Paper TT02

Working with Omics data – A Statistical Programmers perspective

Adrian Czaban, Novo Nordisk, Copenhagen, Denmark

ABSTRACT

This paper provides a comprehensive overview of experiences in the integration and utilization of omics data in a Clinical Development setting. Omics technologies offer a holistic view of the molecular mechanisms within biological systems by analyzing large datasets. The application of omics data can lead to the identification of critical biomarkers and therapeutic targets, but despite its promise, the mainstream adoption of omics data for primary use in clinical trials faces challenges due to its complexity, the need for specialized knowledge, and infrastructural demands. This paper discusses the current landscape of CDISC and BioCompute standards for omics data handling and stresses the necessity of robust IT infrastructure and scientific expertise to leverage omics technologies effectively in clinical research and drug development. It also tries to highlight that the work with Omics data needs to be handled in a systematic way, considering the regulatory space as well as a GCP compliant process.

INTRODUCTION

Omics is a broad term used to describe the comprehensive study of a set of related molecules within a biological sample. The suffix "omics" is derived from the Greek word "ome," which means "all" or "complete." Omics technologies aim to collectively characterize and quantify pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. These technologies are often high throughput, allowing for the analysis of large datasets. This is quite different from what a traditional data a Statistical Programmer is used to. Instead on looking at an individual protein measurement from a blood sample, we're looking at up to tens of thousands of proteins instead.

There are many types of Omics data, for example: genomics (study of the complete sets of genomes), transcriptomics (study of complete set of RNA transcripts produced by the genome under specific circumstances or in a specific cell), proteomics (set of proteins expressed by a genome, cell, tissue or organism), metabolomics (complete set of metabolites within a biological sample), epigenomics (study of a complete set of epigenetic modifications on the genetic material of a cell) and others (1).

All these data types are linked together by the central dogma of biology, which enables the multi-omics analyses. This refers to the integrative analysis of multiple omics datasets to gain a comprehensive understanding of biological systems. By combining data from different omics layers, researchers can achieve a more holistic view of the molecular mechanisms and interactions within cells, tissues, or organisms. The sum of the combined results will give more meaningful and nuanced biologically and clinically relevant information than individual parts.

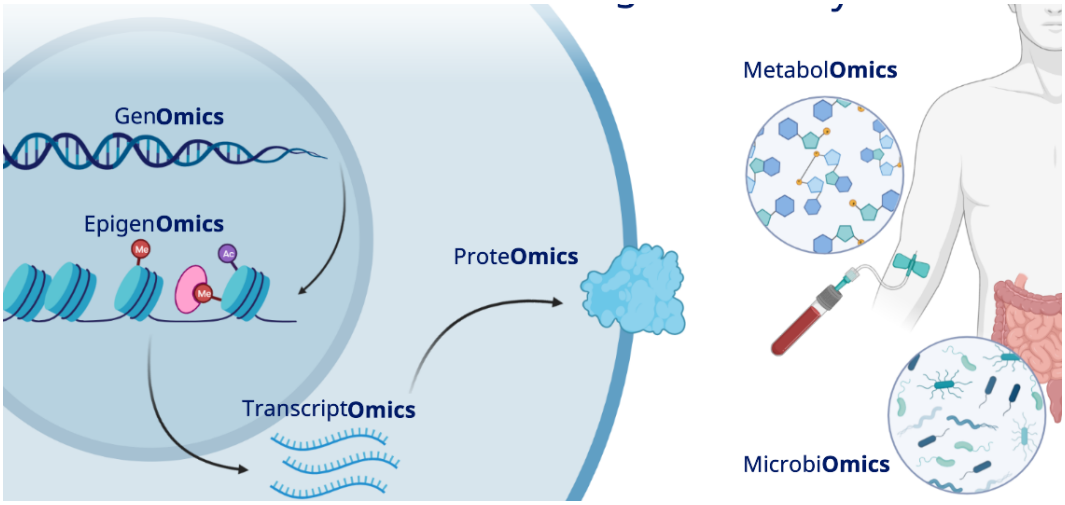

Figure 1. Examples of Omics data types

EXAMPLES OF USE IN THE PHARMACEUTICAL INDUSTRY

Use of Omics data enables multiple opportunities in the pharmaceutical space. One of them is the concept of precision medicine. It is an innovative approach to tailoring medical treatment to the individual characteristics of each patient. It considers factors such as genetics, environment, and lifestyle to design more effective and targeted therapies. The goal of precision medicine is to improve the accuracy of diagnosis, predict disease risk, and optimize treatment strategies for better patient outcomes.

Another high potential field that can be enabled using omics data is targeted drug discovery. It is a modern approach to developing new medications that focuses on identifying and targeting specific molecules or pathways involved in disease processes. This method contrasts with traditional drug discovery, which often relied on broad-spectrum agents that affected multiple pathways and could lead to numerous side effects.

Some notable examples of omics usage in the pharmaceutical industry include:

- identification of the BRCA1 and BRCA2 genes, which are linked to breast cancer, has led to the development of PARP inhibitors like Olaparib (Lynparza) for treating patients with BRCA mutations (2)
- RNA sequencing has been used to identify overexpressed genes in cancer, leading to the development of targeted therapies like HER2 inhibitors for HER2-positive breast cancer (3)
- the use of next-generation sequencing (NGS) to profile tumors in the NCI-MATCH trial has enabled the matching of patients to targeted therapies based on their specific genetic alterations (4)
- identification of prostate-specific antigen (PSA) as a biomarker for prostate cancer has led to the development of PSA tests for early detection (5)
- FDA recommends genetic testing for the CYP2C9 and VKORC1 genes before prescribing Warfarin to determine the appropriate dosage and reduce the risk of adverse effects (6)
- Connectivity Map (CMap) project uses gene expression profiles to predict the potential toxic effects of new compounds by comparing them to known toxicants (7)

Additionally, we can certainly see growing interest in this data from the regulators, especially the US FDA. They have a dedicated Omics Working Group and have very recently hosted the next iteration of the FDA Omics Days Symposium discussing the future of precision medicine (8).

WHY IS THE USE OF OMICS DATA IS NOT MAINSTREAM (YET)?

As presented above, clearly there are numerous advantages to using Omics data during the drug discovery and clinical development. However, it certainly is a niche outside of the treatment area of Oncology, and a vast majority of Statistical Programmers/Statisticians/Clinical Data Scientists have not been exposed to it. There are multiple reasons for this.

To start off, this is still a developing field, and we have only begun scratching the surface of the possibilities that it enables. The technologies used for generation of Omics data are quickly evolving, but in some cases still need to mature. Secondly, working with this data is presenting many challenges due to its complexity – both of in terms of scientific understanding as well as purely technical aspects of its handling. Data coming from Omics analyses can be characterized as overwhelming. This goes both for the actual disc size (where the input files for Whole Genome Sequencing of multiple subjects can reach hundreds of terabytes) and the number of discrete data points that all represent some biological information. To correctly interpret this information, one needs to understand what it means and how it's internally interconnected.

The complexity is highlighted when we realize that each Omics data type has different common standard formats and tools used for analysis. To give some examples, when working with genomics we will encounter the following data formats used at different stages of the analysis pipeline: *FASTA, FASTQ, SAM/BAM, VCF/BCF, BED, CRAM* or *BEDGRAPH*. For proteomics, we will encounter *ADAT, mzML, mzIdentML, mzTab, APL, pepXML or proBed*. While some of them are quite simple, they still need to be learned and well understood.

We at Novo Nordisk have recently started using various types of Omics data in our Clinical Studies. This includes both the analysis of samples for future research as well as protocol endpoints. As we all know from experience, data and results are not worth much unless they can be used in a relevant context and as part of a larger system. The road from data to evidence also requires the implementation of proper standards, validating the software we're using, making sure our IT infrastructure is tailored to the task, ensuring that we follow the regulatory requirements and using our existing scientific knowledge to ensure a proper analysis. All this need to be done in a Good Clinical Practice (GCP) compliant way if it is to be used as part of drug development pipeline and in a regulatory context. *Figure 2* represents a simplified way of what is needed to go from 'data' to 'evidence'.

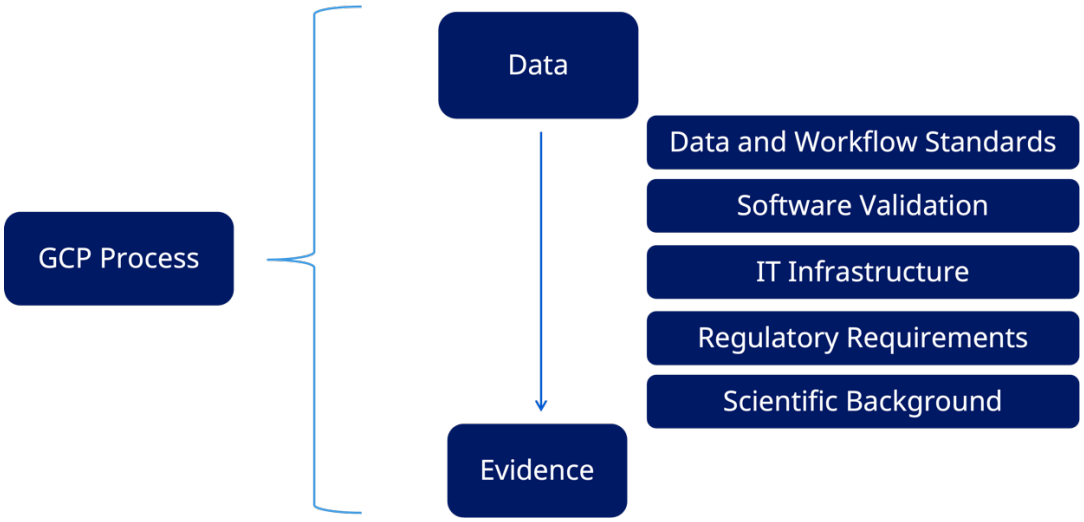

Figure 2. From 'Data' to 'Evidence'

Once we began working with Omics data, it quickly became evident that many aspects in these categories were unclear, lacking, or required more work and investigation compared to traditional data types. The familiar system was no longer present, necessitating an evaluation of where existing components could be adapted and where new ones needed to be implemented.

DATA AND WORKFLOW STANDARDS - CDISC

CDISC standards form the foundation of how we work with clinical study data. As conformance to them is required by regulatory agencies, and they cover a vast majority of data types and modalities, it was the first place we turned to in search of standards to be used.

SDTM IG version 3.4 (9) contains certain domains which are relevant when working with some of the omics data types, and it has made some big changes in comparison to previous versions. Namely, it has deprecated SDTMIG-PGx v1.0 in the following way:

- Provisional PF domain superseded by the GF domain
- Biospecimens domains BE and BS published
- Provisional PG, PB, and SB domains deprecated with re-instantiation considered if valid use cases are found

The new domains proved very helpful in out work and will be described shortly.

BIOSAMPLE METADATA: SDTM.BS AND SDTM.BE

Before any Omics data types are generated, a relevant biosample must be collected for example a biopsy, or a blood sample. This is kept in storage and sent for analysis afterwards. The description of the sample and how it was handled is our biosample metadata. The BS (Biospecimen Findings) and BE (Biospecimen Events) are two domains that are crucial in the generation of biosample metadata. Information present in that data is crucial when making decisions about analysis or performing initial QC.

SDTM.BE is a new domain that can be used to store biomarker data in SDTM format. It is an events domain that documents actions taken that affect or may affect a specimen (e.g., specimen collection, freezing and thawing, aliquoting, transportation). BE domain can also be linked to other SDTM domains, such as DM, LB, or SE, to provide additional information about the biomarker data. Example of data in the BE domain is presented below.

Example 1: Blood Aliquoting Event

STUDYID	DOMAIN	USUBJID	BESEQ	BEDECOD	BETERM	BEREFID	BEPARTY	BESTDTC	BEENDTC
123	BE	01-001	1	COLLECTING	Collecting	32167	SITE	2024-07-20T12:49	

Breakdown of the Table Fields:

- **STUDYID**: Unique identifier for the study.
- **DOMAIN**: Domain abbreviation (BE for Biosample Events).
- **USUBJID**: Unique subject identifier.
- **BESEQ**: Sequence number for the event, which helps in ordering events if multiple occur.
- **BEDECOD:** Dictionary-derived text description of BETERM or BEMODIFY, if applicable
- **BETERM:** Topic variable for an event observation, which is the verbatim or pre-specified name of the event
- **BEREFID:** Internal or external identifier for the specimen created or affected by the event.

- **BEPARTY:** Party accountable for the transferable object (e.g. specimen) as a result of an activity performed in the associated BETERM variable.
- **BESTDTC**: Start date/time of the event
- **BEENDTC:** End date/time of the event

SDTM.BS is a new domain that can be used to store biospecimen data in SDTM format. The BS domain captures the data related to biospecimen characteristics. The BS domain includes variables such as the biospecimen identifier, type, quality metrics, amount or size. Like BE, the BS domain can also be linked to other SDTM domains, such as DM, EX, LB, or GF, to provide additional information about the biospecimen data. The BS domain supports the traceability and quality control of the biospecimens.
An example of BS domain is presented below.

Example 2: RNA Quality Control in SDTM.BS Domain

STUDYID	DOMAIN	USUBJID	BSSEQ	BSTESTCD	BSTEST	BSORRES	BSCAT	BSSPEC	BSBLFL
123	BS	01-001	1	RIN	RNA INTEGRITY NUMBER	8.5	QUALITY CONTROL	rRNA	Y

Breakdown of the Table Fields:

- **STUDYID**: Unique identifier for the study.
- **DOMAIN**: Domain abbreviation (BS for Biospecimen).
- **USUBJID**: Unique subject identifier.
- **BSSEQ**: Sequence number for the biospecimen, which helps in ordering multiple samples if collected.
- **BSTESTCD**: Test code
- **BSTEST**: Descriptive name of the test or biospecimen attribute, specified here as "RNA Integrity."
- **BSORRES**: Original Result, which in this case is the RNA integrity number (RIN)
- **BSCAT**: Category of the topic-variable values
- **BSSPEC**: Specification of the biospecimen, noted here as "rRNA."
- **BSBLFL**: Indicator used to identify a baseline value

In our company we have decided to generate the biosample metadata in two ways. The choice is depending on whether the Omics data is to be generated and analyzed as part of the protocol, or it is generated from biosamples for future research.

- If we are collecting biosamples for future research, the BE domain will as a minimum contain information about Collecting and Receiving events. If it's relevant for the study, information regarding other events such as storing and destroying samples should be requested. BS domain is considered not to be relevant in this scenario and will not be created.
- If omics data is collected in the study, a larger number of information is required. The BE domain will contain information regarding collecting and receiving events, and input should be provided early regarding any other events needed for Biostatistics to do the analysis and submission. The BS domain will be created, and will contain other information regarding the biomarkers, such as RNA integrity or purity. It is also important for Biostatistics to be able to group results in other domains, such as LB or MI by the originally collected biospecimen, include ACCESSIONID in the relevant data specification, e.g. LAB or BIOPSY. The content of the ACCESSIONID data file column should be mapped to the LNKGRP variable in SDTM.

SDTM.GF

This is a special-purpose domain that can be used to store genomics data in SDTM format. I can be used to capture different aspects of the genomics data, such as the reference sequence, the subject sequence, the genomic function, and the genomic results. The GF domain allows for the representation of various types of genomics data, such as single nucleotide polymorphisms (SNPs), insertions/deletions (indels), copy number variations (CNVs), gene expression, and methylation. The GF domain can also be linked to other SDTM domains, such as DM, LB, FA, or CO, to provide additional information about the genomics data. However, the GF domain has some limitations, such as the lack of standard terminologies, the complexity of the data structure, and the large size of the data files.

Example 3: SNP Variation Data in SDTM.GF Domain

STUDYID	DOMAIN	USUBJID	GFSEQ	GFTESTCD	GFTEST	GFORRES	GFSTRESC	GFORREF	GFGENREF	GFGENLOC
123	GF	01-001	1	SNV	Single Nucleotide Variation	A/G	A/G	G/G	GRCh38.p14	4:52918327

Breakdown of the Table Fields:

- **STUDYID**: Unique identifier for the study.
- **DOMAIN**: Domain abbreviation (GF for Genomic Findings).
- **USUBJID**: Unique subject identifier.
- **GFSEQ**: Sequence number for the genomic finding, which helps in ordering multiple findings if collected.
- **GFTESTCD**: Test code
- **GFTEST**: Descriptive name of the test or genomic finding
- **GFORRES**: Original Result
- **GFSTRESC**: Finding in standard format
- **GFORREF:** Reference finding for the result
- **GFGENREF:** Genome Reference used to generate the result. In this case *GRCh38.p14* indicates Genome Reference Consortium Human Build 38 patch 14
- **GFGENLOC:** Genetic location within a sequence for the observed value in GFORRES

In our opinion, the Genomic Findings domain is helpful for very specific types of data, that one would normally generate as a result of a bioinformatic pipeline. It is unfortunately not suitable for raw or pre-processed genomic data, for several reasons. One of them is the size of such datasets, which can easily achieve tens of terabytes. Expanding it from a flat text file into a tabular BDS structure is not feasible. On top of this, a great deal of standards have been established with regards to data formats and analysis procedures, and recreating those in a CDISC format would be a tremendous task. Thus, in our opinion, the GF domain can potentially be used as one of the last steps of analysis and result sharing, but other datasets and standards need to be used. Additionally, there is currently no SDTM domain that would be specific for other types of Omics data, such as transcriptomics, metabolomics or proteomics. In some cases, one can try to fit them into existing domains, such as LB. However, the amount of information contained in input datasets would make it prohibitive for entire datasets and force the use of pre-specified subsets.

DATA AND WORKFLOW STANDARDS - BIOCOMPUTE

BioCompute (10) is the informal name for a framework designed to standardize the documentation and communication of bioinformatics workflows and computational analyses, particularly those used in regulatory submissions. The BioCompute specification emerged from discussions among hundreds of stakeholders, from the U.S. Food and Drug Administration (FDA), private industry, and academia. These discussions highlighted the difficulties in reviewing and validating bioinformatics workflows submitted as part of regulatory applications, particularly for personalized medicine and genomics. Its development was a collaborative effort led by the George Washington University (GWU) in partnership and funding from the FDA. In 2020, the specification was formally standardized and published as IEEE 2791-2020. Standardization through IEEE, a major standards development organization, marked a significant milestone in formalizing the structure and use of BioCompute Objects (BCOs). The full name of the standard is “IEEE 2791-2020: Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication,” though “BioCompute” is still used informally. However, it can also be used to support other bioinformatic analyses, such as proteomics and metabolomics, as well as for other novel data types, such as imaging and machine learning pipelines, as evidenced by examples presented during the recent Biocompute FDA Conference 2024, recorded and publicly available (11).

As a descriptive standard, the goal is to ensure transparency, reproducibility, and interoperability of computational analyses in biomedical research and drug development. Because it has been developed in response to the difficulties in documenting and communicating bioinformatics pipelines, in many aspects it complements the areas that are currently impossible to cover using CDISC standards.

A BioCompute Object (BCO) is a human- and machine-readable record of a bioinformatics pipeline. As such, it contains information about what data was used, which software was executed, and what parameters have been set for the analysis. It can be compared to a chemical experiment, where the input, output, and parameters are being recorded to document and communicate the process. It does not force the use of any specific software, execution engine, or workflow language. Unlike CDISC standards, it does not require the usage of any specific data formats. It also does not mandate the usage of any internal codelists. Instead, it provides tools to describe the computational workflow and acts as an interface for existing standards. *Figure 3* visualizes that concept.

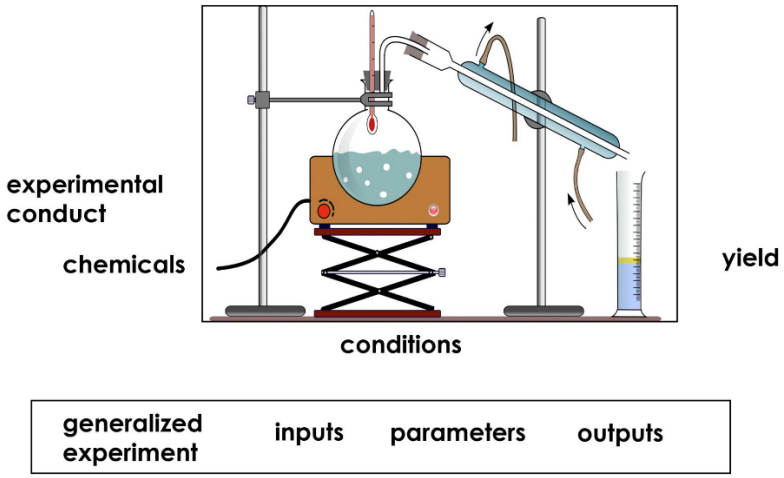

Figure 3 Analogy of a chemical experiment to a bioinformatic pipeline (12)

BioCompute is written in Javascript Object Notation (JSON), which is simply a set of key:value pairs. A BCO encapsulates all the information needed to describe a bioinformatics workflow in a single file. It is characterized by eight domains, 5 of which are mandatory and 3 are optional. They are presented in detail with examples below:

8 Top Level Domains	
Provenance Domain: Metadata describing the BCO	
Usability Domain: Free text field for researcher to explain the analysis and relevant details.	
Extension Domain: User-defined fields	
Description Domain: Steps of the analysis, external resources needed for the steps, and the relationship of I/O objects	
Execution Domain: Information about the environment in which the analysis was run	
Parametric Domain: Records any parameters that were changed from default values	
Input and Output Domain: A list of global input and output files	
Error Domain: Used for describing errors. Can include the limits of detectability, false positives, false negatives, statistical confidence of outcomes, and description of errors	
Required	**Optional**

"created": "2020-08-01T16:29:55-04:00",
"modified": "2020-08-01T16:29:55-04:00",
"contributors": [{
"name": "Eric Donaldson",

[
"Identify baseline single nucleotide polymorphisms (SNPs)[SO:0000694], (insertions)[SO:0000667], and (deletions)[SO:0000045] that correlate with reduced antiviral drug efficacy in (Hepatitis C virus subtype 1)[taxonomy:31646]. Identify treatment emergent amino acid (substitutions) [SO:1000002] that correlate with…

"pipeline_steps": [{
"step_number": 1,
"name": "HIVE-Hexagon",
"version": "2.4.3",

"input_subdomain": [{
"uri": {
"filename": "Hepatitis C virus genotype 1",
"uri":
"https://hive2.biochemistry.gwu.edu/dna.cgi?cmd=objFile&ids=16130"

{
"script_driver": "HIVE",
"software_prerequisites": [
{

{
"param": "_type",
"value": "svc-align-hexagon",
"step": "1"

"algorithmic_error": {
"AthenaFRSCORE_threshold": 0.5,
"AthenaQUALITY": 25,
"AthenaCOVERAGE": 5000

{
"extension_schema":
"https://www.w3id.org/biocompute/extension_domain/1.1.0/dataset/dataset_extension.json",
"dataset_extension": {"additional_license": {

Provenance
Usability
Description
IO
Execution
Parametric
Error
Extension

BioCompute is currently a Supported Standard according to the FDA Standards Catalogue version 10.4 (13). As of that version, it is not a required standard for submissions, and thus there is no there is no 'Requirement Begins' date.

BioCompute can currently be generated in a few ways:

- You can use the standalone "builder" tool available at the BioCompute Portal: [BioCompute Portal (biocomputeobject.org)](biocomputeobject.org).
- When using a platform already supporting BioCompute, you can use tools built into that platform such as DNAnexus/precisionFDA, Galaxy, or Seven Bridges/Cancer Genomics Cloud.
- You can also build an output into your workflow as a JSON file conforming to the standard.

The exact interplay between CDISC and BioCompute is currently not officially established and is something that will certainly be evolving in the future.

The BioCompute Portal also provides many useful tools and documents, such as:

- BCO User Guide
- BCO Documentation
- Frequently Asked Questions
- BCO Curation SOP
- BioCompute database search, providing examples of published BCOs
- Links for cloud-based tools for BioCompute
- Materials from previous workshops

Internally at our company we have started a formal process aimed at integrating the BioCompute standard into the existing procedures. It involves several Subject Matter Experts from potentially impacted skill areas all assessing if and how this new standard should be implemented, with the result being an update of existing ways of working. This ensures streamlined adoption and good communication.

BIOCOMPUTE PILOT

As of the date of publishing of this article, there is a BioCompute pilot ongoing between George Washington University BioCompute team, FDA, and three industry representatives. Its aim is to bring both the sponsors and FDA into agreement around BCO usage (both conceptually and logistically) to streamline and standardize computational workflow submissions and reviews, as well as increase the awareness and adoption of BCO in the industry. As part of the pilot, the sponsors will make a test submission to the FDA without providing any additional context beyond the submission contents. The submission package will contain the input datasets and the BCO. The FDA will review the BCO and share any comments.

The expectations of the outcome include:

- Documentation of any difficulties or barriers when working with BCO
- Update of the 'Best Practices' and 'Security Plan' documents of the BCO
- Finding any potential gaps and streamlining of the submission process
- Exchange of knowledge and experiences.

Public information about the pilot can be found on its official webpage: BCO Pilot Project - BCOeditor Wiki (biocomputeobject.org)

SOFTWARE VALIDATION

The validation of software used in clinical trials is crucial to ensure the accuracy, reliability, and integrity of the data collected and analyzed. It is a broad concept, which has been extensively covered in other materials, and which I do not want to repeat in this paper. However, we've experienced that the validation of software used for bioinformatic analyses is different from the traditional work being done with Clinical Study Data. There are two main reasons for this:

- The software being used is new to the industry
- Bioinformatic pipelines tend to be very standardized and maintained externally

We are currently experiencing a revolution when it comes to Open-Source programming languages, with R being used quite extensively and Python being adopted in an increasing degree. The R White paper on a "Risk-based Approach for Assessing R package Accuracy within a Validated Infrastructure" (https://www.pharmar.org/white-paper/) provides many good recommendations, which we have implemented internally. If the programmers doing the analysis are not familiar with these approaches, this will present additional an additional layer of challenge to them.

Additionally, bioinformatics pipelines use workflow management systems, which can be used for a part or entirety of an analysis. Those workflow management systems, like Nextflow, Snakemake, Cromwell, Galaxy or CWL (Common Workflow Language), provide robust frameworks for defining, executing, and managing complex workflows. They are usually well maintained and people working on them are experts in the industry. Using workflows is also a good way of ensuring reproducibility and documentation an analysis. However, while this provides a high degree of confidence, it does not mean that the Clinical Data Scientist should be using them blindly. An assessment must be made whether a given workflow is the best choice for the data used in the clinical trial. Any changes or small adjustments to an existing workflow might be difficult to implement. In the end, we are responsible for correct analysis, and those tools cannot be treated like end-to-end solutions without consideration.

IT INFRASTRUCTURE

Working with Omics data presents additional challenges with a high impact on the IT infrastructure used. Examples include:

- Volume of Data
- High Computational Demand
- Scalability
- Large Data Transfers
- Data Integration
- Software Compatibility
- Tool Integration

It is highly likely that the current infrastructure being used does not live up to those challenges. Thus, it might be necessary to develop new IT solutions and utilize high-performance computing (HPC) clusters, cloud-based computing services (e.g., AWS, Google Cloud, Azure), or GPU-accelerated computing to meet the computational demands. This is not a task for a Clinical Data Scientist, but it's a problem that has be signaled in the organization at an early stage, to make sure the right resources are allocated, and a long-term strategy can be made.

REGULATORY ASPECTS

Apart from abiding to ICH guidelines, there are a few specific regulations which should be followed when planning to attempt a submission with Omics data. 'Pharmacogenomic Data Submissions' 2023 draft guidance (14) is the latest document from FDA, which aims to clarify the contexts in which pharmacogenomic study findings and data must be included in submissions related to investigational new drug applications (INDs), new drug applications (NDAs), and biologics license applications (BLAs) based on the FDA's regulations. Additionally, it provides recommendations to sponsors and applicants on the format and content of pharmacogenomic data submissions.

Other agencies such as EMA or PMDA have published some high-level guidance documents as well. Since Omics is a novel and fast evolving field, it is very likely that new guidance and documents will be released in the future. One should follow the current developments and pay attention to any new requirements being published. Additionally, when planning a submission with Omics data, relevant regulatory bodies should be contacted at an early stage and presented with the data and proposed analyses. This will ensure both adherence to current requirement, as well as enable a dialogue between the sponsor and the regulators.

SCIENTIFIC BACKGROUND

Working with any type of Omics data requires acquiring new knowledge, both in terms of understanding biological processes and using the correct software solutions for analysis and reporting. While it is possible to learn all those things, the complexity and diversity of biological concepts mean it takes years to fully comprehend them. Ensuring the right competencies are present in the Study Team requires both time and effort. In our experience, the optimal solution for ensuring the correct competencies lies in both training of existing employees as well as hiring/transfer of people with the right knowledge from outside the existing organization. The reasons for this are:

- Attempting to train existing personnel in complex bioinformatics concepts and workflows will take a lot of time and effort. The skills and experience required to make decisions and plan analyses is something that takes years to develop in the research space and it cannot be done on the side of an already running Clinical Study.
- Relying exclusively on new hires with the right bioinformatics background creates a reverse problem: they will be lacking in terms of existing industry standards for Clinical Study conduct and reporting. Additionally, it takes some time to gain clarity on the stakeholder landscape and how to best fit into a currently existing process.

In our experience using a mix of experienced clinical programmers and bioinformaticians allows them to cover each other's weaknesses and allows for the best results. This will likely result with some differences in ways of working for both groups, and an emphasis should be made on good communication and collaboration.

Additionally, working with Omics data brought us closer to the Research part of the organization. This is a new type of a stakeholder, and it required an establishment of solutions as how and when can data be shared, which information can be used and which cannot, as well as the proper channels of communication. This was something that ultimately had to be agreed upon at a higher level of the organization.

GCP PROCESS

All our work needs to be done in a highly compliant GCP environment if it is to be used to create regulatory evidence. If it's not, it might as well have not been done in the first place. The changes introduced to an existing process might prove to have a big impact on a previously existing environment, as new resources, procedures, solutions and IT systems need to be implemented. GCP is a multi-layered concept, involving Ethical Conduct, Scientific Integrity, Compliance with Regulations, Qualified Personnel, Informed Consent, Safety Monitoring and Data Integrity and Management. There is no bullet-point list to tick off to make a process GCP-compliant, and relevant GCP advisors and QA representatives need to be involved at early stages. It took us several months to build solutions which are

GCP-compliant, and I recommend looking into this issue at a very early stage, to make sure that all the work that is being done can be considered reliable and valid in a regulatory setting.

CONCLUSION

Use of Omics data in Clinical Trials brings a lot of promise, but also many challenges to be tackled along the way. The first step to overcoming those challenges is being aware of their existence. We have tried highlighting the current landscape and the steps we needed to take to succeed, and hopefully it gives some information to others who will be facing similar challenges. It is quite clear that a lot of work needs to be done to streamline the process, and we can expect this field to evolve quite rapidly in the coming years.

REFERENCES

1:Krassowski M, Das V, Sahu SK and Misra BB (2020) State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front. Genet. 11:610798.doi: 10.3389/fgene.2020.610798

2: Ragupathi A, Singh M, Perez AM, Zhang D. Targeting the *BRCA1/2* deficient cancer with PARP inhibitors: Clinical outcomes and mechanistic insights. Front Cell Dev Biol. 2023 Mar 22;11:1133472. doi: 10.3389/fcell.2023.1133472. PMID: 37035242; PMCID: PMC10073599.

3:Swain SM, Shastry M, Hamilton E. Targeting HER2-positive breast cancer: advances and future directions. Nat Rev Drug Discov. 2023 Feb;22(2):101-126. doi: 10.1038/s41573-022-00579-0. Epub 2022 Nov 7. PMID: 36344672; PMCID: PMC9640784.

4: Murciano-Goroff YR, Drilon A, Stadler ZK. The NCI-MATCH: A National, Collaborative Precision Oncology Trial for Diverse Tumor Histologies. Cancer Cell. 2021 Jan 11;39(1):22-24. doi: 10.1016/j.ccell.2020.12.021. PMID: 33434511; PMCID: PMC10640715.

6: David MK, Leslie SW. Prostate Specific Antigen. [Updated 2022 Nov 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557495/

6: WARFARIN SODIUM- warfarin tablet [package insert]. Bridgewater, NJ; March 31, 2017. Available from: https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=541c9a70-adaf-4ef3-94ba-ad4e70dfa057.

7: Shah I, Bundy J, Chambers B, Everett LJ, Haggard D, Harrill J, Judson RS, Nyffeler J, Patlewicz G. Navigating Transcriptomic Connectivity Mapping Workflows to Link Chemicals with Bioactivities. Chem Res Toxicol. 2022 Nov 21;35(11):1929-1949. doi: 10.1021/acs.chemrestox.2c00245. Epub 2022 Oct 27. PMID: 36301716; PMCID: PMC10483698.

8: https://www.fda.gov/science-research/scientific-meetings-conferences-and-workshops/fda-omics-days-2024-precision-practice-regulatory-science-best-practices-and-future-directions-omics

9: SDTM:SDTMIG v3.4 | CDISC

10: Simonyan, V., Goecks, J., & Mazumder, R. (2017). Biocompute Objects — A Step towards Evaluation and Validation of Biomedical Scientific Computations. PDA Journal of Pharmaceutical Science and Technology, 71(2), 136–146. doi: 10.5731/pdajpst.2016.006734

11: Biocompute FDA Conference 2024 (youtube.com).

12: Simonyan V, Goecks J, Mazumder R. Biocompute Objects-A Step towards Evaluation and Validation of Biomedical Scientific Computations. PDA J Pharm Sci Technol. 2017 Mar-Apr;71(2):136-146. doi: 10.5731/pdajpst.2016.006734. Epub 2016 Dec 14. PMID: 27974626; PMCID: PMC5510742.

13: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog

14: Pharmacogenomic Data Submissions | FDA

ACKNOWLEDGMENTS

I would like to acknowledge the BioCompute team at George Washington University for their help in preparation of the paper, sharing materials for conference use, as well as their collaboration on the BioCompute pilot. Many thanks to Raja Mazumder, Vahan Simonyan, Hadley King, Tianyi Wang, John Torcivia and Lam Phuc.

RECOMMENDED READING

Comparison of pharmacogenomic information for drug approvals provided by the national regulatory agencies in Korea, Europe, Japan, and the United States - PMC (nih.gov)

Hasanzad M, Sarhangi N, Ehsani Chimeh S, Ayati N, Afzali M, Khatami F, Nikfar S, Aghaei Meybodi HR. Precision medicine journey through omics approach. J Diabetes Metab Disord. 2021 Nov 24;21(1):881-888. doi: 10.1007/s40200-021-00913-0. PMID: 35673436; PMCID: PMC9167178.

CONTACT INFORMATION

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

Author Name: Adrian Czaban
Company: Novo Nordisk

Address: Vandtårnsvej 110, 2880 Søborg
Work Phone: +45 3075 7547
Email: adcz@novonordisk.com

Brand and product names are trademarks of their respective companies.