Engineering Data Architectures for AI/ML Integration in Regulated Manufacturing

Viktoriia Shubina[1], Tuomas Ranti[1], Anne Juppo[2], and Tuomas Mäkilä[1]

[1] University of Turku, Department of Computing, FI-20014 Turku, Finland
{viktoriia.shubina, tuomas.ranti, tuomas.makila}@utu.fi
[2] University of Helsinki, Division of Pharmaceutical Chemistry and Technology,
FI-00014 Helsinki, Finland
anne.juppo@helsinki.fi

Abstract. Life science, i.e. pharmaceutical and medical device, manufacturers are increasingly exploring artificial intelligence (AI) and Machine Learning (ML) to enhance production quality and regulatory compliance. However, current data handling practices result in data fragmentation, and complex regulatory requirements present barriers to wide implementation. In this study, we conducted 20 qualitative interviews with data architects, AI specialists, and regulatory compliance officers. Our aim was to get a better understanding of the current state of the field, challenges and future outlook in regulated manufacturing, employing the Gioia methodology. Our findings highlight data silos and legacy infrastructures as primary technical barriers, while evolving regulatory frameworks and uncertainties in AI validation create significant compliance challenges. Interviewees emphasized the necessity of unified data architectures and platforms, embedded governance mechanisms, enhanced security, and proactive regulatory operations (RegOps) to enable both innovation and compliance. Based on the interview results, we propose a conceptual framework to guide the design of AI-driven data architectures that bridge fragmented systems and support compliant AI/ML lifecycle management. This study is the first phase of research efforts aiming to implement and validate AI/ML solutions grounded in industry needs.

Keywords: AI adoption, data architecture, regulated manufacturing, Gioia method

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly recognized as transformative technologies in regulated manufacturing sectors, including pharmaceuticals and medical devices manufacturing [1]. However, their adoption introduces a dual imperative for the life science manufacturing industry. While AI/ML enable strategic advancements in automation, prediction, and decision support, they also impose significant challenges related to regulatory compliance, particularly concerning system integration, validation, and traceability [2].

The life-science sector is undergoing a profound transformation, as advances in automation and AI/ML reshape research, development, and manufacturing practices across biological domains [3]. Real-world implementation of AI/ML in these settings requires data architectures capable of supporting end-to-end traceability, robust validation, and secure data flow across fragmented and often legacy infrastructures. These environments are typically characterized by data silos, manual processes, and limited interoperability. All conditions that constrain scalable and compliant AI/ML deployment. Designing such architectures requires harmonization of real-time data processing with regulatory demands for auditability, involving careful orchestration of metadata management, interface design, and validation workflows in alignment with Good Practice (GxP) standards.

GxP standards is an umbrella term encompassing regulatory guidelines such as Good Manufacturing Practice (GMP), Good Laboratory Practice (GLP), and Good Clinical Practice (GCP), among others [4]. These frameworks define the principles and procedures necessary to ensure that products in life science sectors are consistently produced, tested, and monitored to the required quality, safety, and efficacy. In the context of data-driven systems, GxP compliance extends beyond physical manufacturing to include rigorous controls over data integrity, traceability, and reproducibility. This necessitates that AI/ML-enabled architectures embed compliance considerations into every stage of the data lifecycle, from acquisition and preprocessing to model deployment and monitoring, ensuring that digital processes meet the same evidentiary and auditability standards as traditional operations.

While the theoretical dimensions of AI explainability, fairness, and validation have received substantial scholarly attention [5], understanding of how these issues are addressed in regulated manufacturing practice remains scarce. In particular, the design and operationalization of compliant data architectures to enable trustworthy and auditable AI/ML systems are underexplored. Recent advances, such as the use of Large Language Models (LLMs) to enhance explainability, offer promising technical pathways [6], yet their applicability in highly regulated domains remains to be systematically investigated. Bridging this gap is essential for advancing AI/ML capabilities in software-intensive systems subject to stringent regulatory oversight.

To investigate these challenges, we conducted a qualitative interview study involving 20 professionals from the pharmaceutical and medical device manufacturing companies, operating in Finland and internationally. Participants were selected based on their expertise in one of three focal domains: data architecture (6 experts), AI/ML applications (6 experts), and regulatory compliance (8 experts). Using the Gioia methodology [7], we systematically analyzed the interview transcripts to construct a grounded theoretical understanding of current practices, obstacles, and emerging trajectories within regulated manufacturing environments.

We organize our study around the following research questions (RQs):

- **RQ1:** What are the current practices and system designs used for data architecture, AI integration, and regulatory compliance in regulated manufacturing?
- **RQ2:** What technical, organizational, and regulatory challenges do regulated manufacturers face in managing data, implementing AI, and maintaining compliance?
- **RQ3:** How do regulated industry professionals envision the future development of data architectures, AI applications, and regulatory practices in their organizations?

This paper makes the three contributions. First, it provides evidence on the technical and regulatory barriers to AI/ML adoption in regulated manufacturing, grounded in 20 expert interviews. Second, it offers overview of the current industry practices. Third, it introduces a conceptual framework for designing AI-driven data architectures that enable compliant AI/ML lifecycle management in regulated industrial contexts.

The remainder of this paper is structured as follows: Section 2 reviews related work on AI and data architecture strategies, especially in regulated domains. Section 3 describes our research design and the Gioia coding process. Section 4 presents the results, including the data structure and conceptual framework. Section 5 discusses implications for system design and regulatory innovation. Section 6 concludes with directions for future research.

2 Related Work

The deployment of AI systems within regulated life-science manufacturing environments necessitates the existence of high-quality and well-governed data infrastructures. Fragmented, siloed, or weakly contextualized data continues to represent a major impediment to the large-scale adoption of AI in GxP-regulated domains. Recent industry frameworks underscore the need for unified data platforms that combine technical robustness with conformance to regulatory and validation requirements [8].

A comprehensive analysis of reference architectures for smart manufacturing demonstrates their critical role in establishing interoperability, modularity, and scalability across heterogeneous production systems [9]. The reviewed frameworks formalize layered architectural structures that specify the interactions among physical assets, data acquisition and integration layers, and analytical components within a unified environment. These abstractions underpin the design of GxP-compliant data fabrics, as articulated by NTT DATA, that focus on the consolidation of distributed data sources while ensuring end-to-end traceability, regulatory compliance,and validation integrity across the manufacturing data lifecycle [8].

In parallel, industry bodies such as the International Society for Pharmaceutical Engineering (ISPE) have proposed governance models for AI/ML in GxP environments [10]. They emphasize the connection between AI lifecycle management and regulatory expectations, including risk-based access control, MLOps integration, and documentation readiness.

While algorithmic auditing frameworks offer structured mechanisms for assessing accountability and transparency in AI systems [11], the practical realization of responsible AI governance within regulated industrial environments remains in its early stages. Recent conceptual advances extend these ideas beyond algorithm-level audits toward organization-wide governance structures that integrate ethical, technical, and compliance perspectives [12]. This emerging research emphasizes the need for multidimensional frameworks that align AI lifecycle management with established governance principles and regulatory expectations. Such connection is particularly critical in life-science manufacturing, where AI-driven decision-making directly intersects with product quality, patient safety, and GxP compliance.

Validation is a particularly complex aspect of AI system deployment in life sciences. Traditional validation approaches are often ill-suited to the dynamic and non-deterministic nature of AI models. Emerging strategies, including those from the ISPE AI Maturity Model [13], propose risk-based model based on the control design tailored to AI. Scholars have also begun to distinguish between narrow AI validation (focused on specific tasks) and platform-level validation (covering ML pipelines and tooling) [14]. These strategies are increasingly aligned with recent regulatory initiatives, such as the EU AI Act [15], which introduces tiered risk categories and compliance obligations for AI systems in high-stakes domains.

From a methodological point of view, prior research provides a basis for qualitative research on technological transformation in manufacturing. A study based on 33 semi-structured interviews in incumbent industrial firms developed a dual-mode digital transformation framework that balances exploration and exploitation [16]. Another meta-analysis of 22 manufacturing companies identified a four-dimensional capability model, highlighting strategic, organizational, technical, and process competencies as key determinants of digital maturity [17].

Thus, prior research provides methodological precedents for interview-based studies in complex industrial environments and highlights the value of thematic and comparative analysis in understanding industrial transformation processes. However, empirical understanding on how data architectures are designed and implemented to enable compliant and trustworthy AI in regulated manufacturing remains limited. And this study addresses that gap through a qualitative exploration. The literature review was conducted to establish a conceptual and regulatory foundation, rather than to provide a systematic synthesis, ensuring coherence between the theoretical framework and the empirical focus of the study.

3 Methodology and Research Design

This study adopts a qualitative, exploratory design based on expert interviews to examine the intersection of data architecture, AI and ML integration, and regulatory practices in pharmaceutical and medical device manufacturing. The methodological approach was structured to ensure rigor, transparency, and analytical depth in line with established standards for qualitative research in complex and regulated environments.

Empirical data were collected through semi-structured interviews with 20 experts representing diverse roles and organizational contexts within the life-science manufacturing ecosystem. The sample comprised 6 data architects responsible for digital infrastructure and enterprise data platforms, 6 AI and automation specialists engaged in the design and deployment of ML-based systems, and 8 professionals specializing in regulatory compliance, validation, and quality assurance. Participants were drawn from 6 organizations operating in Finland and internationally, all compliant with related regulatory standards. Their professional experience ranged from 3 to 43 years, providing first-hand insights into both technical and regulatory challenges across varying levels of digital maturity.

Interviews were conducted online between April and June 2025. The protocol was designed to capture detailed accounts of data architecture design, AI/ML integration, and regulatory practices in manufacturing operations. Each session lasted 30-70 minutes and was recorded with participant consent. The interview guide was developed based on insights from the literature and the practical objectives of the LifeFactFuture research project, financed by Business Finland, ensuring relevance to ongoing initiatives in data-driven transformation and AI adoption within regulated manufacturing. It was subsequently refined through pilot testing to enhance clarity, validity, and alignment with the study's research objectives.

All interviews were transcribed verbatim, anonymized, and stored securely in accordance with ethical and data-protection requirements. The resulting dataset was analyzed systematically using NVivo 15 to identify themes and relationships relevant to data management, validation, and regulatory compliance.

The study employed the Gioia methodology [7], a rigorous inductive approach widely recognized in organizational and technology studies for enabling structured, theory-building research from qualitative data. An illustrative example of the coding process and resulting Gioia structure for RQ1 is provided in Table 1.

First-order codes were generated by adhering to participant language and terminology, capturing grounded concepts and practitioner-centric perspectives. These codes were then systematically organized into second-order themes, reflecting higher-level patterns, processes, and challenges articulated across interviews. Finally, the themes were synthesized into aggregate dimensions, aligning with the study's core research questions regarding data architectural strategies, AI integration, and regulatory compliance mechanisms.

Coding and interpretation were conducted collaboratively by the two researchers from our team to ensure reliability and minimize individual bias. Discrepancies were resolved through discussions, and a codebook was maintained to document analytical decisions and facilitate transparency. Reflexivity was maintained throughout, with regular memoing to capture emerging insights, uncertainties, and researcher positioning.

Table 1. Illustrative Gioia data structure for RQ1: Current practices and system designs in regulated manufacturing.

1st-Order Concepts	2nd-Order Themes	Aggregate Dimensions
Everything must be documented	(Fragmented) data collection practices	**Fragmented digital ecosystem in regulated manufacturing**
Local data storage and in-place analysis	Localized processing and limited data integration	
Building centralized cloud data platform	Emergence of centralized repositories for select data	
Ongoing AI proof-of-concept projects	Early and isolated AI/ML experiments	**Limited and cautious experimentation with new technologies**
Use of ML models in product development	Application of ML for design optimization	

All participants received written information about the study objectives, data handling procedures, and their rights, and provided informed consent prior to participation. Identifying details were removed during transcription, and findings are reported in aggregate to preserve confidentiality.

4 Results

This section presents the findings of our qualitative study, structured according to the three research questions. The analysis follows a Gioia-informed coding approach, grouping first-order concepts into second-order themes and aggregate dimensions. Each thematic area is supported by illustrative quotations from participants to highlight contextual insights and lived practices within regulated manufacturing environments.

4.1 RQ1: Current Practices and System Designs

Participants described the current state of data architectures and AI usage in regulated manufacturing as fragmented but slowly evolving. Legacy systems and siloed practices dominate, yet early attempts at integration and selective modernization suggest a gradual shift toward more connected infrastructures. Overall, current practices remain compliance-driven and reactive, with limited strategic planning for long-term digital transformation.

Aggregate dimension: Fragmented Digital Ecosystem in Regulated Manufacturing

This dimension illustrates how historical system choices continue to shape digital infrastructures, resulting in parallel data environments with limited interoperability. Participants consistently highlighted fragmentation as a structural rather than temporary issue.

2nd-Order Theme: (Fragmented) Data Collection Practices. Participants described in general their companies' data practices, alongside with challenges as highly fragmented data storage practices, with isolated systems for production, quality, and analytics. Data acquisition remains compartmentalized across Manufacturing Execution System (MES), Laboratory Management System (LIMS), Enterprise Resource Planning (ERP), and equipment-specific systems, resulting in silos that hinder data accessibility and transparency. Moreover, participants emphasized that every step in data handling must be documented, further increasing the complexity of managing dispersed systems. As one Data Architect reported, "*Data siloing, especially in document management, stems from old organizational structures. That's exactly why there's a push to rethink and redesign how things are done*".

2nd-Order Theme: Localized Processing and Limited Data Integration. Many organizations continue to rely on localized, on-premise data processing, with minimal cross-functional integration. Integration efforts are often reactive and project-specific, rather than systematic. One Data Architect Expert explained, "*We have a MES at the core of our production environment–it's essentially the heart of manufacturing operations. In parallel, our laboratories operate with a LIMS, while environmental monitoring systems continuously collect data, such as temperature to ensure that manufacturing conditions remain within acceptable thresholds. Each of these systems stores and processes data locally, primarily for operational monitoring and compliance*".

2nd-Order Theme: Emergence of Centralized Repositories for Select Data. Notably, there is a trend towards the development of centralized repositories–often in the form of data lakes or hybrid cloud platforms: for select, high-priority data streams. However, these are frequently constrained to specific use cases (such as regulatory reporting or AI initiatives), rather than being organization-wide solutions. One interviewee observed, "*The core task of our product management is to build the product structure into our ERP system–this includes the materials needed for manufacturing, the required process operations, and also the production masters compiled from customer-provided documentation*".

Aggregate dimension: Limited, Cautious Experimentation with New Technologies

Alongside fragmented infrastructures, organizations are beginning to explore AI/ML. However, experimentation tends to remain narrowly scoped and pilot-driven, with projects serving as both technical trials and organizational learning exercises rather than large-scale deployments.

2nd-Order Theme: Early and Isolated AI/ML Experiments. While regulatory frameworks are still evolving, many companies have already initiated pilot projects to explore the potential of AI technologies in regulated manufacturing settings. Reported use cases ranged from predictive maintenance, inventory optimization, and documentation support to regulatory chatbots and coding assistants. Several firms also experimented with AI in product development and testing, molecule screening, and non-GMP applications, while stressing that deployment into GMP-critical processes is not yet a top priority. As one participant noted, these efforts are often tied to preparing underlying data platforms and MLOps practices, laying the groundwork for future integration. Another interviewee observed, "*Earlier this year, we launched an AI proof-of-concept to validate a specific use case and assess whether the expected outcomes can be achieved with AI. At the same time, it's also about maturing the organization towards an AI culture: understanding data quality, risk management, and the potential benefits like cost savings or new services*".

However, one large global company reported initiating a project to implement a digital twin, highlighting the increasing relevance of such technologies in regulated manufacturing. As one Data Architect noted, digital twins offer a valuable opportunity to design data and system architectures from scratch, free from legacy and other constraints. Beyond process simulation, digital twins provide a structured environment for continuous data collection, model validation, and scenario testing, thereby creating a robust foundation for AI integration and lifecycle management. This dual role, as both a mirror of production processes and a controlled space for validating AI-driven insights, positions digital twins as an enabler for compliant and scalable AI adoption in the regulated industries [18].

4.2 RQ2: Challenges in Data Management, AI Implementation, and Compliance

The barriers identified by participants reveal how technical, regulatory, and organizational factors are tightly interwoven. Challenges extend beyond technology to encompass validation practices, compliance demands, and cultural resistance, illustrating the systemic nature of transformation obstacles.

Aggregate dimension: Systemic Barriers to Digital and AI Transformation

This dimension highlights the intertwined obstacles that collectively slow down the adoption of integrated data architectures and AI. Beyond isolated technical limitations, participants emphasized that regulatory complexity and organizational culture form a reinforcing cycle that constrains innovation.

2nd-Order Theme: Need for Data Standards and Poor Interoperability. Disparate data formats, lack of harmonized taxonomies, and proprietary vendor solutions persist as major obstacles to integrated analytics and AI deployment. Participants pointed out that metadata is often used to address these issues by adding contextual information and supporting integration across systems. Yet, metadata practices are inconsistent and not standardized, which limits their impact. Alongside these technical challenges, interviewees highlighted explainability as a critical concern: many AI models still operate as “black boxes”, producing outcomes without sufficient transparency. For regulated industries, this is particularly problematic, as understanding and justifying model decisions is essential for compliance, trust, and acceptance. As one interviewee reflected, “*Machine learning systems are not explainable. They might give you really good results, but you might not be able to understand*”. These findings align with broader initiatives in the scientific community that emphasize the need for standardized metadata and data governance frameworks, most notably the FAIR data principles (Findable, Accessible, Interoperable, and Reusable) [19]. Adherence to these principles facilitates data traceability, reuse, and semantic interoperability across organizational and regulatory boundaries. In parallel, the integration of *explainable AI* (XAI) techniques (such as SHAP, LIME, and counterfactual reasoning) has been increasingly recognized as essential for ensuring model transparency, interpretability, and accountability in regulated environments [20]. Together, these approaches establish a foundation for technically robust and governance-aware AI systems that support regulatory compliance and stakeholder trust.

2nd-Order Theme: Burden of Documentation and Validation. Participants cited the weight of documentation and validation processes as a limiting factor in system evolution and AI adoption. Regulatory requirements for comprehensive audit trails, validation protocols, and versioning impose significant overhead. As one AI professional stated, “*Regulations are constantly evolving, and with the new AI Act already being implemented, it’s crucial to integrate it into the development process for machine learning systems–something that’s still largely missing today*”.

Another limiting factor identified by participants is the manual workload associated with regulatory compliance and documentation. In particular, the need to generate, maintain, and verify logs in accordance with strict regulatory standards imposes a significant burden on personnel, often diverting resources away from innovation and system optimization efforts.

2nd-Order Theme: Organizational Resistance and Skills Gaps. Technical transformation is further constrained by cultural resistance and insufficient in-house expertise for regulated environments, particularly in advanced analytics and digital validation. One interviewee noted, “*People aren’t using AI at work as much as they should–or could–and the reason might simply be resistance to anything new*”.

4.3 RQ3: Future Visions for Data Architecture, AI, and Compliance

Looking ahead, participants described strategic directions that emphasize unification, scalability, and regulatory readiness. Future visions reflect a shift from reactive compliance and isolated pilots toward proactive, systemic planning of data and AI infrastructures.

Aggregate dimension: Strategic Digital Transformation for Intelligent Manufacturing

This dimension reflects the ambition to transition from fragmented systems to unified, real-time, and intelligent data architectures that can simultaneously deliver operational efficiency and regulatory trustworthiness.

2nd-Order Theme: Aspirations for Unified and Real-Time Data Architectures. There is broad consensus on the value of unified data architectures capable of supporting both real-time and historical analytics. Envisioned systems emphasize modularity, interoperability, and seamless data lineage across all manufacturing functions. One interviewee noted, "*We need some common data architecture, including data definitions. I would even call it common data model*".

2nd-Order Theme: Cloud and Edge Computing as Platforms for Future Growth. Respondents see hybrid cloud and edge architectures as foundational for scalable, secure, and regulatory-compliant operations. This includes deploying AI models at the edge for local inferencing and leveraging cloud services for centralized analytics and model management. As one expert noted, "*Our architecture should evolve in this direction: whatever makes sense to run at the edge, should be at the edge*".

Aggregate dimension: Building Organizational Readiness for AI-Driven Compliance

Finally, participants emphasized the importance of preparing organizations for continuous regulatory adaptation and AI integration. Beyond technology, this requires evolving compliance strategies, workforce skills, and governance structures.

2nd-Order Theme: Proactive Adaptation to Emerging Regulatory Guidance. Organizations increasingly anticipate regulatory shifts related to digital and AI technologies, highlighting the importance of proactive, forward-looking compliance strategies. Rather than waiting for finalized regulatory texts, several participants described efforts to interpret draft guidance, engage in early dialogue with authorities, and integrate continuous validation into development pipelines. As one professional reflected, "*The requirements keep getting stricter, or at least it feels that way.*" This sentiment reflects a broader shift from reacting to new rules after they appear toward staying ahead of them by monitoring changes and updating validation practices as part of everyday work.

5 Discussion

This study shows that while pharmaceutical and medical device manufacturers increasingly explore AI and ML to improve production and compliance, progress is constrained by two interconnected factors: fragmented data infrastructures and regulatory uncertainty. Our analysis highlights how legacy system silos limit data accessibility, and how evolving validation requirements create friction for AI adoption. At the same time, interviewees identified pathways forward, emphasizing the need for unified data architectures, embedded governance mechanisms, and compliance-by-design approaches such as RegOps. These findings suggest that advancing AI in regulated manufacturing requires not only technical innovation but also systematic integration of RegOps into digital architectures.

5.1 Regulation as a Structuring Force in AI-Driven Systems

Our findings show that regulation is not only a constraint but also shapes system design and governance, aligning with studies highlighting how regulatory expectations influence AI-enabled healthcare solutions [21]. Regulatory frameworks such as GxP/GMP, GAMP 5, and ISO 13485 influence architectural decisions, validation workflows, and organizational governance structures. Interviewees described how system modularity, auditability, and data lineage are often tailored explicitly to meet regulatory requirements.

Some participants viewed regulations as slowing AI development, while others saw them as enablers for digitalization. Aligning model lifecycle management with regulatory expectations was seen as essential for trustworthy, scalable AI [21]. Yet, AI integration challenges traditional validation in traceability, reproducibility, and change control, leading organizations to pilot solutions in low-risk settings before broader adoption.

5.2 Data Architecture as a Prerequisite for Compliant AI

A central theme emerging from the interviews is the foundational role of data architecture in enabling compliant AI systems. Participants highlighted the challenges posed by fragmented data landscapes, undocumented pipelines, and inconsistent metadata standards. To address these issues, several organizations are implementing enterprise-wide data platforms (such as validated data lakes or industrial data fabrics) that support data integrity, traceability, and contextualization.

These architectures form the basis for AI readiness, allowing for the integration of machine learning pipelines into GxP-compliant environments. However, the development and maintenance of such infrastructures require sustained coordination across quality, IT, data engineering, and domain-expert functions. Interviewees described the difficulty of harmonizing GxP documentation requirements with the iterative nature of machine learning, underscoring the need for architectural frameworks that bridge compliance and agility. Our results align with prior work on data fabrics and smart manufacturing architectures [22], extending them by emphasizing the need for embedded validation traceability and audit-ready provenance in regulated environments.

5.3 Toward Compliance-by-Design and RegOps

The convergence of AI adoption and digital transformation is prompting a shift toward *compliance-by-design*, which is a proactive strategy in which regulatory considerations are embedded into the design and operation of technical systems from the outset. Emerging practices observed in the field include:

- Modular system design to isolate validated components from experimental subsystems [23];
- Integration of automated documentation (e.g., model cards, data lineage trackers) [24];
- Use of version-controlled pipelines to capture training data, hyperparameters, and performance metrics [25].

These practices align with the emerging paradigm of *Regulatory Operations (RegOps)* [26], which seeks to operationalize compliance as an integrated part of system development lifecycles. Several organizations reported initial steps toward formalizing RegOps practices, often through cross-functional teams that span regulatory affairs, quality assurance, and digital engineering. Nevertheless, there is limited methodological guidance or standardized tooling in this area, representing a critical area for further development.

5.4 Proposed Layered Architecture Framework

The proposed framework in Figure 1 illustrates a validation-aware, layered architecture for compliant AI/ML deployment in regulated manufacturing. It integrates data ingestion, storage, lifecycle management, deployment, and compliance logging within a unified structure that ensures traceability and continuous oversight. Cross-cutting traceability and compliance layers embed auditability into every stage, supporting adaptation to evolving regulations.

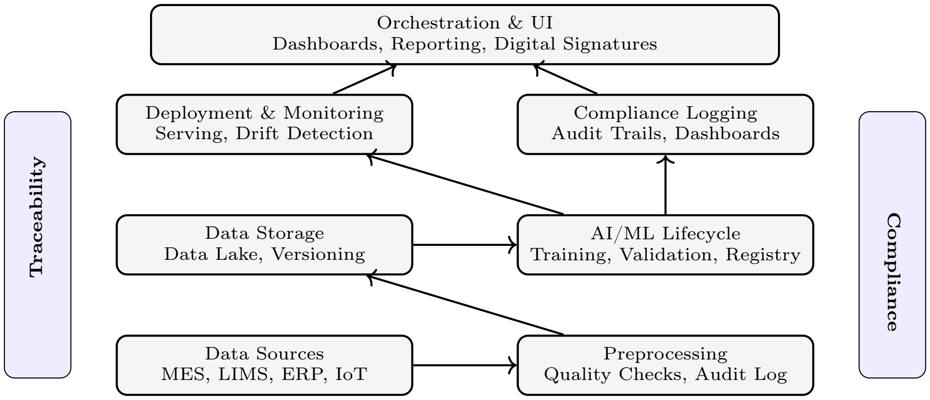

Fig. 1. Validation-aware data architecture for AI in regulated manufacturing. Cross-cutting compliance and traceability layers ensure end-to-end validation and auditability.

Table 2 complements the architectural model by comparing common data architectures used in regulated manufacturing. Figure 1 explains the internal logic of a validation-aware system, while the table shows how different architectures handle governance, validation, and scalability. Together, they outline the design space for building AI-ready infrastructures that ensure both performance and regulatory compliance. Both artefacts are based on findings from the literature and expert interviews.

Table 2. Comparison of data architecture paradigms in regulated manufacturing, focusing on their principles and implications for software business and AI governance.

Architecture Type	**Core Principles / Keywords**	**Implications for Software Business and AI Governance**	**Ref.**
Centralized Data Lake / Warehouse	Unified data hub, ETL pipelines, role-based access, analytics integration.	Facilitates centralized analytics capabilities and enterprise-wide AI initiatives but limits local innovation and agility across sites.	[27]
Operational Data Historian	Time-series storage, SCADA integration, contextual metadata, event traceability.	Acts as a high-integrity data backbone supporting traceable software updates and predictive maintenance services.	[28]
Medallion (Layered / Lakehouse)	Layered curation (Bronze-Silver-Gold), schema evolution, version control.	Enables modular software pipelines, transparent data refinement, and staged AI model deployment with reproducibility.	[29]
Data Mesh	Domain ownership, federated governance, APIs, data-as-product concept.	Promotes cross-organizational collaboration and scalable data product ecosystems, but increases coordination and compliance complexity.	[30]
Data Fabric	Semantic layer, active metadata, automation, orchestration.	Supports interoperability and explainability across business units; useful for AI lifecycle monitoring and governance automation.	[31]
Hybrid / Federated Architecture	Multi-cloud connectors, governance overlay, distributed nodes.	Balances global compliance control with local autonomy, enabling platform-based business models and continuous AI validation.	[14]

Both frameworks emphasize that compliant AI adoption is not solely a technical challenge but a systemic design problem. The layered model demonstrates how traceability and validation can be embedded directly into data and model workflows, while the comparative analysis shows alternative pathways for achieving this integration across organizational boundaries. Together, they provide a foundation for developing RegOps-oriented architectures that align AI innovation with the regulatory expectations of the life science industry.

Collectively, these comparisons position our study within the growing research connecting regulatory frameworks, data architecture design, and AI governance [31]. By extending these ideas into the manufacturing context, we show how RegOps principles can turn *compliance-by-design* from a conceptual goal into a practical framework for continuous validation. This highlights that trustworthy AI in regulated industries relies not only on technical accuracy but also on architectural and organizational readiness.

5.5 Limitations of the Study

This study has several limitations that should be acknowledged. First, due to the highly regulated nature of pharmaceutical and medical device manufacturing, participants were constrained in what they could share regarding internal systems, data architectures, validation workflows, and AI implementation strategies. These confidentiality requirements limited the granularity of some responses, particularly those concerning proprietary tools and processes, leading to higher-level descriptions rather than detailed technical disclosures.

Second, the participating companies represented a range of organizational sizes and geographic contexts. While this diversity offered a broad view of the sector, it also introduced variation in digital maturity, infrastructure, and regulatory environments. Smaller and regionally focused firms tended to be more reserved in discussing ongoing AI or data initiatives, whereas larger multinational organizations often operate with greater internal resources and established digital strategies.

Notably, large pharmaceutical companies such as Pfizer and Siemens, although not part of this interview study, have openly reported their adoption of AI/ML and digital technologies in manufacturing. Pfizer, for instance, applies AI/ML across its global production network, supporting use cases such as automated visual inspection, real-time anomaly detection, and AI-driven root cause analysis, all underpinned by its Manufacturing Intelligence Platform and Manufacturing Excellence (IMEx) program [32]. Siemens has also emphasized the integration of digital twins, IoT sensors, and robotics into production workflows to improve process reliability, yield, and cycle time efficiency [32]. These examples illustrate that AI/ML implementation is gaining traction in the regulated sector, with leading companies demonstrating both the technical feasibility and organizational value of digital transformation in compliance-intensive environments.

Future research may benefit from broader international sampling and, potentially, even sector-specific segmentation.

6 Conclusion

This study examined how data architectures, AI adoption, and regulatory practices co-evolve in GxP/GMP-regulated manufacturing. Drawing on interviews with industry experts, it shows that organizations are in a transitional phase: data systems remain fragmented, AI initiatives are exploratory, and compliance practices are still largely reactive. Nevertheless, there is a clear shift toward aligning digital transformation with regulatory expectations through validation-aware architectures and emerging RegOps approaches.

The study contributes by empirically linking three domains often addressed separately: data architecture design, regulatory governance, and AI lifecycle management. It positions validation-aware architectures as technical enablers and RegOps as an integration strategy for scalable and trustworthy AI within regulated environments.

From a practical perspective, the proposed framework supports organizations seeking to operationalize compliance-by-design, strengthen data governance under GxP constraints, and automate validation without compromising auditability. From a research perspective, it opens opportunities to evaluate RegOps implementation across industrial contexts, measure its effect on compliance effort, and explore how AI/ML techniques (such as explainable models, probabilistic reasoning, and hybrid approaches) can be engineered to meet lifecycle validation requirements [33]. As regulated industries advance toward data-driven automation, the ability to design AI systems that are transparent, auditable, and compliant will remain central to both industrial practice and research.

References

1. Kodumuru, R., Sarkar, S., Parepally, V., Chandarana, J.: Artificial intelligence and Internet of Things integration in pharmaceutical manufacturing: a smart synergy. *Pharmaceutics* **17**(3), 290 (2025)
2. Niazi, S.K.: Regulatory perspectives for AI/ML implementation in pharmaceutical GMP environments. *Pharmaceuticals* **18**(6), 901 (2025)
3. Cizauskas, C., DeBenedictis, E., Kelly, P.: How the past is shaping the future of life science: the influence of automation and AI on biology. *New Biotechnology* **88**, 1–11 (2025)
4. Shah, M., Kakkar, A., Natarajan, V.: CMC and GxP. In: *Translational Pulmonology*, 437–441. Academic Press (2025)
5. Abhilash, P.M., Luo, X., Liu, Q., Madarkar, R., Walker, C.: Towards next-generation smart manufacturing systems: the explainability revolution. *npj Advanced Manufacturing* **1**(1), 8 (2024)
6. Bilal, A., Ebert, D., Lin, B.: LLMs for explainable AI: a comprehensive survey. *arXiv preprint* arXiv:2504.00125 (2025)
7. Gioia, D.A., Corley, K.G., Hamilton, A.L.: Seeking qualitative rigor in inductive research: notes on the Gioia methodology. *Organizational Research Methods* **16**(1), 15–31 (2013)
8. NTT DATA: A GxP data fabric: the foundation of AI in pharma manufacturing. (2024). https://www.nttdata.com/global/en/insights/focus/2024/data-fabric-for-the-pharmaceutical-industry

9. Moghaddam, M., Cadavid, M.N., Kenley, C.R., Deshmukh, A.V.: Reference architectures for smart manufacturing: a critical review. *Journal of Manufacturing Systems* **49**, 215–225 (2018)
10. International Society for Pharmaceutical Engineering (ISPE): Artificial intelligence. *ISPE.org* (2025). https://ispe.org/topics/artificial-intelligence (accessed 25 July 2025)
11. Raji, I.D., Buolamwini, J., Gebru, T., Mitchell, M., Moran, T., Barocas, S.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 33–44 (2020)
12. Papagiannidis, E., Mikalef, P., Conboy, K.: Responsible artificial intelligence governance: a review and research framework. *Journal of Strategic Information Systems* **34**(2), 101885 (2025)
13. International Society for Pharmaceutical Engineering (ISPE): AI maturity model for GxP applications: a foundation for AI deployment. *Pharmaceutical Engineering*, March–April (2022). https://ispe.org/pharmaceutical-engineering/march-april-2022/ai-maturity-model-gxp-application-foundation-ai (accessed 25 July 2025)
14. Higgins, D.C., Johner, C.: Validation of artificial intelligence-containing products across the regulated healthcare industries. *Therapeutic Innovation & Regulatory Science* **57**(4), 797–809 (2023)
15. European Parliament and Council: Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence ("EU AI Act"). *Official Journal of the European Union* L 1689 (2024)
16. Hoessler, S., Carbon, C.C.: Digital transformation in incumbent companies: a qualitative study on exploration and exploitation activities in innovation. *Journal of Innovation and Entrepreneurship* **13**(1), 46 (2024)
17. Ren, X., Jing, H., Zhang, Y.: Construction of digital transformation capability of manufacturing enterprises: qualitative meta-analysis based on current research. *Sustainability* **15**(19), 14168 (2023)
18. Al Zami, M.B., Shaon, S., Quy, V.K., Nguyen, D.C.: Digital twin in industries: a comprehensive survey. *IEEE Access* **13**, 67890–67925 (2025)
19. Wilkinson, S.R., Aloqalaa, M., Belhajjame, K., Crusoe, M.R., de Paula Kinoshita, B., Gadelha, L., Goble, C.: Applying the FAIR principles to computational workflows. *Scientific Data* **12**(1), 328 (2025)
20. Presciuttini, A., Cantini, A., Portioli-Staudacher, A.: From explanations to actions: leveraging SHAP, LIME, and counterfactual analysis for operational excellence in maintenance decisions. In: *Proceedings of the 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 1–6. IEEE (2024)
21. Ajmal, C.S., Yerram, S., Abishek, V., Nizam, V.M., Aglave, G., Patnam, J.D., Raghuvanshi, R.S., Srivastava, S.: Innovative approaches in regulatory affairs: leveraging artificial intelligence and machine learning for efficient compliance and decision-making. *AAPS Journal* **27**(1), 22 (2025)
22. Armbrust, M., Ghodsi, A., Xin, R., Zaharia, M.: Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: *Proceedings of CIDR*, vol. 8, 28 (2021)
23. Dubey, A., Yang, Z., Anžel, A., Hattab, G.: Protocol for implementing the nested model for AI design and validation in compliance with AI regulations. *STAR Protocols* **6**(2), 103771 (2025)

24. Kalokyri, V., Tachos, N.S., Kalantzopoulos, C.N., Sfakianakis, S., Kondylakis, H., Zaridis, D.I., Tsiknakis, M.: AI model passport: data and system traceability framework for transparent AI in health. *arXiv preprint* arXiv:2506.22358 (2025)
25. Schlegel, M., Sattler, K.U.: Capturing end-to-end provenance for machine learning pipelines. *Information Systems* **132**, 102495 (2025)
26. Toivakka, H., Granlund, T., Poranen, T., Zhang, Z.: Towards RegOps: a DevOps pipeline for medical device software. In: *Product-Focused Software Process Improvement Conference*, 290–306 (2021)
27. Zhao, X., Zhang, C., Guan, S.: A data lake-based security transmission and storage scheme for streaming big data. *Cluster Computing* **27**(4), 4741–4755 (2024)
28. Huang, S., Chen, Y., Chen, X., Liu, K., Xu, X., Wang, C., Brown, K., Halilovic, I.: The next-generation operational data historian for IoT based on Informix. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 169–176 (2014)
29. Strengholt, P.: Building medallion architectures: designing with Delta Lake and Spark. O'Reilly Media, Sebastopol, CA (2025)
30. Machado, I.A., Costa, C., Santos, M.Y.: Data mesh: concepts and principles of a paradigm shift in data architectures. *Procedia Computer Science* **196**, 263–271 (2022)
31. Blohm, I., Wortmann, F., Legner, C., Köbler, F.: Data products, data mesh, and data fabric: new paradigm(s) for data and analytics? *Business & Information Systems Engineering* **66**(5), 643–652 (2024)
32. The Pharmaceutical Post: AI and data in pharmaceutical manufacturing. *The Pharmaceutical Post*, Issue 18, 10–13 (April 2024)
33. Bhat, V.N., Bharati, S., Bothiraja, C., Sangshetti, J., Gaikwad, V.: A review on intervention of AI in the pharmaceutical sector: revolutionizing drug discovery and manufacturing. *Intelligent Pharmacy* (2025)