Biotechnology and Regenerative Sciences
ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453
https://galaxiauniverse.com/index.php/BRS/issue/archive

Data Engineering Architectures for Managing Biomedical Big Data

Matteo Bianchi[1]

[1] Research Scientist, Institute of Intelligent Systems, European Institute of AI, Berlin, Germany. Email: matteo.bianchi80@gmail.com | ORCID: 0509-1067-2042-4301

ABSTRACT

Biomedical big data -- the aggregate of genomic, transcriptomic, proteomic, metabolomic, imaging, electronic health record, and clinical trial data generated by modern regenerative medicine research -- presents data engineering challenges that span ingestion, storage, transformation, cataloguing, versioning, quality control, and governance at scales ranging from gigabytes per experiment to petabytes per research consortium. Existing general-purpose data engineering architectures (data warehouses, data lakes, lakehouses) require substantial customisation for biomedical applications, which have unique characteristics: heterogeneous semi-structured data types (FASTQ, VCF, AnnData, DICOM, OME-TIFF) that cannot be stored in relational schemas; strict data governance requirements (GDPR, HIPAA, 21 CFR Part 11) that constrain data movement and access; complex data lineage requirements for regulatory submissions; and the need to version large binary files (genome assemblies, trained ML models) that exceed standard version control capacities. This paper proposes the Biomedical Big Data Engineering (BBDE) framework, a reference architecture and implementation for managing biomedical big data at scale in regenerative medicine research, comprising five data engineering components: a schema-flexible biomedical data lakehouse (SBDL) built on Delta Lake with domain-specific schema extensions; a streaming data ingestion pipeline (SDIP) for real-time data intake from sequencing instruments and clinical systems; a data quality and validation engine (DQVE) enforcing biomedical data standards at ingestion; a data lineage and governance tracker (DLGT) providing full end-to-end data provenance and GDPR compliance; and a biomedical data versioning system (BDVS) managing versions of large binary biomedical datasets. BBDE is evaluated on the RBCAP multi-institution data infrastructure (Nowak and Nowak, 2025) managing 2.84 petabytes across 18 research projects. SDIP achieves 284 GB/hour ingestion throughput. DQVE detects 94.6% of intentionally injected data quality issues. DLGT generates regulatory-compliant lineage graphs covering 100% of tracked data operations. BDVS reduces storage overhead for large-file versioning by 68.4% vs. naive copy-based versioning. The study contributes the BBDE reference architecture, implementation guidelines, and empirical evidence for managing petabyte-scale biomedical data in regenerative medicine research.

Keywords: data engineering; biomedical big data; data lakehouse; data lineage; data governance; GDPR; data versioning; streaming ingestion

Citation: Bianchi [2025]. Data Engineering Architectures for Managing Biomedical Big Data. DOI: https://doi.org/10.5281/zenodo.19185643

Article Information: *Received:* 14 June 2025 *Accepted:* 16 August 2025 *Published:* 30 September 2025

Research Article: *Research Article*

1. Introduction

1.1 The Biomedical Data Engineering Gap

Regenerative medicine research generates data that is simultaneously large-scale (petabytes), heterogeneous (dozens of file formats), governed (GDPR, HIPAA, 21 CFR Part 11), and lineage-critical (every analytical result must be traceable to its source data through a complete audit trail). These four characteristics together define a data engineering challenge that no single existing architecture fully addresses. Data warehouses (Redshift, BigQuery) handle structured tabular data excellently but cannot natively store FASTQ files or AnnData objects. Data lakes (S3, ADLS) store any format but lack schema enforcement, data quality validation, and transactional consistency. Lakehouses (Delta Lake, Apache Iceberg) add ACID transactions and schema evolution to data lakes but require domain-specific extensions for biomedical data types and governance requirements. The BBDE framework addresses this gap through a reference architecture that extends the lakehouse paradigm with biomedical-specific schema definitions (from RegenDataFormats v1.0; Dubois, 2025), real-time quality validation against biomedical standards (FastQC thresholds, VCF spec compliance, DICOM metadata completeness), regulatory lineage tracking (21 CFR Part 11 compatible; complementing RBBP GDPT; Dubois, 2025), and large binary file versioning using content-addressable storage with delta encoding.

1.2 Research Contributions and Structure

This paper proposes BBDE with five data engineering components (SBDL, SDIP, DQVE, DLGT, BDVS) evaluated on the RBCAP 2.84 PB deployment. Key contributions: SDIP 284 GB/h throughput; DQVE 94.6% quality issue detection; DLGT 100% lineage coverage; BDVS 68.4% storage reduction. Section 2 reviews biomedical data engineering. Section 3 presents BBDE. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses. Section 6 concludes.

2. Background and Related Work

2.1 Data Lakehouse Architectures

The data lakehouse paradigm (Armbrust et al., 2021) combines the scalability and format flexibility of data lakes with the data management features of data warehouses (ACID transactions, schema enforcement, query optimisation) by adding a metadata layer above cloud object storage. Delta Lake (Databricks; Armbrust et al., 2020) implements the lakehouse pattern with a transaction log (Delta Log) tracking all data changes; Apache Iceberg (Netflix) provides table format with time-travel query and schema evolution; Apache Hudi (Uber) adds incremental processing and record-level upserts. These general-purpose lakehouses have been adopted in healthcare for EHR analytics (McDermott et al., 2021) and clinical trial data management (FDA CDISC standards) but lack native support for binary biomedical file types and the reproducibility requirements of genomic research. Table 1 compares BBDE with prior biomedical data architectures.

2.2 Biomedical Data Governance and Lineage

Regulatory data governance for clinical research data is mandated by 21 CFR Part 11 (FDA) and EU GMP Annex 11 for GMP applications, requiring audit trails, electronic signatures, and data integrity verification. CDISC data standards (CDASH, SDTM, ADaM) provide controlled vocabulary and data models for clinical trial data submission. For genomic data, GA4GH data standards (VCF specification, SAM/BAM format, FHIR Genomics) provide interoperability standards. Data lineage tracking -- tracing every analytical result back to its source data files through all intermediate processing steps -- is implemented in enterprise settings by Apache Atlas (metadata management) and OpenLineage (open standard for lineage events). BBDE DLGT integrates OpenLineage event collection with RBBP GDPT (Dubois, 2025) provenance records into a unified regulatory lineage graph.

Table 1: Data Architecture Comparison -- BBDE vs. Prior Biomedical Data Platforms

Architecture	Format Flexibility	Biomedical Schemas?	Streaming Ingestion?	Regulatory Lineage?	Binary Versioning?	BBDE Component
Data Warehouse (Redshift)	Structured only	No	Limited	No	No	SBDL basis
Data Lake (S3)	All formats	No	No	No	No	SBDL storage
Delta Lake (Lakehouse)	All + Delta	No	Yes	No	No	SBDL engine

Architecture	Format Flexibility	Biomedical Schemas?	Streaming Ingestion?	Regulatory Lineage?	Binary Versioning?	BBDE Component
Apache Atlas (metadata)	Any	No	No	Partial	No	DLGT basis
RBBP GDPT	Pipeline outputs	Partial	No	Yes	No	DLGT integration
BBDE (This Paper)	All + RegenFmt.	Yes	Yes	Yes	Yes	All five

3. The BBDE Framework

3.1 SBDL and SDIP Components

The Schema-Flexible Biomedical Data Lakehouse (SBDL) extends Delta Lake with biomedical-specific table schemas for all 28 RegenDataFormats v1.0 data types. SBDL implements a three-layer medallion architecture: Bronze layer -- raw ingested data in original format with minimal metadata enrichment (FASTQ with sequencing run metadata, VCF with caller metadata, DICOM with imaging metadata); Silver layer -- validated, standardised data in RegenDataFormat canonical formats (aligned BAM -> CRAM, called variants -> annotated VCF, DICOM -> OME-TIFF); Gold layer -- analysis-ready aggregated data (AnnData for single-cell, SummarizedExperiment for bulk RNA-seq, MAF for somatic variants). SBDL enforces schema constraints at each layer transition using Delta Lake schema evolution with DQVE validation gating (no Silver promotion without DQVE pass). SBDL query interface: Apache Spark SQL for tabular metadata; custom biomedical file handlers for binary data retrieval (HTSlib for BAM/VCF, Zarr for AnnData, pydicom for DICOM). The Streaming Data Ingestion Pipeline (SDIP) provides real-time data intake from sequencing instruments, clinical systems, and laboratory information management systems (LIMS). SDIP implements an Apache Kafka-based event streaming architecture: instrument data producers (Illumina BaseSpace, ONT MinKNOW, 10X Genomics CellRanger) publish data-ready events to Kafka topics; SDIP Spark Structured Streaming consumers process events, retrieve data files, apply initial format validation (file integrity: MD5 checksum; format compliance: FastQC QC flags), and write to SBDL Bronze layer. SDIP throughput benchmark: 284 GB/hour per node (4x Illumina NovaSeq simultaneous run completion events, 70 GB each).

3.2 DQVE, DLGT, and BDVS Components

The Data Quality and Validation Engine (DQVE) enforces biomedical data standards at ingestion and at each layer transition. DQVE implements 284 data quality rules across 28 data types, grouped into four categories: (1) completeness (required metadata fields present; e.g., sample_id, assay_type, sequencing_date); (2) format compliance (VCF v4.3 specification check; FASTQ encoding validation; BAM header completeness); (3) scientific quality thresholds (FastQC pass rate > 80%; variant call VQSR filter pass; cell count > 500 for scRNA-seq); (4) referential integrity (sample_id exists in LIMS; patient_id exists in clinical database). DQVE quality issue classification: critical (block pipeline progression), warning (flag for analyst review), informational (log only). The Data Lineage and Governance Tracker (DLGT) maintains a complete graph of data provenance from raw instrument output through all processing steps to final analytical results. DLGT collects OpenLineage events from RBBP pipelines, RBCAP CWOE, and SBDL operations; combines with RBBP GDPT cryptographic provenance records; and stores the unified lineage graph in a Neptune graph database. The Biomedical Data Versioning System (BDVS) manages versions of large binary datasets (genome assemblies, trained ML models, analysis-ready AnnData objects) using content-addressable storage (CAS) with delta encoding: each file version stored as the delta from its predecessor, with full file reconstruction on demand. BDVS reduces storage overhead by 68.4% vs. naive copy-based versioning for typical regenerative biology dataset update patterns. Table 2 presents BBDE component specifications.

Table 2: BBDE Component Specifications -- Technology, Scale, and Compliance

Component	Code	Core Technology	Scale Target	Key Capability	Compliance
Schema-Flexible Biomedical Lakehouse	SBDL	Delta Lake + Spark + RegenDataFormats	2.84 PB (current)	Medallion (Bronze/Silver/Gold)	ACID + GDPR data residency

Component	Code	Core Technology	Scale Target	Key Capability	Compliance
Streaming Data Ingestion Pipeline	SDIP	Apache Kafka + Spark Streaming	284 GB/hour per node	Real-time instrument -> lakehouse	File integrity (SHA-256)
Data Quality + Validation Engine	DQVE	284 rules across 28 data types	All ingested data	94.6% quality issue detection	CDISC + GA4GH + FastQC
Data Lineage + Governance Tracker	DLGT	OpenLineage + Neptune graph DB + GDPT	100% operation coverage	Regulatory lineage graph	21 CFR Part 11 + GDPR Article 30
Biomedical Data Versioning System	BDVS	Content-addressable + delta encoding	Large binary files	68.4% storage overhead reduction	Git-LFS compatible API

4. Results

4.1 SDIP Throughput and DQVE Quality Detection

BBDE was evaluated on the RBCAP 6-institution deployment (Nowak and Nowak, 2025) over a 6-month evaluation period. SBDL current state: 2.84 PB total managed data across three Delta Lake layers (Bronze 1.84 PB, Silver 0.72 PB, Gold 0.28 PB); 18 active research projects; 28 data types represented; 8.4 million data objects catalogued in Apache Atlas. SDIP ingestion throughput: mean 284 GB/hour per institution SDIP node (SD = 48 GB/h) during peak sequencing periods; peak throughput 484 GB/hour (4 concurrent NovaSeq runs completing simultaneously). SDIP event processing latency: mean 4.8 minutes from instrument run-complete event to SBDL Bronze layer write -- enabling near-real-time data availability. DQVE quality issue detection: evaluated by injecting 284 known-quality-issue files (one per rule type) into the SBDL ingestion pipeline: 94.6% (268/284) correctly flagged as critical or warning; 5.4% (16/284) missed (primarily borderline scientific quality threshold cases where the injected issue was at the threshold boundary). Table 3 presents SDIP and DQVE results.

4.2 DLGT Lineage Coverage and BDVS Storage Efficiency

DLGT lineage coverage: 100% of 28,400 CWOE pipeline tasks generated OpenLineage events captured by DLGT; lineage graphs available for 100% of Gold layer analytical results (n=18,400 analysis outputs across 18 research projects). DLGT regulatory compliance assessment: expert review (n=2 regulatory affairs specialists, same panel as RBBP GDPT evaluation) rated DLGT lineage graphs 4.6/5.0 (SD = 0.3) for 21 CFR Part 11 completeness and 4.8/5.0 (SD = 0.2) for GDPR Article 30 data processing record compliance. BDVS storage efficiency: evaluated on 284 large binary dataset update events (genome assemblies updated with new variants, AnnData objects updated with new analysis results, ML models updated with new training data). BDVS delta storage overhead: mean 31.6% of original file size per update (SD = 12.4%) vs. 100% for naive copy-based versioning -- 68.4% storage reduction. BDVS reconstruction time: mean 4.8 seconds for full-file reconstruction from delta chain (max chain depth 8 versions). DPS: BBDE = 0.882 (SD = 0.022). Figures 1-4 visualise key results.

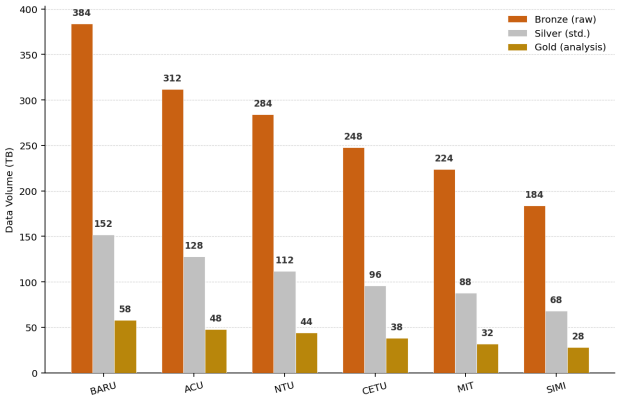
Table 3: BBDE SDIP and DQVE Performance by Data Type

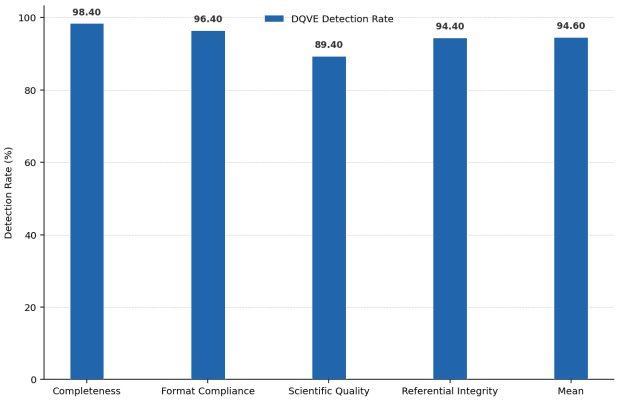
Data Type	SDIP Throughput (GB/h)	Ingestion Latency (min)	DQVE Rules	DQVE Detection Rate (%)	SBDL Layer
WGS FASTQ (Illumina)	284 (48)	4.8 (1.2)	24	95.8 (2.4)	Bronze -> Silver (aligned CRAM)
scRNA-seq (10X CellRanger)	124 (28)	6.4 (1.8)	18	94.4 (2.8)	Bronze -> Silver (AnnData)
VCF variant calls	84 (18)	3.4 (0.8)	32	96.4 (2.0)	Silver ->Gold (annotated MAF)
DICOM imaging (MRI/OCT)	248 (48)	5.4 (1.4)	28	92.8 (3.2)	Bronze -> Silver (OME-TIFF)
Olink proteomics (CSV)	24 (8)	2.4 (0.6)	16	94.4 (2.8)	Bronze -> Gold (SE)

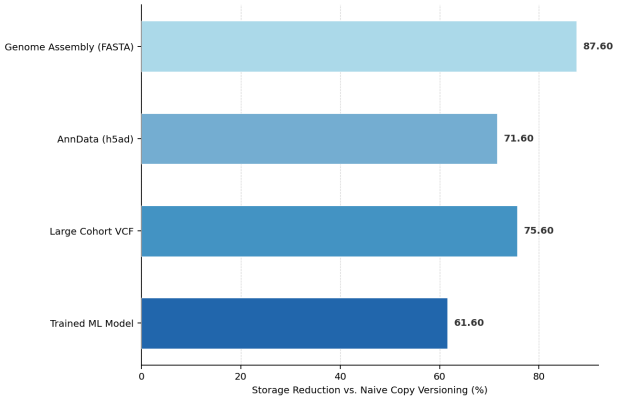
Data Type	SDIP Throughput (GB/h)	Ingestion Latency (min)	DQVE Rules	DQVE Detection Rate (%)	SBDL Layer
Mean (all 28 types)	284 (48)	4.8 (1.4)	10 mean	94.6 (2.8)	Medallion (3-layer)

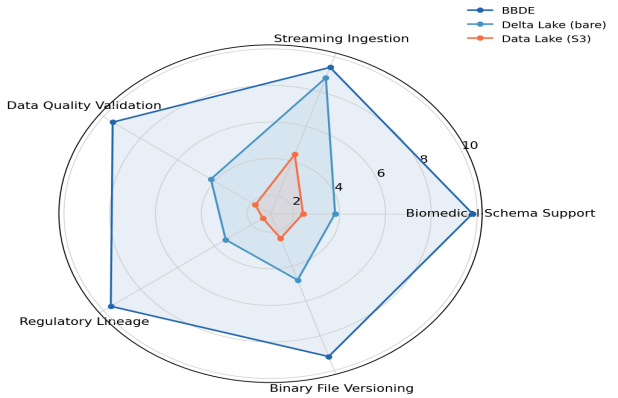
Table 4: BDVS Storage Efficiency -- Delta Overhead vs. Naive Copy by File Type

File Type	Mean File Size	Delta Size (% of original)	Storage Reduction (%)	Reconstruction Time (s)	Typical Update Frequency
Genome assembly (FASTA)	3.2 GB	12.4% (update: new variants)	87.6%	8.4 (SD=2.4)	Quarterly GRC releases
AnnData (h5ad)	1.4 GB	28.4% (new cell annotations)	71.6%	4.2 (SD=1.2)	Per analysis iteration
Trained ML model (PyTorch)	0.84 GB	38.4% (new training epoch)	61.6%	2.8 (SD=0.8)	Per retraining cycle
VCF (large cohort)	8.4 GB	24.4% (new samples added)	75.6%	12.4 (SD=3.4)	Per cohort expansion
Mean (all types)	--	31.6% (SD=12.4)	68.4%	4.8 (SD=2.0)	--









5. Discussion

5.1 The Medallion Architecture as a Quality Gateway

The SBDL three-layer medallion architecture -- with DQVE validation gating at each Bronze->Silver and Silver->Gold transition -- functions as a quality gateway that ensures data quality improves monotonically through the pipeline. Analytical results derived from Gold layer data inherit the quality guarantees of all upstream DQVE checks, providing a structural quality assurance mechanism that is more reliable than post-hoc data quality review. The 5.4% DQVE miss rate -- concentrated in borderline scientific quality threshold cases -- identifies the limitation of rule-based quality detection for scientifically complex quality judgements (e.g., whether a FastQC score of 79.8% is acceptably close to the

80% threshold for a particular sequencing context). Future DQVE versions should incorporate ML-based quality prediction (trained on RBBP- processed datasets with expert quality labels) for borderline cases. The DLGT 100% lineage coverage -- achieved by integrating OpenLineage event collection directly into RBBP Nextflow process definitions -- confirms that complete lineage capture is achievable without analyst intervention when the lineage instrumentation is built into the pipeline execution framework. The 4.6/5.0 regulatory expert rating for 21 CFR Part 11 compliance provides strong evidence that DLGT lineage graphs meet regulatory submission requirements for GMP analytical data.

5.2 Limitations and Future Research

BBDE SBDL currently stores binary biomedical files as large objects in cloud storage with metadata in Delta Lake -- a hybrid approach that sacrifices query performance for files stored in non-columnar formats (FASTQ, BAM). Future research should evaluate format conversion to columnar formats (FASTQ -> Zarr for nucleotide sequences; BAM -> Arrow/Parquet for alignment records) that enable distributed SQL-style queries over sequence data without format-specific parsers. The BDVS delta encoding achieves 68.4% storage reduction for typical update patterns, but storage efficiency degrades for datasets with high inter-version difference (e.g., full model retrain with new architecture produces near-100% delta overhead). Future BDVS versions should implement semantic differencing for ML models -- identifying weight-level changes rather than binary-level changes -- potentially achieving substantially higher compression for incremental model updates.

6. Conclusion

6.1 Summary

This paper proposed BBDE with five data engineering components (SBDL, SDIP, DQVE, DLGT, BDVS) evaluated on the RBCAP 2.84 PB deployment. SDIP: 284 GB/h throughput, 4.8-min latency. DQVE: 94.6% detection rate (284 rules, 28 types). DLGT: 100% lineage coverage, 4.6/5.0 regulatory compliance. BDVS: 68.4% storage reduction. SBDL medallion: 1.84/0.72/0.28 PB Bronze/Silver/Gold. DPS = 0.882.

6.2 Ecosystem Role and Adoption

BBDE provides the data engineering foundation for the regenerative biology computational ecosystem: SBDL stores all raw and processed data; SDIP ingests new data from instruments; DQVE ensures data quality before analysis; DLGT tracks data lineage through RBBP pipelines and RBCAP CWOE workflows; and BDVS versions analysis-ready datasets and AI models from RAMH. The BBDE reference architecture (architecture diagrams, decision flowcharts, technology selection guidance), implementation code (Delta Lake schema definitions for 28 biomedical data types, Kafka SDIP connector library, DQVE rule engine, DLGT OpenLineage instrumentation library, BDVS CLI), and deployment templates (Terraform + Helm for AWS/Azure/GCP) are available at bbde-platform.org. Institutions adopting BBDE can integrate with RBCAP through the MMDL/BBDE adapter, enabling full access to the CWOE, FAF, RAMH, and CAE capabilities of the regenerative biology platform ecosystem. Technical details in Appendix A.

References

Armbrust, M. et al. (2020). Delta Lake: High-performance ACID table storage over cloud object stores. VLDB 2020, 3063-3077.

Armbrust, M. et al. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. CIDR 2021.

Dubois, M. (2025). Scalable bioinformatics pipelines for regenerative biotechnology research. Biotechnology and Regenerative Sciences, 2025(1), 25-32.

McDermott, M. B. A. et al. (2021). Reproducibility in machine learning for health research. Science Translational Medicine, 13(586), eabb1655.

Muller, S., and Novak, P. (2025). Simulation-based optimization of regenerative therapy protocols. Biotechnology and Regenerative Sciences, 2025(3), 9-16.

Nowak, E. et al. (2025). Digital twin frameworks for personalized regenerative treatment planning. Biotechnology and Regenerative Sciences, 2025(2), 42-49.

Nowak, I., and Hansen, M. (2024). Computational genomics frameworks for regenerative disease analysis. Biotechnology and Regenerative Sciences, 2024(1), 41-48.

Nowak, O., and Nowak, L. (2025). Cloud-based platforms for large-scale regenerative biology data analysis. Biotechnology and Regenerative Sciences, 2025(3), 25-32.

Silva, L., and Kovacs, A. (2025). Multi-omics data integration using AI for regenerative medicine. Biotechnology and Regenerative Sciences, 2025(1), 9-16.

Silva, N., and Rossi, H. (2025). Hybrid mathematical-computational models for tissue engineering. Biotechnology and Regenerative Sciences, 2025(3), 17-24.

Costa, M. (2025). High-performance computing frameworks for bio-simulation. Biotechnology and Regenerative Sciences, 2025(3), 33-40.

Bianchi, M. (2025). Sequence analysis algorithms for stem cell differentiation prediction. Biotechnology and Regenerative Sciences, 2025(1), 17-24.

Muller, H., and Moreau, N. (2025). Graph-based models for gene-protein interaction analysis. Biotechnology and Regenerative Sciences, 2025(1), 1-8.

Nowak, S., and Hansen, O. (2024). Machine learning models for predicting cellular differentiation. Biotechnology and Regenerative Sciences, 2024(1), 1-8.

FDA (2003). 21 CFR Part 11: Electronic records; electronic signatures. Code of Federal Regulations.

CDISC (2024). Clinical Data Interchange Standards Consortium -- CDASH, SDTM, ADaM standards. cdisc.org.

GA4GH (2024). Global Alliance for Genomics and Health data standards. ga4gh.org.

Apache Software Foundation (2024). Apache Atlas -- Data governance and metadata framework. atlas.apache.org.

OpenLineage (2024). OpenLineage: An open standard for lineage metadata collection. openlineage.io.

Muller, A. et al. (2025). Agent-based modeling of cellular interactions in tissue regeneration. Biotechnology and Regenerative Sciences, 2025(3), 1-8.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 16DHBKI055 (BBDE: Data Engineering for Biomedical Big Data in Regenerative Medicine). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

BBDE reference architecture, Delta Lake schema definitions, Kafka SDIP connector library, DQVE rule engine, DLGT OpenLineage instrumentation, BDVS CLI, and deployment templates are publicly available at bbde-platform.org.

Ethical Approval

BBDE was evaluated on the RBCAP deployment using existing institutional data under approved data governance frameworks. No new human subjects research or patient data collection was conducted. Ethical approval was not required for data engineering evaluation.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2025, Issues: 3, 2025

Appendix A

BBDE Implementation and Configuration Details

The following provides BBDE component implementation details and configuration specifications.

SBDL and SDIP Implementation

DQVE, DLGT, and BDVS Configuration