Paper DH04

Framework for Integrating Omics Data with Other Clinical and Non-Clinical Data Sources

Mehdi Harek, Bristol Myers Squibb, Basel, Switzerland
Jonathon Keeney, The George Washington University, Washington, USA
Bron Kisler, Independent, Global and Genomics Standards Expert, Texas, USA
Adrian Czaban, Novo Nordisk, Copenhagen, Denmark
Ashwini Yermal Shanbhogue, Massachusetts, USA

ABSTRACT

The integration of omics data into clinical research is increasingly critical for advancing precision medicine, yet current industry standards, such as CDISC®, do not fully address the complexity of omics datasets. This presentation introduces a Data Integration Framework designed to merge omics data with clinical and non-clinical sources, offering a comprehensive view of patient health and treatment outcomes. Key challenges will be discussed, including the heterogeneity of omics data formats, ensuring data quality and reproducibility across platforms, managing large-scale data storage, and the complexity of downstream bioinformatics analysis. The proposed framework emphasizes standardization, interoperability, scalable data management, and reproducible analytics to support reliable integration and interpretation. Practical examples will illustrate how this approach enhances clinical insights and supports translational research efforts, ultimately paving the way for a more integrated and patient-centric data ecosystem in clinical development.

INTRODUCTION

This document addresses a critical challenge in the biomedical development: the integration of omics data with clinical data. While some solutions are already in place, the complexity and diversity of these data types continue to hinder seamless reconciliation.

The aim is to not only present existing approaches but also to foster cross-industry collaboration and research dialogue—bringing together stakeholders from pharmaceutical companies, clinical research organizations, academic institutions and genomics databases / biobanks to advance data harmonization and interoperability practices.
Ultimately, this document serves as a starting point for researchers, clinicians, and data scientists, who are considering and striving to unlock the full potential of multi-dimensional health data (where multi-dimensional data means multiple high dimensional data sets). By doing so, it supports the evolution of personalized medicine, also known as precision medicine, and contributes to improved patient outcomes.

The integration of clinical and biomarker data requires structured methodologies or frameworks capable of merging different types of human biology and DNA sequencing — such as genomics, proteomics, metabolomics, and transcriptomics—with clinical, and non-clinical data. This enables richer, data-driven insights into biomedical research, personalized medicine, and healthcare delivery.

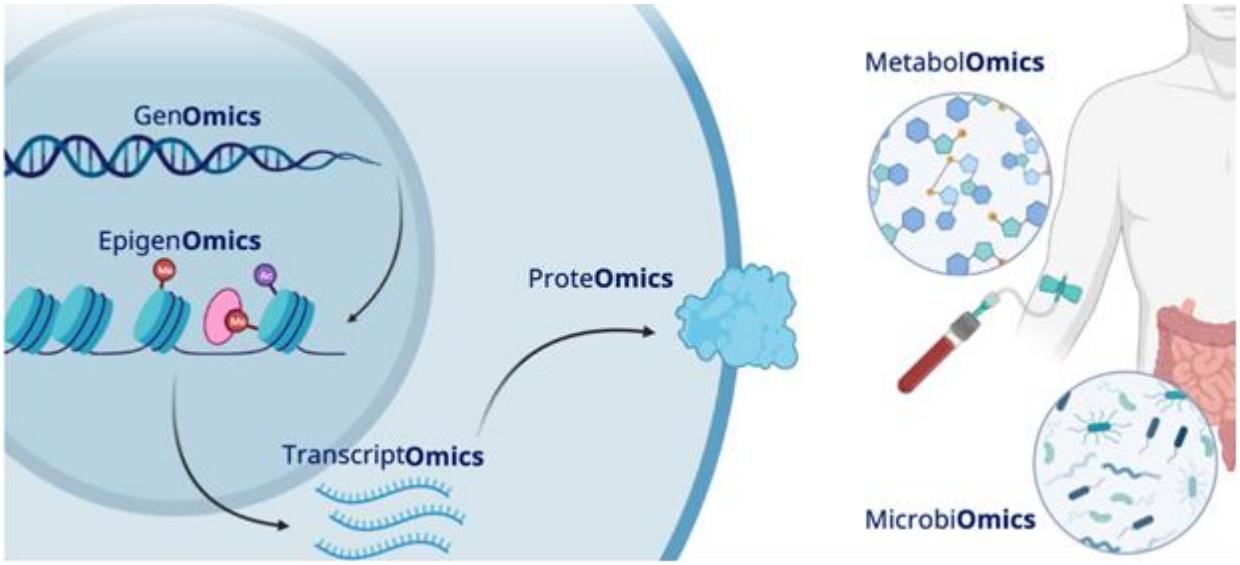

Figure 1: Examples of Omics data types

Although current clinical trial protocols may not heavily rely on biomarkers as primary or secondary endpoints, the industry is increasingly shifting towards more targeted and personalized approaches on a patient's biology and DNA. This trend underscores the growing importance of multi-omics data in clinical development.

As the volume and complexity of multi-modal healthcare (healthcare delivered through a plurality of modalities) data expands, there is a pressing need for standardized methodologies and interoperable solutions to support seamless integration.

Just as standards like CDISC®, HL7, ISO Genomics Informatics Sub-Committee (SC1) [Ref 1], SNOMED, etc exist for omics as well as clinical data, a robust framework should address the integration of:

- **Omics Data**: Genomics (DNA), Transcriptomics (RNA), Proteomics (proteins), Metabolomics (metabolites), Epigenomics (DNA modifications)
- **Clinical Trial and Research Data**
- **Clinical Data**
- **Non-Clinical Data**

Establishing such standards across the industry would:

- **Promote Interoperability**: Harmonize data standards, ontologies, and formats across omics, clinical (EHR, imaging, lab tests), and non-clinical sources (lifestyle, environmental, social determinants of health).
- **Facilitate Technical Integration**: Provide guidance on data warehousing, pipeline development, linking techniques, and the use of AI/ML for extracting insights.
- **Share Best Practices**: Offer real-world examples, tools, and resources from existing integration efforts to inform future projects and collaborations.

Ultimately, these efforts will accelerate the delivery of innovative therapies to patients by enabling more efficient and insightful research based in their pharmacogenomics.

Additionally, cross Standard Development Organisations (SDOs) coordination needs to be considered to establish interoperable "data bridges" from Genomics to Drug Discovery…to Clinical Research…to Regulated Clinical Trials…and ultimately to personalized Healthcare Delivery. As the *Figure 2* depicts, there are many moving pieces across the SDO landscape. ISO SC1, GA4GH and HL7 are already working to collaborate and align standards. For instance, genomics standards are being aligned with HL7 FHIR.

The Scope of this paper focuses more on aligning standards activities and building data bridges between Genomics (e.g., ISO SC1) and Clinical Trials Research (e.g., FDA, IEEE BioCompute Standard [Ref 2])

DEFINITIONS

Omics Data: High-dimensional biological datasets characterizing molecular components of organisms (collectively referred to as "omics"). This includes genomics (DNA sequence data), transcriptomics (RNA expression profiles), proteomics (protein expression profiles), metabolomics (metabolite profiles) and epigenomics (DNA Modification)

Clinical Trial Data: Clinical trial data are structured datasets collected under controlled conditions during clinical studies to evaluate the safety and efficacy of medical interventions, following regulatory standards such as CDISC® SDTM and ADaM. CDISC® standards are required by the **FDA** (U.S.) , **PMDA** (Japan) and are increasingly adopted or recommended by the **EMA** (Europe) and **NMPA** (China) to support regulatory submissions and ensure data consistency, interoperability, and review readiness.

Clinical Data: Health-related information collected during routine medical care, such as diagnoses, treatments, laboratory results, and imaging, typically recorded in electronic health records (EHRs).

Non-Clinical Data: For purpose of this paper, it refers to data *not* collected as part of conventional clinical trial CRFs or EHRs. This includes external and other contextual information. Examples include medical imaging data, laboratory or biomarker data obtained outside the trial protocol, etc.

Multi-dimensional Data: Two or more high dimensional datasets (for example, genomics and transcriptomics data). When analyses query multiple high dimensional datasets, complexity can quickly explode. Solutions are needed to manage complexity spread across multiple datasets.

Personalized medicine: This term is synonymous with "precision medicine." However, some clinicians and researchers feel the latter term has been oversold in recent years while little actual progress has been realized, we use "personalized medicine" in this document and follow the US FDA definition – "The right medication for the right patient at the right time" [Ref 3]. As truly interoperable large datasets come online and analyses become advanced enough to handle them, we feel it's important not to lose sight of this goal. The term is meant to describe a medical approach that tailors treatment and prevention strategies to individual patient characteristics, such as genetics, environment, and lifestyle, improving outcomes and minimizing unnecessary interventions.

Electronic Health Record Data:
EHR data are digital records of a patient's health information generated during clinical care. This includes demographics, medical history, diagnoses, medications, lab results, imaging, and clinical notes. Standards like **HL7 FHIR** (Fast Healthcare Interoperability Resources) define how EHR data are structured and exchanged to support interoperability across healthcare systems.

Regulated Clinical Research Data: Datasets generated during clinical trials that are submitted to health authorities for regulatory review. The BioCompute Standard (IEEE 2791-2020) provides a structured framework for documenting bioinformatics workflows used in generating such data, ensuring transparency, reproducibility, and regulatory compliance—especially for high-throughput sequencing and omics analyses.

PROBLEM STATEMENT

Each of the SDOs and standards mentioned above are understandably focused on their own stakeholders and funders. There are no "Data Bridges" to tie these critical areas of translational research together in order to achieve Personalized Medicine. There does not currently exist a project / program or funding mechanism for necessary cross-SDO coordination, alignment of standards, and ultimately interoperability between genomics, drug discovery, clinical research, and healthcare. At a recent GA4GH meeting, a keynote speaker referred to the gap between genomics information and other uses as the "valley of death". Everyone in the audience was in clear agreement. The Food & Drug Administration, the National Cancer Institute, and other NIH Institutes have definitively expressed interest in their area.

Furthermore, there exists no interoperability between clinical and research systems. EHR systems, genetic testing labs, clinical research file formats, public databases, and disease-based registries (e.g., cancer, rare diseases). All use different formats to represent genomics data. Each one is owned and governed independently. If this continues, the current interoperability breakdown will only get worse. In order to unlock the potential of these data for R&D purposes, we need to develop new ways of integrating and analysing data sets more efficiently.

BACKGROUND

Omics data, such as genomics, transcriptomics, proteomics, and others, are widely collected in national biobanks around the world and used in basic research settings. These data also have extraordinary potential for driving translational research, personalized medicine, and supporting clinical decision making, an application that is rapidly gaining. For example, the ARPA-H funded project called "FEAST," now in its second year, aims to integrate a broad variety of data types from several healthcare settings. Pharmaceutical companies can use this data as part of Regulated Clinical Trials, both to investigate efficacy of the drugs, as well as the Mode of Action (MoA) [Ref 4].

The potential benefits of a multi-omics approach from a variety of data sources are quite large, and range from rare disease detection, diagnosis, and treatment, as well as stronger insights into cancer [Ref 5].

Navigating the logistical and legal frameworks for these data is a major milestone. However, technical challenges to integrating data from a variety of omics data sources remain and are a barrier to widespread usage. Specifically, we define five categories of challenges:

1. the heterogeneity of omics data formats
2. ensuring data quality and reproducibility across platforms
3. collecting multi-omics information in clinical trials, managing large-scale data storage such as National Biobanks
4. the complexity of downstream bioinformatics analysis connected with large scale upstream databases and national biobanks
5. And secondary use of genomics / omics information to improve drug discovery, clinical research and patient care

1. The heterogeneity of omics data formats typically involves complicated, often bespoke data processing solutions, and can include correlation, covariance-based techniques, matrix decomposition, probabilistic, and/or Bayesian strategies [Ref 6] [Ref 7]. As many non-clinical resources work to unify and standardize their data to make them much more amenable to integration (for example, the "Workflow Playbook" [Ref 8] and UniProt API [Ref 9]), unifying datasets is becoming a more manageable task. However, regulatory data formats like CDISC® SDTM are not well designed to capture complex, multi-omics analyses such as these, which can create burdensome challenges to moving data from the research space to the applied space.

2. Data quality is a reflection of one's confidence that a particular piece of data represents reality. The way in which that confidence is represented can vary by data type. Further, data models for assessing quality are highly divergent by resource. There is no single, unifying confidence metric that can be applied across all data types.

Further, reproducibility remains challenging. Although many workflow solutions have been developed to handle execution tasks reliably [Ref 10] [Ref 11] [Ref 12], these do not communicate meaning, context, purpose, or other important information for decision making in the context of scientific review. The lack of supporting information has resulted in formal regulatory submission being rejected for lack of context, despite being reproducible. Further, none of the workflow engines are officially supported by the US FDA. Because of this, most reviewers do not invest the time in learning them.

3. Collecting multi-omics information in clinical trials and managing large-scale data storage, such as National Biobanks, introduces significant challenges compared to traditional clinical datasets. FDA submission requirements highlight this contrast: the Technical Conformance Guide [Ref 13] mandates splitting datasets larger than 5 gigabytes (GB) into smaller segments no greater than 5 GB, and the Electronic Submission Gateway limits uncompressed files to 100 GB [Ref 14]. Meanwhile, modern Next Generation Sequencing can generate raw data in the tens of terabytes. Even if pharmaceutical companies or CROs can store these volumes, processing them efficiently is often impractical. This drives the need to modernize IT infrastructure or implement new solutions capable of handling such data. These solutions must then undergo GxP validation and be integrated into clinical study workflows—a substantial effort for both IT systems and validated processes.

4. The complexity of downstream bioinformatics analysis connected with large-scale upstream databases and national biobanks means that more kinds of data can often lead to deeper insights into human health. However, it also requires far more complicated data processing. For example, the Velsera "Neoantigen workflow" includes over 100 steps in the data processing pipeline, drawing from multiple data types. Although incredibly powerful and capable of finding novel insights, pipelines this complex can be challenging to reproduce and especially difficult for scientific and regulatory reviewers to navigate.

5. Secondary Use of Genomics Data has been identified as a priority by many countries and government organizations through the International Standards Organization (ISO) Genomics Informatics Sub-Committee (SC1) which is formally comprised of 24 countries. Secondary use of genomics data refers to the use of genomics data generated for a single purpose, which is also needed for other purposes e.g., research, clinical decision making, population health management. This approach can lead to more efficient and effective use of data, which can ultimately improve patient outcomes. However, unique challenges will need to be addressed such as data standardization and interoperability; privacy and security; as well as informed consent.

EVOLVING DATA STANDARD LANDSCAPE

Within SC1 a landscape analysis and survey have been conducted, which will lead to publication of a Technical Report on the findings. This initial work is leading to new ISO standards projects being proposed from across Asia.

National Genomics Programs now exist all over the world. This has now led to significant investments in Precision Medicine Programs across Asia in particular. For instance, Singapore has made major investments and are moving into the third phase of collecting DNA sequencing from Singaporeans and aligning with the National Health Record System [Ref 15]. Although the investments are not as significant, Genome Canada has established a Precision Medicine Program [Ref 16]. To follow Middle Eastern countries will be making even greater investments.

Even with these advancements there exist data silos between genomics databases / biobanks, clinical research data, clinical trials data, and healthcare data. Multiple SDOs are addressing different critical areas to truly achieve personalized medicine; Genomics Standards à Clinical Research / Clinical Trials Standards à Healthcare Standards as well as various terminology standards and data models. This landscape continues to evolve. Several SDOs have been actively coordinating for a number of years, particularly ISO/TC 215 SC1, GA4GH and HL7 FHIR. Coordination with IEEE 2791-2020 (BioCompute) is now underway. But this will take more work and focused collaboration to establish the necessary "data bridges" to achieve interoperability across the life sciences and translational research landscape.

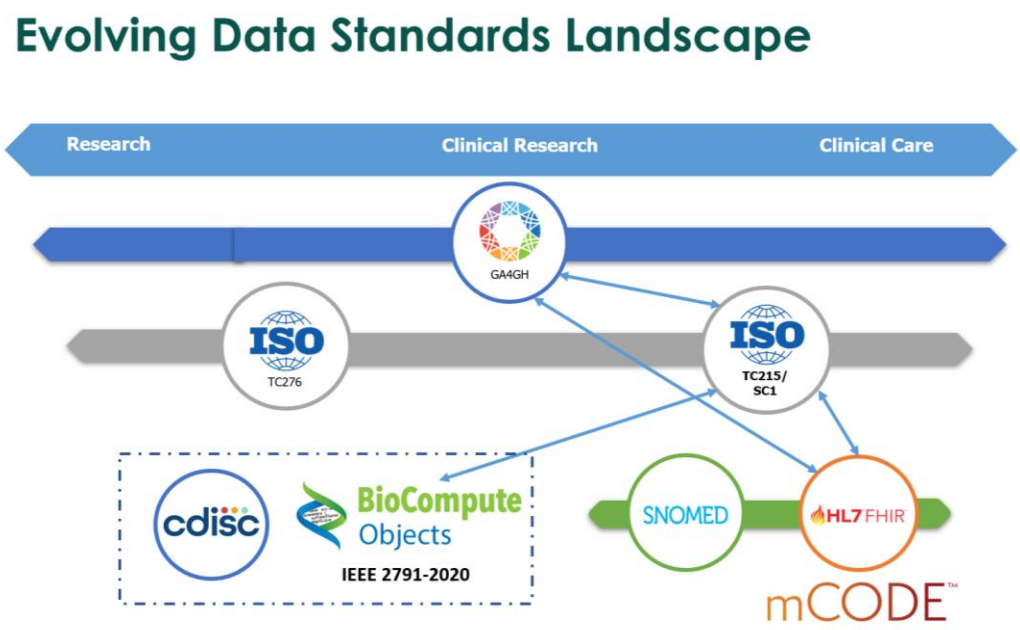

Figure 2: Evolving data standards Landscape

As existing standards are not well adapted to, one potential solution is "BioCompute." The IEEE 2791-2020 standard, which is colloquially known as "BioCompute," was designed for documenting workflows and carries many relevant benefits. This global standard is a JSON-based standard for describing workflows. It standardizes metadata for workflows and helps manage complexity. Because it is written in JSON, it can accommodate nested or hierarchical relationships very well. An instance of a workflow which conforms to the standard is called a BioCompute Object (BCO).

The initial conception of what would become BioCompute was proposed in 2012 at the United States National Center for Biotechnology Information (NCBI) as they were planning their data model for what would ultimately become the BioProject. Rather than simply recording an assertion in a database, it was proposed that a standardized way of documenting a workflow needed to be developed in order to record provenance for pieces of data. For instance, a paper trail where researchers can investigate and repeat, if needed.

The idea gained traction at the US Food and Drug Administration (FDA), where miscommunications around workflows in regulatory submissions from sponsors are common. The BioCompute project aimed to improve clarity between the FDA and regulated industry by establishing a set of guidelines for what information needed to be communicated, and how it should be communicated.

The project rapidly gained momentum from groups both inside and outside of the FDA. By the time the project led to a global standard in 2020, hundreds of individuals from dozens of organizations had already contributed. During this time, several use cases were proposed, including documentation for medical devices, genomics and other 'omics workflows, and standards harmonization, in addition to the original data provenance use case.

After official standardization through IEEE, documentation of workflows in genomics was the use case which gained the most traction. The FDA officially supported the standard across three centers [Ref 17]. BCOs have since started to be submitted for regulatory reviews for this purpose. A pilot project was recently launched in partnership with The George Washington University to map out the process and create guidance for usage of BCO. As of the date of this document, the results have not yet been published.

A BCO is written in JavaScript Object Notation (JSON), a flexible, hierarchical structure that is well adapted to representing complex data, like nested relationships or annotations, and which is both human and machine readable. The schema for BioCompute [Ref 18] breaks a workflow down into 8 "domains," 5 of which are required. These domains capture specific portions of workflows and can capture external data resources used in an analysis, including domain-specific identifiers, timestamps, and versions. A BCO does not include data within it but acts like a manifest of the data by using URIs, which are flexible identifiers.

RESULTS AND SOLUTIONS

BIOCOMPUTE

BioCompute has been used in a small pilot submission to the US FDA in partnership with pharmaceutical companies, and has started to be used in regulatory submissions. The standard was written to be relatively easy to implement, so as not to excessively increase the burden on regulated industry. Since standardization, other tools have been developed to work with the standard, including an R package [Ref 19], an R Shiny app [Ref 20], platform-specific tools like the BCOnexus for DNAnexus, the BioCompute tool on Velsera's platforms (Cavatica, Cancer Genomics Cloud, Seven Bridges, and BioData Catalyst), the Nextflow engine [Ref 21], and the BioCompute export on Galaxy, and even an LLM- based tool to help with generating the plain text descriptions from a workflow [Ref 22].

Figure 3 Overview of a BioCompute Object (BCO). BCOs are organized into 8 top level domains, 5 of which are required, 3 are optional. The 5 required domains are: Provenance (records information about the provenance of the workflow itself (who carried out the workflow itself (who carried out the work, what their role was, what the timestamp was)), Usability (high level description of the purpose and other contextual information), Description (details about each step, including its scientific rationale), IO (manifest of inputs and outputs used for the workflow), and Execution (information about the environment that the workflow was executed in). The three optional domains are: Parametric (parameters associated with each step, only required if the uses non-default values), Error (information related to error in the pipeline, such as limits of detection, margins, etc.), and Extension (user-defined mechanism for making the standard extensible).

A "Portal" was also developed as a repository to store BCOs [Ref 23]. The Portal uses "prefixes" in the unique identifiers of BCOs. Users are registered to one or more prefixes that allow them to read, write, publish, delete, and share a BCO, and administer other users within that prefix. The "BCO" prefix is open to the public, and examples can be seen there (e.g. https://biocomputeobject.org/viewer?https://biocomputeobject.org/BCO_000277/4.0; although these will typically be from academic papers and not regulatory submissions). This Portal is a clone of the same Portal that's running inside the FDA firewall, but which is not yet open to the public.

The harmonization of data standards is particularly relevant for the integration of omics data with clinical and non-clinical data sources. More recently, there has also been an additional use case as a security implementation. The Federated Ecosystem for Analytics and Standardization Technologies (FEAST) project is an ARPA-H funded project meant to bridge healthcare data from a plurality of healthcare organizations to drive insights into healthcare data. A major goal of the project is to standardize clinical data from all of the participating institutions, and another goal is to ensure the integrity of the computational analyses being run. Essentially, how can we guarantee that any bioinformatic tool that's being used to analyze sensitive clinical data hasn't been tampered with? BioCompute workflow hashes are used and distributed to all of the participating sites to ensure that vetted analyses are executing as expected.

BioCompute is a flexible, powerful way to document a workflow in a way that sets a common set of expectations through the official standard, and which can accommodate complex workflows like omics workflows.

MULTIASSAYEXPERIMENT (MAE)

In addition to the metadata captured in a BCO, other specific tools might be considered for the integration of omics data with clinical and nonclinical data. One of these potential solutions is the **MultiAssayExperiment (MAE)** R package. MAE utilizes Bioconductor [Ref 24] software and framework to represent, integrate, manage, analyze, and visualize multiomics data [Ref 25]. It provides a robust infrastructure for managing and harmonizing multiple assays performed on overlapping sets of biological specimens. This capability is essential for biomarker-driven research, where diverse omics data types must be combined with clinical and phenotypic information. MAE not only facilitates the integration of **multiomics** data with other **clinical** and **non-clinical** data, it also does so while overcoming the challenges of **heterogeneity**, managing **large-scale data storage**, the complexity of **downstream bioinformatics analysis**, and ensuring **reproducibility**.

MAE can ingest heterogeneous omics data like genomics, transcriptomics, and proteomics data for a set of specimens. The data to be ingested by MAE can be stored in memory, on disk, or remotely in the case of very large datasets, thus solving the problem of large-scale data storage, which is one of the most common problems associated with the management of data generated by omics assays.

MAE accomplishes the task of representing, managing, and integrating multiomics data by introducing a Bioconductor object-oriented S4 data class, which has three key components:

- **colData**, a data frame containing patient, cell-line, or other biological unit related data (i.e., clinical or non-clinical data) with rows for each unit and columns for each characteristic variable (histological, pathological, demographic, etc.)
- **ExperimentList,** a list of assays, where each list element or assay is a dataset, with rows for variables like genes or genomic ranges and columns for assay results. The elements of a list do not have to have the same number of rows and columns, allowing for replicates, time series data, or missing assays
- **sampleMap**, a data frame containing a column for assay names, a column for identifiers of biological units, and a column for identifiers of assay results. sampleMap relates each column in ExperimentList to exactly one row in colData; however, one row of colData may map to one column per assay or, in cases of missing assays or replicates, to zero or multiple columns per assay. The sampleMap thus elegantly represents the integration and alignment of each biological unit with the assays related to it.

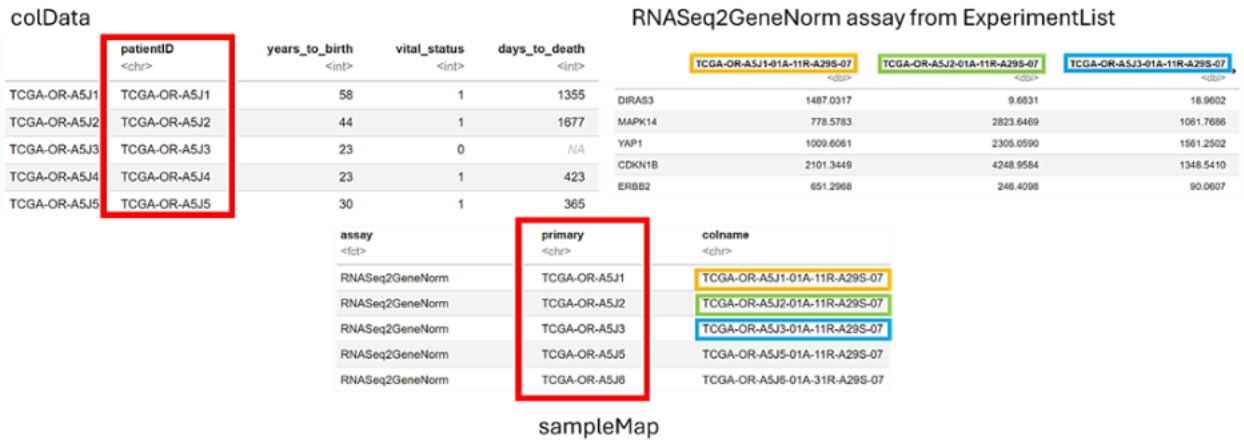

Figure 4: Screenshots illustrating colData, RNASeq2GeneNorm dataset from the ExperimentList, and sampleMap from miniACC, an MAE object provided along with the MAE package for demonstration purposes (code provided in the appendix).

The **ExperimentList** alone or in combination with the **colData** and **sampleMap** components of the data class, along with the optional inclusion of metadata, can be used to create a **MultiAssayExperiment** object that stores omics data along with clinical or non-clinical data. The **MultiAssayExperiment** object can then be subset, reshaped, or transformed into forms that are compatible with downstream analysis and visualization.

Finally, the R script used to design the process above can be shared and replicated by another researcher, thus overcoming the challenge of reproducibility.

Key Advantages

- **Standardisation**: MAE enforces a consistent structure for multi-omics datasets, reducing complexity and improving reproducibility.
- **Interoperability**: Integrates seamlessly with other Bioconductor packages for downstream analysis, including survival modelling and machine learning.
- **Scalability**: Handles large datasets across multiple platforms, making it suitable for clinical development pipelines.

CONCLUSION

In high dimensional omics data, a data point may be related to many other data elements, some of which may be conditional. Flat, tabular models are not ideal for capturing and representing data in this way. JavaScript Object Notation (JSON) is a better option, because it can represent concepts in nested, hierarchical relationships.
BioCompute, a colloquial name for the IEEE-2791 standard, is a JSON-based standard for describing workflows. BioCompute excels at recording data provenance and data harmonization. A BioCompute Object (BCO – an instance of a workflow that conforms to the standard) can capture external data resources used in an analysis, including domain-specific identifiers, timestamps, and versions. It can also represent a mapping of concepts between domains for data harmonization.
Because of these strengths and because BioCompute is used for managing complexity by representing relationships in JSON, a recommended framework for integrating omics data with other clinical and non-clinical data sources is a combination of MAE with BioCompute.
In this scenario, MAE is used for formal analysis, and a BCO is used to document the workflow, data provenance, and harmonization of data sources.
Depending on the use case, it might be beneficial to explore existing CDISC® SDTM.GF (Genomic Findings) examples [Ref 26]. While the nature of tabular format limits the amount of information that can be stored there, the structured and well recognised architecture might prove beneficial when communicating the results to the Regulators. A pragmatic solution would be to use the above-described solutions to get from raw data to intermediate results and then use CDISC® as the last step for generating final analysis and/or tables, listings figures.
Another high-level recommendation is to reach out to the relevant Regulatory Authority early in the study lifecycle and establish a dialogue. A high-level overview of the tools and workflows used should be in place for such a discussion, and the specifics can be a part of the agenda. This reduces the uncertainty on the Sponsors side and gives the Agency an insight into the challenges faces by the industry and possible way of overcoming them.
SAS® has been the dominant programming language for a vast majority of companies for Clinical Data Analysis. With the emergence of new types of data, such as Omics into the Clinical space, many new Open-Source Software Languages are being adopted as part of a wider industry trend. This brings a lot of promise in terms of innovation, but it also comes with its own challenges. Those include:

- Validation of external packages
- Maintenance of programming environments
- IT security risk assessments

A centralised strategy needs to be established on how to deal with Open-Source Software in a regulated environment. A good reference for this kind of work can be found here [Ref 27].

APPENDIX

```r
# Load {MultiAssayExperiment} and miniACC, a MultiAssayExperiment object provided with the package for
demo purposes.

library(MultiAssayExperiment)
data(miniACC)

# Extract colData from miniACC, coerce it to a dataframe and display as paged table.

col <- colData(miniACC)
col_data <-  as.data.frame(col)
rmarkdown::paged_table(col_data)

# Extract one of the assays (RNASeq2GeneNorm) from miniACC, coerce it to a dataframe and display as paged
table.

assay <- as.data.frame(assay(miniACC))
rmarkdown::paged_table(assay)

# Extract sampleMap from miniACC, coerce it to a dataframe and display as paged table.

samp_map <- as.data.frame(sampleMap(miniACC))
rmarkdown::paged_table(samp_map)
```

Appendix 1: R code used to generate the tables used in Figure 3.

REFERENCES

[1] https://www.iso.org/committee/7546903.html
[2] https://standards.ieee.org/ieee/2791/7337/
[3] https://www.fda.gov/medical-devices/in-vitro-diagnostics/precision-medicine
[4] https://www.cdisc.org/sites/default/files/2025-05/6B%20CDISC%202025-6B-Fitting%20multi-omics%20into%20SDTM.pdf
[5] https://zenodo.org/me/uploads?q=&f=shared_with_me%3Afalse&l=list&p=1&s=10&sort=newest
[6] https://arxiv.org/pdf/2501.17729
[7] https://pmc.ncbi.nlm.nih.gov/articles/PMC7003173/
[8] https://doi.org/10.1371/journal.pcbi.1012901
[9] https://doi.org/10.1093/nar/gkaf394
[10] http://www.nature.com/nbt/journal/v35/n4/full/nbt.3820.html
[11] https://www.commonwl.org/
[12] https://doi.org/10.7490/f1000research.1114634.1
[13] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document
[14] https://www.fda.gov/industry/create-esg-account/frequently-asked-questions
[15] https://www.npm.sg
[16] https://genomecanada.ca/challenge-areas/canadian-precision-health-initiative/
[17] https://www.federalregister.gov/documents/2020/07/22/2020-15771/electronic-submissions-data-standards-support-for-the-international-institute-of-electrical-and
[18] https://opensource.ieee.org/2791-object/ieee-2791-schema/
[19] https://github.com/NovoNordisk-OpenSource/whirl/blob/main/R/biocompute.R
[20] https://cran.r-project.org/web/packages/biocompute/index.html
[21] https://github.com/nextflow-io/nf-prov
[22] https://biocompute-objects.github.io/bco-rag/
[23] https://biocomputeobject.org/
[24] https://doi.org/10.1038/nmeth.3252
[25] https://doi.org/10.1158/0008-5472.can-17-0344
[26] https://biocomputeobject.org/
[27] https://pharmar.org/white-paper/

ACKNOWLEDGMENTS

We would like to express our gratitude to **Jeffrey Long** and **Sireesha Krosuri** for their contributions throughout the development of this work. We extend our appreciation to **Rashad Rasul** for initiating the connection that enabled this project to begin and for supporting its early development.

This paper was developed as part of the PHUSE Working Group on "*Integration of Omics Data into Clinical Drug Development*", whose collaborative efforts have been instrumental in shaping the framework presented here.

We also acknowledge the PHUSE EU Connect Stream Leader from the Data Handling & Data Engineering stream, for their leadership and support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Author Name: **Mehdi Harek**
Company: Bristol Myers Squibb
Email: mehdi.harek@bms.com - harek.mehdi@gmail.com

Author Name: **Jonathon Keeney**
Company: Georges Washington University
Email: keeneyjg@gwu.edu

Author Name: **Bron Kisler**
Company: Independent, Global and Genomics Standards Expert
Email: bronkisler@icloud.com

Author Name: **Adrian Czaban**
Company: Novo Nordisk
Email: adcz@novonordisk.com

Author Name: **Ashwini Yermal Shanbhogue**
Email: ash23shan@yahoo.com