Integrating Clinical Trial Data to Accelerate Drug Development with an Enterprise Data Fabric

WHITE PAPER

early adopter research

CAMBRIDGE SEMANTICS
THE SMART DATA COMPANY®

CONTENTS

early adopter research

Biopharma and healthcare industry leaders are recognizing the potential value of using vast collections of patient data from multiple clinical trials to accelerate research and development of new drugs and improve patient outcomes.

INTRODUCTION

Cambridge Semantics brings deep experience to helping global companies in many industries tackle the challenges of integrating large, complex datasets from across the enterprise. Its product, Anzo, uses graph technology to accelerate data integration initiatives.

This Early Adopter Research field report explains Cambridge Semantics' approach to solving one of the most challenging and urgent problems facing pharmaceutical companies: using data to accelerate development of lifesaving medicines. Data strategy leaders across all industries should read this field report to appreciate how Cambridge Semantics can help tackle their toughest data integration problems behind critical digital transformation initiatives.

THE UNTAPPED POTENTIAL OF MASSIVE COLLECTIONS OF CLINICAL TRIAL DATA

Biopharma and healthcare industry leaders are recognizing the potential value of using vast collections of patient data from multiple clinical trials to accelerate research and development of new drugs and improve patient outcomes.

When aligned around a common model, integrated data from multiple trials can be used to improve foundational research, trial design, site selection, patient recruitment, safety, and drug development success.

However, practical and technical barriers stand in the way, including differences in how and what data are collected across trials, the need for deep understanding of the data to enable accurate integration, and the sheer number of datasets. These combine to make integration of large-scale cross-clinical trial datasets practically impossible in many organizations.

Cambridge Semantics works with leading biopharma and healthcare organizations to unleash the value of decades of accumulated clinical trial data.

To accurately combine data from multiple trials, researchers and others with deep understanding of clinical trial data require detailed visibility into the content of each dataset to accurately map data into the target model.

THE THORNY NATURE OF CLINICAL TRIAL DATA

Aligning and combining data from multiple clinical trials is complex. Differences in how data is recorded across studies, which domains and variables are included in each clinical trial set, and changes over time to standards from the Clinical Data Interchange Consortium (CDISC) mean that the data from any one study must be thoroughly analyzed and compared to data from other studies before merging can occur.

When combining data from multiple clinical trials, the integration process presents special challenges. To accurately combine data from multiple trials, researchers and others with deep understanding of clinical trial data require detailed visibility into the content of each dataset to accurately map data into the target model. They need integration tools that define, try out, modify, and apply highly customized and sophisticated transformation and mapping rules against potentially hundreds or thousands of distinct fields of data.

Researchers need the ability to represent data through multiple technical and business-oriented models simultaneously—for example models that represent the original dataset from a single clinical trial, that same data transformed toward a target model, and then the data's final form as part of an SDTM (Study Data Tabulation Model) data collection. These models need to be persisted and connected over time using end-to-end lineage and provenance tracking. Finally, to efficiently handle the volume of datasets that need to be integrated, these integration capabilities must be applied automatically using intelligent rules, QA processes, and high-performance scalable data management architectures and platforms.

Until now, many life sciences organizations have taken manual approaches to this problem, enlisting clinical trial domain experts to comb through and map individual datasets into a common model. Such approaches do not scale and divert skilled resources from higher value tasks. Further, these approaches tend to rely on data management and data preparation tools designed to operate on just one or two clinical trial datasets at a time.

In recent years, a new generation of data integration technologies—emerging out of the era of enterprise-scale analytics and digital transformation—has enabled fresh approaches to integrating cross-clinical trial data.

Attempting to scale beyond these manual approaches, data integration solutions, which grew from business intelligence and relational database modeling technologies, also struggle to meet these challenges. Built around a fixed, linear, stepwise approach to extracting, transforming, and loading data from source systems into a predefined target schema, traditional approaches lack the agility and scale required to integrate data from multiple clinical trials into a common harmonized model. Additionally, because traditional tools are typically managed by IT teams, using such tools largely prevents people in the business with deep subject matter expertise, such as pharma informatics experts and researchers, from participating in the data exploration, interpretation, and integration process.

In recent years, a new generation of data integration technologies—emerging out of the era of enterprise-scale analytics and digital transformation—has enabled fresh approaches to integrating cross-clinical trial data. Cambridge Semantics has helped several companies build enterprise-scale collections of cross-clinical trial data and leverage those collections to modernize and transform the drug development and commercialization process.

FROM CROSS-CLINICAL TRIAL DATA TO ENTERPRISE DATA FABRIC: BIOPHARMA AND HEALTHCARE DATA USE CASES OVER TIME

Before explaining key steps in modern data integration for cross-clinical trial data, it's worth examining the roadmap that many projects take as they mature.

As mentioned, the drivers for cross-clinical trial data harmonization initiatives include escalating drug development costs and an increasingly small set of unexploited therapeutic domains against which to target new initiatives.

Faced with these challenges, many life sciences organizations begin building harmonized, blended datasets from past trials to fuel faster, more efficient, productive, and impactful drug development.

Other clinical data management organizations find their entry point into modern data integration through a different, but equally urgent need: the development of integrated metadata to support mandated delivery of information on new drugs to regulatory agencies.

As new drugs move through the approval for commercialization process, metadata accumulates, documenting the lineage, evolution, business meaning, and ownership of various datasets. The metadata are recast from one data model to another as the drug approval process moves forward. Tracking the metadata across the process is essential to allowing drug companies to produce highly detailed and specific datasets in response to regulated reporting requirements.

Regardless of the entry point, integration initiatives often expand over time with the addition of new data domains, user groups, and projects. In a life sciences organization, expansion can include the addition of new datasets and use cases originating upstream in R&D or product development or downstream in safety, medical affairs, or marketing functions. When a large number of datasets are integrated in this way, an enterprise data fabric emerges.

As more data from different domains is added into the fabric; as datasets are integrated and blended to expose connections across siloed applications; and as additional business-oriented data definitions and concepts are added using semantic graph data models, users from across the organization begin leveraging the data fabric as the trusted source for enterprise data for a wide variety of analytic, reporting, and AI/ML applications.

With this context in mind, we can now get to the core mission of this field report: explaining how Anzo blends and harmonizes clinical trial data sustainably and at scale.

KEY TACTICS FOR HARMONIZING CLINICAL TRIAL DATA

Anzo facilitates four stages to building blended and harmonized repositories of clinical trial data:

- Onboard, in which the description of the data and the relationships between data elements, data sources, and models are collected in a metadata catalog, and the data itself is prepositioned in a raw graph format ready for integration.
- Model, in which modelers, working with domain experts, define semantic models that describe the concepts and relationships within the data, not in relational form but in semantic form. Typically these models are based on **CDISC SDTM** standards.

- Blend, in which users transform the data within layers that are described by the models.
- Access, in which the data is distilled, analyzed, and delivered as needed to provide insights and other forms of value.

The details below reference specific implementations of enterprise data fabrics used to integrate data from multiple clinical trials, highlighting key capabilities specific to integrating data from individual clinical trial datasets into a cross-clinical trial data collection, targeting the SDTM data model.

Onboard

During the initial onboarding stage, clinical trial datasets are imported into Anzo, starting with metadata and data profiling, and ultimately loading all data points into the graph data model. The onboarding stage focuses on understanding the content of each separate clinical trial dataset in preparation for subsequent mapping into the SDTM data model.

Key capabilities in the onboard stage include:

- Ability to ingest all types of data, including specialized clinical data. It is critical to bring in data from all the various source systems relevant for clinical data, including CSVs, SAS files, and files from clinical data management systems. The solution should also support all common data types and file types, including time series data from sensors.
- The ability to onboard datasets from multiple clinical trials quickly and automatically without writing code. Automated pipelines are needed to accelerate this process and allow more people to participate, regardless of their technical skills. The process transforms incoming data from its original format into a graph data model.
- Access to metadata, including schema, column names, data types, and sample data, from each source during onboarding. Such metadata allows SMEs to quickly see and understand fields of data in each clinical trial dataset.

During the onboarding stage, the metadata catalog grows steadily with the addition of new metadata from each clinical trial dataset. First metadata from the source system is added, then profiling metadata, sample data, and known mappings. If transformations are applied to the data during onboarding, they are saved as new metadata as well. (This is addressed in the Blend section.)

When creating an enterprise data fabric of any type, the model phase instantiates the target model, capturing and blending data from multiple sources. In the case of clinical trial data, this model is often an enhanced or customized version of an SDTM or model.

In a clinical trial context, the metadata catalog created during the onboard stage maps all the data across all clinical trials. By allowing SMEs to review and explore that metadata, they can begin understanding the data and envisioning how it should be ideally mapped to the target SDTM model. In the next two stages, modeling and blending, the metadata catalog is further used to capture models and the blending and transformation steps used to evolve the clinical trial data from their siloed original form into an integrated harmonized cross-clinical trial data collection.

Model

When creating an enterprise data fabric of any type, the model phase instantiates the target model, capturing and blending data from multiple sources. In the case of clinical trial data, this model is often an enhanced or customized version of an SDTM or model. This model comprehensively describes what all the siloed incoming clinical trial datasets should look like once conformed and harmonized to align with the target SDTM model.

Key capabilities in the model stage include:

- Ability to represent an intact or customized version of the SDTM model. The SDTM data model must be represented in the platform so that each clinical trial dataset can be mapped to it. Ideally the platform allows the model to be imported or created from scratch. In addition, it's helpful if data modelers can extend or adapt the model to meet organization-specific requirements.
- Robust tools to explore the SDTM data model. These tools allow data modelers and SMEs to quickly understand the SDTM model as a whole, explore relationships between different domains, and examine specific classes and properties.
- Integration of controlled vocabularies (including synonyms) and code lists to accelerate integration and standardize terms.

Anzo supports these three capabilities with a standards-based semantic graph modeling toolset. The model stage builds upon automatic transformation of each incoming set of clinical trial data from the source format into the semantic graph data model. This allows each dataset to be more easily understood in business terms and lays the foundation for easier and faster transformation, blending, and harmonization of that dataset into the SDTM data model during the blend stage.

Blend

The blend stage transforms, integrates, and maps datasets from multiple clinical trials into the SDTM data model.

Key capabilities in the blend stage include:

- Visualization of each incoming clinical trial dataset as a graph data model. SMEs can quickly ascertain which domains and variables are present. This enables SMEs to begin understanding how to map each dataset to the SDTM data model accurately and efficiently.
- Creation of a statistical profile for each clinical trial dataset. This profile information allows SMEs to understand the content of each field and begin planning how to map that into the target model more quickly. Profile information can also identify fields that should comply with a controlled vocabulary, but which have missing or incorrect values and therefore require cleansing and standardization before integration.
- Suggested mappings for which data fields in an incoming clinical trial dataset may be strong candidates to map into specific variables in the target model. These suggestions can be based on a comparison of source and target metadata and statistical profile data to flag possible matches.
- Automatic application of transformation and mapping steps against values within each clinical trial dataset. This capability must be extremely agile to allow for rapid development of sophisticated transformation rules that address the often-nuanced demands associated with mapping data from individual trials into the target model.

Automatic transformation is critical to moving quickly. Essential capabilities in this area include:

- Extremely fast and agile development of transformation and mapping rules. This allows SMEs to develop, insert, remove, tweak, reorder, and apply multiple transformation steps to datasets quickly to arrive at the ideal “recipe” to map data into the target model in a matter of minutes or hours.
- Suggestion and automation of mapping rules that leverage insights gained from metadata, data profiling, and suggested data mappings to optimally map variables in individual clinical trial datasets into the target model at scale.

Anzo's high-performance underlying graph database supports integration of data from thousands of clinical trials, interactively. This is essential to meeting the data volume and analytic loads associated with the harmonizing of clinical trial data.

- Automatic application of mapping rules to each incoming clinical trial dataset to map specific fields into particular properties in the target model at scale. Developing a large set of sophisticated mapping rules and applying them to all incoming datasets automates harmonization and allows many datasets to be aligned quickly with minimal human intervention.
- Ability to apply transformations that conform fields that are subject to a controlled vocabulary or code list to eliminate non-standard values and improve quality.

Anzo's high-performance underlying graph database supports integration of data from thousands of clinical trials, interactively. This is essential to meeting the data volume and analytic loads associated with the harmonizing of clinical trial data.

The blended datasets in a graph-based data fabric deliver more insights than those constrained in traditional relational databases. Relational databases are developed with specific questions in mind. As the questions change, relational databases struggle to keep up. Often this means that new questions require complex JOINs over multiple tables, which are not performant, if they run at all. With Anzo, users create dashboards and conduct analytics across many related concepts in a highly flexible and scalable platform.

The result of these efforts is that new connections are established between data elements, and data sources are uniform and consistent. As this occurs, the data layers in the data fabric improve so that users can build their own blended datasets more easily and quickly.

The blending and modeling phases interact until you have a blended data product: integrated clinical trial data.

Access

The ultimate goal of integration of large-scale cross-clinical trial datasets is accessing and analyzing the data. Anzo includes exploratory dashboards as well as API endpoints for accessing the data from external tools.

The users driving the integration have incredible visibility throughout the process. These users access metadata and data within the underlying catalog that documents and manages the overall cross-clinical trial harmonization platform.

QC dashboards take advantage of the underlying graph data model to let the user navigate from data in the target model back to the raw data.

Key capabilities in the access stage include:

- The ability to see and explore all data and metadata in charts, graphs, and other visualizations. Anzo affords users powerful tools to understand and examine all the data in incoming clinical trials datasets, in the blended harmonized SDTM data model, or data at any point in the data transformation, blending, and harmonization process. Users can build dashboards that show:
  - Raw data and metadata from individual incoming clinical trial datasets
  - Profile data and sample data on individual clinical trial datasets
  - Controlled vocabularies, code lists, and synonyms
  - Blended data in the final combined cross-clinical trial target model
  - Metadata including transformations applied to each dataset throughout the data preparation and blending process
- Implementation of quality check (QC) rules and QC dashboards that allow QC and data domain experts to verify that transformations and mappings applied to each dataset are accurate and correct. This essential step ensures that the resulting cross-clinical trial dataset is trustworthy, transparent, and useful. QC dashboards take advantage of the underlying graph data model to let the user navigate from data in the target model back to the raw data.
- Creation of new blended analytic-ready datasets that can be published for use with leading analytic, visualization, or BI tools and used by applications or machine learning algorithms. These datasets can include data that was originally in potentially thousands of separate clinical trial datasets, but which has been blended and harmonized into a single SDTM model.

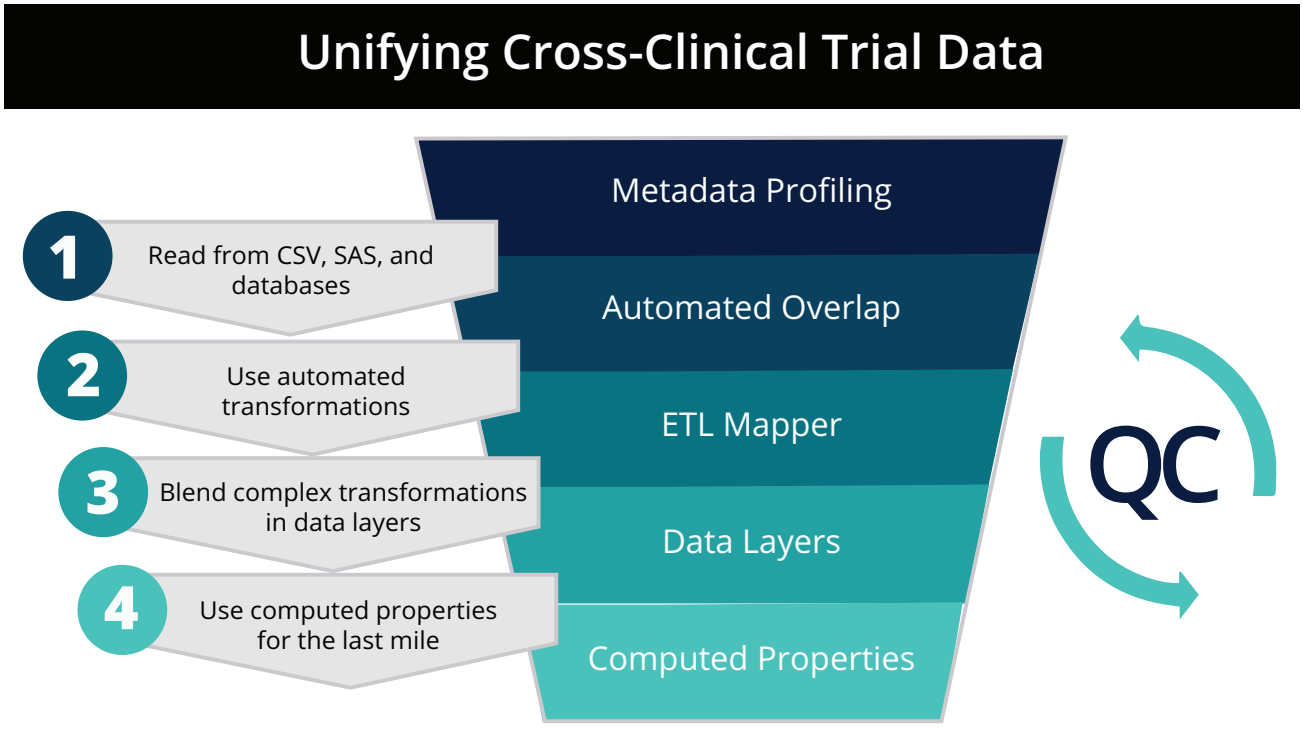
AUTOMATION, SCALE, AND QUALITY: PRACTICAL CONSIDERATIONS

With data from potentially thousands of past trials available for harmonization and a multitude of other data sources available for blending (e.g., clinical case, genomic, PubMed, or adverse event reports), automation is essential for enabling scale in the integration process. However, automation can't come at the cost of accuracy and quality.

Whatever approach is used must be flexible and intelligent so that SMEs can define and apply sophisticated logic to accurately detect and align related fields of data. Automated approaches must be transparent and auditable so that quality control experts can see exactly what transformations were applied to the data and verify that no inaccuracies inadvertently corrupted the datasets as they were brought together.

One way to bring automation and scale to the integration process while not sacrificing quality is outlined below and shown in the following diagram. In a typical complex integration, the workflow looks like this:

1. Start with metadata and data profile from CSV, SAS, and public databases.
2. Apply automated ETL-style mappings to find common fields and create simple transformations, covering a substantial portion of the overall transformations – up to 60 percent in some cases.
3. Use data layers to blend more complicated transformations.
4. The computed properties perform the last 5 to 10 percent of the transformations.



ANZO: PROVEN FOR HANDLING CROSS-CLINICAL TRIAL DATA UNIFICATION

Anzo, Cambridge Semantics' solution, is well positioned for the most complex use cases, and certainly cross-clinical trial data unification falls into that category. Anzo is used by life sciences companies to deliver analytics-ready metadata and data that solve high-value enterprise scale problems across the drug development and commercialization processes. To learn more about Anzo and the value it delivers for life sciences companies, visit **www.cambridgesemantics.com**.

This paper was written by Early Adopter Research and sponsored by Cambridge Semantics. Learn more about **Cambridge Semantics** or request a **demo**.

Connect with us

early adopter research