Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

A PHARMACEUTICAL DATA MESH THAT PRESERVES OWNERSHIP WHILE ENABLING SEMANTIC INTEROPERABILITY, MODEL DEVELOPMENT, AND EVIDENCE REUSE

William Davis[1]*, Anna Kowalski[2], Hannah Brown[3], Jeffrey Collins[4]

1. ***Department of Pharmaceutical Data Mesh and Ownership, College of Pharmacy, Pennsylvania State University, University Park, United States.***
2. ***Department of Semantic Interoperability and Model Development, Faculty of Pharmaceutical Sciences, University of Wrocław, Wrocław, Poland.***
3. ***Department of Evidence Reuse and Data Governance, Faculty of Pharmacy, McGill University, Montreal, Canada.***
4. ***Department of Federated Data Architecture, Faculty of Pharmacy, University of Sydney, Sydney, Australia.***

ARTICLE INFO

Received:
11 March 2025
Received in revised form:
19 June 2025
Accepted:
21 June 2025
Available online:
28 June 2025

Keywords: Pharmaceutical data architecture, Data mesh, Domain ownership, Data products, Semantic interoperability, Evidence lineage

ABSTRACT

Pharmaceutical research increasingly depends on the combined interpretation of chemical, biological, preclinical, clinical, safety, manufacturing, and real-world evidence. These data are commonly distributed across scientific functions, technology platforms, legal jurisdictions, and stages of the product lifecycle. Centralized architectures attempt to overcome this fragmentation by consolidating data into shared repositories, but consolidation alone does not preserve scientific context, resolve semantic incompatibility, establish fitness for reuse, or reconcile competing ownership and access obligations. This original pharmaceutical data architecture article develops a proposed ownership-preserving pharmaceutical data mesh in which scientific domains remain accountable for versioned data products while participating in federated semantic, lineage, access, computational, and governance services. The approach integrates evidence from pharmaceutical data science, FAIR data practice, biomedical interoperability, distributed computation, provenance engineering, model governance, and translational implementation. Its principal contribution is a conceptual architecture that distinguishes physical data location from semantic interoperability and separates data availability from scientific suitability. Domain-owned data products are connected through explicit semantic contracts, distributed evidence-lineage records, federated stewardship, policy-aware access control, governed compute-to-data services, and qualified model-reuse mechanisms. The architecture is evaluated through multiple dimensions rather than a single performance measure, including semantic validity, lineage completeness, policy correctness, reproducibility, model transportability, evidence-reuse fidelity, and organizational sustainability. The proposal remains conceptual: it does not establish privacy guarantees, empirical superiority, regulatory acceptability, clinical utility, or deployment readiness. Its value lies in organizing the scientific and governance conditions under which pharmaceutical evidence may become reusable without being detached from its ownership, provenance, uncertainty, and intended context.

To Cite This Article: Davis W, Kowalski A, Brown H, Collins J. A Pharmaceutical Data Mesh That Preserves Ownership while Enabling Semantic Interoperability, Model Development, and Evidence Reuse. Pharmacophore. 2025;16(3):96-105. https://doi.org/10.51847/eFLYQytugp

Introduction

Artificial intelligence and computational modeling now contribute to activities extending from target identification and molecular design to biomarker development, safety assessment, clinical development, and evidence generation. These activities rely on heterogeneous chemical structures, assay measurements, omics profiles, preclinical observations, clinical variables, manufacturing records, and external knowledge resources. Their value does not arise merely from data volume. Pharmaceutical interpretation depends on how observations were generated, which entities they describe, which experimental or clinical context applies, what transformations occurred, and which limitations accompany the evidence. Reviews of machine-learning applications across pharmaceutical research therefore depict a distributed evidentiary landscape rather than

Corresponding Author: William Davis; Department of Pharmaceutical Data Mesh and Ownership, College of Pharmacy, Pennsylvania State University, University Park, United States. E-mail: william.davis@psu.edu.

a single homogeneous analytical resource [1]. A data architecture intended to support this landscape must coordinate evidence without erasing the scientific distinctions on which valid interpretation depends.

The expansion of deep learning has intensified this architectural requirement. Learned representations can support molecular property prediction, virtual screening, de novo design, and other discovery tasks, but their performance is conditional on the composition, quality, scale, representation, and task definition of the underlying data [2]. A model may reproduce a benchmark label while relying on hidden dataset artifacts, incompatible assay definitions, duplicated compounds, unrecognized temporal leakage, or a population that differs from the intended application. Consequently, predictive performance cannot serve as a sufficient measure of architectural success. The architecture must also make it possible to establish which data were used, how they were represented, whether they were suitable for the task, which uncertainty remains, and whether reuse crosses a scientifically meaningful boundary. The central problem is therefore not simply how to make more pharmaceutical data available to more models, but how to make appropriately governed evidence reusable without allowing access convenience to substitute for scientific validity.

FAIR data principles provide an important foundation because they emphasize that data and related digital objects should be findable, accessible under defined conditions, interoperable, and reusable. In biopharmaceutical research, however, FAIR implementation requires coordinated metadata practices, standards, identifiers, repositories, services, responsibilities, and organizational processes rather than the introduction of a repository alone [3]. A centralized platform may make files technically discoverable while leaving experimental context incomplete, terminology inconsistent, ownership unclear, access decisions repetitive, and downstream transformations insufficiently traceable. Conversely, a highly controlled local repository may preserve context but prevent legitimate cross-domain analysis. The unresolved design problem is to combine local scientific accountability with enterprise-level interoperability and governed reuse without assuming that every dataset should be physically consolidated or interpreted through a universal schema.

Drug-discovery datasets illustrate why this problem is simultaneously semantic, scientific, and organizational. Relevant resources differ in identifiers, entity definitions, relation types, evidence coverage, licensing conditions, curation procedures, and update frequencies, while knowledge-graph integration can expose relationships only to the extent that these differences are explicitly represented and managed [4]. This article therefore proposes an ownership-preserving pharmaceutical data mesh as an original conceptual architecture. The construct treats scientific domains as accountable producers of versioned data products and connects those products through federated stewardship, semantic contracts, distributed evidence lineage, policy-aware access control, compute-to-data services, and qualified model and evidence reuse. The objective is not to claim that a mesh is universally preferable to centralization. It is to define the conditions under which distributed ownership and shared interoperability services may be combined, specify how the resulting architecture should be evaluated, and establish the boundaries beyond which architectural availability cannot be interpreted as pharmaceutical readiness.

Why Centralized Pharmaceutical Data Architectures Struggle at Scale

Centralization is often presented as the direct remedy for fragmented pharmaceutical evidence: source data are copied into a shared lake, warehouse, platform, or knowledge environment, after which common analytics are expected to become easier. This strategy can provide valuable infrastructure, especially when harmonized datasets and shared computational services are required. Its limitations emerge when physical consolidation is treated as equivalent to legitimate, context-preserving reuse. Pharmaceutical data may contain sensitive patient information, proprietary molecular structures, unpublished assay results, manufacturing knowledge, partner-controlled evidence, or jurisdiction-specific restrictions. Aggregation and linkage can increase the possibility of re-identification and can separate secondary uses from the expectations, permissions, and governance conditions under which the data were collected [5]. A central repository may therefore simplify storage while concentrating privacy, confidentiality, authorization, and accountability risks.

Access friction is also not reducible to the location of the files. Cross-institutional research frequently depends on data-use agreements that define permitted purposes, responsibilities, security controls, publication rights, derived-data ownership, and liability. The negotiation and execution of these agreements can delay research even where the technical transfer itself is straightforward [6]. Within pharmaceutical organizations, analogous friction occurs across therapeutic areas, development programs, acquired companies, external partnerships, and regulated or non-regulated environments. Copying data into an enterprise platform does not automatically resolve whether a proposed user is authorized, whether the requested purpose is compatible with the original agreement, whether a derived model may be shared, or whether the receiving environment satisfies applicable controls. When these conditions remain external to the data architecture, access decisions must be reconstructed repeatedly and may become inconsistent, overly permissive, or unnecessarily restrictive.

A second scaling problem concerns semantic heterogeneity. Molecularly driven research requires integration across clinical, genomic, phenotypic, pharmacological, and operational data, but interoperability depends on harmonized representations and common data models rather than colocation alone [7]. The same compound may appear under registry identifiers, internal codes, salt forms, stereochemical variants, batch-specific identifiers, or structure-derived keys. Assay endpoints may share a label while differing in protocol, matrix, detection technology, normalization, units, or biological interpretation. Clinical concepts can vary across coding systems and study designs, while manufacturing variables may be meaningful only within a particular process, site, equipment train, or method version. Centralized pipelines commonly address this heterogeneity through repeated extraction, transformation, and harmonization. At scale, those pipelines can detach transformations from the

domain experts who understand their assumptions, create parallel mappings for similar concepts, and propagate silent semantic changes into downstream models.

Distributed computation demonstrates that collaboration does not always require raw-data pooling, but it also shows why decentralization cannot be treated as a complete solution. Federated learning can coordinate model development across institutions while data remain local, yet its reliability depends on data comparability, participant governance, security assumptions, validation design, communication protocols, and the handling of non-identically distributed data [8]. The architectural implication is that centralized and distributed patterns should not be evaluated through a single measure such as query speed, storage cost, model accuracy, or the number of integrated datasets. A scalable pharmaceutical architecture must instead be assessed across ownership preservation, semantic validity, evidence traceability, policy correctness, computational reproducibility, model transportability, and organizational maintainability. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why centralized pharmaceutical data architectures struggle at scale within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for why centralized pharmaceutical data architectures struggle at scale

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Physical consolidation	Does moving data into one environment preserve the scientific meaning and conditions of reuse?	Data retain dependencies on source systems, study designs, assays, populations, transformations, and governance conditions.	Distinguishes storage topology from pharmaceutical interoperability.	Assuming that a central repository creates a coherent evidence base.	Colocation can support access and computation but does not establish semantic or scientific comparability.
Privacy and secondary use	Can aggregated evidence be reused without violating confidentiality, consent, purpose, or jurisdictional constraints?	Aggregation, linkage, and secondary use can amplify privacy and governance risks [5].	Establishes why custody and purpose restrictions must remain visible within the architecture.	Interpreting controlled storage as a privacy guarantee.	Architectural controls may reduce defined risks but cannot eliminate re-identification, misuse, or legal uncertainty.
Contractual and institutional access	Are authorization, permitted purpose, derived-data rights, and institutional responsibilities represented in reusable form?	Data-use agreements can create substantial legal and administrative delays [6].	Supports policy-aware access services and reusable decision records.	Treating technical access as equivalent to legitimate access.	Automated policy evaluation must remain subject to applicable agreements, legal review, and accountable exception handling.
Semantic heterogeneity	Do common labels, identifiers, and fields carry equivalent scientific meaning across domains?	Molecular and clinical integration requires harmonized representations and common data models [7].	Motivates explicit semantic contracts rather than implicit schema alignment.	Silent equivalence among compounds, assays, endpoints, populations, or process variables.	Semantic alignment supports interpretation but does not establish biological equivalence, mechanism, or causality.
Central harmonization pipelines	Who owns mappings, transformation assumptions, quality checks, and compatibility decisions?	Repeated central transformation can separate decisions from domain knowledge and create duplicated or opaque mappings.	Connects the limits of centralization to the proposed domain-product model.	Treating standardized output as evidence that transformations are correct or universally reusable.	Harmonized products require versioned assumptions, accountable owners, and context-specific validation.
Distributed computation	Can approved analyses be performed without default transfer of raw data?	Federated learning provides a precedent for coordinated computation over locally retained data [8].	Supports compute-to-data as one architectural service within the proposed mesh.	Equating federation with privacy, security, or model validity.	Distributed execution requires task-specific threat models, comparability assessment, reproducibility controls, and external validation.
Multi-dimensional architectural evaluation	Which properties must be demonstrated before the architecture can support qualified reuse?	Pharmaceutical usefulness depends on scientific meaning, governance, provenance, model behavior, and operating context.	Rejects optimization around a single technical or predictive metric.	Creating a composite readiness score that conceals failure in a critical dimension.	No single performance measure can establish that an architecture is scientifically, organizationally, or regulatorily fit for use.

Domain Ownership, Data Products, and Federated Stewardship

Domain ownership is proposed here as accountable scientific responsibility rather than unrestricted control over isolated data silos. A domain is a coherent area of pharmaceutical work—such as discovery biology, medicinal chemistry, preclinical pharmacology, clinical development, pharmacometrics, pharmacovigilance, or manufacturing—in which teams possess the expertise required to interpret how evidence was generated and what its limitations mean. The domain owner is responsible for maintaining the scientific identity, context, quality characterization, version history, and support pathway of a data product. Enterprise interoperability still requires shared standards, repositories, and policies, but community-based registries demonstrate that these resources are interdependent and benefit from explicit stewardship rather than unmanaged proliferation

[9]. In the proposed architecture, ownership therefore remains local enough to preserve interpretation while stewardship is federated enough to maintain discoverability, compatibility, and enterprise obligations.

A pharmaceutical data product is defined as a versioned, discoverable, documented, quality-characterized, and access-governed package of data and interfaces prepared for specified forms of reuse. It may expose files, query endpoints, graph relations, feature sets, analytical views, or controlled compute functions, but it must also expose the metadata needed to judge what the payload means. Minimum product content includes an accountable owner, scientific context, canonical identifiers, schema, units, provenance, quality profile, missingness description, access conditions, update policy, compatibility information, and a statement of intended and excluded uses. FAIRness metrics provide useful ways to assess whether digital objects are discoverable, accessible, interoperable, and reusable, but such metrics do not establish whether a product is scientifically suitable for a particular pharmaceutical question [10]. A product can be highly findable and technically interoperable while containing biased observations, incompatible assay conditions, insufficient population coverage, unstable labels, or evidence that is too weak for the intended inference.

Federated stewardship coordinates the parts of product management that cannot be resolved independently within each domain. Domain stewards maintain scientific definitions, local quality controls, source-system knowledge, and product support. Shared semantic stewards maintain identifier policies, ontology profiles, mapping procedures, unit conventions, and compatibility rules. Security, privacy, legal, quality, and regulatory functions establish minimum controls and adjudicate uses that cross contractual, jurisdictional, or regulated boundaries. A cross-domain governance body resolves conflicts, approves shared standards, manages exceptions, and establishes escalation procedures. This allocation is necessary because effective translational data curation is a lifecycle activity involving planned metadata, standards, quality controls, responsibilities, and continuing maintenance rather than an isolated downstream-cleaning step [11]. Federated stewardship is therefore neither decentralized autonomy without coordination nor centralized control without domain accountability.

The resulting relationship among ownership, products, and stewardship is recursive. Domains publish products through shared contracts; consumers identify incompatibilities, missing context, and new requirements; stewards evaluate whether changes are local or enterprise-wide; and versioned updates are returned to the product without rewriting its historical lineage. Standards-based custodianship platforms show that metadata, controlled access, and translational reuse can be coordinated while explicit responsibility for governed data assets is retained [12]. The proposed data mesh extends this logic by treating ownership, semantic compatibility, policy evaluation, and evidence reuse as linked but separable architectural functions. Domain ownership preserves accountability, the data product makes evidence consumable, and federated stewardship establishes the minimum shared conditions under which products can interact. None of these components alone establishes pharmaceutical usefulness. Their contribution is to make the assumptions, responsibilities, and limitations of reuse sufficiently explicit that model developers and evidence consumers can determine when additional qualification, transformation, validation, or human review is required.

Semantic Contracts, Interoperability, and Evidence Lineage

A semantic contract is proposed as a versioned agreement specifying how a pharmaceutical data product should be interpreted across domain boundaries. It should define canonical identifiers, entity types, schema fields, units, allowable values, ontology bindings, missing-data conventions, scientific context, compatibility behavior, and version history. Biomedical identifier mappings require particular scrutiny because apparent equivalence may conceal differences in salt form, stereochemistry, biological entity, disease concept, assay definition, or source scope. Curated mapping resources demonstrate that identifier alignment should record provenance and confidence rather than silently asserting identity [13]. A semantic contract therefore makes interpretive assumptions inspectable, but it cannot establish biological equivalence or scientific validity by itself.

Evidence lineage complements the semantic contract by recording how evidence moved from its source state to a reusable product, model input, analytical result, or decision artifact. Required lineage events may include source-system and product versions, transformations, exclusions, quality checks, approvals, consent or licensing conditions, computational environments, model use, and downstream interpretation. Distributed provenance models show that these records can be linked across complex environments without requiring every provenance event to reside in one central store [14]. The proposed architecture therefore preserves locally generated lineage while enabling cross-domain reconstruction. Lineage completeness supports auditability and reproducibility, but it does not prove that the source evidence was reliable or that a transformation was scientifically justified.

Semantic interoperability emerges when multiple data products can be combined without concealing differences that remain scientifically important. Biomedical knowledge graphs illustrate how drugs, targets, diseases, pathways, phenotypes, and other entities can be integrated into a queryable representation [15]. Within the proposed mesh, however, the semantic layer should not replace domain products with one decontextualized enterprise graph. It should provide resolvable links among domain-held entities while preserving source attribution, evidence type, mapping method, uncertainty, and update history. This distinction is essential because graph connectivity may reflect curated association, database co-occurrence, inferred relation, experimental observation, or model-derived prediction, each of which supports a different strength of claim.

The interpretive value of semantic integration depends on whether integrated relations remain connected to their evidentiary origins and context. Knowledge-graph approaches can help situate molecular measurements within broader disease, drug, pathway, and clinical knowledge, but such integration remains bounded by source coverage, graph design, curation decisions, and the meaning assigned to each relation [16]. The article therefore treats semantic contracts, interoperability, and evidence

lineage as mutually dependent components. Contracts define how products may be interpreted, interoperability permits qualified connection, and lineage preserves how each connection and transformation arose. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop semantic contracts, interoperability, and evidence lineage within the article's central argument.

Table 2. Components, relationships, uncertainties, and validation needs within Semantic contracts, interoperability, and evidence lineage

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Canonical entity identification	Source identifiers, structures, sequences, study codes, terminology codes	Distinguish and align pharmaceutical entities across products	Versioned canonical identifiers with source mappings	Identity may depend on salt, stereochemistry, formulation, biological context, or terminology scope	Expert-adjudicated mapping tests and unresolved-identity reporting
Schema and unit contract	Field definitions, data types, unit systems, allowable values	Make structural and quantitative meaning explicit	Machine-readable schema with unit and value constraints	Schema conformance may coexist with incompatible scientific protocols	Conformance testing plus domain review of protocol comparability
Ontology binding	Domain terms, controlled vocabularies, relation types	Connect local concepts to shared semantic references	Traceable ontology mappings and relation definitions	Ontologies may be incomplete, overlapping, or revised	Competency questions, curator review, and version-compatibility testing
Context-of-use statement	Study design, assay protocol, population, species, site, process, intended application	Define the conditions under which the product may be interpreted or reused	Explicit intended, permitted, and excluded uses	Context may be incompletely documented or change over time	Domain-owner approval and task-specific fitness assessment
Mapping provenance	Mapping source, curator, algorithm, evidence, date, confidence	Make semantic alignment decisions inspectable	Reviewable mapping record linked to contract version	Confidence scores may not represent scientific equivalence	Independent mapping audit and error analysis
Distributed evidence lineage	Source versions, transformations, quality checks, approvals, policies	Reconstruct how evidence was created, modified, and reused	Linked provenance path across domain and shared services	Missing events, broken links, excessive granularity, or record tampering	Lineage reconstruction tests and completeness requirements
Interoperability service	Contract metadata, mappings, schemas, lineage references	Enable compatible discovery, query, and integration across products	Resolved cross-domain connections with retained source context	Technical connection may be mistaken for scientific comparability	Cross-domain challenge cases and expert adjudication
Compatibility and change control	Product versions, contract versions, dependency records	Detect breaking changes and communicate their downstream impact	Versioned compatibility decisions and migration requirements	Silent changes may propagate into models and analyses	Automated compatibility tests and accountable release review
Evidence-strength annotation	Source type, study design, curation status, model-derived status	Preserve the basis and limits of each represented relation	Relation-level evidence category and uncertainty statement	Categories may be applied inconsistently across domains	Inter-rater assessment and claim-to-evidence audit
Semantic exception process	Unresolved mappings, conflicts, domain-specific terms	Escalate cases that cannot be safely standardized	Documented exception, local extension, or rejected mapping	Excessive exceptions can recreate fragmentation	Governance review, usage monitoring, and periodic consolidation

Architecture of the Pharmaceutical Data Mesh

The proposed architecture separates domain custody from shared interoperability. Centralized data commons demonstrate that common ingestion, harmonization, controlled access, and analytical services can make large biomedical collections more usable [17]. The pharmaceutical data mesh retains these useful service principles but does not require every source payload to be transferred into one enterprise-owned repository. Each scientific domain maintains accountable data-product nodes, while shared services provide catalog discovery, semantic validation, evidence-lineage resolution, policy evaluation, federated query, model registration, and monitoring. Central repositories may still exist where scientifically and operationally appropriate; the mesh governs how centralized and locally retained products participate in the same evidence environment.

The architecture is organized into four interacting planes. The domain-product plane contains versioned evidence and accountable owners. The semantic and lineage plane defines meaning, compatibility, transformation history, and evidentiary context. The policy and governance plane evaluates identity, purpose, authorization, licensing, jurisdiction, and applicable quality controls. The federated service plane supports discovery, approved query, compute-to-data execution, model registration, and evidence-reuse recording. Frameworks for interoperable biomedical cloud platforms similarly indicate that portability, FAIRness, security, and governance must be designed together rather than treated as independent platform attributes [18]. **Figure 1** presents the federated pharmaceutical data mesh, showing how the article's key components and boundaries are connected within architecture of the pharmaceutical data mesh.

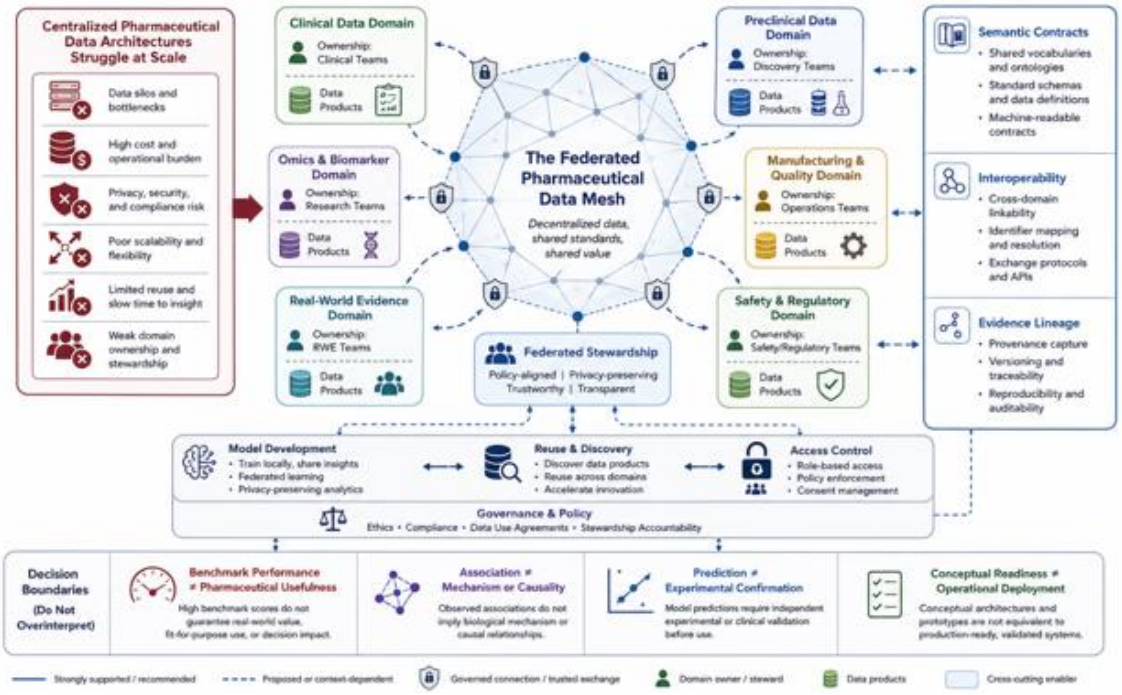

Figure 1. Federated Pharmaceutical Data Mesh.

The figure is an original conceptual synthesis that organizes the article's central contribution across centralized pharmaceutical data architectures struggle at scale, domain ownership, data products, and federated stewardship, domain ownership, data products, federated stewardship, semantic contracts, interoperability, and evidence lineage. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

The mesh should expose products through standardized metadata and governed service endpoints while leaving payload location conditional on scientific, contractual, security, and operational requirements. Decentralized FAIR platforms in the life sciences provide a precedent for making distributed datasets globally discoverable while preserving local control over the underlying resources [19]. In the proposed pharmaceutical architecture, a catalog entry would identify the product, owner, scientific context, contract version, available interfaces, access requirements, and status. Discovery would not imply authorization, and authorization would not imply unrestricted export. Depending on the product, an approved consumer might receive a controlled extract, a remote query result, a secure workspace, or permission to execute a registered model against domain-held data.

The architecture also supports evidence products organized around pharmaceutical questions rather than only source systems. Integrated target-discovery platforms demonstrate how heterogeneous genetics, omics, drug, disease, and literature evidence can be normalized into reusable services while retaining source attribution and evidence-type distinctions [20]. The mesh generalizes this approach by allowing domains to publish purpose-specific products without surrendering ownership or erasing source context. A model-development workflow may consume several approved products, generate a versioned model package, and return lineage and validation records to the mesh. An evidence-synthesis workflow may reuse prior analyses while recording the new question and interpretive boundary. The architecture therefore connects evidence without implying that all connected evidence is equivalent, independent, causal, or ready for pharmaceutical decision-making.

Model Development, Reuse, Access Control, and Governance

Model development within the mesh begins with a declared question, intended context of use, required evidence domains, and prespecified validation plan. The model package should record code, parameters, software dependencies, product versions, feature definitions, exclusions, training lineage, validation evidence, calibration information, uncertainty, responsible owner, and permitted uses. Guidance on deep-learning development emphasizes that data curation, task specification, validation, generalization, and workflow integration are necessary alongside algorithm selection [21]. Accordingly, registration within the mesh does not certify model quality. It makes the evidence and assumptions needed for qualification visible and allows later users to determine whether retraining, external validation, or rejection is appropriate.

Access control must govern more than file download. A pharmaceutical data product may allow metadata discovery while restricting row-level retrieval, model training, partner access, commercial use, or export of derived information. The policy decision service should consider identity, role, project purpose, legal basis, consent or license, jurisdiction, product sensitivity, computational method, and requested output. Privacy-preserving and federated machine-learning literature shows that federation alone does not prevent information leakage and that safeguards require explicit threat models and technical controls

[22]. The proposed mesh therefore treats secure aggregation, confidential execution, output filtering, and privacy mechanisms as optional, risk-based controls rather than universal guarantees.

Compute-to-data services provide one mechanism for developing models across products that cannot be physically pooled. Registered code or a model package may be executed in approved local environments, after which permitted summaries or model updates are returned. Multi-institutional federated-learning studies demonstrate the technical feasibility of collaborative training without central transfer of patient data in defined settings [23]. Such feasibility does not establish transportability to medicinal chemistry, toxicology, manufacturing, or all clinical-development tasks. Each use requires assessment of label compatibility, population or assay differences, node reliability, computational reproducibility, security exposure, aggregation behavior, and whether the distributed result is scientifically interpretable.

Governance links product ownership, access control, model qualification, monitoring, and retirement. An accountable lifecycle should specify who approves the question, who verifies data fitness, who reviews model behavior, who authorizes reuse, who monitors change, and who can suspend or retire an artifact. Governance models for health-care artificial intelligence emphasize that accountability extends from design and validation through implementation and monitoring [24]. The proposed pharmaceutical extension also requires domain owners, model owners, semantic stewards, security and privacy reviewers, legal and quality functions, and decision users. Human oversight is not a final signature applied after technical development; it is distributed across the decisions that determine what evidence enters the model, what claim the output may support, and when reuse is no longer justified.

Adoption Barriers and Organizational Implementation

Adoption is likely to fail when the mesh is treated as a technology installation rather than a change in scientific accountability and operating practice. Pharmaceutical modeling fields have repeatedly shown that technically capable methods may remain uncommon when evidence expectations, stakeholder roles, infrastructure, training, workflow integration, and incentives are not aligned [25]. Domain teams may resist ownership responsibilities if stewardship work is unfunded, while central teams may resist federation if it reduces direct control over schemas and pipelines. Implementation must therefore define decision rights, service expectations, staffing, escalation routes, and the resources required to maintain products after their initial publication.

A second barrier is the disruption of established workflows. Scientists may bypass shared services when product registration, access requests, semantic review, or lineage capture adds delay without visible benefit. Conversely, governance teams may over-engineer controls around low-risk products because the architecture lacks differentiated assurance levels. Research on the delivery of clinical impact from artificial intelligence emphasizes prospective evaluation, workflow redesign, logistics, user acceptance, and sociotechnical change rather than model development alone [26]. A pharmaceutical implementation should therefore begin with bounded use cases where ownership conflicts, semantic incompatibility, or repeated evidence reconstruction are already material problems and where improvement can be assessed without claiming enterprise-wide readiness.

Ethical and professional responsibilities also require explicit design. Machine-learning implementation raises concerns involving confidentiality, bias, transparency, accountability, and changes in professional roles [27]. Within the mesh, these concerns affect which products are created, how populations and assay systems are represented, who can challenge mappings, how excluded evidence is documented, and whether model outputs are allowed to influence decisions. Automation should not obscure the responsible scientist, data custodian, model owner, or final decision-maker. The architecture should preserve contestability by allowing users to inspect lineage, report incompatibilities, request correction, and identify when an output exceeds the product's or model's intended use.

Implementation should proceed as a staged learning process. Responsible machine-learning roadmaps emphasize multidisciplinary problem formulation, rigorous validation, transparent implementation, monitoring, and mechanisms for identifying and mitigating harm [28]. The first stage should establish ownership, contract, lineage, and catalog capabilities for a limited number of high-value products. A second stage may add policy-aware access and controlled cross-domain queries. Federated computation and model reuse should follow only when semantic, security, and validation prerequisites are demonstrated. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop adoption barriers and organizational implementation within the article's central argument.

Table 3. Evaluation, implementation, and research priorities arising from A Pharmaceutical Data Mesh That Preserves Ownership while Enabling Semantic Interoperability, Model

Evaluation dimension	What must be tested	Suitable evidence or assessment	Failure signal	Interpretive limitation	Decision relevance
Semantic validity	Whether identifiers, units, schemas, and ontology bindings preserve intended meaning	Curated challenge cases, competency questions, mapping audits, domain-expert review	False equivalence, unresolved identity, unit mismatch, incompatible context	Conformance does not establish biological or causal validity	Determines whether products can be combined without silent semantic distortion
Data-product fitness	Whether evidence is suitable for the proposed analytical or scientific use	Quality profile, missingness analysis, population or assay coverage, bias	Reuse outside context, hidden selection, incompatible protocol, stale evidence	FAIRness and availability do not prove fitness	Determines whether a product should enter a model or evidence synthesis

		assessment, intended-use review			
Lineage completeness	Whether source, transformations, approvals, policies, and downstream use can be reconstructed	Provenance-event audit, version resolution, workflow reconstruction	Broken lineage, unverifiable transformation, missing source version	Complete lineage does not prove correctness	Supports audit, reproducibility, investigation, and qualified reuse
Policy correctness	Whether identity, purpose, contract, consent, jurisdiction, and risk are correctly evaluated	Approved and prohibited test cases, legal review, audit logs, exception analysis	Unauthorized grant, inappropriate denial, policy drift, inconsistent exception	Passing tests cannot exclude undiscovered legal or security conflicts	Determines whether access or computation may proceed
Privacy and security	Whether queries, updates, outputs, or metadata expose prohibited information	Threat modeling, penetration testing, red-team exercises, leakage assessment	Inference attack, poisoning, collusion, insider misuse, metadata disclosure	No finite assessment establishes absolute privacy or security	Determines required safeguards and whether a use remains acceptable
Federated reproducibility	Whether approved distributed analyses can be repeated across nodes and versions	Environment verification, version locks, execution logs, repeated runs	Dependency drift, node inconsistency, unstable aggregation, partial failure	Reproducibility does not establish scientific validity	Determines whether distributed outputs can be independently checked
Model transportability	Whether model performance and uncertainty remain appropriate in a new domain or product version	External and temporal validation, calibration analysis, subgroup evaluation, shift diagnostics	Degradation, overconfidence, subgroup error, label incompatibility	Benchmark success does not establish pharmaceutical usefulness	Determines whether a model may be reused, retrained, restricted, or rejected
Evidence-reuse fidelity	Whether prior evidence retains its meaning, assumptions, limitations, and dependency structure	Claim-to-evidence audit, lineage review, original-versus-new context comparison	Citation laundering, duplicated evidence, unsupported claim escalation	Traceability cannot strengthen weak evidence	Determines the permissible strength of a reused scientific claim
Governance effectiveness	Whether ownership, escalation, monitoring, correction, and retirement responsibilities function	Decision logs, issue-resolution analysis, product freshness, monitoring records	Orphaned products, unresolved conflicts, silent changes, delayed retirement	Fast governance is not necessarily scientifically sound governance	Determines whether accountability remains operational after launch
Organizational sustainability	Whether domains and shared services can maintain the architecture over time	Workload analysis, staffing, service reliability, reuse quality, cost and incident review	Shadow pipelines, decaying products, duplicated platforms, user bypass	High utilization may indicate convenience rather than scientific value	Guides staged investment, resourcing, and continuation decisions

Limitations and Boundary Conditions

The proposed data mesh is a conceptual architecture derived from selected pharmaceutical, biomedical-informatics, semantic, provenance, federated-learning, and governance literature. It has not been empirically compared with centralized lakes, warehouses, data commons, or hybrid architectures. Its suitability is likely to vary by evidence type, organizational structure, jurisdiction, intellectual-property environment, quality-system status, and computational capability. Federated methods may be unnecessary where legitimate central harmonization is efficient, while highly distributed arrangements may impose excessive maintenance and coordination costs. Architecture selection should therefore remain conditional on the scientific problem and operating environment rather than on adherence to a universal design doctrine.

The architecture cannot guarantee semantic correctness, privacy, security, reproducibility, or valid evidence reuse. Federated learning may reduce raw-data movement while remaining vulnerable to leakage, poisoning, or incompatible local data [22]. Distributed lineage can improve traceability while still containing missing, incorrect, or misleading events [14]. Interoperable cloud and platform services can support portability and governance, but they do not remove the need for context-specific assurance [18]. Similar boundaries apply to model outputs: a registered and reproducible model may remain poorly calibrated, scientifically implausible, non-transportable, or inappropriate for the intended decision.

The construct also cannot establish regulatory acceptability, clinical utility, experimental confirmation, causality, or organizational benefit. A semantic relation may represent association rather than mechanism, a model prediction may not survive experimental testing, and an integrated evidence product may contain correlated sources that should not be treated as independent confirmation. Domain ownership may preserve accountability but can also reinforce organizational silos if shared stewardship is weak. Conversely, enterprise governance may recreate centralized bottlenecks if minimum controls expand into universal approval of local scientific decisions. These tensions are not implementation defects that can be removed completely; they are boundary conditions that must be made visible and evaluated.

Research Agenda and Implementation Priorities

The first research priority is to develop and test a minimum pharmaceutical data-product contract. Candidate fields should include scientific context, canonical identity, schema, units, quality profile, missingness, lineage, ownership, access policy, versioning, compatibility, and intended-use boundaries. FAIRness metrics can assess important properties of digital objects, but they should be combined with task-specific fitness evaluation rather than converted into a single readiness score [10]. Prospective studies should test whether users can correctly identify suitable and unsuitable products, detect semantic conflicts, reproduce transformations, and understand what claims a product cannot support.

A second priority is validation of semantic and lineage services across representative pharmaceutical domains. Identifier mappings should be evaluated with curator-adjudicated cases that distinguish chemical identity, biological entity, assay

context, disease concept, study variable, and manufacturing meaning [13]. Distributed provenance profiles should be tested to determine which events are necessary for reconstructing model inputs, analytical transformations, approvals, and evidence reuse [14]. Security research should include explicit threat models, adversarial testing, policy-conflict scenarios, and reproducibility assessment across heterogeneous nodes. Federated model studies should prespecify intended use, external validation, calibration, subgroup behavior, and failure handling rather than report distributed feasibility alone [23].

Implementation research should examine the operating model as carefully as the technology. Comparative case studies are needed to determine how ownership, stewardship, semantic governance, legal review, access decisions, and model oversight should be allocated across domain and enterprise teams. Governance evaluation should test whether responsibilities remain clear through product updates, model reuse, incidents, personnel changes, and retirement decisions [24]. Initial implementation should prioritize bounded workflows with measurable integration burden and clear accountable users. Expansion should depend on demonstrated semantic validity, traceability, policy correctness, maintainability, and scientific value, not on the number of products registered or models connected. Regulatory or quality-system use should be considered only through separate, context-specific qualification pathways.

Conclusion

The ownership-preserving pharmaceutical data mesh is proposed as a conceptual response to a recurring architectural tension: pharmaceutical evidence must become interoperable and reusable without being detached from the domains that understand, maintain, and govern it. The construct separates physical data location from semantic interoperability and links domain-owned data products through federated stewardship, semantic contracts, distributed evidence lineage, policy-aware access, governed computation, model registration, and evidence-reuse records. Its principal contribution is not a new optimization target but a multi-dimensional account of architectural fitness in which availability, FAIRness, model performance, and technical integration remain insufficient without scientific suitability, traceability, policy correctness, uncertainty characterization, human accountability, and organizational sustainability. The architecture remains proposed rather than validated. It does not guarantee privacy, model validity, experimental confirmation, clinical utility, regulatory acceptance, or operational readiness. Its practical value will depend on whether future work can demonstrate that ownership and interoperability can be combined without reproducing the rigidity of centralization or the fragmentation of unmanaged decentralization.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23(6):1241-50. doi:10.1016/j.drudis.2018.01.039
3. Wise J, Grebe de Barron A, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019;24(4):933-8. doi:10.1016/j.drudis.2019.01.008
4. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Brief Bioinform. 2022;23(6). doi:10.1093/bib/bbac404
5. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25(1):37-43. doi:10.1038/s41591-018-0272-7
6. Mello MM, Triantis G, Stanton R. Waiting for data: Barriers to executing data use agreements. Science. 2020;367(6474):150-2. doi:10.1126/science.aaz7028
7. Lee JSH, Kibbe WA, Grossman RL. Data harmonization for a molecularly driven health system. Cell. 2018;174(5):1045-8. doi:10.1016/j.cell.2018.08.012
8. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119. doi:10.1038/s41746-020-00323-1
9. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019;37(4):358-67. doi:10.1038/s41587-019-0080-8
10. Wilkinson MD, Sansone SA, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Sci Data. 2018;5:180118. doi:10.1038/sdata.2018.118

11. Gu W, Hasan S, Rocca-Serra P, Satagopam VP. Road to effective data curation for translational research. Drug Discov Today. 2021;26(3):626-30. doi:10.1016/j.drudis.2020.12.007
12. Emam I, Elyasigomari V, Matthews A, Pavlidis S, Rocca-Serra P, Guitton F, et al. PlatformTM, a standards-based data custodianship platform for translational medicine research. Sci Data. 2019;6(1):149. doi:10.1038/s41597-019-0156-9
13. Hoyt CT, Hoyt AL, Gyori BM. Prediction and curation of missing biomedical identifier mappings with Biomappings. Bioinformatics. 2023;39(4). doi:10.1093/bioinformatics/btad130
14. Wittner R, Mascia C, Gallo M, Frexia F, Müller H, Plass M, et al. Lightweight distributed provenance model for complex real-world environments. Sci Data. 2022;9(1):503. doi:10.1038/s41597-022-01537-6
15. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. doi:10.1038/s41597-023-01960-3
16. Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol. 2022;40(5):692-702. doi:10.1038/s41587-021-01145-6
17. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood. 2017;130(4):453-9. doi:10.1182/blood-2017-03-735654
18. Grossman RL, Boyles RR, Davis-Dusenbery BN, Haddock A, Heath AP, O'Connor BD, et al. A framework for the interoperability of cloud platforms: Towards FAIR data in SAFE environments. Sci Data. 2024;11(1):241. doi:10.1038/s41597-024-03041-5
19. Vazquez P, Hirayama-Shoji K, Novik S, Krauss S, Rayner S. Globally accessible distributed data sharing (GADDS): A decentralized FAIR platform to facilitate data sharing in the life sciences. Bioinformatics. 2022;38(15):3812-7. doi:10.1093/bioinformatics/btac362
20. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong CK, Fumis L, Karamanis N, et al. Open Targets Platform: New developments and updates two years on. Nucleic Acids Res. 2019;47(D1). doi:10.1093/nar/gky1133
21. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-9. doi:10.1038/s41591-018-0316-z
22. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2(6):305-11. doi:10.1038/s42256-020-0186-1
23. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598. doi:10.1038/s41598-020-69250-1
24. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27(3):491-7. doi:10.1093/jamia/ocz192
25. Darwich AS, Ogungbenro K, Vinks AA, Powell JR, Reny JL, Marsousi N, et al. Why has model-informed precision dosing not yet become common clinical reality? Lessons from the past and a roadmap for the future. Clin Pharmacol Ther. 2017;101(5):646-56. doi:10.1002/cpt.659
26. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
27. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981-3. doi:10.1056/NEJMp1714229
28. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6