# Bringing Together Multi-Omics and Real-World Data to Accelerate Insight Delivery for Biomarker Discovery and Drug Development

[1]Rancho Biosciences, LLC, PO Box 7208, Rancho Santa Fe, CA 92067
[2]Sapient Bioanalytics, LLC, 10421 Wateridge Circle, San Diego, CA 92121

Marjolein Spronk[1], Tao Long[2], Marwan Elkholy[1], Alena Fedarovich[1], Aishwarya Pathak[1], Kabenla Armah[1], Eric Gonzalez[1], Yaw Nti-Addie[1], Jamie Cho[2], D'Angelo Aguirre[2], Adam Large[2], Felicia Sabatino[2], Nicholas Fiorentini[2], Mo Jain[2], Shahraz Niwaz[1], Tania Khasanova[1]

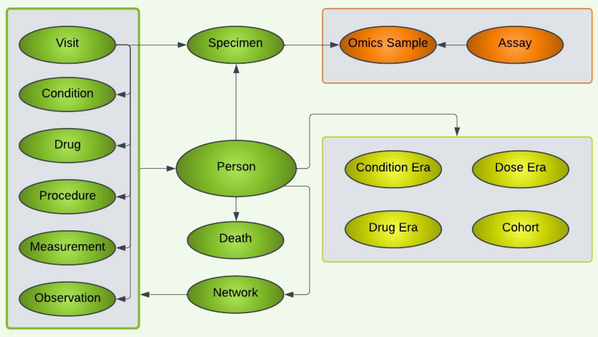
## Background

Despite the vast amount of life science data at our disposal, integrating and utilizing it remains challenging due to the diversity in data origins, formats and data types. Here we describe the collaboration between Rancho Biosciences and Sapient Bioanalytics to build a data lakehouse that encompasses Sapient's proprietary multi-omics and real-world data (**RWD**) and enables novel approaches to improving patient treatments and outcomes.

- A **data lakehouse** represents a hybrid data architecture that integrates the scalability and flexibility of data lakes – capable of storing large volumes of raw, unstructured data – with the robust performance and data management capabilities of traditional data warehouses, which are optimized for complex analytical queries.
- This architectural paradigm is particularly well-suited for the integration and analysis of **multi-omics** data, such **as proteomics, metabolomics, and genomics measures, and RWD**, which comprises electronic health records (EHR) as well as clinical outcomes, treatments, and lab measures.
- Bringing these complex and disparate datasets together allows for more comprehensive understanding of the biological underpinnings of disease and drug response, within the context of patients' real-world experience and outcomes. Leveraging the data lakehouse infrastructure, we are able to **significantly enhance the integration, visualization, analysis, and utilization** of Sapient's multi-dimensional biomarker and clinical datasets to inform treatment strategies.

## Common Data Model

Integrating RWD with multi-omics data can present significant challenges due to differences in data structure and contents, and the longitudinal nature of the data. To address this, we developed a data model that aligns with the Observational Medical Outcomes Partnership (**OMOP**) CDM, enabling integration and analysis of diverse biomedical datasets. The resulting data model enables large-scale, multimodal analyses to support biomarker discovery and drug development.




- **Conceptual model** – High level framework defining the key RWD and omics entities.
- **Logical model** – Structured schema specifying all entities, fields and their relationships in a database-ready format.

***Simplified, person-centered conceptual model aligned to OMOP to integrate RWD and omics data.***

## Data Harmonization and Curation

The unprocessed, unstructured clinical information coming from multiple sources and several databases including EHR and survey data was harmonized and aligned with seven CDM categories. Unique terms for 16 fields in person, condition, procedure, drugs, observation, and measurement were extracted, cleaned, and aligned with OMOP. Both manual and automated curation approaches were used. Rancho's TMS tool, applied to map 1,875 drug generic names to OMOP, reduced manual curation time by 50% with an output of 65% high quality matches (by similarity score). TMS mapping results were manually reviewed confirming the accuracy of the mapped ingredients to their corresponding generic drug names and 45.6k unique drug descriptions.

**Terminology Management Solution (TMS) Mapping**

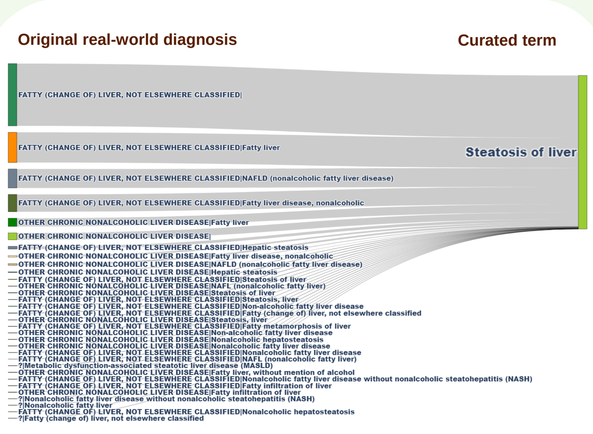
1.Extract unique terms from sources

| Raw drug description | Raw generic drug |
|---|---|
| TRILIPIX | FENOFIBRIC ACID (CHOLINE) |
| TRILIPIX DR 135 MG CAPSULE | FENOFIBRIC ACID (CHOLINE) |
| TRILIPIX 135MG | FENOFIBRIC ACID (CHOLINE) |
| TRILIPIX CAP 135MG | FENOFIBRIC ACID (CHOLINE) |
| FENOFIBRIC CAP 135MG DR | FENOFIBRIC ACID (CHOLINE) |
| FENOFIBRIC ACID DR 135 MG CAP | FENOFIBRIC ACID (CHOLINE) |
| FENOFIBRIC 135MG DR CAPSULES | FENOFIBRIC ACID (CHOLINE) |
| FENOFIBRIC ACID DR 45 MG CAP | FENOFIBRIC ACID (CHOLINE) |
| FENOFIBRIC ACID 135 MG CAPSULE DR | FENOFIBRIC ACID (CHOLINE) |
| FENOFIBRIC 135MG DR CAP | FENOFIBRIC ACID (CHOLINE) |
| ABELCET 5 MG/ML SUSP | AMPHOTERICIN B LIPID COMPLEX |
| ABELCET 100 MG/20 ML VIAL | AMPHOTERICIN B LIPID COMPLEX |
| CARBIDOPA-LEVODOPA-ENTA 200 MG | CARBIDOPA/LEVODOPA/ENTACAPONE |
| CARBIDOPA-LEVODOPA-ENTA 150 MG | CARBIDOPA/LEVODOPA/ENTACAPONE |
| BUTALBITAL/ASPIRIN/CAFFEINE CAPS | BUTALBITAL/ASPIRIN/CAFFEINE |
| BUTALBITAL/ASPIRIN/CAFFEINE 50-325-40 MG CAPS | BUTALBITAL/ASPIRIN/CAFFEINE |

2.Use TMS to map terms to OMOP

3.Manual review and QC, finalize

| TMS Mapped to OMOP | TMS Similarity Score | drug_generic | drug_generic_concept_id |
|---|---|---|---|
| choline fenofibrate | 1 | choline fenofibrate | 43560137 |
| amphotericin B | 1 | amphotericin B | 1717240 |
| carbidopa/entacapone/ levodopa Oral Gel | 0.75 | carbidopa;entacapone; levodopa | 740560;782211; 789578 |
| aspirin 325 MG / butalbital 50 MG / caffeine 40 MG Oral Capsule | 0.9 | aspirin;butalbital; caffeine | 1112807;1139699; 1134439 |

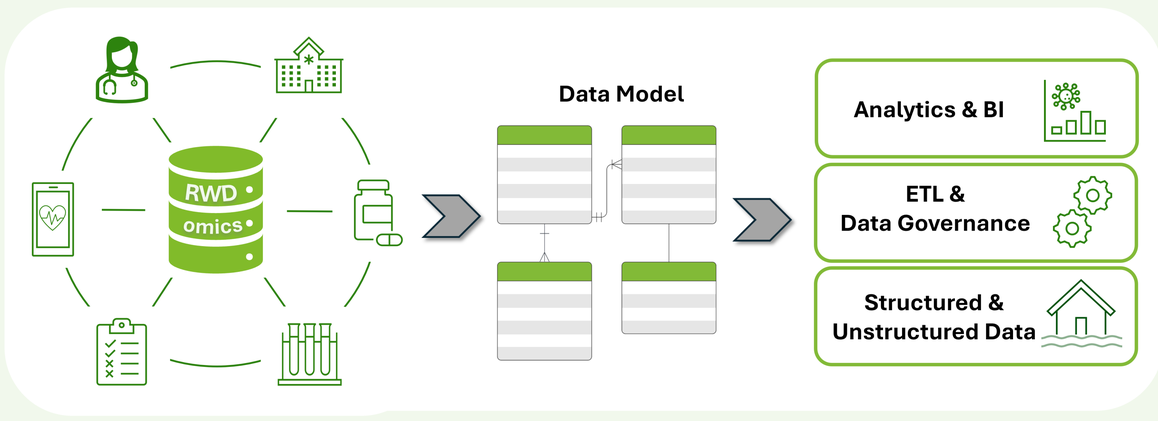



A significant reduction of redundancy and variability in terminology was achieved through semi-automated curation efforts, as can be seen in this example for diagnosis. The 32 original RWD terms could be harmonized to one standard term "Steatosis of liver" (SNOMED). Diagnoses were further grouped in classes (not shown here) to make cohort stratification more user-friendly.

## Data Lakehouse and ETL

Our approach to integrate RWD and omics datasets leverages the data lakehouse architecture, which allows us to ingest clean and harmonized data while also preserving the original metadata. A critical feature of the data lakehouse is its ability to incorporate longitudinal clinical data, facilitating dynamic and up-to-date analyses.

Data harmonization and curation are empowered by automated approaches, utilizing novel algorithms that have been incorporated into a workstream to efficiently combine various data types into a comprehensive dataset aligned to a Common Data Model (CDM). The robust ETL automates the data extraction and transformation and ensures data consistency and interoperability across datasets from different sources. This framework accelerates insight delivery by providing quick access to high quality integrated data, which are furthermore visualized via a BI dashboard.




***Framework illustrating how RWD and omics data are integrated into the data lakehouse while adhering to a data structure that is aligned with OMOP standards.***

Data from different sources is ingested at scale into the data lakehouse and transformed into a unified structure ready for analysis and visualization to generate new insights:

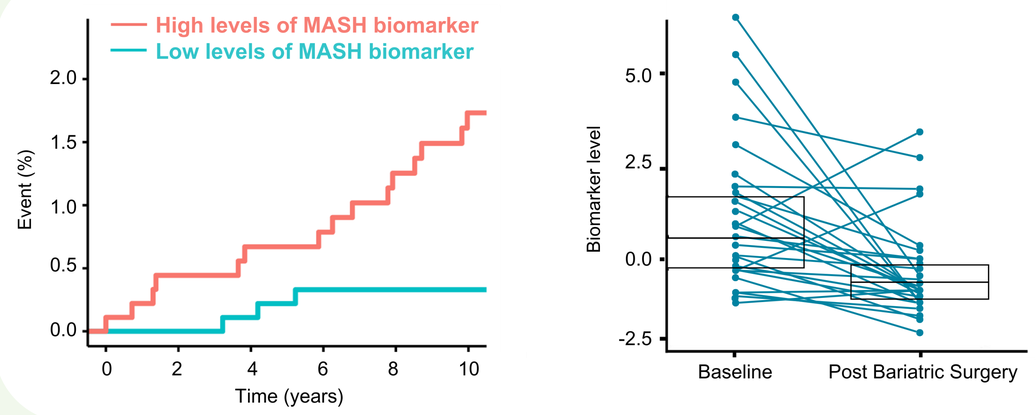
- Prepare schemas based on predefined data model
- Identify all source file locations
- Extract data from diverse sources (health visits, survey data, labs, etc.)

- Align fields to OMOP-compliant data model
- Data harmonization: apply mappings and data transformation rules
- Deduplication and standardization of data

- Write processed data into data lakehouse
- Optimize storage
- Analysis and BI-ready data

## Use Case: identifying an early diagnostic biomarker

Sapient leveraged the data lakehouse to rapidly analyze metabolomics data paired with RWD in >20,000 human samples across a diverse population to identify an early diagnostic biomarker which predicts metabolic dysfunction-associated steatohepatitis (MASH).




***The biomarker identified through nontargeted metabolomics analysis was found to be elevated in individuals >10 years prior to formal diagnosis of MASH.***

***The biomarker was found to be stable in healthy individuals over time, but changes dynamically in MASH patients at 2 weeks post bariatric surgery.***

Through this collaborative effort to make data more accessible and standardized for scientists, Rancho Biosciences and Sapient have demonstrated the data lakehouse's **transformative impact** in accelerating biomarker discovery and drug development.

Rancho BioSciences

SAPIENT
sapient.bio

Download Poster