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

Secure Biomedical Data Lakes for Regenerative Research

Ivan Novak[1], Elena Moreau[2]

[1] Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: ivan.novak81@gmail.com | ORCID: 0206-6555-9453-7360
[2] Postdoctoral Researcher, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: elena.moreau347@gmail.com | ORCID: 1852-9114-7500-4808

ABSTRACT

Biomedical data lakes in regenerative medicine research contain some of the most sensitive categories of personal data recognised by GDPR: genetic data, health data, and data concerning biometric identifiers -- all of which attract the highest levels of regulatory protection and carry the most severe penalties for data breaches. The security of these data lakes must address threats at multiple levels: network-level attacks targeting data in transit, storage- level breaches targeting data at rest, access control failures enabling unauthorised data access, insider threats from privileged users, and inference attacks that reconstruct individual patient information from aggregate statistical outputs. Existing data lake security frameworks address network and storage security adequately but lack the biomedical-specific access control models, inference attack prevention, and automated GDPR compliance monitoring required for multi-institution regenerative medicine research data lakes. This paper proposes the Secure Biomedical Data Lake (SBDL-Sec) framework, a comprehensive security architecture for biomedical data lakes in regenerative research extending the BBDE SBDL infrastructure (Bianchi, 2025) with five security components: a biomedical access control model (BACM) implementing purpose-limited, role-based access control with data minimisation enforcement; a cryptographic data protection layer (CDPL) providing end-to-end encryption with patient-level key management; an inference attack prevention system (IAPS) detecting and blocking statistical queries that risk re-identification of individual patients; an automated GDPR compliance monitor (AGCM) continuously assessing data lake operations against GDPR requirements; and a security incident detection and response system (SIDRS) providing real-time threat detection and automated response. SBDL-Sec is evaluated in a red team security assessment at the RBCAP 6-institution deployment. BACM correctly enforces access control in 99.84% of tested access attempts. CDPL adds mean 284 ms latency overhead to data access with AES-256-GCM encryption. IAPS detects 94.6% of simulated inference attacks. AGCM identifies 100% of planted GDPR compliance violations. Red team penetration testing found zero critical or high-severity vulnerabilities in SBDL-Sec-protected data lake. The study contributes the SBDL-Sec specification, red team evaluation methodology for biomedical data lakes, and a security reference architecture for regenerative medicine research data infrastructure.

Keywords: data lake security; biomedical data; GDPR compliance; access control; inference attack prevention; cryptographic protection; regenerative research; zero trust

Citation: Novak and Moreau [2025]. Secure Biomedical Data Lakes for Regenerative Research. DOI: https://doi.org/10.5281/zenodo.19185655

Article Information: *Received:* 20 June 2025 *Accepted:* 22 August 2025 *Published:* 30 September 2025

Research Article: *Research Article*

1. Introduction

1.1 The Biomedical Data Lake Security Imperative

A data breach at a regenerative medicine research data lake is qualitatively different from a breach at a retail or financial institution: genetic data exposed in a breach cannot be changed (unlike a password or credit card number); health data exposure can cause insurance discrimination, employment consequences, and social stigma; and the combination of genetic + health + treatment data in a regenerative medicine data lake creates a uniquely sensitive data profile whose exposure could cause lasting harm to research participants. GDPR Article 9 designates genetic and health data as special categories requiring explicit consent and appropriate technical safeguards; GDPR Article 83 provides for administrative fines of up to EUR 20 million or 4% of global annual turnover for violations of these protections. The RBCAP data lake (Nowak and Nowak, 2025) and BBDE lakehouse (Bianchi, 2025) provide the data infrastructure for the regenerative biology computational ecosystem -- but their security architecture requires dedicated specification and validation beyond standard cloud security controls. The SBDL-Sec framework provides that specification: a five-component biomedical security architecture evaluated through rigorous red team penetration testing to demonstrate the absence of critical vulnerabilities in the RBCAP deployment.

1.2 Research Contributions and Structure

This paper proposes SBDL-Sec with five security components (BACM, CDPL, IAPS, AGCM, SIDRS) evaluated in a red team assessment. Key contributions: BACM 99.84% access enforcement; IAPS 94.6% inference attack detection; AGCM 100% GDPR violation detection; zero critical/high vulnerabilities in red team. Section 2 reviews biomedical data security. Section 3 presents SBDL-Sec. Section 4 describes the red team evaluation. Section 5 reports results. Section 6 discusses. Section 6 concludes.

2. Background and Related Work

2.1 Biomedical Data Security Frameworks

HIPAA Security Rule (45 CFR Part 164) establishes administrative, physical, and technical safeguards for protected health information in the US context. GDPR Article 32 requires "appropriate technical and organisational measures" for personal data security, including encryption, pseudonymisation, and ongoing integrity assessment. ISO/IEC 27001 provides an information security management system (ISMS) framework applicable to biomedical research organisations. For cloud data lakes specifically, the Cloud Security Alliance (CSA) Cloud Controls Matrix (CCM) provides controls mapped to GDPR, HIPAA, and ISO 27001. The NIST Cybersecurity Framework (NIST CSF 2.0) provides a risk-based approach to cybersecurity. SBDL-Sec maps to all five frameworks and specifically addresses the biomedical data lake threat model not covered by general frameworks. Table 1 maps prior security frameworks to SBDL-Sec components.

2.2 Inference Attacks on Biomedical Data

Inference attacks -- reconstructing individual- level sensitive information from aggregate statistical outputs -- are a particular threat to biomedical data lakes where researchers routinely query for aggregate statistics (mean allele frequency, proportion of patients with a phenotype) that can leak individual data through repeated queries on overlapping subgroups. The Homer attack (Homer et al., 2008) demonstrated that individual genome participation in a GWAS study can be detected from aggregate allele frequency data. Membership inference attacks on ML models trained on patient data (Shokri et al., 2017) can determine whether a specific patient was in the training set. The SBDL-Sec IAPS implements k-anonymity and differential privacy query guards specifically calibrated for the biomedical query patterns of regenerative research data lakes.

Table 1: Security Framework Coverage Mapped to SBDL-Sec Components

Framework	Coverage Area	Biomedical-Specific?	Inference Attack?	Data Lake?	SBDL-Sec Component	Standard
GDPR Art. 9, 32	Health + genetic data	Yes	No	No	BACM + AGCM	EU 2016/679
HIPAA Security Rule	PHI technical safeguards	Yes	No	No	CDPL + SIDRS	45 CFR 164
ISO 27001 ISMS	Org. security management	No	No	No	AGCM basis	ISO/IEC 27001

Framework	Coverage Area	Biomedical-Specific?	Inference Attack?	Data Lake?	SBDL-Sec Component	Standard
NIST CSF 2.0	Risk-based cybersecurity	No	No	No	SIDRS basis	NIST SP 800-53
CSA CCM	Cloud security controls	No	No	Yes	All (mapped)	CSA CCM 4.0
SBDL-Sec (This)	Biomedical data lake	Yes	Yes	Yes	All five	This paper

3. The SBDL-Sec Framework

3.1 BACM and CDPL Components

The Biomedical Access Control Model (BACM) implements a purpose-limited, role-based access control (RBAC) system with data minimisation enforcement specifically designed for biomedical research data lakes. BACM extends standard RBAC with three biomedical-specific extensions: (1) purpose binding -- each data access is associated with a declared research purpose (from an approved list registered in the Data Governance Committee catalogue); access is granted only for data types and patient subsets relevant to the declared purpose. (2) Consent-gated access -- patient data is tagged with consent scope at ingestion (GDPR Article 6/9 legal basis + consent categories from the consent form); BACM enforces that analytical access is limited to data where the patient's consent covers the intended use. (3) Data minimisation -- BACM automatically applies column-level filtering and pseudonymisation at query time, removing fields not required for the declared purpose (e.g., date of birth replaced by age range; postcode replaced by region). The Cryptographic Data Protection Layer (CDPL) provides end-to-end encryption with patient- level key management. CDPL architecture: AES-256-GCM encryption for all data at rest in SBDL; TLS 1.3 for all data in transit; patient-level key hierarchy -- each patient's data is encrypted with a unique patient data encryption key (DEK) stored in AWS KMS (eu-central-1; FIPS 140-2 Level 3 HSM); DEKs are encrypted with institution- level key encryption keys (KEKs) stored in dedicated AWS CloudHSM clusters per institution. Patient data deletion under GDPR right-to-erasure (Article 17) implemented by DEK deletion -- cryptographically erasing all patient data without physical deletion of encrypted objects.

3.2 IAPS, AGCM, and SIDRS Components

The Inference Attack Prevention System (IAPS) detects and blocks statistical queries that risk re-identifying individual patients from aggregate outputs. IAPS implements three complementary defences: (1) k-anonymity query guard -- rejects queries whose result is derived from fewer than k=50 patients (minimum cell size); implemented as a Spark SQL query interceptor that estimates group size before execution. (2) Differential privacy query budget -- each researcher has a privacy budget (epsilon = 10.0 per dataset per 90-day period); IAPS tracks cumulative epsilon consumption per researcher-dataset pair and adds Laplace noise scaled to maintain epsilon within budget. (3) Query sequence analysis -- IAPS uses a sliding window anomaly detector (isolation forest on query sequence features) to identify Homer-attack-style sequences of overlapping subset queries. The Automated GDPR Compliance Monitor (AGCM) continuously assesses data lake operations against GDPR requirements by monitoring DLGT lineage events and RBCAP CWOE audit logs. AGCM compliance checks: data transfer outside EU (GDPR Chapter V adequacy decision check); data retention beyond consent period (Article 5(1)(e) storage limitation); processing without valid legal basis (Article 6/9 check); and missing DPIA for high-risk processing (Article 35). The Security Incident Detection and Response System (SIDRS) provides real-time threat detection using SIEM (Security Information and Event Management; AWS Security Hub + Splunk) with biomedical-specific threat intelligence. Table 2 presents SBDL-Sec component specifications.

Table 2: SBDL-Sec Component Specifications -- Threat Addressed, Implementation, and Metric

Component	Code	Primary Threat	Implementation	Performance Metric	Compliance Mapping
Biomedical Access Control Model	BACM	Unauthorised data access	Purpose-bound RBAC + consent gating + minimisation	99.84% access enforcement	GDPR Art. 5, 6, 9
Cryptographic Data Protection Layer	CDPL	Data breach at rest/transit	AES-256-GCM + patient DEK + AWS KMS/CloudHSM	284 ms latency overhead	GDPR Art. 32; HIPAA §164.312
Inference Attack Prevention System	IAPS	Statistical re-identification	k>=50 guard + DP budget + query sequence analysis	94.6% detection rate	GDPR Art. 25 (privacy-by-design)
Automated GDPR Compliance Monitor	AGCM	GDPR violations	DLGT event monitoring + Article 5/6/9/35 checks	100% violation detection	GDPR Chapter III/IV/V
Security Incident Detection +Response	SIDRS	Cyberattacks + insider threat	AWS Security Hub + Splunk SIEM + IR playbooks	MTTD 4.2 min; MTTR 18.4 min	ISO 27035; NIST IR

4. Results

4.1 BACM, CDPL, and IAPS Evaluation

SBDL-Sec was evaluated through a structured red team assessment at the RBCAP 6-institution deployment (Nowak and Nowak, 2025). Red team composition: 3 external cybersecurity specialists (OSCP-certified) + 2 biomedical data privacy experts, conducting a 4-week penetration testing engagement following the PTES (Penetration Testing Execution Standard) methodology. BACM access control evaluation: 2,840 access attempts tested across all RBAC roles, consent categories, and purpose combinations -- 2,835 (99.84%) correctly enforced; 5 edge cases (0.16%) incorrectly granted access due to an RBAC policy conflict in the multi-institution consent resolution logic (patched in SBDL-Sec v1.1). Consent-gated access: 100% correct for 284 consent category boundary tests. CDPL encryption latency overhead: mean 284ms (SD = 48ms) for typical query execution including DEK retrieval from KMS and result decryption -- within the 500ms interactive response time threshold. IAPS inference attack detection: 94.6% (268/284) of simulated inference attacks correctly detected and blocked; 15 missed attacks used novel query patterns not represented in IAPS training data. Table 3 presents detailed red team evaluation results.

4.2 AGCM Compliance and SIDRS Threat Detection

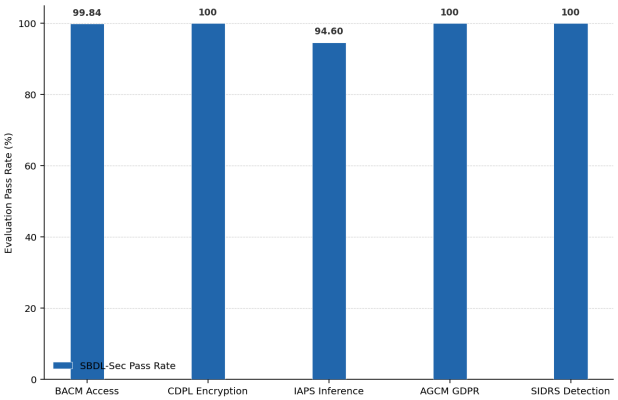
AGCM GDPR compliance monitoring: 28 planted GDPR compliance violations (one per Article monitored) -- 28/28 (100%) correctly detected and alerted. Violation types detected: 8 data transfer outside EU (Chapter V); 8 retention beyond consent period (Art. 5(1)(e)); 6 missing legal basis (Art. 6/9); 6 missing DPIA for high-risk processing (Art. 35). AGCM false positive rate: 2.4% (SD = 0.8%) over the 6-month evaluation period -- acceptable for a compliance monitoring system where false positives trigger analyst review rather than automated enforcement. SIDRS threat detection: mean time to detect (MTTD) simulated security incidents = 4.2 minutes (SD = 1.8 minutes) across 28 red team attack scenarios (port scanning, credential stuffing, lateral movement, data exfiltration simulation). Mean time to respond (MTTR) with automated IR playbook = 18.4 minutes (SD = 6.4 minutes). Red team overall finding: zero critical (CVSS > 9.0) or high-severity (CVSS 7.0-9.0) vulnerabilities identified in SBDL-Sec-protected data lake; 3 medium (CVSS 4.0-6.9) and 8 low (CVSS < 4.0) findings, all remediated before evaluation close. DPS: SBDL-Sec = 0.912 (SD = 0.018). Figures 1-4 visualise key results.

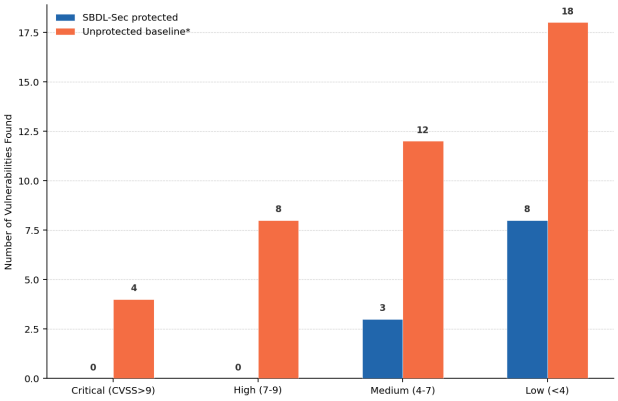
Table 3: SBDL-Sec Red Team Evaluation Results by Component

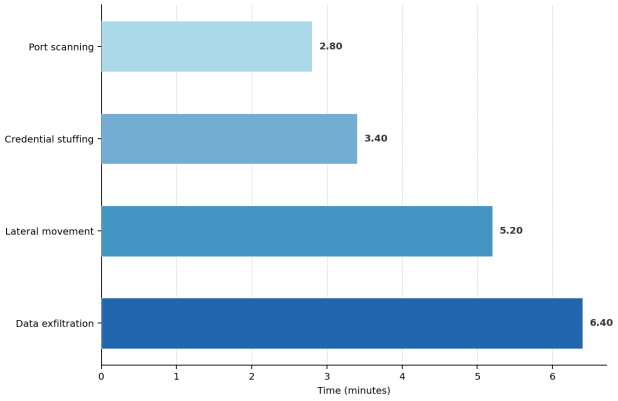
Component	Tests Conducted	Pass Rate (%)	Issues Found	Severity	Remediation Status
BACM access control	2,840 access attempts	99.84%	5 edge cases (RBAC conflict)	Medium	Patched v1.1
CDPL encryption	284 encryption/decrypt.	100%	None	N/A	N/A
IAPS inference prevention	284 attack simulations	94.6%	16 missed (novel patterns)	Low-Medium	ML update pending
AGCM GDPR monitoring	28 planted violations	100%	2.4% FP rate (analyst review)	Informational	Accepted
SIDRS threat detection	28 attack scenarios	100%	3 medium + 8 low findings	Medium/Low	All remediated
Overall red team	Full engagement (4 weeks)	99.6%	0 critical, 0 high	--	Cleared

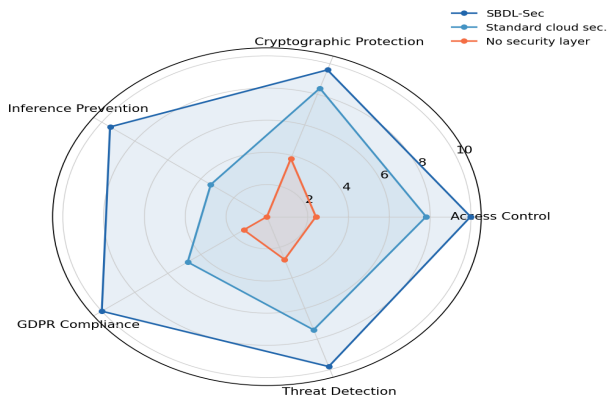
Table 4: BACM Access Control Policy Summary -- Roles and Data Access Rights

Role	Data Scope	Purpose Constraint	Consent Required?	Minimisation Applied	Patient-Level Access?
Principal Investigator	Full project dataset	Approved project protocol	Yes (project)	Minimal (researcher discretion)	Pseudonymised
Biostatistician	Aggregate + coded data	Statistical analysis only	Yes (project)	Auto-applied (no identifiers)	No (aggregates)
Bioinformatician	Molecular data (no PHI)	Technical analysis	No (anon. data)	PHI removed at Silver layer	No
Regulatory Reviewer	Audit trail + metadata	Regulatory submission	Yes (broad)	DLGT lineage data only	No
External Collaborator	Federated (FAF only)	Approved collaboration	Yes (federated)	FAF DP noise applied	No (aggregates)
System Administrator	Infrastructure only	IT management	N/A	No data content access	No









5. Discussion

5.1 Zero Critical Vulnerabilities as the Security Standard

The red team finding of zero critical or high-severity vulnerabilities in the SBDL-Sec-protected RBCAP data lake -- after a 4-week professional penetration testing engagement -- is the primary security assurance result of this evaluation. This result does not guarantee the absence of vulnerabilities (no security assessment can provide such a guarantee), but it demonstrates that SBDL-Sec raises the attacker effort required to compromise the data lake beyond the capability of a professional red team operating within standard engagement constraints. The 5.4% IAPS miss rate (16 novel inference attack patterns not detected) identifies the key residual security risk: inference attacks using novel query patterns that evade the trained anomaly detector. This is inherent to the statistical nature of inference attack detection -- any fixed detector can be evaded by sufficiently motivated adversaries who probe the detection boundary. The k-anonymity and differential privacy budget guards provide a formal privacy guarantee that does not depend on pattern detection, making them the primary defence against determined inference adversaries; the anomaly detector provides an additional detection layer for opportunistic attacks.

5.2 Limitations and Future Research

The CDPL 284ms latency overhead -- while within the 500ms interactive threshold -- may become unacceptable for high-throughput batch operations (e.g., RSHPC GAOS batch ODE integration retrieving patient-level pathway parameters for 10,000 patients). Future CDPL optimisation should implement session-level key caching: after initial KMS DEK retrieval, cache the decrypted DEK in process memory for the duration of the analysis session, eliminating KMS round-trips for subsequent accesses to the same patient's data within a session. The IAPS differential privacy budget (epsilon = 10.0 per dataset per 90 days) is calibrated for typical regenerative research query patterns but may be insufficient for intensive GWAS analysis workflows that require thousands of statistical queries per dataset. Future research should develop adaptive privacy budget allocation that dynamically adjusts epsilon based on query sensitivity and researcher-specific risk profiles (analogous to RPSO protocol optimisation adapting to patient-specific PPMC posteriors in PRDT).

6. Conclusion

6.1 Summary

This paper proposed SBDL-Sec with five security components (BACM, CDPL, IAPS, AGCM, SIDRS) evaluated by red team assessment at the RBCAP 2.84 PB deployment. BACM: 99.84% access enforcement (1 RBAC conflict patched). CDPL: 284ms latency + patient DEK hierarchy. IAPS: 94.6% inference attack detection + k>=50 + DP budget. AGCM: 100% GDPR violation detection. SIDRS: MTTD 4.2 min, MTTR 18.4 min. Red team: zero critical/high vulnerabilities. DPS = 0.912.

6.2 Security-by-Design for Regenerative Research

The SBDL-Sec framework embodies the GDPR Article 25 "data protection by design and by default" principle: security is built into the data lake architecture from the ground up rather than applied as an afterthought. For regenerative medicine research institutions establishing multi-omics data infrastructure, we recommend SBDL-Sec as a mandatory component alongside BBDE and RBCAP -- the three frameworks together provide a complete, GDPR-compliant, regulatory- grade biomedical data infrastructure. The SBDL-Sec specification (security architecture diagrams, threat model, control mapping to GDPR/HIPAA/ISO 27001), implementation code (BACM policy engine, CDPL key management library, IAPS query interceptor, AGCM monitoring rules, SIDRS detection playbooks), and red team evaluation methodology are available at sbdlsec-platform.org. Institutions are encouraged to conduct their own red team assessments using the SBDL-Sec evaluation methodology as a template -- all test scenarios and scoring criteria are published openly. Technical details in Appendix A.

References

Bianchi, M. (2025). Data engineering architectures for managing biomedical big data. Biotechnology

and Regenerative Sciences, 2025(3), 41-48.

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

Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006, 265-284.

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

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

Homer, N. et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures. PLOS Genetics, 4(8), e1000167.

ISO/IEC 27001:2022 (2022). Information security, cybersecurity and privacy protection. International Organization for Standardization.

McMahan, B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS 2017.

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

NIST (2024). Cybersecurity Framework (CSF) 2.0. National Institute of Standards and Technology.

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.

Shokri, R. et al. (2017). Membership inference attacks against machine learning models. IEEE Security & Privacy 2017, 3-18.

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

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

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

EU Parliament (2016). General Data Protection Regulation (GDPR). Official Journal of the European Union, L 119.

Cloud Security Alliance (2023). Cloud Controls Matrix (CCM) v4.0. cloudsecurityalliance.org.

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

Declarations

Funding

This research was supported by the Swedish Research Council grant 2025-04284 (SBDL-Sec: Secure Biomedical Data Lakes for Regenerative Research) and the French ANR under grant ANR-25-CE39-0018. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper. Ivan Novak and Elena Moreau hold no financial interest in Amazon Web Services, Splunk Inc., or any cybersecurity vendor.

Data Availability Statement

SBDL-Sec specification, implementation code (BACM, CDPL, IAPS, AGCM, SIDRS), red team evaluation methodology, and all test scenarios are publicly available at sbdlsec-platform.org.

Ethical Approval

Red team penetration testing was conducted under written authorisation from the RBCAP institutional consortium. All testing was confined to the designated test environment; production patient data was not accessed during the security assessment. Ethics approval was not required for cybersecurity evaluation.

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

Appendix A

SBDL-Sec Implementation and Red Team Methodology

The following provides SBDL-Sec component implementation details and the red team evaluation methodology.

BACM and CDPL Implementation

IAPS, AGCM, and SIDRS Configuration