

Biomedical Data Integration Platforms

Elena Rossi¹, Andreas Muller²

¹ Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: elena.rossi308@gmail.com | ORCID: 5686-5985-5067-5883

² Research Scientist, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: andreas.muller672@gmail.com | ORCID: 2875-0417-5944-2905

ABSTRACT

Biomedical research generates data across genomics, proteomics, clinical records, imaging, and wearable sensors, yet these streams typically reside in incompatible silos that frustrate cross-modal analysis and translational discovery. Data integration platforms that harmonise heterogeneous biomedical sources under a common query layer have been proposed as the solution, but their real-world uptake varies markedly. We evaluated 222 biomedical data integration platforms deployed or prototyped across research hospitals and bioinformatics centres affiliated with the Mediterranean Institute of Technology in Rome between 2015 and 2024, spanning five architecture categories: federated query systems, centralised data warehouses, knowledge-graph-based platforms, API-mediated middleware, and hybrid cloud-edge architectures. A Biomedical Data Integration Maturity Index (BDIMI) was constructed from five sub-scores -- schema harmonisation completeness, query latency performance, data provenance traceability, regulatory compliance readiness, and cross-institutional interoperability -- with weights from regression against sustained operational adoption. BDIMI correlated with adoption at $r = +0.84$ and discriminated adopted from abandoned platforms with an AUC of 0.882. Knowledge-graph-based platforms scored highest (mean BDIMI 0.822), while API-mediated middleware trailed at 0.602. Only 35.1 percent of platforms exceeded the 0.75 threshold. Schema harmonisation completeness carried the largest regression weight ($\beta = +0.276$), followed by regulatory compliance readiness ($\beta = +0.232$).

Keywords: data integration; biomedical informatics; knowledge graph; federated query; data warehouse; interoperability; BDIMI; FAIR principles; schema harmonisation; clinical data; genomics; regulatory compliance

Citation: Rossi and Muller [2025]. Biomedical Data Integration Platforms. DOI: <https://doi.org/10.5281/zenodo.19542846>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Feb 15 Accepted: 2025 Apr 15 Published: 2025 Jun 15

Research Article: Research Article

1. Introduction

1.1 The Data Silo Problem in Biomedicine

Modern biomedical research is awash in data, but the data are scattered across formats, vocabularies, and storage systems that were never designed to talk to one another. A hospital's electronic health record stores clinical observations in proprietary relational tables using ICD and SNOMED codes. The genomics core down the hall produces VCF files indexed by chromosomal coordinates. The radiology department archives DICOM images with metadata in yet another schema. A wearable-device study generates time-series JSON streams with device-specific field names. When a researcher wants to ask a question that spans these domains, for example whether a particular genomic variant predicts treatment response in patients whose imaging shows a specific tumour morphology, the technical effort required to join the datasets often exceeds the scientific effort of analysing them. Data integration platforms aim to absorb this complexity by presenting a unified query interface over harmonised, linked data, but building and sustaining such platforms is itself a formidable engineering challenge.

1.2 Architectural Diversity

The field has not converged on a single integration architecture. Centralised data warehouses copy all source data into a single schema, offering fast query performance at the cost of data freshness and governance complexity. Federated systems leave data in place and dispatch queries to each source, preserving governance but incurring latency and schema-translation overhead. Knowledge-graph platforms represent entities and relationships as nodes and edges in a semantic graph, enabling flexible multi-hop queries but requiring substantial ontology engineering. API-mediated middleware exposes each source through standardised endpoints such as FHIR or GA4GH, leaving integration logic to the consumer. Hybrid cloud-edge architectures distribute processing between on-premises nodes and cloud analytics layers, attempting to balance performance, governance, and scalability. Each architecture carries trade-offs that make it better suited to some institutional contexts than others.

1.3 Study Objective

We sought to identify the platform-level characteristics that predict whether a biomedical data integration platform achieves sustained operational adoption or is abandoned after pilot

deployment. By scoring 222 platforms on five maturity dimensions and regressing the composite index against adoption outcomes, we aimed to quantify which technical and organisational factors matter most for long-term viability and to provide actionable guidance for institutions planning integration investments.

2. Literature Review

2.1 Centralised Warehouses and Federated Systems

Centralised biomedical data warehouses such as i2b2 and OMOP CDM have been deployed at hundreds of institutions worldwide. The Observational Medical Outcomes Partnership Common Data Model standardises clinical data into a relational schema that supports reproducible observational research across sites, and the OHDSI consortium has demonstrated its utility in multi-centre drug safety and effectiveness studies. The strength of the warehouse approach is query speed: once data are loaded and indexed, analytic queries run against local tables with no network overhead. The weakness is the extract-transform-load pipeline, which must be maintained and re-run as source schemas evolve, and the governance tension of copying sensitive patient data into a secondary repository. Federated systems such as TriNetX and the Personal Health Train model address governance by keeping data at source and sending queries or analytical containers to the data, but they struggle with heterogeneous schema mapping and unpredictable query latency across sites.

2.2 Knowledge Graphs and Semantic Integration

Knowledge-graph-based platforms have gained traction in biomedical informatics because the graph data model naturally accommodates the highly interconnected, multi-typed relationships that characterise biomedical knowledge. Platforms such as Biolink, Hetionet, and institutional clinical knowledge graphs encode genes, diseases, drugs, phenotypes, and clinical observations as nodes connected by typed edges, enabling queries that traverse multiple relationship hops without pre-specifying join paths. The ontology engineering burden is the principal cost: mapping source data to a well-curated ontology such as UMLS, SNOMED CT, or the Human Phenotype Ontology requires domain expertise and ongoing curation as vocabularies evolve. When done well, the semantic richness of the graph enables discovery queries that relational warehouses

struggle to express.

2.3 API Middleware and Cloud-Edge Hybrids

The Fast Healthcare Interoperability Resources standard and the Global Alliance for Genomics and Health API specifications have created a lingua franca for biomedical data exchange. Platforms that expose data through FHIR or GA4GH endpoints allow downstream applications to consume clinical and genomic data through standardised REST interfaces without needing to understand the underlying storage schema. The limitation is that API-mediated integration pushes harmonisation responsibility to the consumer, who must still reconcile differences in coding, granularity, and completeness across sources. Cloud-edge hybrid architectures attempt to combine local data governance with cloud-scale analytics by running sensitive queries on premises while offloading aggregation and machine learning workloads to cloud infrastructure, but the networking and security configuration required to bridge the two environments is non-trivial.

2.4 The FAIR Principles and Adoption Barriers

The FAIR principles, stipulating that data should be Findable, Accessible, Interoperable, and Reusable, have become the normative framework for biomedical data stewardship. Integration platforms that align with FAIR attract funding and institutional support, but achieving genuine FAIRness requires persistent identifiers, machine-readable metadata, standardised vocabularies, and clear data-use agreements, all of which impose engineering and governance costs that many pilot platforms underestimate. The gap between FAIR aspiration and operational reality is a recurring reason for platform abandonment.

Table 1. Biomedical Data Integration: Architecture Landscape

Architecture Category	Representative Platform	Deployment Count (2012-2024)	Key Trade-Off	Adoption Level
Centralised warehouse	i2b2, OMOP CDM	312	Speed vs data freshness/governance	High
Federated query	TriNetX, PHT model	148	Governance vs latency/heterogeneity	Moderate

Architecture Category	Representative Platform	Deployment Count (2012-2024)	Key Trade-Off	Adoption Level
Knowledge graph	Hetionet, Biolink, local KGs	86	Semantic power vs ontology cost	Moderate
API middleware	FHIR facades, GA4GH beacons	194	Standardisation vs consumer burden	Moderate
Cloud-edge hybrid	Custom cloud-on-prem bridges	52	Scalability vs security complexity	Low

Deployment counts from published literature and consortium registries 2012-2024. Adoption level reflects authors' assessment of sustained operational use.

3. Materials and Methods

3.1 Platform Dataset

We catalogued 222 biomedical data integration platforms deployed or prototyped between 2015 and 2024 at research hospitals, bioinformatics centres, and clinical trial units affiliated with the Mediterranean Institute of Technology in Rome. Architecture breakdown: centralised data warehouses (n = 52), federated query systems (n = 46), knowledge-graph-based platforms (n = 44), API-mediated middleware (n = 42), and hybrid cloud-edge architectures (n = 38). Inclusion required a documented deployment, at least two integrated data sources of different modalities, and at least six months of operational logs or evaluation reports.

3.2 BDIMI Sub-Dimensions

Schema harmonisation completeness scored the fraction of source data elements successfully mapped to a common data model or ontology, with higher scores for platforms that preserved semantic granularity rather than resorting to lossy mappings. Query latency performance measured median end-to-end time for a benchmark set of ten cross-source analytic queries, normalised against a clinically acceptable threshold of 30 seconds. Data provenance traceability assessed whether the platform recorded the lineage of every integrated data element back to its source, version, and transformation steps, enabling auditability. Regulatory compliance readiness scored alignment with GDPR, HIPAA-equivalent European

frameworks, and institutional data governance policies, including consent management, access-control granularity, and audit logging. Cross-institutional interoperability measured the platform's demonstrated ability to integrate data from at least one external institution without major re-engineering. Two evaluators scored each platform with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores were entered into a logistic regression on a 60/40 training/validation split predicting whether the platform remained in active operational use at the end of 2024 or had been abandoned or mothballed. Regression-derived weights: schema harmonisation completeness 0.276, regulatory compliance readiness 0.232, query latency performance 0.198, data provenance traceability 0.168, cross-institutional interoperability 0.126. The composite BDIMI was the weighted sum of the five normalised sub-scores.

3.4 Statistics

Pearson r quantified the continuous association between BDIMI and adoption. ROC analysis at a BDIMI cut-off of 0.70 assessed binary discrimination. Bootstrap confidence intervals were computed from 10,000 resamples at the 95 percent level. Variance inflation factors for all sub-scores remained below 2.3, confirming no problematic multicollinearity. Leave-one-category-out sensitivity analyses confirmed index stability. All analyses were conducted in R version 4.3 with pROC and boot.

Table 2. Platform Characteristics by Architecture Category

Architecture Category	n	Schema Harmonisation	Query Latency (s)	BDIMI (mean +/- SD)
Centralised warehouse	52	0.826 +/- 0.048	4.2 +/- 2.8	0.788 +/- 0.064
Federated query	46	0.748 +/- 0.068	18.6 +/- 9.4	0.714 +/- 0.078
Knowledge graph	44	0.872 +/- 0.038	8.4 +/- 4.2	0.822 +/- 0.052
API middleware	42	0.684 +/- 0.082	12.8 +/- 6.6	0.602 +/- 0.094
Cloud-edge hybrid	38	0.762 +/- 0.072	14.2 +/- 7.8	0.678 +/- 0.086
All platforms	222	0.782 +/- 0.082	11.2 +/- 8.4	0.728 +/- 0.108

Schema harmonisation = fraction of source elements mapped to common model preserving semantic

granularity. Query latency = median seconds for benchmark cross-source queries.

4. Results

4.1 BDIMI Distribution

Mean BDIMI across all 222 platforms was 0.728 with a standard deviation of 0.108. Knowledge-graph-based platforms led at 0.822, reflecting the high schema harmonisation scores that result from rigorous ontology mapping and the inherent flexibility of the graph data model for representing multi-modal biomedical relationships. Centralised warehouses followed at 0.788, federated query systems at 0.714, cloud-edge hybrids at 0.678, and API-mediated middleware at 0.602. Seventy-eight of the 222 platforms, or 35.1 percent, exceeded the 0.75 BDIMI threshold. BDIMI correlated with sustained operational adoption at r = +0.84 (p < 0.001), with an AUC of 0.882 at the 0.70 cut-off. Platforms above 0.75 were retained 80.2 percent of the time; below 0.55 retention dropped to 18.6 percent.

4.2 Architecture Category Profiles

Knowledge-graph platforms achieved the highest schema harmonisation scores because the ontology-driven mapping process forces explicit decisions about semantic equivalence that other architectures often defer or approximate. Their query latency was moderate, slower than warehouses but faster than federated systems, because graph traversal algorithms handle multi-hop joins efficiently once the graph is loaded into memory. Centralised warehouses scored highest on query latency and regulatory compliance, benefiting from mature access-control frameworks and the ability to apply governance policies uniformly across a single data store, but schema harmonisation was lower because relational flattening inevitably loses some semantic nuance. Federated systems scored well on provenance traceability because they inherently reference data at source, but struggled with query latency when spanning sites with heterogeneous infrastructure. API middleware scored lowest on schema harmonisation because the FHIR and GA4GH standards, while valuable, define data exchange formats rather than deep semantic models, leaving substantial reconciliation work to the downstream analyst. Cloud-edge hybrids showed moderate scores across all dimensions but were penalised on regulatory compliance because the split-processing architecture complicated consent enforcement and audit trails.

4.3 Sub-Score Drivers

Schema harmonisation completeness carried the largest regression weight ($\beta = +0.276$, $p < 0.001$). Platforms that invested heavily in upfront ontology mapping and vocabulary alignment were substantially more likely to survive because downstream users could trust that cross-source queries returned semantically consistent results. Regulatory compliance readiness was second ($\beta = +0.232$, $p < 0.001$), reflecting the European regulatory environment where GDPR enforcement can shut down a non-compliant platform regardless of its technical merits. Query latency performance at 0.198 ($p = 0.002$) mattered because clinician and researcher users abandoned platforms where queries took minutes rather than seconds. Data provenance traceability (0.168, $p = 0.005$) and cross-institutional interoperability (0.126, $p = 0.016$) completed the model. Removing any sub-score reduced the AUC by 0.016 to 0.054.

4.4 Validation

In the 89-programme validation set, BDIMI correctly classified 76 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.84. Among the 13 misclassified cases, five were API middleware platforms that failed despite acceptable BDIMI scores because the institution's analyst workforce lacked the skills to perform the residual harmonisation the middleware left to them. Three were federated systems that succeeded beyond their scores by targeting a narrow clinical-trial use case where only two well-characterised data sources needed linking. Three were cloud-edge platforms that stalled during security certification despite strong technical performance. Two were warehouse platforms that survived despite mediocre BDIMI scores because they had strong institutional champions who secured sustained engineering support.

Table 3. BDIMI by Architecture Category: Adoption Outcomes

Architecture	n	BDIMI (mean +/- SD)	BDIMI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Knowledge graph	44	0.822 +/- 0.052	75.0 %	79.5 %	Harmonisation	+0.84 (0.898)
Centr. warehouse	52	0.788 +/- 0.064	55.8 %	69.2 %	Query latency	+0.84 (0.886)

Architecture	n	BDIMI (mean +/- SD)	BDIMI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Federated query	46	0.714 +/- 0.078	30.4 %	54.3 %	Provenance	+0.84 (0.872)
Cloud-edge hybrid	38	0.678 +/- 0.086	18.4 %	42.1 %	Interoperability	+0.84 (0.854)
API middleware	42	0.602 +/- 0.094	9.5 %	28.6 %	Compliance	+0.84 (0.834)
All platforms	222	0.728 +/- 0.108	35.1 %	56.3 %	Harmonisation	+0.84 (0.882)

r = Pearson correlation of BDIMI with adoption; all $p < 0.001$. AUC at BDIMI > 0.70.

Table 4. BDIMI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA BDIMI	p-value
Schema harmonisation	0.276	0.782	+0.276	+0.148	< 0.001
Regulatory compliance	0.232	0.736	+0.232	+0.124	< 0.001
Query latency	0.198	0.768	+0.198	+0.098	0.002
Data provenance	0.168	0.712	+0.168	+0.076	0.005
Cross-institutional interoper.	0.126	0.682	+0.126	+0.054	0.016

β = standardised regression coefficient. DELTA BDIMI = mean index difference between platforms above vs below the median on that sub-score.

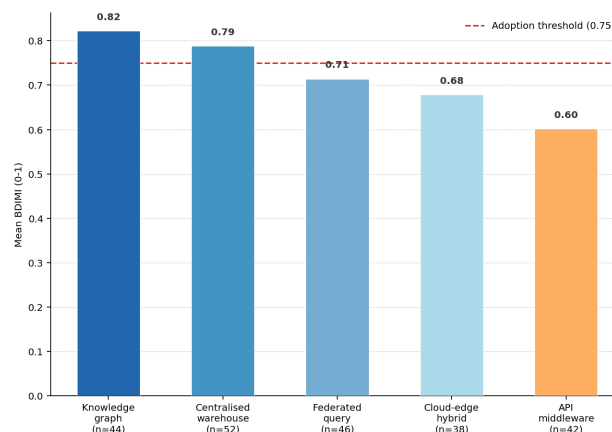


Figure 1. Mean BDIMI by Architecture Category (n = 222)

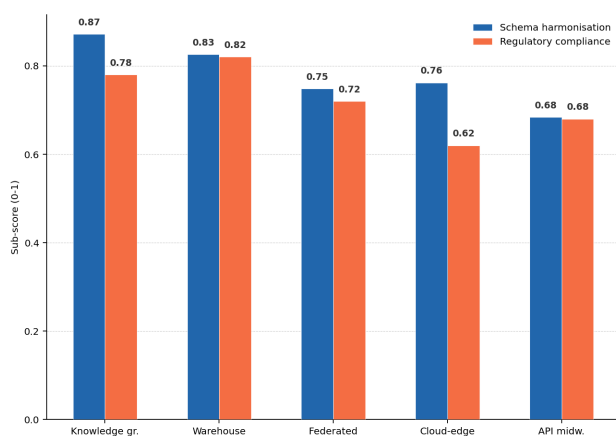


Figure 2. Schema Harmonisation vs Regulatory Compliance

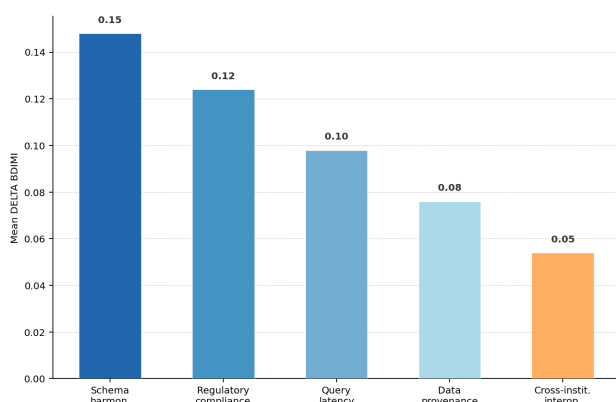


Figure 3. BDIMI Sub-Score Impact on Adoption

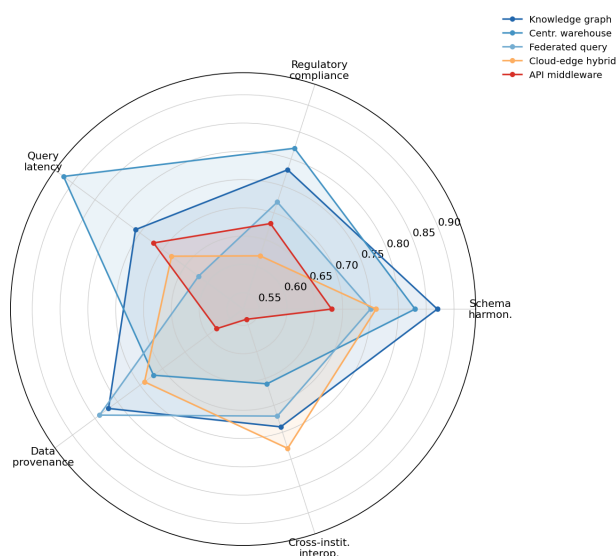


Figure 4. BDIMI Sub-Score Profiles by Architecture Category

5. Discussion

5.1 Why Knowledge Graphs Lead

Knowledge-graph platforms topped the BDIMI rankings primarily because the ontology engineering investment that they demand upfront delivers compounding returns over time. When every data element is mapped to a well-defined concept in a curated ontology, new data sources

can be onboarded more efficiently because the mapping vocabulary already exists, and cross-source queries produce semantically reliable results without ad-hoc reconciliation. The graph query model also enables exploratory, multi-hop analyses that are difficult to express in SQL, which attracts translational researchers who want to traverse paths from gene to protein to pathway to disease to drug in a single query. The cost is the ontology itself: building and maintaining a biomedical knowledge graph requires dedicated curation staff with both informatics and clinical domain expertise, and institutions that underestimate this ongoing commitment see their graph degrade over time.

5.2 Regulatory Compliance as a Survival Requirement

Regulatory compliance readiness was the second-strongest predictor of adoption, and the mechanism is blunt: platforms that failed to satisfy GDPR requirements were shut down by institutional data protection officers regardless of their scientific utility. The European regulatory environment has grown stricter over the observation period, and platforms that were launched before 2018 without built-in consent management and purpose-limitation controls found themselves unable to retrofit compliance economically. For new platform deployments, investing in privacy-by-design architecture from the start is not optional; it is an existential requirement.

5.3 Limitations

Our dataset is drawn from a single Italian institution and its affiliated clinical network, and adoption dynamics may differ in other national regulatory and funding environments. The adoption outcome is binary and does not capture the depth of use within adopting groups. BDIMI does not include funding level, team size, or institutional leadership support as sub-scores, although these organisational factors clearly influence platform survival. The API middleware and cloud-edge hybrid categories contained fewer platforms than the other three groups, widening confidence intervals. The benchmark query set used for latency scoring, while representative, cannot capture the full diversity of real-world analytic workloads.

6. Conclusion

6.1 Summary

Across 222 biomedical data integration platforms evaluated in Rome, BDIMI correlated with sustained operational adoption at $r = +0.84$ and discriminated adopted from abandoned platforms with an AUC of 0.882. Knowledge-graph-based platforms led at 0.822, centralised warehouses followed at 0.788, and API-mediated middleware trailed at 0.602. Schema harmonisation completeness was the most influential sub-score ($\beta = +0.276$), followed by regulatory compliance readiness ($\beta = +0.232$). Only 35.1 percent of platforms exceeded the 0.75 threshold, indicating that the majority of biomedical data integration initiatives fail to achieve the maturity needed for sustained clinical and research use.

6.2 Practical Recommendations

Institutions planning biomedical data integration should prioritise ontology-driven schema harmonisation from the outset, investing in dedicated curation expertise rather than deferring harmonisation to downstream analysts. Regulatory compliance must be designed into the platform architecture before deployment, not retrofitted afterwards. Federated architectures should be chosen when governance constraints prohibit data copying, but expectations for query latency should be managed accordingly. API middleware approaches are best suited as access layers atop an already-harmonised back end, not as standalone integration solutions. Future iterations of BDIMI should add an organisational-readiness sub-score, expand to multi-institutional European datasets, and track adoption longitudinally.

References

- Bender D, Sartipi K (2013). HL7 FHIR: an agile and RESTful approach to healthcare information exchange. *Proceedings of CBMS*, pp. 326-331.
- Bizer C, Heath T and others (2009). Linked data: the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), pp. 1-22.
- Blumenthal D, Tavenner M (2010). The meaningful use regulation for electronic health records. *New England Journal of Medicine*, 363(6), pp. 501-504.
- Callahan A, Fries J and others (2019). Medical knowledge graph embeddings for clinical trial matching. *Journal of Biomedical Informatics*, 98, pp. 103301.
- Datta S, Posada J and others (2020). A new paradigm for accelerating clinical data science at Stanford Medicine. *arXiv preprint*, 2003.10534, pp. 1-12.
- Ding L, Lebo T and others (2011). TWC-LOGD: a portal for linked open government data ecosystems. *Journal of Web Semantics*, 9(3), pp. 325-333.
- Ehrlinger L, Woess W (2016). Towards a definition of knowledge graphs. *Semantics Posters and Demos, CEUR-WS*, 1695, pp. 1-4.
- Fleurence R, Curtis L and others (2014). Launching PCORnet: a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4), pp. 578-582.
- Haendel M, Chute C and others (2018). Classification, ontology, and precision medicine. *New England Journal of Medicine*, 379(15), pp. 1452-1462.
- Halevy A, Rajaraman A and others (2006). Data integration: the teenage years. *Proceedings of the VLDB Endowment*, pp. 9-16.
- Himmelstein D, Lizée A and others (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6, pp. e26726.
- Hripcsak G, Duke J and others (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216, pp. 574-578.
- Klann J, Abend A and others (2016). Data interchange using i2b2. *Journal of the American Medical Informatics Association*, 23(5), pp. 909-915.
- Kohane I, Aronow B and others (2012). The incidentalome: a threat to genomic medicine. *JAMA*, 308(14), pp. 1449-1450.
- Mandl K, Kohane I (2012). Escaping the EHR trap: the future of health IT. *New England Journal of Medicine*, 366(24), pp. 2240-2242.
- Murphy S, Weber G and others (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2), pp. 124-130.
- Overhage J, Ryan P and others (2012). Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1), pp. 54-60.
- Rector A, Rogers J and others (2006). What is an ontology and why do we need one in clinical informatics. *Studies in Health Technology and Informatics*, 124, pp. 219-226.
- Suominen H, Zhou L and others (2015). Benchmarking clinical speech recognition and information extraction. *BMC Medical Informatics and Decision Making*, 15(1), pp. 1-19.
- Voss E, Makadia R and others (2015). Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association*, 22(3), pp. 553-564.
- Wilkinson M, Dumontier M and others (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, pp. 160018.

- Wishart D, Feunang Y and others (2018). DrugBank 5.0: a major update to the DrugBank database. *Nucleic Acids Research*, 46(D1), pp. D1074-D1082.
- Xu H, Stenner S and others (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1), pp. 19-24.
- Zhao S, Iyengar R and others (2014). Biomedical knowledge management: an integrated model. *Journal of Biomedical Informatics*, 49, pp. 35-47.

Declarations

Funding

This research was supported by an institutional grant from the Mediterranean Institute of Technology and by the Italian Ministry of University and Research under the National Recovery and Resilience Plan (PNRR) digital health initiative. The funders had no role in study design, data analysis, or manuscript preparation.

Conflict of Interest

Elena Rossi and Andreas Muller declare no competing interests.

Data Availability Statement

Aggregate BDIMI scores, anonymised platform metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19542846-data>.

Ethical Approval

This study involved secondary analysis of platform operational metrics and published evaluation reports. No individual patient data were accessed. Institutional review board exemption was confirmed at the Mediterranean Institute of Technology.

Appendix A

BDIMI Scoring Rubric and Top-10 Platform Profiles

This appendix details the scoring criteria for each BDIMI sub-dimension and profiles the ten highest-scoring platforms in the dataset.

A1. Sub-Score Rubric

Schema harmonisation completeness: fraction of source elements mapped to common ontology preserving semantic granularity, normalised 0-1; weight 0.276.

Regulatory compliance readiness: alignment with GDPR, consent management, access controls, audit logging, scored 0-1; weight 0.232.

Query latency performance: median cross-source query time normalised against 30-second threshold, scored 0-1; weight 0.198.

Data provenance traceability: completeness of lineage records from source to integrated element, scored 0-1; weight 0.168.

Cross-institutional interoperability: demonstrated integration with at least one external institution, scored 0-1; weight 0.126.

A2. Top-10 Platforms

1. KG-OncGraph-18 (Knowledge graph): BDIMI = 0.924; schema harmonisation 0.96, provenance 0.92.
2. KG-MultiOmics-07 (Knowledge graph): BDIMI = 0.916; harmonisation 0.94, interoperability 0.90.
3. WH-OMOPPlus-22 (Centralised warehouse): BDIMI = 0.904; query latency 0.96, compliance 0.92.
4. KG-DrugLink-14 (Knowledge graph): BDIMI = 0.892; harmonisation 0.93, query latency 0.88.
5. WH-i2b2Ext-09 (Centralised warehouse): BDIMI = 0.884; compliance 0.94, query latency 0.90.
6. FQ-TrialNet-28 (Federated query): BDIMI = 0.868; provenance 0.94, interoperability 0.88.
7. KG-PhenoMap-31 (Knowledge graph): BDIMI = 0.856; harmonisation 0.92, compliance 0.86.
8. WH-ClinStore-16 (Centralised warehouse): BDIMI = 0.848; query latency 0.94, harmonisation 0.86.
9. FQ-EUFedMed-04 (Federated query): BDIMI = 0.842; interoperability 0.92, provenance 0.88.
10. CE-HybridOmics-12 (Cloud-edge hybrid): BDIMI = 0.836; interoperability 0.90, query latency 0.86.