards.
- **Cross-Source Validation:** Integrate additional sources (e.g., claims data or registries) to fill gaps and improve completeness.
- **Temporal Trends:** Analyze consistency of data collection over time to ensure longitudinal completeness.

4 Comparing Datasets

Assessing completeness often requires comparing datasets based on key factors such as data depth, provider details, insurance types, and clinical coverage. A scorecard can help structure this evaluation. Since no dataset is perfect, trade-offs must be made between factors such as **population coverage**, **data linkability**, and the availability of **detailed individual-level information**. Therefore, a thorough review of the source, time frame, and structure of each dataset is essential for an informed choice [3].

5 Challenges

Ensuring **completeness** in *RWD* involves facing several common challenges [5].

- **Data Entry Errors:** Inaccuracies during manual input or from faulty equipment can result in missing or incomplete data.
- **Data Integration Issues:** Combining data from diverse sources may lead to compatibility problems.
- **Data Quality Control:** To deal with persisten errors, continuous monitoring is needed.
- **Obsolete Data Systems:** Obsolete systems may not support modern data formats, resulting in lost or unreadable information.
- **Lack of Data Governance:** Ambiguity in roles and ownership can affect data accuracy.
- **Insufficient Feedback Loops:** It is necessary to implement user feedback to deal with some errors.
- **Systematic Missingness:** Structured patterns of missing data can bias analysis.
- **Lack of Transparency:** Inadequate documentation of data processes reduces trust and reproducibility.

6 Conclusion

A comprehensive completeness assessment combines exploratory analysis, domain expertise, and detailed comparison. This approach ensures that the data set is suitable for its intended purpose while guiding data cleaning efforts and optimizing study design strategies.

References

[1] Atlan (2023). “What is Data Completeness? Examples, Differences & Steps“. https://atlan.com/what-is-data-completeness/.

[2] Blue Health Intelligence. “5 Key Indicators of Real-World Data Quality“. https://bluehealthintelligence.com/blog/five-key-indicators-of-real-world-data-quality-in-healthcare/.

[3] McKinsey & Company (2023). “Real-world data quality: What are the opportunities and challenges?”. https://www.mckinsey.com/industries/life-sciences/our-insights/real-world-data-quality-what-are-the-opportunities-and-challenges.

[4] TelmAI. “Ensuring data completeness: building trust in your data“. https://www.telm.ai/blog/how-to-measure-data-completeness-a-step-by-step-guide/.

[5] Hevo Academy. “What is Data Completeness? Examples, Challenges Steps“. https://hevoacademy.com/data-analytics-resources/data-completeness/.