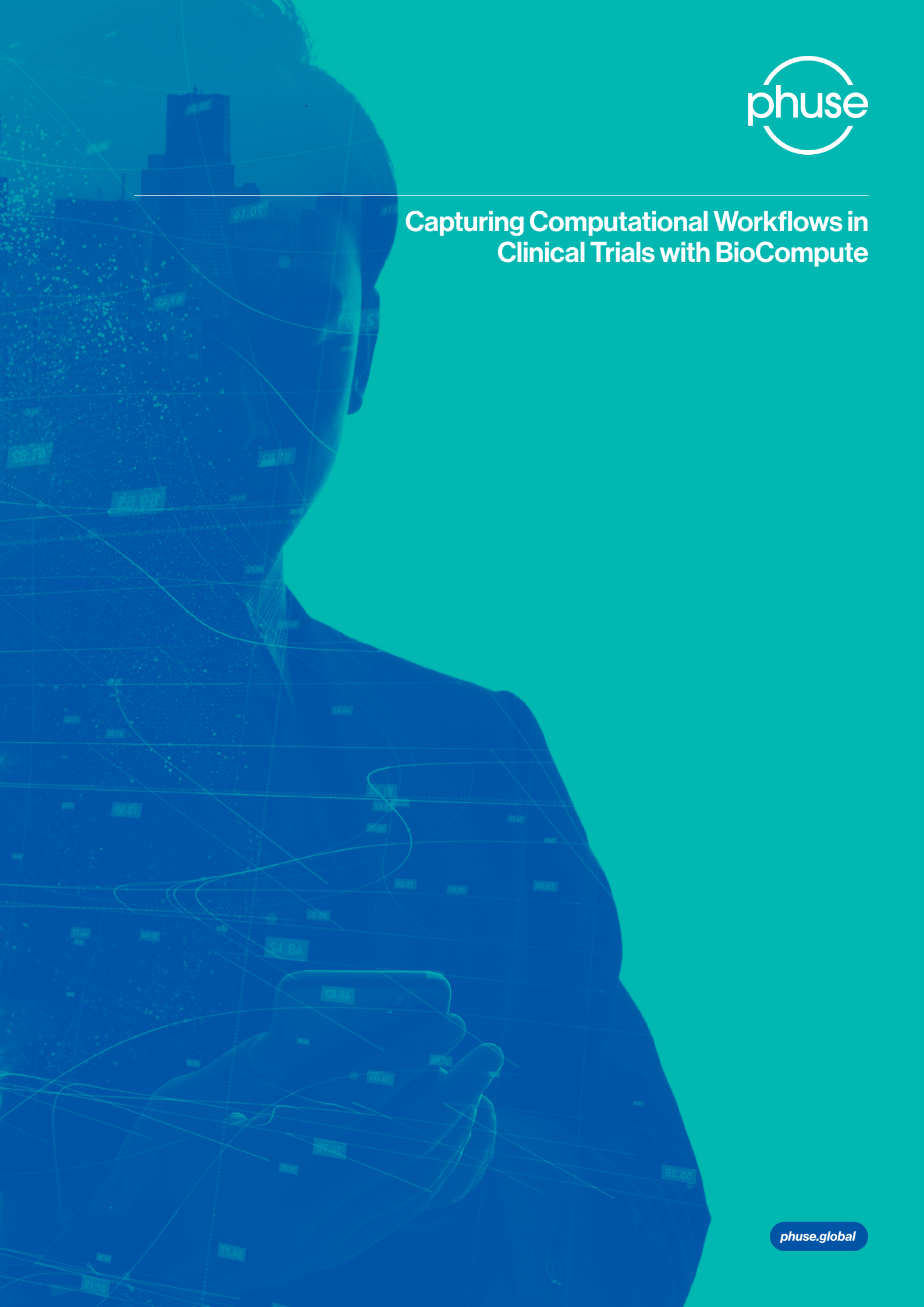

Capturing Computational Workflows in Clinical Trials with BioCompute

phuse.global

Contents

1: Overview: Purpose of this Document 1

2: Scope 1

3: Definitions 1

4: Problem Statement 2

5: Background 2
Introduction of the BioCompute Framework 2
Structure of a BioCompute Object 2
BioCompute Object Database (BCODB) 3
Tools and Platform Integrations 3
Pilot Project and FDA Adoption 3

6: Case Studies 4
Multi-Omics Cancer Prognosis Model 4
Clinical Trials with Transcriptomics and Proteomics Data 4
Clear Cell Renal Cell Carcinoma Glycoproteomic Risk Prediction 4
Clinical Trials with Imaging Protocol Endpoints 5

7: Recommendation 5

8: Disclaimer 5

9: References 6

10: Project Contact Information 6

Revision History

Version	Date	Summary
1.0	2025-DEC-17	Original Version

1: Overview: Purpose of this Document

This white paper explores the emerging role of the BioCompute standard (IEEE 2791-2020, [1]) in clinical trials. It presents practical applications through real-world case studies, with the aim of enhancing understanding among data scientists, regulatory agency personnel, and other professionals involved in designing, executing and reviewing computational workflows in clinical research.

The document showcases how BioCompute supports complex data types – such as omics and imaging – across clinical and computational contexts and how it integrates into diverse computational platforms. It illustrates how BioCompute enhances reproducibility, transparency and collaboration by enabling structured, auditable documentation of computational methods.

The aim is to foster continued engagement with the BioCompute framework and support its broader adoption and evolution within clinical research.

2: Scope

This document examines applying the BioCompute framework to clinical research, with a focus on:

- **Clinical trial integration:** For improving transparency and repeatability in data analysis workflows
- **Regulatory submissions:** For documenting computational procedures, to improve quality and efficiency in submissions to regulatory authorities
- **Complex data processing:** For documenting analysis pipelines for omics and imaging data
- **Tools and platforms:** For highlighting which platforms and open-source packages facilitate BioCompute implementation.

The target audience includes clinical data scientists, regulatory agency personnel, and other people with an interest in standardising workflow communication.

3: Definitions

Table 1 Definitions

Term	Definition
API	Application programming interface
BCO	BioCompute object
BCODB	BioCompute Object Database
BRCA	Breast cancer gene
CBER	Center for Biologics Evaluation and Research
ccRCC	Clear cell renal cell carcinoma
CDISC	Clinical Data Interchange Standards Consortium
CDER	Center for Drug Evaluation and Research
Cox PH	Cox proportional hazards
CWL	Common workflow language
eCTD	Electronic common technical document
FDA	Food and Drug Administration
GEO	Gene Expression Omnibus
GxP	Good practice within a particular field, such as GCP (good clinical practice)
HCC	Hepatocellular carcinoma
HFP	Human Foods Program
HIVE	High-performance Integrated Virtual Environment
JSON	JavaScript Object Notation
LASSO	Least Absolute Shrinkage and Selection Operator
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
Omics	The collective characterisation and quantification of entire sets of biological molecules, such as genomics, proteomics, metabolomics and transcriptomics
ROI	Region of interest
SDTM	Study Data Tabulation Model
TCGA	The Cancer Genome Atlas
TMT	Tandem mass tag
URI	Uniform resource identifier

4: Problem Statement

Modern bioinformatics workflows are increasingly complex, involving diverse data types, evolving tools, and multi-step analyses. Yet, documentation practices remain inconsistent, making it difficult to reproduce results, assess data provenance or meet regulatory expectations. The existing CDISC standards for structuring clinical trial data do not provide mechanisms for documenting computational workflows associated with complex biomedical data types.

Consistent workflow communication requires a shared understanding and approach between industry and regulatory authorities. BioCompute provides a structured framework well suited for this purpose, enabling reproducible documentation of computational analyses.

5: Background

Introduction of the BioCompute Framework

The increasing complexity of bioinformatics pipelines has made it difficult to consistently document and reproduce analyses. Without a standardised approach, workflow descriptions are often fragmented, ad hoc, and insufficient for review or replication.

The BioCompute standard for workflow documentation addresses these challenges. Starting from an idea conceived in 2012, the standard has been developed as a collaboration between George Washington University and the US Food and Drug Administration (FDA), with input from contributors from public and private sectors. In 2020, BioCompute was formally recognised as an IEEE standard (IEEE 2791-2020) [2–4].

BioCompute provides a formalised approach for capturing the full context of a computational workflow, including metadata, inputs, parameters, execution environment, and outputs. Rather than replacing standards, BioCompute complements them by acting as a communication layer that improves transparency and reproducibility. Detailed BioCompute guidance and resources are available via the BioCompute Portal [5].

Structure of a BioCompute Object

A BioCompute Object (BCO), an instance of the BioCompute standard, is a machine-readable record of workflow documentation in JSON format [6]. The JSON structure is organised into sections known as domains, which serve specific purposes in documenting the workflow.

The domains include:

- **Provenance Domain:** BCO metadata, including version, status and contributors
- **Usability Domain:** Free-text field for describing the purpose and context of the analysis
- **Description Domain:** Analysis step outline with step inputs and outputs
- **Parametric Domain (optional):** List of parameters for customising the computational flow
- **Input and Output Domain:** List of global input and output files
- **Execution Domain:** Computational environment description
- **Extension Domain (optional):** User-defined information that is not defined in the standard BioCompute structure
- **Error Domain (optional):** Definition of workflow-specific empirical and algorithmic error tolerances.

A BCO example is shown in **Figure 1.**

Figure 1 *BCO example with the required domains, plus the optional Error Domain in collapsed view*

BioCompute Object Database (BCODB)

To support the storage, management and sharing of BCOs, a centralised BioCompute Object Database (BCODB) has been established [7]. This secure, access-controlled system allows organisations to manage BCOs under assigned, organisation-specific prefixes, with configurable user permissions such as reading, writing, publishing and sharing.

Resources within the BCODB are available through well-documented application programming interface (API) endpoints to facilitate automation and integration with external applications [8]. This includes endpoints for creating, retrieving, updating and validating BCOs. See **Figure 2.**

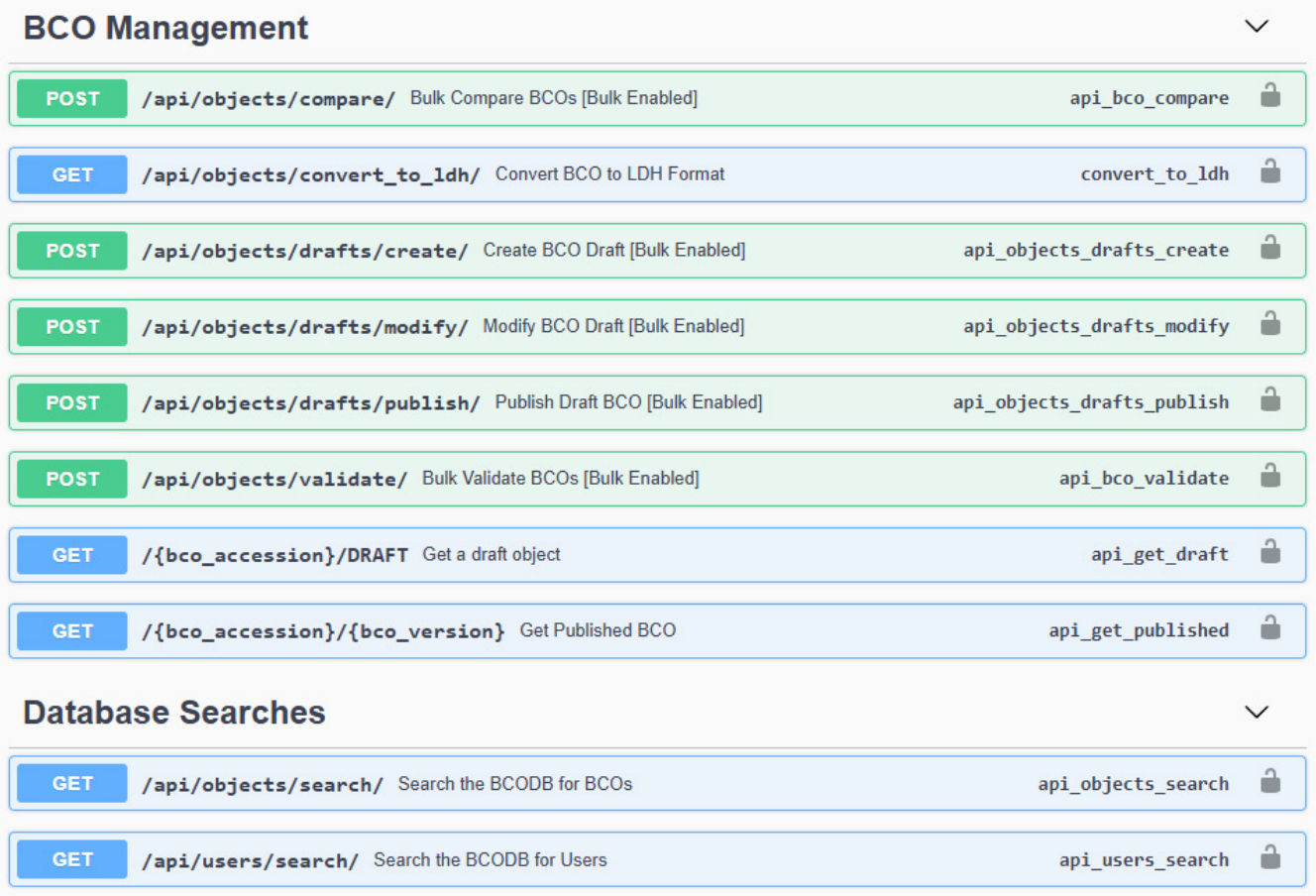

Figure 2 *A subset of the BCODB API endpoints*

Tools and Platform Integrations

BioCompute has been integrated into a range of bioinformatics platforms and tools, enabling users to generate, export and manage BCOs directly from their workflows:

- **BioCompute Builder:** A platform-free, form-based editor for BioCompute creation [9].
- **Galaxy:** An open-source, web-based platform for data-intensive biomedical research that allows users to export workflows as BCOs [10].
- **Seven Bridges Cancer Genomics Cloud:** Provides the BCO app, which supports BCO generation from user workflows, with visualisation support and direct submission to the BCODB [11,12].
- **DNAnexus:** Offers the BCOnexus app, which can generate BCOs from DNAnexus or precisionFDA workflows. The app supports common workflow language (CWL) and includes API connectivity to the BCODB for saving, sharing and importing BCOs [13].
- **HIVE:** The High-performance Integrated Virtual Environment computing environment supports workflow-based BCO creation [14].
- **Nextflow:** A scientific workflow system that enables BCO creation from Nextflow pipelines [15,16].
- **The whirl package:** An open-source R package developed by Novo Nordisk that enables users to create BCOs in connection with script execution using input from a YAML configuration file [17].

Pilot Project and FDA Adoption

To evaluate and expand the usability of BCOs, the FDA's Center for Biologics Evaluation and Research (CBER) and Center for Drug Evaluation and Research (CDER) conducted a pilot project with three pharmaceutical companies. The project addressed logistical aspects of including BCOs in the electronic common technical document (eCTD) submission package and explored how BCOs could streamline the submission and review process. It resulted in a comprehensive FAQ covering topics such as BCO content, formatting, and placement within regulatory submissions [18].

As of today, CBER, CDER and the Human Foods Program (HFP) accept BCOs as part of regulatory submissions [19].

6: Case Studies

This section presents selected case studies that demonstrate how BCOs have been applied to document complex bioinformatics workflows in clinical contexts. These examples illustrate how BioCompute can be integrated into diverse data environments – from multi-omics and imaging to machine learning pipelines.

Each case highlights a use scenario, offering practical insights into how BioCompute supports workflow documentation, regulatory readiness, and collaborative research.

Multi-Omics Cancer Prognosis Model

This case study demonstrates how BioCompute was used to document the application of DeepProg [20] for cancer prognosis modelling across 32 cancer types. DeepProg is an ensemble framework of deep-learning and machine-learning approaches that predicts patient survival subtypes using multi-omics data. By transforming RNA-seq, miRNA, and DNA methylation data into survival-associated features, the model enables patient clustering and outcome prediction.

The BCO provides a versioned record of the entire workflow, including modelling framework, model tuning, data sources and outputs. This example demonstrates how BioCompute facilitates translating advanced modelling techniques into clinical practice by providing a framework that supports replicability and compliance review.

The BCO can be explored in JSON format [21] or via a viewer [22]. It ties together:

- **What ran:** Python3 with scikit-optimize v0.8.1, Ray tune v2.9.3, lifelines v0.28.0, GitHub repository *(https://github.com/lanagarmire/DeepProg)*
- **How it was tuned:** The parametric domain captures key settings, a training set of 100 samples, normalisation (0–1 range), 10 epochs, 50% dropout rate, and log-rank p-values of <0.01 for survival association
- **Data sources:** 32 TCGA cancer types via TCGA-Assembler v2.0.5, validation on GEO datasets (GSE4922/GSE1456/GSE3494/GSE7390), and METABRIC (syn1688369) with 1981 breast cancer samples
- **What artifacts were produced:** Performance tables with Cox PH log-rank p-values and C-indices for HCC/BRCA validation cohorts, clustering stability scores for 2–5 clusters across all 32 cancers (MOESM1_ESM.xlsx, MOESM3_ESM.xlsx)

Clinical Trials with Transcriptomics and Proteomics Data

BioCompute was used to capture multi-stage analysis pipelines applied to single-cell transcriptomics and proteomics data in a Phase 3 clinical trial. The trial protocol specified a primary endpoint based on single-cell transcriptomics and an exploratory endpoint based on proteomics.

The transcriptomics dataset, approximately 2TB in size, exceeded the capacity of traditional statistical computing environments, prompting GxP validation and using a cloud-based system for data storage and preprocessing. For the preprocessing, scrnaseq [23] was executed in a Nextflow pipeline, resulting in an extensive results folder that included gene count matrices per cell, which were used in the subsequent processing steps. As part of the Nextflow pipeline execution, a BCO was generated via an embedded plugin.

The output of the NextFlow pipeline served as input for R-based processing scripts. A Seurat object [24] was created following standard processing steps, including quality control, filtering, normalisation, scaling, dimensionality reduction, clustering, doublet detection, and harmony-based integration [25]. This object was subsequently used to generate the predefined primary endpoint. A second BCO was created for the R-based analysis using the whirl package, which logged the full R environment and analysis steps. Manual curation was applied to the second BCO to add provenance information.

Proteomics data was analysed in a separate R environment, with a BCO created to support internal reproducibility and traceability. While the proteomics BCO was not intended for regulatory submission, the transcriptomics BCOs are planned for inclusion in Module 5.3.5.4 of the electronic eCTD.

Clear Cell Renal Cell Carcinoma Glycoproteomic Risk Prediction

A machine-learning workflow was developed to estimate five-year disease progression risk in clear cell renal cell carcinoma (ccRCC) using tumour glycoproteomic profiles. The analysis was based on 10,814 intact glycopeptide abundances measured across 183 resected ccRCC tumour and adjacent normal tissue samples. A multilayer perceptron model was trained using tumour data to assign high- or low-risk progression labels, with clinical metadata used to derive outcome categories. The workflow was designed to support clinical interpretation by linking molecular features to patient prognosis.

A BCO was created to capture all key components of the workflow and to enable external reviewers to replicate the analysis. The BCO can be explored in JSON format [26] or via a viewer [27]. It covers:

- **Runtime and code (Execution Domain):** Python 3.11.5 with scikit-learn 1.3.2, pandas 2.1.4, xgboost 2.0.3, joblib 1.3.2 and numpy 1.26.3. Scripts include ccrcc_classifier.py, ccrcc_preprocessing.py and ccrcc_predict.py (manual script driver). Repository: *https://github.com/GW-HIVE/PredictMod*.
- **Configuration and preprocessing (Description Domain):** The workflow uses an 80/20 train–test split, standardises features within the training split and applies the same transform to test/prediction data, performs LASSO feature selection, imputes missing values with DreamAI for glycopeptides quantified in >50% of samples, and restricts training to tumour samples.
- **Cohort and labels (Usability Domain and Description Domain):** High-/Low-risk labels are derived from recorded clinical fields: vital status, days to last known disease status, disease response, last known disease status, days to death, and tissue type. Label assignment occurs in the classifier preprocessing step.
- **Inputs with links (Description Domain and I/O Domain):** The BCO lists the TMT-labelled intact-glycopeptide abundance matrix and the associated biospecimen, clinical, exposure, and follow-up TSV files, plus an example single-patient CSV file used by the predictor. All filenames and URIs are enumerated.

- **Outputs and performance (Usability Domain and Description Domain):** Pipeline execution produces a serialised classifier ccrcc_classifier.pkl (URI recorded) and a prediction utility that consumes the example patient CSV. Held-out test metrics are recorded as AUC 0.943, Accuracy 0.917, Sensitivity 1.00, Specificity 0.800.

Clinical Trials with Imaging Protocol Endpoints

The final case study is an example of planned BCO usage for documenting an image processing pipeline. It involves a Phase 3a clinical trial with an exploratory imaging endpoint to evaluate changes in cardiac tissue texture and inflammation between the investigational treatment and placebo arms. Raw echocardiography images are transferred from the imaging vendor to the sponsor, along with metadata to support image selection.

A cardiovascular disease expert identifies high-quality images based on predefined criteria and annotates region of interest (ROI) using Medis software [28]. The subsequent computation of textural and quantitative features is performed with Python scripts. The computational results form part of the SDTM data for the trial.

For documentation, a BCO records the computational steps, inputs and parameters. While manual image selection and annotation cannot be captured directly, they are documented via an audit trail and described in the BCO.

If used for a regulatory submission, the raw image files along with the image metadata are provided to the regulatory agency alongside the BCO file and the SDTM data.

7: Recommendation

To improve the transparency, reproducibility and auditability of computational workflows in clinical research, it is recommended that organisations adopt the BioCompute framework for capturing bioinformatics analyses. BCOs provide a structured, standardised format that supports consistent communication of workflow details across teams and platforms. By incorporating BioCompute into documentation practices, organisations can prepare for regulatory review and strengthen confidence in the integrity of submitted analyses.

For those interested in learning more about BioCompute, the following resources are available:

- BioCompute Object Portal: *https://biocomputeobject.org/*
- IEEE 2791-2020 Standard: *https://standards.ieee.org/ieee/2791/7337/*
- BioCompute User Guide: *https://wiki.biocomputeobject.org/User_guide*
- BioCompute Cheat Sheet: *https://wiki.biocomputeobject.org/Cheatsheet*
- FAQ: *https://wiki.biocomputeobject.org/FAQs*
- BCO Resources: *https://wiki.biocomputeobject.org/BCO_Resources*

Active involvement in the ongoing development of open-source BioCompute tools is strongly encouraged. Contributions to community-driven packages – such as whirl R – not only strengthen the ecosystem but also ensure BioCompute remains adaptable to emerging scientific and regulatory needs. Developers, researchers and platform providers are encouraged to advocate for BioCompute integration in widely used bioinformatics environments to promote interoperability and streamline workflow documentation.

The case studies presented in this white paper demonstrate how BioCompute enhances workflow interoperability and accountability across diverse domains of clinical research. By showcasing real-world applications, platform integrations and pilot project outcomes, this document highlights the transformative potential of BioCompute in bridging the gap between complex bioinformatics processes and regulatory expectations.

Continued collaboration between industry, academic institutions and regulatory agencies will be essential to realise the full potential of BioCompute and embed its principles into the future of clinical data science.

8: Disclaimer

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PHUSE members, respective companies/organisations or regulators' views or policies. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities.

9: References

1. IEEE Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication, IEEE 2791-2020. (2020) https://standards.ieee.org/ieee/2791/7337/

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

3. Alterovitz G., Dean D., Goble C. et al. (2018). Enabling precision medicine via standard communication of HTS provenance, analysis, and results. PLoS Biol., 16(12):e3000099. doi: 10.1371/journal.pbio.3000099

4. https://docs.biocomputeobject.org/

5. https://biocomputeobject.org/

6. https://wiki.biocomputeobject.org/Bco-domains

7. https://biocomputeobject.org/bcodbs

8. https://biocomputeobject.org/api/docs/

9. https://biocomputeobject.org/builder

10. https://galaxyproject.org/use/biocompute-object/

11. https://www.cancergenomicscloud.org/

12. https://github.com/sbg/bco-app

13. https://hub.docker.com/r/dnanexus/bconexus

14. https://hive.biochemistry.gwu.edu/dna.cgi?cmd=main

15. https://www.nextflow.io/

16. https://github.com/nextflow-io/nf-prov

17. https://github.com/NovoNordisk-OpenSource/whirl

18. https://wiki.biocomputeobject.org/FAQs

19. https://www.federalregister.gov/documents/2020/07/22/2020-15771/electronic-submissions-data-standards-support-for-the-international-institute-of-electrical-and

20. Poirion O.B., Jing Z., Chaudhary K. et al. (2021). DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Med., 13(1):112. doi: 10.1186/s13073-021-00930-x

21. https://biocomputeobject.org/BCO_000453/2.0

22. https://biocomputeobject.org/viewer?https://biocomputeobject.org/BCO_000453/2.0

23. https://nf-co.re/scrnaseq/4.0.0/

24. https://satijalab.org/seurat/

25. https://github.com/immunogenomics/harmony

26. https://biocomputeobject.org/BCO_000501/1.1

27. https://biocomputeobject.org/viewer?https://biocomputeobject.org/BCO_000501/1.1

28. https://aacrjournals.org/cancerres/article/77/21/e39/662607/Software-for-the-Integration-of-Multiomics

10: Project Contact Information

- PHUSE Working Groups: workinggroups@phuse.global
- Mehdi Harek: mehdi.harek@bms.com
- Jonathon Keeney: keeneyjg@gwu.edu
- Bron Kisler: bronkisler@icloud.com
- Adrian Czaban: adcz@novonordisk.com
- Jeffrey Long: jeffrey.x.long@gsk.com
- Ashwini Yermal Shanbhogue: ash23shan@yahoo.com
- Addie Nina Olsen: aieo@novonordisk.com
- Amaya Zaratiegui: azeu@novonordisk.com