Paper PP26

Data Curation Beyond Today's Standards

Catharina Dahlbo, Capish, Malmö, Sweden
Eva Kelty, Capish, Malmö, Sweden

ABSTRACT

Clinical trial data is inherently complex and sourced from various systems, including electronic health records, patient surveys, compliance tools, and standard or specialty laboratories. This diverse origin and storage across different databases pose significant challenges to data consistency. Therefore, data curation is essential for maximizing the utility of clinical trial data. While standards like CDISC are utilized for specific purposes, challenges persist due to tabular formats and inconsistencies across domains.

To fully leverage clinical trial data for scientific exploration and knowledge enhancement, additional processing is necessary. Adopting a model optimized for exploring data stored in a graph database offers significant benefits. Graph databases provide flexibility with an easily expandable data model, organizing data as a mind map, thus facilitating comprehensive data exploration and visualization.

We will delve into examples of further curated data and how it can empower researchers with its findability, accessibility, interoperability, and reusability.

INTRODUCTION

Clinical trials generate vast amounts of data from a variety of sources, for instance electronic health records, laboratory results, adverse events and patient-reported outcomes. These diverse datasets are essential for assessing the safety and efficacy of new treatments, but their complexity often leads to significant challenges in managing and analyzing the information.

Standards such as CDISC have been developed to streamline the organization of clinical trial data for regulatory purposes, offering a uniform structure for data submission. However, these standards alone do not solve many of the challenges that arise from handling complex, heterogeneous data, particularly when aiming for deeper scientific insights.

The limitations of current standards lie in their reliance on static tabular formats, which can lack the flexibility needed to handle real-world variability in clinical trial data. Additionally, inconsistencies across different studies, such as variations in how visits or outcomes are recorded, further complicate data harmonization and analysis. As the scope of clinical trials continues to expand with the integration of digital tools and real-world data sources, it becomes clear that a more advanced model for data curation is necessary.

This paper examines the challenges posed by current data curation standards and explores how moving beyond these constraints, particularly using graph databases, can unlock the full potential of clinical trial data.

DATA CONSUMPTION

Data consumption in clinical trials refers to how the data, collected during the study, is accessed, analysed, and used by various stakeholders. This data includes critical information such as patient demographics, treatment outcomes, adverse events, lab results, and other measurements taken throughout the trial.

The main goal of consuming the data is to extract valuable insights that help researchers and decision-makers in several ways: tracking the trial's progress, ensuring that patient safety is maintained, and determining whether the treatment being tested is effective. For data consumption to be efficient, it requires the data to be well-organized, carefully curated, and easily accessible.

For an end-user, clinical trial data consumption is not just about accessing data - it is about being able to interpret and analyse the data in meaningful ways that lead to actionable insights. This process can be complex due to the volume, variability, and nature of clinical trial data, but with the right tools and curation, end-users can transform raw data into knowledge that benefits both scientific advancement and patient outcomes.

CHALLENGES WHEN CONSUMING DATA

Within the CDISC framework, models like SDTM and ADaM serve specific roles in structuring clinical trial data. SDTM organizes raw data into a standardized format that facilitates collection, management, and reporting, making the data easier to navigate. ADaM, on the other hand, prepares datasets for statistical analysis, ensuring that the data is structured for accurate and thorough evaluation. Even though SDTM and ADaM address regulatory and research needs, inconsistencies may still arise across domains or trials, impacting analysis accuracy. These standards focus primarily on regulatory compliance, and additional curation is often needed to harmonize data from diverse sources and maximize its scientific value.

Despite their importance, CDISC standards alone are not the best for broader scientific exploration. Researchers often encounter challenges when integrating data from different studies or when conducting cross-study comparisons due to mismatched data structures and inconsistent coding. These limitations reduce the utility of the data for deeper scientific inquiry and knowledge generation.

A significant challenge in clinical trial data consumption is ensuring **data quality**. Incomplete, inconsistent, or erroneous data—such as incorrect dates, duplicate entries, or missing information—can lead to inaccurate conclusions or delays in regulatory approval. For example, tracking patient outcomes over time becomes difficult when visit records or timelines are inconsistent, making it hard to establish accurate cause-and-effect relationships.

Data complexity adds another layer of difficulty. Clinical trials generate vast amounts of data from diverse sources, including electronic health records, lab reports, and patient-reported outcomes. The integration of structured data, like coded values, and unstructured data, such as clinical notes or medical images, requires advanced tools and expertise. The lack of standardized formats often complicates this process.

Data integration is also a challenge, especially when comparing multiple trials or data from different institutions. Despite the use of common standards, inconsistencies still arise, requiring extensive manual curation before analysis. Additionally, tracking events over time can be difficult when dates or visit records are inaccurate, distorting trial results and complicating analyses.

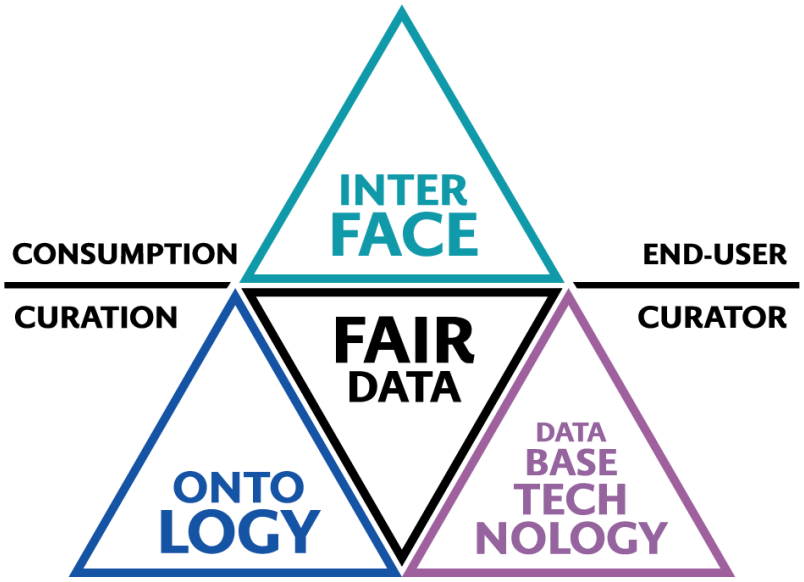


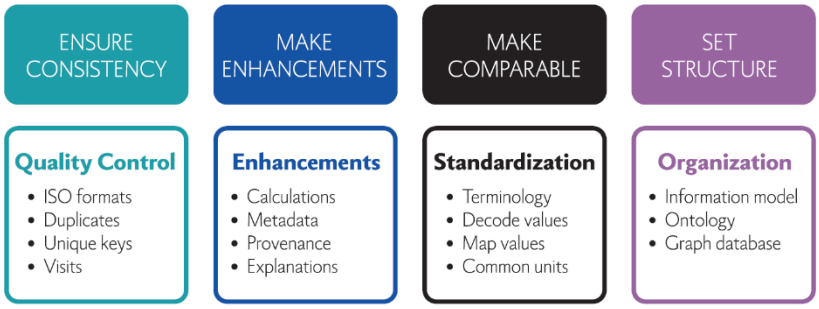
Figure 1. The data should be curated to let the end-user focus on the data through an intuitive interface.

DATA CURATION

On the way to good looking data for the end-user, the curation plays a pivotal role in maximizing the utility and value. It involves cleaning, standardizing and organizing the data to ensure its accuracy, consistency, and accessibility for various stakeholders, including researchers, regulatory bodies, and healthcare professionals. Given the complexity of clinical trial data - ranging from patient demographics and treatment outcomes to lab results and adverse events - effective data curation is essential for transforming raw data into a structured format that can be easily utilized for different purposes, such as analysis, regulatory submissions, and long-term research.

FAIR DATA

To address the challenges and improve the way clinical trial data is consumed, there must be a focus on facilitating data consumption through more advanced data curation techniques and the use of modern technologies. Making data FAIR (findable, accessible, interoperable, reusable) is nowadays a matter of course and serves as a foundation but achieving data reusability requires a practical approach and solution.
Several key aspects contribute to more efficient and effective data consumption.



The first step, including quality control of the data, is to **ensure consistency**.

- Duplicate entries – adjust to only one record.
- Conflicting answers – include only one answer handled by documented rules.
- Comments in result fields – move comments into another field so the result field only contains relevant information.
- Values spelled in different ways –map to one common spelling.
- Unique keys – ensure unique keys.
- Date corrections – ensure correct dates.
- Visits – harmonize inconsistent visits.

The second step is to make enhancements to **make data useable.**

- Calculations – add calculations that add value, e.g., aggregate measures
- Data & Metadata – combine metadata and data.
- Provenance – add provenance.
- Comments – add comments for explanations.

The third step includes efforts to **make data comparable**.

- Standardization –map data according to e.g. ISO 8601.
- Terminology –map data to a uniform terminology.
- Common units – ensure measurement units to be common for the same measurement.
- Decode values – decode according to value lists e.g., F/M to Female/Male.
- Map values – map e.g., Yes/No to understandable terms within the medical concept.

The fourth step is to **set a structure** for effective data modelling.

- Information Model – describes how the data is modelled.
- Ontology – describes the structure and holds the rules for the model.
- Graph database – stores the data like a mind-map and link data via relationships.

Together, these elements streamline data consumption, improve decision-making, and ultimately lead to better outcomes in clinical trials. Although extensive data curation is potentially a time-consuming activity, it pays major dividends in the subsequent analysis of the data.

ONTOLOGY & DATABASE TECHNOLOGY

One of the key reasons data curations is so important is that it directly influences how well clinical trial data can be leveraged for scientific exploration and knowledge enhancement. Researchers rely on curated data to conduct in-depth analyses, identify patterns, and make evidence-based decisions. For this process to be successful, it is essential that the data is intuitive to understand and interpret. This requires not only basic curation but also harmonization, ensuring that data collected under different circumstances, from multiple studies, or across various regions, is standardized and comparable. Without this, inconsistencies may arise, making it difficult to draw reliable conclusions or perform cross-study analyses.

Many clinical trial datasets are presented in tabular formats, which have inherent limitations. For example, tabular data may fail to capture the complexity of relationships between different variables or events, and inconsistencies

may occur across different domains or categories within the data. These inconsistencies can complicate data processing, requiring additional steps to clean, harmonize, or reformat the data to ensure consistency and reliability.

Using a graph database provides flexibility with an easily expandable data model, and organizing data as a mind map, thus facilitating comprehensive data exploration and visualization. An ontology is holding the rules for the information model and describes the structure of the content in the graph. It defines how the nodes of the graph relate to each other, which fields are included in each node, the datatypes of the fields, units to be used, etcetera. The applied ontology also provides a terminology controlling the naming of all the entities.

By ensuring intuitive data understanding and addressing inconsistencies across datasets, stakeholders can fully unlock the potential of clinical trial data, ultimately leading to better conclusions, improved patient outcomes, and advancements in evidence-based medicine. By utilizing the ontology, data will be easier to access, easier to explore, easier to share, and above all, easier to understand.

A FAIR repository is a good starting point when trying to get a holistic view of clinical trials. By storing data in a graph database and adding a powerful interface, the end-user has the opportunity to interact with the database in the background. If the primary goal is simply interacting with the data, a basic graph database will suffice. However, for exploring complex relationships and correlations, more advanced technology is required.

BENEFITS OF A MODEL OPTIMIZED FOR DATA EXPLORATION

Facilitating data consumption involves a combination of adequate technology, user training, and continuous process improvements to make data more accessible, usable, and valuable for all users. Letting the curated data with its metadata then be explored in a tailor-made intuitive application, that takes advantage of advanced ways of modelling the data, allows users to move from detailed patient information to analysis of the entire population and vice versa.

Data curation has numerous benefits that can help businesses and organizations in various ways. One of the primary advantages of data curation is improved data quality**.** By properly curating their data, companies can ensure that they have accurate, complete, objective and consistent information to work with. Data curation is all about efficiency, both from the perspective of the curator as well as the perspective of the data consumer.

When a graph database is combined with a conceptual ontology, it mirrors the human brain's organization of information. Adopting this method of data modelling removes the need for a user to have knowledge about the database structure in order to access and explore data.

Interoperability between systems and collaboration across teams allow for better data sharing and enhancing teamwork. Adhering to regulatory compliance and maintaining security ensures that sensitive patient data is handled properly, avoiding legal issues and maintaining trial integrity.

CONCLUSION

When data doesn’t need to be connected, minor inconsistencies in well-curated data may not pose significant problems. However, when the goal is to relate and compare data, consistency becomes crucial. A well-curated dataset, accessible through a robust interface, greatly enhances the end-user's ability to derive meaningful insights. By consuming data without worrying about how it's organized or whether it's consistent across studies, users can focus more on patient safety, correlations, and improvements.

Proper curation and harmonization make clinical trial data more valuable, as they allow for seamless integration into larger datasets, facilitate comparative analyses, and contribute to the advancement of medical research and healthcare delivery. Facilitated data consumption in clinical trials aims to improve how easily and efficiently stakeholders can access, explore, and extract insights from trial data.

Investing in data curation allows stakeholders to unlock the full potential of trial data, leading to better decision-making in pharma/healthcare and the advancement of medical knowledge and treatments. Additionally, effective post-trial data curation plays a key role in optimizing future trial designs and furthering scientific understanding. By ensuring that trial data remains valuable and actionable, such investments ultimately contribute to improved patient care and the progress of evidence-based medicine.

ACKNOWLEDGMENTS

We would like to thank Patric Moreau for his dedication to the poster design.

RECOMMENDED READING

Graph vs Table – Take your Pick, C.Dahlbo & E. Kelty, PP14 PHUSE US CONNECT 2023

Make the Most of Your Data - Explore Different Perspectives, C. Dahlbo & E. Kelty, PP01 PHUSE US CONNECT 2019

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Catharina Dahlbo
Work Phone: +46733800191
Email: catharina.dahlbo@capish.com

Author Name: Eva Kelty
Work Phone: +46702416375
Email: eva.kelty@capish.com

Company: Capish
Address: Lokgatan 8 211 20 Malmö

Website: capish.com