Paper DH02 – PHUSE EU Connect 2025

The Value of Clinical Trial Data: From Submission to Innovation

Anne Katrine Alsing, Novo Nordisk, Søborg, Denmark

ABSTRACT

In the evolving landscape of clinical trials, the data contributed by subjects holds immense potential beyond regulatory submission. This presentation emphasizes the crucial value of acknowledging this data, advocating for its secondary use to enhance future studies and unearth unmet medical needs. We will explore the ethical considerations and data governance frameworks necessary for responsible utilization of clinical data, ensuring integrity and patient trust. Furthermore, we discuss the importance of harmonizing core data domains to foster transparency and interoperability across different studies and clinical programs. Robust and well-documented datasets serve as foundational resources for leveraging artificial intelligence (AI) insights, ultimately advancing patient care through enhanced clinical decision-making.

INTRODUCTION

This paper aims to share an approach to facilitate greater value gains from the historical clinical trial data. Traditionally, this data has been largely restricted to submission packages, limiting its potential for broader application. By advocating for a more expansive use of clinical trial data, this paper will touch on the necessary governance frameworks and ethical considerations that must be addressed to facilitate such an initiative. Further we will share our experiences at Novo Nordisk, where we recognized the challenge of data availability and transparency across projects and how we have built a cross project wide pool of harmonised data to overcome this challenge. Fostering an environment that encourages exploratory research, responsible secondary use of data and ultimately enhances patient outcomes.

DATA GOVERNANCE AND ETHICS

In exploring how to govern the use of clinical trial data beyond its primary use, it is essential to acknowledge the existing legal and ethical frameworks that guide the use. Clinical trial subjects provide their consent for data usage primarily for the purposes of the study; however, we also collect consent for the broader use of this data to help improve medical care and science. Each study is unique, necessitating a nuanced approach to ensure that any secondary exploration of the data adheres to the stipulations laid out in the unique informed consent.
To successfully broaden the application of clinical trial data three key components should be addresses:

1. **Consulting the Informed Consent**: It is imperative to have a clear alignment between the informed consent and the desired secondary purpose. As the complexity of consent forms can vary over time, we have centralized this effort to relief the burden from the individual users and align assessments.
2. **Data Ownership**: Clearly defined roles within the data management hierarchy are essential. Each data product should have designated data owners and stewards who determine the appropriate usage of the data. Understanding ownership will also address potential issues related to company disclosure, ensuring that internal access to data is managed responsibly.
3. **Conforming to legislation**: Use of clinical trial data for secondary use will often be associated with constraints coming from external legislation or internal ethical requirements. It is essential to govern this in a general fashion to enable data users equal and correct access and guidance.

While the hurdles associated with data governance and ethical use may seem significant, establishing a robust framework for clinical trial data utilization can unlock untapped potential and foster a culture of collaborative research that ultimately benefits patients.

DATA ENGINEERING – SYSTEM DESIGN AND DATA RELEASE

To enhance the accessibility of clinical trial data within Novo Nordisk, we have undertaken two approaches to harmonize and organize our data effectively.

Study-Level Metadata: Harmonising study metadata so that users can search for studies based on general descriptions, populations, and therapeutic areas. Enhanced with metadata on available data modalities. This strategy provides an initial layer of searchability, allowing users to understand the scope of data available.

Subject-Level Data Harmonization: Recognizing the limitations of merely searching through metadata, we prioritized the alignment of core clinical data domains across all studies. By analyzing actual study reports and transforming the subject-level data into a cohesive format, we enable users to perform more granular queries across datasets. This not only allows for the identification of specific subpopulations but also offers deeper insights for exploratory analysis, Figure 1.

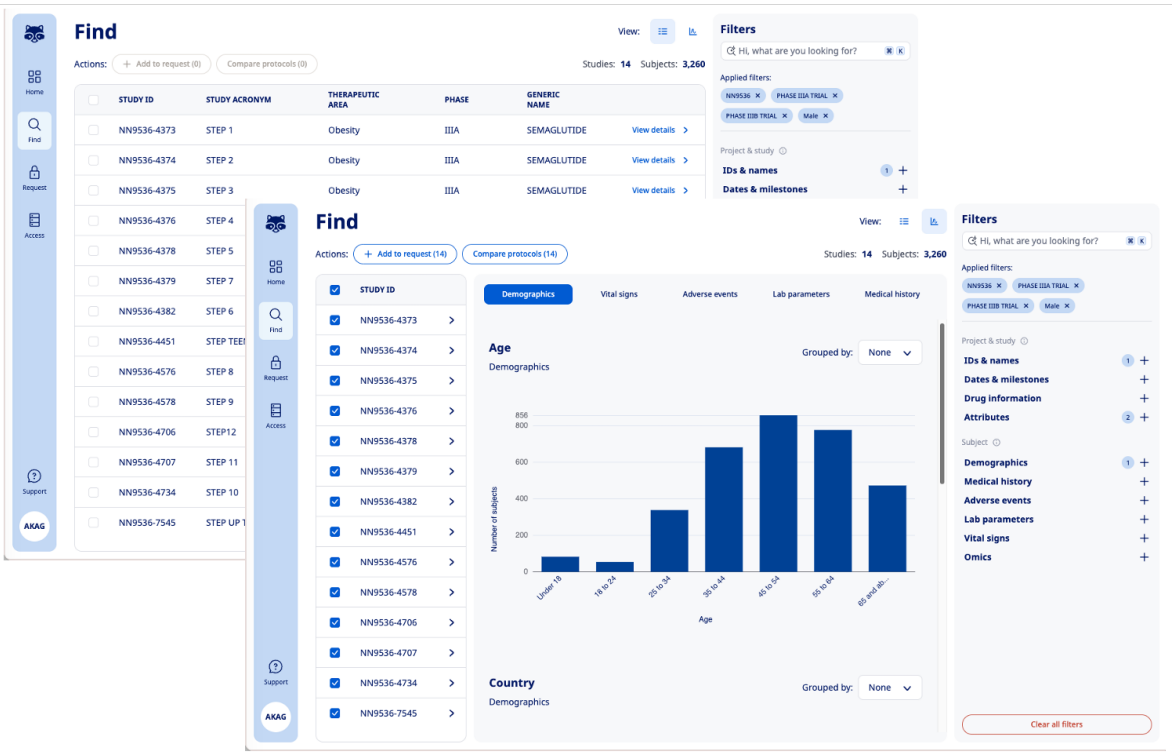

Figure 1 FounData interface enabling the users to filter historic studies on both study and subject-level parameters. The harmonisation of core domains facilitates a further exploration of available sub-populations of special interest.

Technically we are approaching the harmonisations using a combination of coding and customized configuration tables. Generic code orchestrates the transformations of source data, and an elaborate set of configuration tables manage study specific logic (See subsequent sections on data handling for examples).

To ensure robust version control of both code and configuration tables the project utilizes Azure DevOps Git repository integration into CI/CD pipelines. The code is deployed as asset bundles through DEV, VAL and PRD environments ensuring all quality checks and reviews are met.
Similarly the configuration tables of each new study ingestion are examined and updated within a TEST environment using the same code base as released to PRD. Once all configurations tables have passed the data standard expert review the configuration tables are released into PRD.

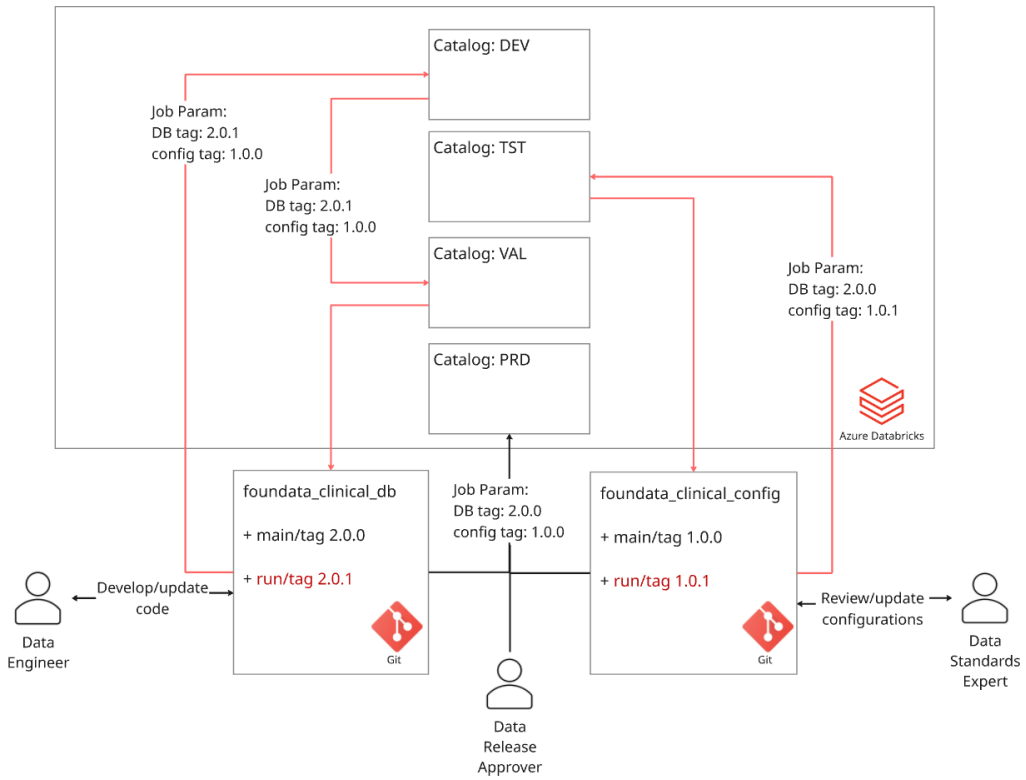

Figure 2 Component flow to support 1) controlled updates and versioning of both code and configurations 2) Separation of stable production logic from experimental or in-progress development and mapping configuration and 3) Full auditability and traceability of code and configuration changes through git history and tagging and related release notes.

Each release into PRD will result in a new data product release accompanied by a release note detailing the semantic version of both code and configuration ADO repository, guaranteeing stability, traceability, and reproducibility of the harmonization and transformation logic. This approach ensures that only validated and approved code and configurations are used in production data processing, Figure 2.

For non-production workflows (development, testing, validation), mapping configuration files can be sourced from a specific branch of the ADO repository or semantic version tag. This enables developers and data standard experts to test, review, and iterate on mapping logic before promoting changes to production via a new semantic version tag.

DATA HANDLING – EXEMPLIFIED BY THE CLINICAL HARMONISATION PIPELINE

Data is ingested in batches from clinical trial datasets, handling various file types such as SAS7BDAT, XPT, and Parquet, with the bronze layer serving as the raw data repository and the source of truth. The data pipeline employs the Databricks medallion architecture comprised of three layers: bronze, silver, and gold, see Figure 3. Progressively refining the data for quality and usability through the layers.

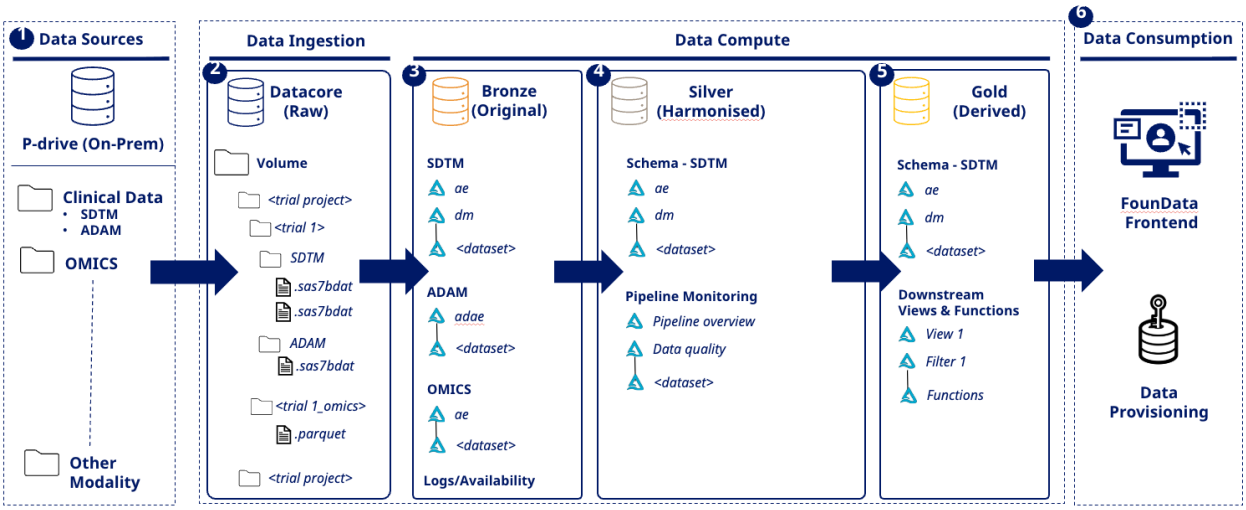

Figure 3 Medallion architecture, exemplified on the clinical data pipeline, progressively parsing the data through transformations of data harmonisation and enrichments. (1) Legacy data sources. (2) Azure storage point establishing one source of truth (3) In the Bronze layer the data is parsed into stacked delta tables by domain name, in reality preparing a (heterogeneous) pool of studies across therapeutic areas. (4) In the silver layer, the pipeline harmonizes a selection of SDTM domains using ontologies to ensure quality and standardization. (5) The gold layer includes enriched data products for consumption by downstream systems and end user analysis (6). Delta tables are utilized for all three layers, facilitating data management like a data warehouse, including time travel and transaction logging for version control. Spark is used for data processing, ensuring compatibility and efficiency in data handling.

BRONZE LAYER

The bronze layer contains non-harmonized data which has been stacked per domain in delta table format. The data is read from source, then converted to delta table format and initially stored in string datatype. To ensure no loss of information from source, automated data integrity checks are performed (e.g.: check the number of columns and non-null values counts).

In the **Data Parsing & Stacking** step, the converted and mapped study datasets are all stacked per domain, resulting in one single dataset per domain. Where columns do not exist for a given study, their values are simply set to null.

> *Since SAS® has no concept of null (it uses empty strings for missing character data), this leads to a difference in the significance of null and empty string in the Spark data:*
> - *Null values in a column mean that the data item simply did not exist in the original dataset for that study.*
> - *Empty string values mean that the column was collected in the original clinical data, but that the value was never populated, i.e. it was "unknown" in the true sense of the word.*

Appropriate **Spark datatypes** are assigned to the String-type columns in the stacked datasets leveraging a configuration table to accurately discern the datatypes associated with each selected dataset. The configuration table links SDTM variable names to their corresponding datatypes, including a set of regular expressions for ADaM variables to facilitate accurate identification.

Finally the **supplemental-qualifier datasets are joined** onto its base domain dataset. This is done to address the challenge of multiple IDVARs for a single QNAM, ensuring no data is lost and preparing for consistent use of the data.

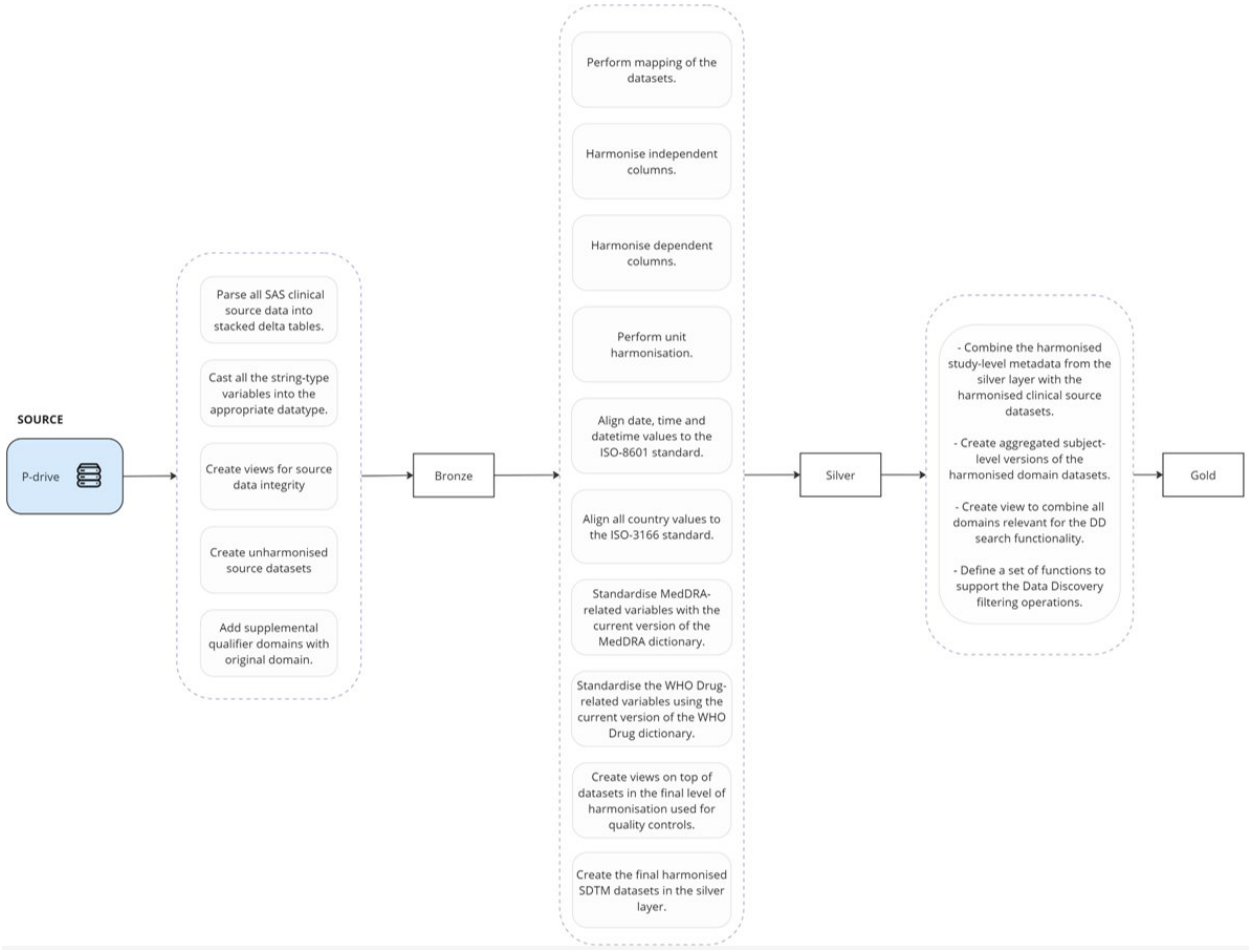

Figure 4 Schematic overview of the transformations addressed across the clinical data pipeline, from the Statistical computing environment (P-drive) to the core source domains harmonised and enriched in the golden layer. The golden layer also includes data transformations to enable the study and subject filters of the data discovery part of the FounData Application (DD). Across the nodes configuration tables control the data transformations.

SILVER LAYER

In the Silver Layer, the raw non-harmonized data from the bronze layer is processed, standardized, harmonized and enriched to share data with the users in a meaningful and efficient way. The data transformation steps of the silver layer are outlined in Figure 4 with few more details provide below.

The **Domain and Variable Mapping** step aims to map the naming convention of the variables and domains as per the Clinical Data Interchange Standards Consortium (CDISC®) model. This step is performed for the 14 core source domians currently included in our harmonisation pipeline:

ae, cm, dm, ds, ex, lb, mh, se, sv, ta, te, ts, tv, vs

The target domain and column mapping allow source variables from different datasets to be mapped to a target domain dataset. When multiple source datasets are mapped to the same target dataset as DM, AE, etc., the rows from all contributing tables are concatenated (i.e. appended) using the SQL union operation. Both domain and column mappings are specified by Clinical Data Specialists in configuration tables.

The **Independent Harmonization** step consists of a review of the values within a certain variable and its standardization using CDISC® controlled terminology codelists or input provided by data standard experts in configuration tables. When values are identified as conflicts and their corresponding standardized values are not available in the related CDISC® codelists, the expert investigates each conflict using their knowledge and other resources to provide input on the new standardized value.
Columns from Findings-type domains which have an interdependency upon one another are excluded from this routine, e.g.: LBTEST, LBTESTCD, LBSPEC, LBCAT and LBSTRESU in the LB domain.

The **Dependent Harmonization** ensures that there is a unique and standardized combination of values for key variables[12] of certain domains[8] (e.g.: LBTEST – Laboratory Test Name, LBTESTCD – Laboratory Test Code, LBSPEC – Laboratory Test Specimen).
The data transformations are controlled by configuration tables governed by data standard experts, ensuring consistency in both column alignment and content integrity. An example of the configuration metadata is the dependent harmonisation mapping sheet where LBTEST, LBTESTCD and LBSTRESU is mapped to consistent values across a multitude of historical capture values.

LBTEST	LBTESTCD	LBSPEC	LBSTRESU	mapped_LBTEST	mapped_LBTESTCD	mapped_LBSPEC	mapped_LBSTRESU
Albumin	ALB	BLOOD	g/L	Albumin	ALB	BLOOD	g/dL
Albumin	ALBS	SERUM	g/dL	Albumin	ALB	SERUM	g/dL
Albumin_serum	ALBS	SERUM		Albumin	ALB	SERUM	g/dL
ALBUMIN	ALB	URINE	mg/L	Albumin	ALB	URINE	mg/L

Table 1 Part of configuration metadata sheet governing the harmonisation of dependent variables LBTEST, LBTESTCD, LBSPEC and LBSTRESU. In addition to the shown values, the dependent harmonisation also updates LBMETHOD to enable a comprehensive description of the assessment. Data standards expert will often consult both data, study protocol and lab specifications to assign the mapped values.

During the first pass through in the Dependent content harmonization, each set of dependent variables are matched against previous mappings and controlled terminology. If matches are found, they are harmonized accordingly. If no match exists, the routine prepares a qualified suggestion for the experts to review. Each suggestion is based on an ordered set of logical matches and rules, and assigned an expected level of review by the data standard experts.

In the **Unit Harmonization** step, the standard unit for a given Laboratory Test (LBTEST) or Vital Sign Test (VSTEST) are identified alongside their correspondent unit conversion factor on the unit cofiguration tables. Next, the conversion calculations are implemented to ensure all values are converted according to its correspondent standard unit.

The **STRESC Harmonization** maintain consistency across categorical values. As part of a clinical study process, multiple tests are conducted on subjects, results of which are captured as both numerical and categorical values. Due to the global nature of clinical study execution, there can be multiple synonyms for a particular categorical value captured in STRESC. To maintain consistency across data, the STRESC variable is harmonized according to the CDISC® codelist values, so that all the synonyms can be narrowed down to a single standard categorical codelist value.
For values not diretly mapped to synonyms on the CDISC® codelist the data standard experts provide configuration table input to ensure consistent mapping of all values.

In the Medical Dictionary for Regulatory Affairs **(MedDRA) standardization** step, all the Adverse Event (AE) and Medical History (MH) terms are standardized to the most recent version of a common dictionary of MedDRA allowing seamlessly finding and browsing through the adverse events and medical history captured in the clinical study data. The MedDRA harmonization routine follows a hierarchical decision-making process to align clinical dataset values with the MedDRA Dictionary.

- Check the --LLT variable for a case-insensitive match; if found, all 10 MedDRA values are populated. If no match exists, it examines the MODIFY and --TERM variables, followed by the --LLTCD.
- Check the Preferred Term level (PT or DECOD) with the same process.
- Check the Higher-Level Term (HLT) level with the same process.
- Check the Higher-Level Grouping Term (HLGT) with the same process.
- Finally, ascend to the System Organ Class (SOC) level.

For each successful match, the corresponding variables are populated, while those at lower levels are set to null accordingly. Any values which cannot be matched with the dictionary are addressed by data standard experts in a configuration table.

In the World Health Organization Drug Dictionary **(WHO-DD) standardization** step, all Drug Names in the CM (Concomitant Medications) domain are standardized to the most recent version of WHO Drug Dictionary which gives access to the most up-to-date and comprehensive drug information. This ensures that our data reflects the latest global standards and classifications for pharmaceutical products, supporting accurate and consistent representation within our system.

In the **ISO Countries Standardization** step, the country codes across all datasets are formatted to ISO 3166 standard format for consistency across all the country codes.

In the **ISO Dates Standardization** step, the dates across all studies are formatted to ISO 8601 standard format for consistency across all the dates and datetimes.

GOLD LAYER

In the Gold Layer the core source domains are enriched with additional variables commonly used for analysis Figure 5.
Data Enrichment involves adding variables to the datasets that would usually be added as part of user analyses (e.g., pseudo baseline for lab results, and time since first exposure in event domains). This helps users get started on their analysis and "fail fast" as they test their hypothesis with the data. The end users can always re-derive these variables once they have tested their initial hypothesis.

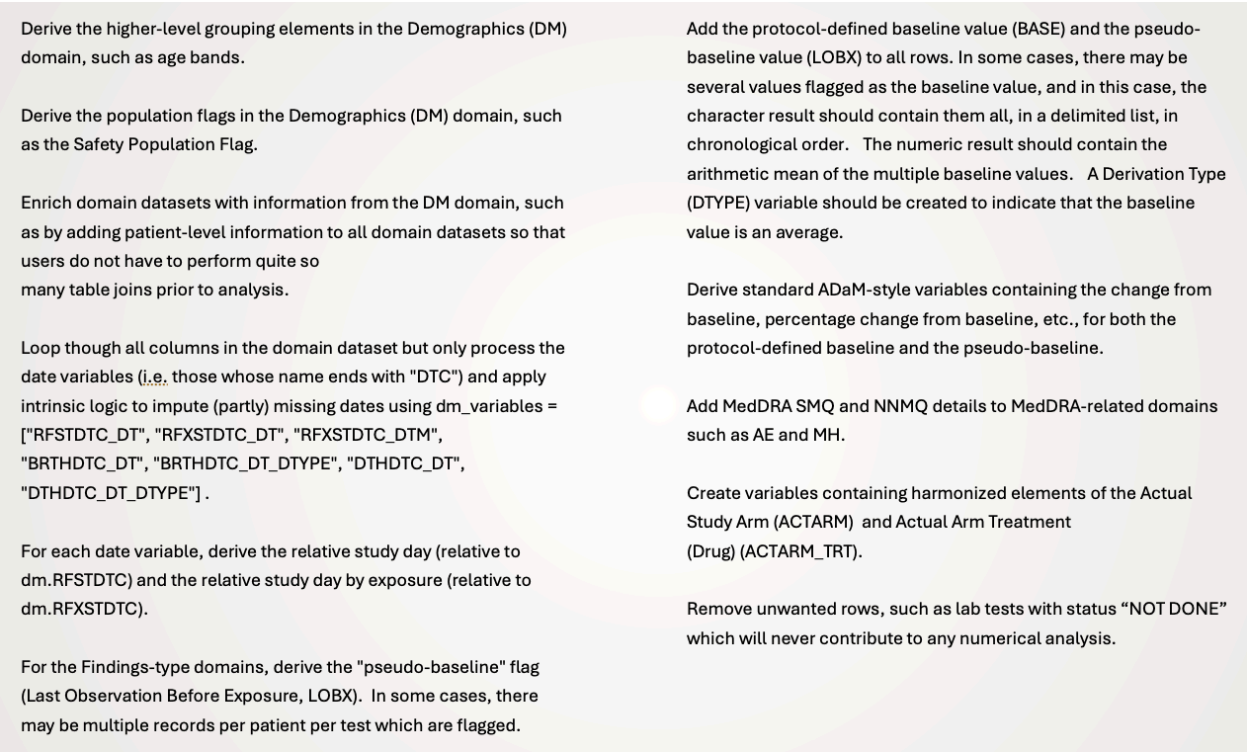

Figure 5 Examples of enrichments added to the core domains and available in the golden layer of the pipeline. One enrichment that has proved particular useful is the standardization of the ACTARM values including population of missing values logically using information from Reason for Null Arm (ARMNRS) and ds.DSDECOD. Further a new value ACTARM_TRT was assigned to provide a consistent variable to identify the drug of the actual treatment across studies.

Throughout the harmonisation pipeline data integrity and quality is monitored and assessed as part of the data release. Across *the stacked clinical source domains the user-facing data schema is ordered to present columns appropriately, and tags and comments are applied to the tables and columns to guide users who are less familiar with the CDISC® standards.*

DATA USAGE – ACCESS MANAGEMENT, USER SUPPORT, AND USE OF AI

While harmonizing the data is crucial, enabling user access to this data is equally significant. We recognize that although much of the clinical trial data is now more organized and findable, access must be user-friendly and guided. Our approach encompasses several key components:

1. Access Governance: We are establishing clear governance by the implementation of a semi automated risk based framework to access informed consent compatability, internal confidentiality concerns and legal recquirements in a fair and equal maner across end users.
2. Data Connection: Enabling pathways for colleagues to connect to data, utilizing both a dedicated Posit server and providing access through Databricks. This dual approach ensures that users with varying technical skills can engage with the data efficiently.
3. User Guidance: Understanding that not all users are data scientists, we will provide comprehensive guidance on how to navigate and utilize the available data effectively. This will include creating user manuals and hosting training sessions to empower users with the knowledge to draw insights from the data.
4. Integration of AI Tools: To further support users with limited data analysis experience, we are incorporating AI-driven tools designed to generate insights from our harmonized datasets. By employing AI agents, users can pose questions and receive data-driven answers without needing to delve deeply into complex data analysis processes.

CONCLUSION

Historically, access to clinical trial data has been limited to specific teams involved in processing the data for regulatory submissions. We advocate for a more inclusive approach that allows a wider range of colleagues to use the data for exploratory analyses as soon as possible after study submission.

We are preparing our clinical trial data for findability and usability by harmonisation of core doamins while maintaining clear traces back to the original clinical trial data sources. All transformations of content are captured in configurations files and both transformation code and configuration files are governed through CICD repository releases. In this way we ensure clear traceability an visibility of the entire data lineage.

Harmonizing data is vital, but user-friendly access is equally essential. We are establishing clear pathways for colleagues to access our clinical trial data via a dedicated Posit server and Databricks, catering to varying technical skills. To further support those with limited analysis experience, we are integrating AI-driven tools that allow users to generate insights from harmonized datasets without requiring advanced data analysis skills.

In conclusion, our comprehensive strategy focuses not only on harmonizing and making clinical trial data findable but also on ensuring that users can access and leverage this data effectively. By combining structured data access with user support and AI integration, we aim to facilitate informed decision-making and enable a culture of innovation within the organization.

ACKNOWLEDGMENTS

Big thank you to my wonderful colleagues providing vital input on both paper and presentation.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:
Anne Katrine Alsing
Nove Nordisk A/S
Work Phone: +45 3075 0346
Email: akag@novonordisk.com