LUISS

Degree Program in Data Science and Management

Course of Data Mining I

Data Governance: A Case Study in a Multinational Pharmaceutical Organization

Prof. Finocchi Irene

SUPERVISOR

Prof. Kazemargi Niloofar

CO-SUPERVISOR

Tartaglia Camilla 786951

CANDIDATE

Academic Year 2024/2025

Contents

Chapter 1

Introduction

This chapter introduces the motivation behind the thesis and defines the key research objectives. As artificial intelligence (AI) becomes increasingly integrated into enterprise systems, the need for robust and scalable data governance grows. **Section 1.1** explores the foundational tension between innovation and control—highlighting why data governance must evolve to support ethical, explainable, and compliant AI.

Section 1.2 provides an overview of the pharmaceutical sector, with a particular emphasis on its data-intensive nature. It discusses structural characteristics such as R&D complexity, global regulatory constraints, and the increasing reliance on AI-driven systems across domains like clinical trials, pharmacovigilance, and supply chain management.

This section establishes why pharma is a particularly suitable context for exploring the future of data governance.

Finally, in **Section 1.3**, the methodology adopted throughout the thesis is outlined. The research is grounded in a qualitative analysis of existing governance frameworks, complemented by domain-specific studies in pharmaceutical AI deployments.

The structure of the thesis is also detailed, offering a roadmap through the subsequent chapters, which cover theoretical foundations, sector-specific analysis, challenges, and proposed governance solutions.

1.1 Background, Motivation, and Research Objectives

In recent years, the integration of artificial intelligence (AI) into business processes has transformed how organizations operate, make decisions, and deliver value. With access to increasingly large and complex datasets, AI systems have grown more powerful — but also more opaque and potentially risky. Among the most pressing concerns is the presence of bias in AI models, which can lead to unfair, discriminatory, or unreliable outcomes. These risks are heightened when models are trained on unbalanced, incomplete, or poorly governed data.

As a response to these challenges, data governance has emerged as a critical field that defines how data is collected, managed, and used across its lifecycle. Effective governance frameworks ensure not only compliance with legal and regulatory standards but also promote ethical data use. They are essential for improving data quality, traceability, and accountability — all of which are foundational for developing fair and transparent AI systems.

This issue is particularly relevant in the pharmaceutical industry, where decisions made by AI can have direct implications for public health, research outcomes, and regulatory compliance. The sector handles large volumes of sensitive data, often originating from heterogeneous sources such as clinical trials, supply chains, or real-world evidence. These complexities require robust governance strategies to mitigate

the risk of bias and maintain stakeholder trust.

Against this backdrop, the thesis explores how data governance can serve as a tool to prevent bias in AI systems operating on large-scale datasets, with a specific focus on pharmaceutical organizations.

The research is guided by the following question:

Which data governance techniques are most effective in preventing bias in AI models operating on large volumes of data, ensuring fairness and transparency?

To address this question, the thesis sets out the following objectives:

- Review the core principles of data governance and identify their relevance to AI systems.
- Analyze the current state of data governance practices in the AI context and identify key limitations.
- Investigate the specific governance challenges within the pharmaceutical sector.
- Propose actionable strategies and best practices for bias prevention, AI integrity, compliance support, and cross-functional alignment in complex enterprise settings through data governance.
- Apply these insights to a real-world case study to assess how governance techniques can be implemented effectively.

1.2 Overview of the Pharmaceutical Industry: Structure and Data Management Challenges

The pharmaceutical industry is a highly regulated and data-intensive sector that operates at the intersection of science, healthcare, and global commerce. Its structure is typically composed of several core functions, including research and development (R&D), manufacturing, regulatory affairs, supply chain management, and commercial operations. Each of these areas generates and relies on vast amounts of data, ranging from clinical trial results and patient safety records to production data, distribution logs, and real-world evidence collected post-market.

Data plays a critical role in enabling pharmaceutical companies to innovate, ensure compliance with strict regulatory standards, and bring safe and effective products to market. However, the management of such data is complex due to its volume, sensitivity, and variety. Information often originates from multiple sources—internal systems, third-party providers, healthcare professionals, and patients—and exists in different formats and levels of quality. This makes integration and standardization particularly challenging[1].

Moreover, the industry faces increasing pressure to accelerate development timelines and reduce costs while maintaining rigorous oversight. These demands have intensified the need for robust data governance frameworks that ensure data is accurate, traceable, and fit for use across different functions and jurisdictions. Effective governance is essential not only for meeting compliance requirements, such as those set by the FDA or EMA, but also for supporting advanced analytics and AI applications[2].

[1]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. Communications of the ACM, 53(1), 148–152.

[2]Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.

Despite the growing adoption of digital technologies, many pharmaceutical organizations continue to struggle with fragmented data architectures, inconsistent master data, and siloed governance practices. These issues can lead to inefficiencies, reduced data trust, and risks of bias when data is used in AI-driven decision-making[3]. Addressing these challenges requires a coordinated effort to align governance policies, systems, and tools—making data governance a strategic priority for the industry[4].

1.3 Methodology and structure of the thesis

This thesis adopts a qualitative research approach, combining theoretical analysis with a practical case study to investigate the role of data governance in mitigating bias in AI systems. The methodology is structured around two main components:

- **Theoretical Analysis**: A literature review of academic sources, industry frameworks, and regulatory guidelines is used to explore the relationship between data governance and bias in artificial intelligence. This analysis provides the conceptual foundation for identifying relevant governance techniques and assessing their impact on fairness and transparency.

- **Case Study**:The theoretical findings are applied to a real-world context through a case study of a global pharmaceutical company that uses enterprise tools such as SAP S/4HANA and SAP Master Data Governance (MDG) for data management. Rather than evaluating the current system's effectiveness, the case study focuses on how data governance practices should be adjusted and opti-

[3]Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning.* Available at `https://fairmlbook.org/`.

[4]World Economic Forum. (2020). *Responsible Use of Technology: The IBM Case Study.* Retrieved from `https://www.weforum.org/whitepapers`.

mized in light of the existing technological infrastructure. It explores how tools like SAP MDG can support a governance model that actively mitigates bias in AI applications, by improving data traceability, oversight, and representativeness.This mixed approach allows for a comprehensive understanding of the research problem, combining general insights with sector-specific depth.

The structure of the thesis is organized as follows. This first chapter introduces the research context, outlines the motivation behind the study, and presents the main research question and objectives, followed by an overview of the methodology. Next, in Chapter 2, the thesis provides a detailed examination of data governance principles, with particular attention to master data management and the key components required for effective governance frameworks. This is followed by an exploration of the intersection between data governance and artificial intelligence, focusing on how bias arises in AI systems and identifying current governance gaps, especially within the pharmaceutical sector. The subsequent section presents a set of practical strategies and best practices aimed at mitigating bias through governance measures, both at a general level and within the specific context of pharmaceutical companies. An in-depth case study of a leading multinational pharmaceutical company is then presented, analyzing how its use of tools like SAP S/4HANA and SAP Master Data Governance (MDG) could support improved data governance strategies, and proposing adjustments to better address bias in AI applications. The thesis concludes with a summary of the key findings, a discussion of the study's contributions and limitations, and suggestions for future research directions.

Chapter 2

Understanding Data Governance

This chapter lays the conceptual groundwork for analyzing data governance in the context of AI-driven enterprises.

Section 2.1 introduces formal definitions of data governance and distinguishes it from adjacent concepts such as data management and data stewardship. **Section 2.2** outlines the core components of a governance framework, including data quality, metadata, access control, and stewardship structures.

Section 2.3 explores the critical role of master data in enabling consistency and traceability across enterprise systems.

Finally **Section 2.4** examines the sector-specific challenges faced by pharmaceutical organizations, with particular attention to compliance requirements, regulatory complexity, and the demands of high-integrity data environments.

2.1 Defining Data Governance

Data governance is a strategic framework that defines the authority and control exercised over the management of data assets. It establishes how data is created, maintained, accessed, and used across an organization to ensure accuracy, security, and regulatory compliance. Unlike operational data management—which focuses on technical aspects such as data storage, processing, and movement—data governance addresses broader organizational elements, including accountability, oversight, compliance, value realization, and issue resolution[1].

A robust data governance program formalizes policies and procedures, encourages stewardship practices across departments, and supports organizational change management. These efforts define the rules for how data is treated throughout its lifecycle (creation, usage, transformation, and archival) ensuring that it remains trustworthy and aligned with both business goals and regulatory obligations.

It is important to distinguish data governance from IT governance. Whereas IT governance focuses on technology investments, application lifecycle, and infrastructure management, data governance is exclusively concerned with treating data as a corporate asset. It spans the enterprise and governs all data-related activities, whether tied to projects or operational workflows[2].

In many organizations, data governance arises reactively, often triggered by data quality challenges or the need to implement Master Data Management (MDM). For example, a company pursuing better customer intelligence may adopt a Customer MDM program, only to realize that governance is essential to ensure its success.

[1]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148–152.

[2]Ladley, J. (2019). *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program.* Academic Press.

From a business standpoint, effective data governance enables better customer understanding, process optimization, and scalable analytics or AI capabilities. In pharmaceutical supply chains, for instance, high-quality and standardized data supports real-time inventory management and forecasting. In research and development (R&D), governed data improves confidence in decisions regarding clinical trial design, patient cohort identification, and candidate selection.

On the regulatory front, data governance helps ensure internal control mechanisms comply with key frameworks, such as:

- **GDPR** (General Data Protection Regulation) — data subject rights, transparency, and accountability;
- **21 CFR Part 11** — U.S. FDA rules for electronic records and signatures;
- **ICH E6 (R2) and GxP** — clinical and manufacturing guidelines for data integrity;
- **ISO Standards** — especially those concerning information security (e.g., ISO 27001) and quality management (e.g., ISO 9001).

Governance frameworks define protocols for data classification, retention, lineage tracking, and access control. These processes are essential for audit readiness, regulatory inspections, and legal compliance. Poor governance can lead to delayed product approvals, financial penalties, and reputational damage[3].

Yet compliance is only part of the value proposition. High-quality, governed data enables faster and more reliable decision-making, strengthens innovation capacity, and increases organizational agility. In contrast, the absence of governance leads to

[3]World Economic Forum. (2020). *Responsible Use of Technology: The IBM Case Study*. Retrieved from `https://www.weforum.org/whitepapers`.

data silos, inconsistency, and duplication—undermining the effectiveness of strategic initiatives and increasing risk.

To realize both the compliance and strategic benefits of governance, organizations must develop clearly defined roles, robust frameworks, and long-term investment strategies. Data governance is not a standalone function but a foundational capability that drives value creation and risk reduction[4]. A governance program may take the form of a formal office or a virtual group, but it must embed responsibility across all relevant actors.

Moreover, establishing a governance culture requires a shift in mindset. A data-centric organization treats data not as a secondary outcome of digital processes, but as a core driver of business value. This cultural shift mandates that data quality and usability be central to process design and that stakeholders across business and IT collaborate around shared goals.

Finally, in an era increasingly shaped by AI, the importance of effective governance cannot be overstated. Poorly governed data introduces the risk of training AI models on incomplete, inconsistent, or biased inputs—potentially resulting in decisions that are not only inaccurate but also ethically questionable[5]. As such, data governance is essential for supporting ethical AI development and safeguarding trust in data-driven enterprises.

[4]DAMA International. (2017). *DAMA-DMBOK2: Data Management Body of Knowledge.* Technics Publications.

[5]Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning.* Available at `https://fairmlbook.org/`.

2.2 Core Components of Data Governance

An effective data governance program is built upon several interconnected components, each playing a distinct role in ensuring the reliability, quality, and ethical use of data.[6] Structural governance mechanisms define the organizational framework by establishing reporting lines, governance bodies, and areas of accountability.

These mechanisms include the assignment of roles and responsibilities, along with the distribution of decision-making authority across relevant stakeholders. Clearly defined responsibilities are essential for managing specific datasets or domains, ensuring both ownership of data and accountability for its proper use throughout the organization.[7]

Key roles and governance bodies typically include the **executive sponsor, data governance leader, data owner, data steward, data governance council, data governance office, data producer, and data consumer**.

- **The executive sponsor** provides strategic direction, ensures alignment with business objectives, and secures funding for data governance initiatives. Ideally, this individual holds a high-ranking position—such as a member of the C-suite—and serves as a visible champion of the program across the organization.

- **The data governance leader** is responsible for the day-to-day coordination and oversight of the governance program. This role involves guiding the design, implementation, and continuous improvement of data policies and standards.

  Data policies are organization-wide rules that support data standards and define expected behaviors related to the management and use of data. While

[6] DAMA International. (2017). *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK2)*. Technics Publications.

[7] Ibid.

these policies can vary significantly across organizations, they typically define the "what" of data governance—what must be done and what must be avoided—while standards and procedures describe the "how", outlining specific steps to implement and enforce the policies.

Ideally, policies should be few, clearly stated, and easy to interpret. They often address critical areas such as privacy protection, consent management, and the ethical use of data, which are especially important in the context of AI systems and personal health information.[8]

Within this context, the data governance leader ensures policy compliance, coordinates the work of data steward teams, and reports on the performance and progress of the governance program.

- **Data owners** are typically line-of-business executives accountable for the quality and proper use of data assets within their domain. They define high-level data requirements, assess risks, and ensure that data governance practices are integrated into business processes.

- **Data stewards** are often subject matter experts with a deep understanding of both the data and the business processes it supports. They translate business needs into technical data requirements and are responsible for various operational tasks. These include monitoring data quality, resolving data issues, documenting rules and standards, executing daily data governance activities, and managing core metadata.

  Metadata refers to information that describes data—its origin, structure, con-

[8]Ghosh, A., Saini, A., & Barad, H. (2022). Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions. *AI and Ethics, 2*, 205–228. https://doi.org/10.1007/s43681-022-00146-9

text, and transformation history.

Effective metadata management improves transparency and enables organizations to track data usage across systems, which is essential for regulatory audits and algorithmic accountability.[3]

- **Business data stewards** are drawn from functional areas such as marketing, finance, or supply chain, where they apply their domain-specific expertise. In contrast, technical data stewards are IT professionals who collaborate with business stewards to ensure system integration and enforce data standards across platforms.

- **The data governance council** is a cross-functional, hierarchical body that defines the strategic direction of the data governance program. It ensures alignment with broader organizational objectives, oversees the development of policies, monitors performance, and drives continuous improvement.

- **The data governance office** acts as the operational core of the governance structure. It supports the council and stewardship teams by coordinating communication, organizing meetings, facilitating issue resolution, documenting decisions, and providing stakeholder education and training.

- **Data producers** are individuals or systems that generate or aggregate data. They are responsible for ensuring data accuracy and consistency at the point of creation.

- **Data consumers**, on the other hand, use data for analytical, operational, or strategic purposes. They define data requirements, report quality issues, and provide feedback to enhance data assets over time.

Together, these roles and structures form the foundation of an effective data governance model, enabling organizations to manage data as a strategic asset while ensuring compliance, accountability, and long-term value.

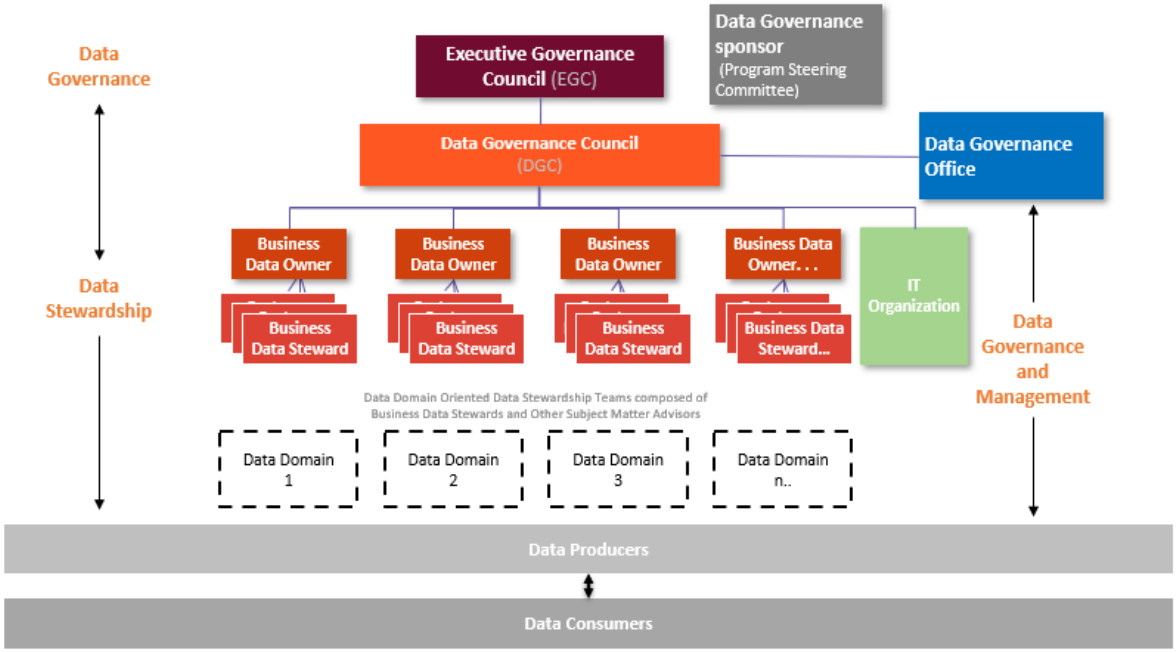

Figure 2.1: Enterprise Data Governance Operating Model: Roles and Responsibilities Across Council, Stewards, and Domains

2.3 Master Data and Its Role in Governance

Master Data refers to the core entities that are essential for the operation of a business, such as customers, products, employees, vendors, and locations. Unlike transactional data, which captures specific events (e.g., purchases, shipments, payments), master data provides the stable, consistent context required across multiple systems and business processes.[9] It plays a critical role in ensuring that the same information — for example, a product ID or customer profile — is used uniformly in various departments, from sales and finance to supply chain and analytics.

According to *Chisholm's taxonomy*, master data can be understood as an aggregation of three distinct data types:[10]

- **reference data**
- **enterprise structure data**
- **transaction structure data**

Reference data includes standardized values such as **codes and classifications** (e.g., country codes, product categories); enterprise structure data represents the **organizational setup** (e.g., charts of accounts or business units); and transaction structure data includes the **key entities required for a transaction** to occur, such as customer or product identifiers.

While master data is distinct from transactional, audit, and metadata, it depends on these types for contextual accuracy and traceability. The management of master data—commonly referred to as Master Data Management (MDM)—focuses on

[9]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

[10]Plotkin, D. (2020). *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press.

resolving inconsistencies in how these entities are represented across systems and ensuring a "single source of truth."[11]

A major challenge in MDM is entity resolution, which involves identifying and aligning instances of the same entity (e.g., a customer) that appear differently in different systems. This is often enabled through enterprise platforms such as SAP Master Data Governance (MDG), which provides tools for data harmonization, validation, and approval workflows.

In the context of AI, reliable master data ensures that machine learning models are trained on consistent and relevant input variables, reducing the likelihood of biased or incorrect predictions.[12]

In the pharmaceutical sector, master data may include critical information such as drug identification codes, molecule characteristics, manufacturing site data, and healthcare provider records.

Given the complexity of pharmaceutical operations, this data is often duplicated, inconsistent, or incomplete across key functions such as research and development, regulatory affairs, supply chain, and commercial operations. This fragmentation can result in significant operational inefficiencies and regulatory risks. For example, if a product is classified differently across departments, discrepancies may occur in reporting to regulatory authorities, potentially affecting compliance with standards set by bodies such as the FDA or EMA.[13]

To address these challenges, maintaining high-quality master data is essential—particularly

[11] Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management, 49*, 424–438.

[12] Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health, 2*(10), e486–e488.

[13] European Medicines Agency (EMA). (2021). *Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.*

for effective data integration, which is often required when consolidating clinical, operational, and commercial datasets for analytics or AI applications.[14] Without harmonized data structures and consistent definitions, the outputs of AI models can be inaccurate, misleading, or unreliable, ultimately compromising data-driven decision-making and regulatory adherence in pharmaceutical organizations.[15]

2.4 Sector-Specific Considerations in Pharma

In the pharmaceutical industry, data governance takes on a heightened level of importance due to the sector's complexity, strict regulatory environment, and direct impact on public health. Every stage of the pharmaceutical lifecycle—from molecule discovery to post-market surveillance—involves the generation and use of sensitive data that must be accurate, traceable, and compliant with global standards. Regulatory frameworks such as the FDA's 21 CFR Part 11, the EU's General Data Protection Regulation (GDPR), and Good Automated Manufacturing Practice (GAMP) require companies to maintain detailed audit trails, validate systems, and protect personal health data.[16] [17]

Pharmaceutical companies are increasingly adopting artificial intelligence (AI) to accelerate drug discovery, optimize clinical trials, improve demand forecasting, and personalize patient treatments.[18] However, the use of AI also introduces new chal-

[14]McKinsey & Company (2022). *Winning with Data: How Pharma Companies Can Gain a Competitive Edge.*

[15]Ciani, O., et al. (2021). Real-World Evidence in the EU: An Overview of Opportunities and Challenges. *Journal of Comparative Effectiveness Research, 10*(12), 901–913.

[16]FDA (2023). *Framework for Regulatory Use of Real-World Evidence.*

[17]European Medicines Agency (EMA). *Guideline on Computerized Systems and Electronic Data in Clinical Trials (2023 Draft).*

[18]McKinsey & Company (2022). *Winning with Data: How Pharma Companies Can Gain a Competitive Edge.*

lenges, particularly in ensuring the quality and integrity of training data. Models trained on biased, incomplete, or poorly documented data may produce inaccurate or non-compliant outputs.[19]

In addition to regulatory pressures, pharma organizations face operational challenges related to data fragmentation. Data is often collected across a decentralized global network that includes research laboratories, clinical trial partners, contract manufacturers, and regulatory agencies. These actors frequently rely on different formats, taxonomies, and systems, making data harmonization and governance coordination particularly difficult. Legacy IT infrastructure, third-party vendors, and outsourcing models add further complexity to data ownership and standardization efforts.[20]

To address these issues, leading organizations are increasingly adopting cross-functional governance models, investing in master data management, and aligning with international frameworks such as the FAIR data principles (Findability, Accessibility, Interoperability, and Reusability).[21] These principles provide a structured approach to data stewardship, supporting interoperability, reuse, and transparency—particularly when integrating real-world data (RWD) and real-world evidence (RWE) into regulatory submissions.[22]

Ultimately, robust data governance in the pharmaceutical sector is not only a regulatory necessity but also a strategic enabler. It supports innovation through trustworthy AI, enhances patient safety by ensuring data accuracy, and allows for consistent, efficient decision-making across highly complex, multi-stakeholder environments.

---

[19]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health, 2*(10), e486–e488.

[20]Deloitte (2020). *Pharma's Data Revolution: Unlocking the Value of Real-World Data.*

[21]Taylor, C. F., et al. (2018). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1). DOI: 10.1038/sdata.2016.18

[22]Ciani, O., et al. (2021). Real-World Evidence in the EU: An Overview of Opportunities and Challenges. *Journal of Comparative Effectiveness Research, 10*(12), 901–913.

Chapter 3

Data Governance and Artificial Intelligence

In this chapter, we explore the growing intersection between data governance and artificial intelligence (AI), especially within data-intensive and regulated environments.

In **Section 3.1**, we describe how modern AI systems challenge traditional governance structures and outline the importance of operational governance for ensuring trustworthy AI.

Section 3.2 focuses on the specific governance challenges that arise in AI systems, such as bias, unpredictability, accountability gaps, and regulatory lag.

In **Section 3.3**, we analyze the current state of governance implementation, highlighting both emerging practices and critical gaps.

Finally, **Section 3.4** narrows the focus to the pharmaceutical sector, examining AI use cases and real-world governance vulnerabilities across the industry.

3.1 Challenges in Data Governance for AI Systems

While artificial intelligence presents considerable potential for operational efficiency and innovation, its effective and responsible deployment depends heavily on robust data governance.[1]

However, aligning data governance frameworks with the specific demands of AI systems in complex, multinational environments presents a number of persistent **challenges**.

These difficulties extend beyond data management and ethics, touching on *issues of coordination, infrastructure, regulatory ambiguity, and organizational behavior.* Understanding these challenges is essential for designing effective governance strategies in high-stakes domains such as pharmaceuticals, public health, and public administration.[2]

Fragmented Accountability and Siloed Ownership

A persistent obstacle in aligning data governance with AI systems is the **fragmentation of responsibilities** across organizational silos.[3]

In large enterprises, different departments — such as IT, analytics, compliance, and business units — often manage and consume data according to their own priorities,

[1] Ghosh, A., Saini, A., & Barad, H. (2023). Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions.

[2] Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

[3] Otto, B. (2011). Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. *Communications of the Association for Information Systems*, 29(3).

standards, and definitions. This *decentralized approach* leads to inconsistent governance practices, particularly when AI development spans across teams that lack shared objectives or coordinated oversight mechanisms.

This governance misalignment reflects a broader collective action dilemma: stakeholders act based on local incentives rather than a unified goal, resulting in suboptimal outcomes for the organization as a whole. Without institutional arrangements that facilitate collaboration and clarify roles, governance efforts risk becoming disjointed or symbolic.[4]

As a result, data governance is frequently viewed not as a strategic enabler but as a compliance exercise, with limited perceived value by those expected to uphold it.

This undermines consistency in data stewardship, weakens trust in enterprise-wide data initiatives, and ultimately impairs the accountability and traceability that AI systems require to function responsibly.

Unclear Value Perception and Free-Rider Dynamics

A central challenge in implementing effective data governance is the difficulty many organizations face in demonstrating its **tangible value**.[5]

The benefits of governance—such as improved traceability, auditability, risk reduction, and regulatory compliance—are typically **long-term and indirect**, making them difficult to quantify in the short term.

As a result, data governance is often perceived as a **cost center** rather than a strategic asset, leading to underinvestment in processes such as metadata maintenance,

[4]Brous, P., Janssen, M., & Herder, P. (2020). The Dual Effects of Data Governance: A Systems Theory Perspective on Smart City Data Governance. *Information Systems Frontiers*, 22, 1109–1127.

[5]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

quality assurance, and documentation.[6]

This perception problem is amplified by a classic **"free-rider" dynamic**.

While *all units* within an organization benefit from high-quality, well-governed data, *few* are willing to take responsibility for sustaining the infrastructure that enables it.

When **governance contributions are optional or unevenly distributed**, participation declines and the overall system degrades.

In AI contexts, this can have significant consequences that can impair model accuracy, transparency, and trustworthiness:[7]

- inconsistent data labeling
- poor version control
- missing documentation

Overcoming this challenge requires not only tools and standards but also a shift in mindset—recognizing governance as an enabler of organizational intelligence rather than a compliance overhead.

Capacity Constraints: Skills, Infrastructure, and Organizational Readiness

Effective data governance for AI requires not only robust frameworks and policies but also the organizational capacity to implement them consistently.

[6]Deloitte (2021). *Data Governance: Driving Value While Mitigating Risk*.

[7]Ghosh, A., Saini, A., & Barad, H. (2023). Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions.

In practice, many organizations face significant constraints in both human and technical resources. These limitations often hinder the deployment of governance practices across the AI lifecycle, from data preparation and model validation to deployment and monitoring.[8]

A major obstacle is the **heterogeneity** of technical capabilities and data literacy across departments and geographies.

While some teams have the expertise to manage data quality and ensure compliance with governance standards, others **lack the foundational skills** required to participate meaningfully in stewardship activities.

This variation leads to inconsistent implementation of policies, fragmented accountability, and uneven data quality — all of which can compromise the reliability and auditability of AI systems.

Additionally, AI initiatives typically involve collaboration among diverse stakeholders, including data scientists, legal teams, compliance officers, and business leaders. However, **organizational silos and hierarchical boundaries** often limit effective coordination, further exacerbating governance fragmentation.[9]

These human resource gaps are often compounded by **legacy infrastructure and technological debt**.

Many organizations — particularly in the public sector — continue to operate outdated IT systems that are incompatible with modern governance requirements such as automated policy enforcement, real-time metadata tracking, or cross-platform

[8] Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

[9] Otto, B. (2011). Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. *Communications of the Association for Information Systems*, 29(3).

integration.[10]

Even when governance policies are well designed, these systems may lack the capabilities to enforce them effectively or to scale governance activities across distributed environments.

Budgetary constraints and talent shortages further limit the ability of organizations to modernize their governance ecosystems.

Investments in data governance tools, training programs, and cross-functional governance roles are frequently deprioritized in favor of short-term technological gains or operational efficiency. As a result, governance is too often implemented reactively, applied only after risks or compliance failures have materialized — rather than embedded proactively into the design and execution of AI systems.[11]

In combination, these limitations in skills, infrastructure, and readiness form a **significant barrier** to operationalizing responsible and scalable data governance.

Addressing them requires long-term organizational commitment and sustained investment in people, processes, and technology.

Unpredictability and Non-Determinism of AI

AI systems — particularly those based on machine learning and deep learning — introduce inherent **unpredictability** due to their complex, data-driven, and often opaque decision-making processes.[12]

Unlike rule-based systems, their outputs are *non-deterministic*, meaning the same input may not always yield the same outcome, especially as models evolve with new

[10]McKinsey & Company (2022). *Transforming Pharma with Data and AI.*

[11]Gartner (2021). *Seven Must-Have Foundations for Modern Data and Analytics Governance.*

[12]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health*, 2(10), e486–e488.

data.

This undermines the foundational goals of data governance, such as transparency, reliability, and reproducibility.

Even when the input data are well-curated and the lineage is documented, the internal workings of the model may remain **untraceable**, making it difficult to explain how specific decisions are made.[13]

This challenge is compounded by model drift over time, where changing data distributions gradually erode accuracy and consistency.

In high-stakes domains such as healthcare, finance, or autonomous mobility, the inability to fully audit or predict model behavior poses serious governance concerns — not only in terms of performance but also in terms of safety, accountability, and regulatory compliance.[14]

Accountability and Liability Gaps

As AI systems operate with increasing autonomy, assigning responsibility for their outcomes becomes significantly more complex.[15]

Traditional legal and governance frameworks are built on the assumption that human actors are ultimately accountable for system behavior.

However, when AI models make decisions with limited human oversight, this assumption becomes tenuous. In cases of error or harm, it is often **unclear** whether

[13]European Medicines Agency (EMA). (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.

[14]Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

[15]Batool, A., et al. (2025). Mapping accountability across the AI lifecycle: A governance-oriented framework.

liability lies with the developer, the deploying organization, the data provider, or the end user — especially in multinational contexts where legal standards diverge.[16]

This **diffusion of responsibility** poses serious risks for organizations, including regulatory penalties, reputational damage, and a loss of stakeholder trust.

To address this, governance frameworks must evolve to embed mechanisms such as decision provenance, risk profiling, and traceability throughout the AI lifecycle.[17]

Without clear structures to map accountability across the supply chain and usage scenarios, organizations may find themselves legally and ethically unprepared for the consequences of AI-driven decisions.

Bias, Fairness, and Ethical Trade-offs

Bias in AI models remains one of the most visible yet persistently unresolved challenges in data governance.[18]

When models are trained on **historical or unbalanced** datasets, they often replicate - and sometimes amplify -existing *social inequities*.

This is particularly concerning in high-stakes applications such as hiring, lending, or healthcare, where biased outcomes can lead to systemic discrimination and erode public trust.[19]

Even when governance frameworks include audits and fairness checks, these mechanisms are frequently applied inconsistently or lack the enforcement power needed to

[16]Daly, A., Hagendorff, T., & Renda, A. (2019). Governance of Artificial Intelligence: Global Strategies and Emerging Gaps.

[17]Ghosh, A., Saini, A., & Barad, H. (2023). Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions.

[18]Ghosh, A., Saini, A., & Barad, H. (2023). Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions.

[19]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health*, 2(10), e486–e488.

ensure equitable outcomes.

Governance oversight often **fails to extend to early stages** of the AI pipeline — such as data selection, labeling, and feature engineering — where many biases originate. Moreover, the trade-off between fairness and performance is rarely addressed explicitly, especially in private-sector contexts where predictive accuracy and efficiency are prioritized.[20]

Without comprehensive and proactive governance strategies, organizations remain vulnerable to ethical, reputational, and regulatory consequences stemming from biased AI behavior.

Privacy, Surveillance, and Consent Management

AI systems often rely on **large volumes of personal and sensitive data**, placing substantial pressure on data governance to ensure compliance with privacy regulations such as GDPR, manage user consent effectively, and prevent unauthorized data sharing or profiling.[21]

However, governance frameworks frequently **lag behind the evolving data demands** of AI, leaving organizations exposed to privacy *risks and legal uncertainty*.

A growing area of concern is the use of AI for surveillance — in workplaces, public services, or urban monitoring — which raises questions about civil liberties, transparency, and democratic oversight.[22]

Without strong, proactive governance, such deployments can lead to function creep,

[20]Daly, A., Hagendorff, T., & Renda, A. (2019). Governance of Artificial Intelligence: Global Strategies and Emerging Gaps.

[21]European Medicines Agency (EMA). (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.

[22]Ghosh, A., Saini, A., & Barad, H. (2023). Artificial intelligence in governance: recent trends, risks, challenges, innovative frameworks and future directions.

loss of user control, or misuse of data beyond its original intent.

These risks are further compounded in multinational organizations, where data often crosses legal jurisdictions, introducing additional complexity in ensuring consistent privacy protections.

Governance-by-design approaches must therefore address privacy, consent, and transparency not as afterthoughts, but as embedded elements of AI system architecture.[23]

Regulatory Lag and Standardization Gaps

The rapid pace of AI innovation continues to outstrip the ability of regulatory frameworks to respond effectively.[24]

This **regulatory lag** creates a governance vacuum in which organizations are left to interpret *ambiguous rules or rely on voluntary, inconsistent standards.*

In the absence of clear, enforceable guidance, many institutions resort to **"checkbox compliance"** — implementing governance as a formal exercise rather than embedding it as a structural capability that ensures transparency, accountability, and ethical oversight.[25]

At the same time, the **lack of universally adopted technical standards** for data interoperability, explainability, and auditability hinders the implementation of responsible AI practices across sectors.[26]

Without common benchmarks, it becomes difficult to assess system behavior, ensure

[23]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148–152.

[24]Daly, A., Hagendorff, T., & Renda, A. (2019). Governance of Artificial Intelligence: Global Strategies and Emerging Gaps.

[25]European Medicines Agency (EMA). (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.

[26]Gartner. (2021). Seven Must-Have Foundations for Modern Data and Analytics Governance.

consistency, or share best practices across organizational or national boundaries. While soft law mechanisms and self-regulation offer flexibility, they often lack the enforcement power needed to prevent misuse or ensure fairness. This fragmented and reactive approach to governance ultimately slows down the adoption of trustworthy and scalable AI systems.

In summary, the governance of AI systems presents a **multi-dimensional challenge** that combines technical, organizational, ethical, and legal complexity.

These challenges are not isolated; they are interconnected and mutually reinforcing.[27] *Fragmented accountability, limited infrastructure, ethical ambiguity, and regulatory uncertainty* all converge to undermine the effectiveness of current data governance models.

Addressing them will require not only updated policies and tools but also a rethinking of how governance is distributed, incentivized, and sustained across increasingly intelligent and interconnected systems.

Table 3.1: Key Challenges in Data Governance for AI Systems

Challenge	Description
Fragmented Accountability	Governance responsibilities are distributed across silos (IT, legal, business units), leading to inconsistent standards, lack of coordination, and symbolic compliance practices.

Table continued on next page

[27]Ghosh, A., Saini, A., & Barad, H. (2023). Artificial Intelligence in Governance: Recent Trends, Risks, Challenges, Innovative Frameworks and Future Directions.

Table continued from previous page

Challenge	Description
Unclear Value Perception	Governance is perceived as a cost center rather than a strategic enabler, resulting in underinvestment and free-rider behavior where benefits are shared but effort is unevenly distributed.
Capacity Constraints	Many teams lack the skills, tools, and infrastructure to support governance processes, especially in cross-functional AI environments. Legacy systems hinder integration and enforcement.
AI Non-Determinism	ML models are inherently non-deterministic and evolve over time, making outputs unpredictable and challenging traditional expectations of auditability, traceability, and reliability.
Liability Gaps	As AI systems gain autonomy, traditional legal frameworks struggle to assign responsibility clearly across developers, users, data providers, and regulators.
Bias and Fairness Trade-offs	Models trained on historical or unbalanced datasets can perpetuate or amplify existing inequalities. Fairness auditing is often ad hoc, inconsistent, or lacking enforcement.

Table continued on next page

Table continued from previous page

Challenge	Description
Privacy and Surveillance Risk	AI relies on sensitive data. Without strong governance, risks include profiling, surveillance, and cross-jurisdictional compliance breaches (e.g., GDPR, HIPAA).
Regulatory Lag	Regulatory bodies struggle to keep pace with AI innovation. In the absence of clear standards, organizations default to checkbox compliance or ambiguous best practices.

3.2 Current State of Data Governance Practices in AI Development

The growing reliance on artificial intelligence across enterprise environments has led many organizations to recognize the strategic importance of data governance.

However, in practice, the **implementation** of governance frameworks for AI development remains **highly uneven**.

While awareness of the risks associated with unregulated AI systems is increasing — particularly in regulated industries such as pharmaceuticals, healthcare, and finance — actual governance practices are often fragmented, reactive, and inconsistently enforced.

This disconnect between governance intent and operational maturity creates vulnerabilities that compromise both technical performance and compliance outcomes.

Principle-Driven Governance: Awareness Without Execution

At a foundational level, many organizations have adopted data governance policies that articulate principles related to data quality, privacy, access control, and ethical use.[28]

These policies are often informed by external regulatory frameworks such as the EU AI Act (draft), GDPR, HIPAA, or sector-specific standards like GxP in the pharmaceutical industry. On paper, these guidelines reflect a **growing awareness** of the importance of **responsible AI** development and **data stewardship**.

However, in practice, they often fall short of delivering meaningful governance outcomes.

A key issue is the **lack of operational enforcement mechanisms**. Governance policies may be formally documented but are rarely embedded into the day-to-day workflows of data scientists, engineers, or compliance teams.[29]

In many AI projects, critical decisions — such as training data selection, preprocessing methods, or model documentation — are made informally or at the discretion of individual teams. This is particularly risky in sectors like pharmaceuticals, where data sensitivity and regulatory oversight demand high levels of traceability, accuracy, and consistency.

The limitations of internal policy frameworks are compounded by reliance on **soft law instruments** and **voluntary standards**.

Many enterprises adopt reference points from global bodies like ISO, the OECD,

[28] Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

[29] Ghosh, A., Saini, A., & Barad, H. (2023). Artificial Intelligence in Governance: Recent Trends, Risks, Challenges, Innovative Frameworks and Future Directions.

or sector alliances such as the Pharmaceutical Innovation AI Consortium.[30] These frameworks provide valuable principles on fairness, explainability, and human oversight — yet they remain non-binding, and uptake varies widely. In the absence of clear implementation guidance, organizations may struggle to translate these values into specific procedures, tools, or governance roles.

Moreover, in multinational environments, organizations must navigate conflicting legal jurisdictions and **inconsistent regulatory expectations**, which further complicates compliance. As a result, policy-driven governance tends to devolve into "checkbox compliance," where formal requirements are met without achieving substantive control, consistency, or accountability across AI systems.

Ultimately, both internal policies and external frameworks reflect a growing governance consciousness, but they lack the integration, enforcement, and operational clarity needed to support robust, scalable AI governance.

Without stronger links between policy and practice, these efforts risk becoming symbolic rather than structural — highlighting the urgent need for governance models that are not only principled but also executable.[31]

Tool-Based and Platform-Centric Practices: Emerging but Incomplete

Some organizations have begun to invest in governance-enabling **technologies**, particularly tools for metadata management, data cataloging, and access control.[32]

Platforms such as *Collibra, Informatica, and Alation* are increasingly used to support centralized oversight of data assets.

[30]Daly, A., Hagendorff, T., & Renda, A. (2019). Governance of Artificial Intelligence: Global Strategies and Emerging Gaps.

[31]Deloitte. (2021). Data Governance: Driving Value While Mitigating Risk.

[32]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program.* Morgan Kaufmann.

In regulated sectors, tools like *SAP Master Data Governance (MDG)* and *Veeva Vault* have become popular for managing structured product and customer information, especially in pharma where regulatory submissions depend on accurate and consistent data.[33]

However, while these tools offer valuable infrastructure, their effectiveness depends on how deeply they are embedded in organizational processes.

Tool-based governance is often deployed in **isolation** — for instance, as part of IT or compliance initiatives — without meaningful integration into AI or data science workflows.

As a result, governance tools may support lineage tracking or role-based access controls, but still **fail to address core AI-specific issues** such as explainability, model monitoring, or fairness auditing.[34]

Moreover, many legacy systems used in large pharmaceutical companies do not natively support interoperability or dynamic governance requirements.

This limits the scalability and responsiveness of governance initiatives, especially when AI systems depend on real-time data flows or integration across R&D, regulatory, and commercial teams.[35]

The Case of AI High Performers

A small but influential subset of large enterprises — often described as AI **"high performers"** — are pioneering a more advanced approach to data governance by

[33]McKinsey & Company. (2022). Winning with Data: How Pharma Companies Can Gain a Competitive Edge.

[34]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health*, 2(10), e486–e488.

[35]Plotkin, D. (2020). *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press.

embedding governance mechanisms directly into the platforms where AI development and deployment occur.[36]

These organizations, which represent **only about 8 percent** of surveyed enterprises, are distinguished not only by their ability to derive significant EBIT from AI (20% or more) but also by the strategic alignment between their AI systems and governance structures.[37]

These enterprises have adopted what can be described as **governance-by-design**, integrating approval workflows, audit checkpoints, and policy controls directly into their data science environments. This includes embedding governance protocols into data pipelines, automating documentation processes, and aligning governance milestones with agile or DevOps cycles.

Many have also institutionalized governance oversight through **MLOps platforms or AI ethics committees** that monitor compliance, fairness, and model reliability throughout the AI lifecycle.[38]

This approach allows them to scale AI effectively, while maintaining consistency, accountability, and trust. High performers typically build modular data architectures that support rapid integration of new applications and standardize processes across development teams. They also automate key functions like data quality checks, lineage tracking, and access control — reducing operational friction and minimizing human error.[39]

Crucially, their governance frameworks extend beyond infrastructure into the man-

[36]McKinsey & Company. (2022). Winning with Data: How Pharma Companies Can Gain a Competitive Edge.

[37]Deloitte. (2021). Data Governance: Driving Value While Mitigating Risk.

[38]Ghosh, A., Saini, A., & Barad, H. (2023). Artificial Intelligence in Governance: Recent Trends, Risks, Challenges, Innovative Frameworks and Future Directions.

[39]Plotkin, D. (2020). *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press.

agement of AI-specific risks.

These organizations are more likely to proactively address fairness, privacy, and explainability concerns. For example, they routinely test the validity of their models, monitor them for drift or bias, and implement policies that support equitable outcomes and transparency.[40]

In this sense, governance is not treated as an afterthought or a constraint, but as a **strategic capability** and **competitive differentiator**.

However, the rarity of these high-performing organizations highlights the governance **maturity gap** across the broader enterprise landscape. Most companies — including those in heavily regulated sectors — continue to rely on fragmented or reactive governance strategies.

Even where some elements of governance are in place (e.g., access control or policy documentation), coverage is often uneven. Algorithmic accountability, post-deployment monitoring, and fairness validation remain weak spots in otherwise well-resourced environments.[41]

This patchwork maturity reveals that achieving robust, scalable, and context-aware governance requires more than tool adoption or policy formalization. It demands a **shift in mindset**: from governance as compliance to governance as a *structural pillar* of enterprise AI strategy. As the few high performers demonstrate, such a shift yields not only ethical and regulatory resilience, but also significant business value.

In summary, platform-integrated governance is emerging among digital leaders, but remains rare. And while voluntary standards offer guidance, they do not yet provide

[40]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health*, 2(10), e486–e488.

[41]Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

the enforceable structure needed to ensure responsible, enterprise-scale AI deployment.

The emergence of a small group of high-performing organizations shows what is possible — but also emphasizes how far most enterprises still have to go.

Building on this gap, the following chapter will explore how these governance shortcomings play out in highly regulated, data-intensive environments, with a particular focus on the pharmaceutical sector.

3.3 Focus on the Pharmaceutical Sector: AI Use Cases and Governance Gaps

The pharmaceutical industry is undergoing a significant digital transformation, driven in part by the rapid adoption of artificial intelligence (AI) across its value chain.

From early-stage drug discovery to post-market surveillance, AI technologies are increasingly used to enhance precision, accelerate timelines, and optimize resource allocation.

However, the benefits of AI in pharma are closely tied to the quality, integrity, and traceability of the data it relies on — areas where governance gaps remain persistent. As such, the pharmaceutical sector offers a compelling lens through which to examine the real-world challenges of operationalizing data governance in support of responsible AI.

AI Use Cases in the Pharmaceutical Industry

AI is currently used in pharma across a variety of domains, each of which places different demands on data governance.

In **drug discovery** and **target identification**, machine learning models are trained on large-scale omics datasets, literature mining, and compound screening data to predict molecular interactions and identify candidate compounds.[42]

These models depend heavily on high-quality, well-annotated biological data — yet many datasets remain fragmented, inconsistently labeled, or poorly documented, especially across global R&D networks.[43]

In **clinical trial** design and optimization, AI is used to simulate trial outcomes, predict patient recruitment success, and detect protocol deviations. Natural language processing (NLP) is also applied to electronic health records (EHRs) to identify eligible patient populations.

These applications introduce heightened sensitivity around privacy, consent, and bias: training data is often demographically skewed, and trial protocols may rely on incomplete or siloed metadata from different institutions or systems.[44]

Pharmacovigilance, another rapidly growing use case, leverages AI to monitor adverse drug events (ADEs) by analyzing unstructured text from social media, clinical notes, and regulatory databases.[45]

Ensuring traceability, standardization, and auditability of the data feeding into these

[42]McKinsey & Company. (2022). Transforming Pharma with Data and AI.

[43]Taylor, C. F., et al. (2018). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). https://doi.org/10.1038/sdata.2016.18

[44]Ciani, O., et al. (2021). Real-World Evidence in the EU: An Overview of Opportunities and Challenges. *Journal of Comparative Effectiveness Research*, 10(12), 901–913.

[45]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health*, 2(10), e486–e488.

systems is essential, particularly given their impact on patient safety and regulatory compliance.

Finally, in **supply chain optimization**, AI models are applied to demand forecasting, inventory management, and logistics routing.

These systems depend on *harmonized master data* — including product identifiers, supplier records, and shipment logs — which are often stored in different formats across enterprise resource planning (ERP) systems.[46]

The lack of centralized governance for such records increases the risk of AI-driven decisions being based on outdated or inconsistent data, especially in times of market volatility or crisis response (e.g., vaccine distribution during the COVID-19 pandemic).[47]

Governance Gaps and Risk Factors

Despite the widespread adoption of AI in the pharmaceutical sector, data governance frameworks frequently struggle to keep pace with the speed and complexity of technological deployment.

A particularly persistent gap lies in **data provenance** and **lifecycle traceability**, which are essential for ensuring reproducibility, regulatory compliance, and ethical reuse of data.[48]

This issue is especially pronounced in AI-enabled R&D initiatives that rely on external data sources, cross-institutional collaborations, or secondary use of health data.

In many such projects, the provenance of datasets is inadequately recorded, making

[46]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

[47]Deloitte. (2020). Pharma's Data Revolution: Unlocking the Value of Real-World Data.

[48]FDA. (2023). Framework for Regulatory Use of Real-World Evidence.

it difficult to verify how data has been collected, processed, or modified over time. This **lack of transparency** not only undermines the reproducibility of findings but also complicates regulatory submissions, where data traceability is a core requirement under frameworks such as GxP.[49]

As emphasized in a policy report by the American Medical Informatics Association (AMIA) — a leading authority in biomedical and health data governance — these shortcomings demand new governance models that extend beyond traditional clinical data management.

Specifically, AMIA highlights the need for approaches that support contextual understanding and continuity of oversight when health data is repurposed for non-clinical domains, including algorithm training, public health, or commercial research.[50]

Without enforceable, cross-cutting stewardship practices, the reuse of data in AI systems risks falling short of both regulatory and ethical expectations.

Another systemic issue is **siloed master data management**. Product, supplier, and regulatory data are often duplicated across global business units with inconsistent formatting or standards, making integration with AI systems error-prone.

These inconsistencies not only reduce model reliability but also complicate compliance with jurisdiction-specific regulations such as the EU's IDMP (Identification of Medicinal Products) standards.[51]

In response to these challenges, many pharmaceutical companies are beginning to adopt enterprise-grade tools for master data governance, such as SAP Master Data

[49]European Medicines Agency. (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.

[50]Brous, P., Janssen, M., & Herder, P. (2020). The Dual Effects of Data Governance: A Systems Theory Perspective on Smart City Data Governance. *Information Systems Frontiers*, 22, 1109–1127.

[51]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148–152.

Governance (MDG). SAP MDG enables centralized management of critical business objects — such as product hierarchies, customers, and suppliers — through rule-based validation, change workflows, and version control.

In the pharmaceutical context, this helps organizations maintain data consistency across manufacturing, regulatory, and commercial systems.

When integrated with AI workflows, platforms like SAP MDG can ensure that training data and model inputs are based on harmonized, audited master data — improving both technical accuracy and regulatory defensibility.[52]

For example, global pharma firms using SAP MDG have reported improvements in data quality for product labeling, faster onboarding of suppliers, and better coordination between regulatory affairs and supply chain teams.

These gains are especially valuable when AI is used to generate insights that depend on consolidated views of product lifecycle data.

However, even in companies that deploy SAP MDG, the full benefits are often limited by **incomplete integration with AI teams** or **disconnected governance policies** at the enterprise level.[53]

The Need for Cross-Domain Governance

One of the most complex governance challenges in the pharmaceutical industry lies in the **cross-functional and cross-domain nature of AI deployments**.[54]

[52]Plotkin, D. (2020). *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press.

[53]Otto, B. (2011). Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. *Communications of the Association for Information Systems*, 29(3).

[54]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148–152.

Data flows seamlessly between departments — from R&D to regulatory affairs, manufacturing, and market access — each with its own systems, standards, and compliance requirements.

Yet governance structures often remain functionally siloed, with **limited coordination** between technical teams, data stewards, and compliance functions. This fragmentation results in inconsistent risk assessments, delayed approvals, and misalignment between experimental AI initiatives and enterprise accountability.[55]

While the sector is generally risk-averse and compliance-driven, traditional governance models are often poorly suited to the **iterative and dynamic nature of AI development**.

Regulatory requirements such as GxP and GDPR are typically enforced through static documentation, retrospective audits, and point-in-time validations — mechanisms that cannot easily accommodate the continuous data flows and evolving model behavior that define AI systems.[56]

This disconnect is further exacerbated by the emergence of new, tech-enabled data sources such as mobile health apps, genomic datasets, and patient-generated data. These inputs are increasingly used in model training and inference, yet they fall outside the scope of conventional governance frameworks.[57]

As emphasized by the American Medical Informatics Association (AMIA), addressing this gap requires the adoption of enforceable stewardship models that support data reuse, cross-context integrity, and public trust.[58]

[55]Deloitte. (2021). Data Governance: Driving Value While Mitigating Risk.

[56]European Medicines Agency. (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.

[57]FDA. (2023). Framework for Regulatory Use of Real-World Evidence.

[58]Brous, P., Janssen, M., & Herder, P. (2020). The Dual Effects of Data Governance: A Systems Theory Perspective on Smart City Data Governance. *Information Systems Frontiers*, 22, 1109–1127.

Without such adaptive governance structures, the pharmaceutical sector risks **stalling innovation** or **compromising safety** as AI continues to expand beyond its traditional clinical boundaries.[59]

[59]Taylor, C. F., et al. (2018). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). https://doi.org/10.1038/sdata.2016.18

Chapter 4

Addressing the Gaps – Techniques and Best Practices

This chapter proposes concrete strategies to address the governance gaps identified in the previous sections.

In **Section 4.1**, we introduce a governance framework based on polycentric control and layered accountability to improve coordination across domains.

Section 4.2 examines how to make governance more visible and investable through metrics, dashboards, and data product thinking.

Section 4.3 presents governance techniques that embed fairness, transparency, and auditability throughout the AI lifecycle.

Section 4.4focuses on the technical and organizational readiness needed to scale governance—including infrastructure, standards, and integration with MLOps.

We conclude in **Section 4.5** by synthesizing these strategies into a coherent model tailored for pharmaceutical enterprises aiming to scale AI responsibly.

4.1 From Fragmented Accountability to Federated Control: A Polycentric and Layered Governance Solution

A persistent governance challenge in AI-enabled pharmaceutical organizations is the fragmentation of data responsibilities across departments.

Data is often **managed in silos**—by clinical R&D, regulatory, pharmacovigilance, or commercial teams—without a unifying framework. This results in inconsistent stewardship, unclear accountability, and difficulty scaling AI systems in a compliant and traceable way.

Additionally, many organizations struggle with **ambiguous responsibility hierarchies**. Even when governance roles exist, it's often unclear who is answerable for model risks, data quality issues, or compliance violations—especially when decisions cross functional or geographical boundaries.

To overcome these issues, organizations must adopt a *comprehensive governance architecture*—one that scales with complexity, respects local autonomy, and embeds accountability across the AI lifecycle.

This thesis proposes a combined model based on two complementary concepts:

1. **Polycentric governance**, which horizontally distributes authority across autonomous but aligned domains[1]

2. **Layered accountability**, which vertically structures responsibilities by role

[1]Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

and risk.[2]

Polycentric Governance: Distributed Authority Across Domains

In a **polycentric governance** system, each data domain (e.g., clinical research, supply chain, regulatory affairs) operates as a *semi-autonomous* center of governance.[3] These units retain control over their data processes—such as access rights, validation workflows, and usage policies—while adhering to **shared enterprise-wide principles**, metadata structures, and compliance thresholds.

This **horizontal** distribution of power ensures flexibility in meeting local regulatory requirements (e.g., GDPR in Europe, HIPAA in the U.S.), aligning with domain-specific needs while contributing to an integrated enterprise data strategy.[4]

Governance nodes remain **responsive to context**, yet are held together by centralized platforms like SAP MDG, MLOps environments, or metadata repositories that enforce interoperability and common standards.[5]

Key features of a mature polycentric model include:

- **Multi-leveled and Diffuse Authority**: Inspired by Scholte's work on governance networks, polycentric governance embraces a layered structure, where local data actors (e.g., product owners, trial leads) operate with **autonomy** but are accountable to **global standards** enforced by **centralized platforms**

[2]Otto, B. (2011). Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. *Communications of the Association for Information Systems*, 29(3).

[3]Brous, P., Janssen, M., & Herder, P. (2020). The Dual Effects of Data Governance: A Systems Theory Perspective on Smart City Data Governance. *Information Systems Frontiers*, 22, 1109–1127.

[4]European Medicines Agency (2021). Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle.

[5]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

like SAP MDG or enterprise MLOps systems.[6]

Stakeholders at different organizational "levels" (e.g., site, region, corporate) participate in framing and enforcing governance, ensuring that AI inputs and outputs are auditable and aligned with risk profiles.

- **Equity and Pluralism in Stewardship**: Polycentric governance encourages the inclusion of **diverse stakeholder perspectives** — not only data scientists and compliance officers but also frontline clinical teams, regulatory liaisons, and even patient engagement units.[7]

  In practice, this can be implemented through *cross-functional governance councils*, where ethical, legal, and technical perspectives are weighed in decisions about AI development, model approval, or sensitive data usage.

- **Fluidity and Adaptability**: Unlike legacy governance structures built around rigid hierarchies or static workflows, polycentric governance supports **continuous learning and adaptation**.[8]

  For example, governance processes can evolve through *question-based deliberation* (e.g., the 100 Questions Initiative), where stakeholders identify emerging challenges and **co-design governance rules** accordingly.

  This deliberative mode of governance allows organizations to respond rapidly to regulatory changes, new AI use cases, or risks such as algorithmic drift.

[6]McKinsey & Company. (2022). Winning with Data: How Pharma Companies Can Gain a Competitive Edge.

[7]Tschandl, P., et al. (2020). Explainable Artificial Intelligence for Medical Applications. *The Lancet Digital Health*, 2(10), e486–e488.

[8]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148–152.

- **Role of 'Bilinguals'**: A vital operational component is the use of so-called "bilinguals" — individuals with **both domain knowledge** (e.g., pharmacovigilance, clinical trial design) and data/AI literacy.[9]

  These actors act as *translators* between business and technical teams, helping shape governance questions, assess model risk, and guide cross-domain implementation.

  Their presence is essential for embedding data governance principles into agile AI workflows without sacrificing domain specificity.

Layered Accountability: Stratified Responsibility by Role and Risk

While governance power is distributed, accountability must remain structured and enforceable. As articulated by Basti and Vitiello (2023), Layered accountability ensures that **roles are clearly defined** across the *strategic, operational, and compliance spectrum*, with responsibilities aligned to risk exposure and decision authority.[10]

- **Strategic Oversight**: At the strategic level, the **Executive Governance Committee (EGC)** acts as the ultimate authority, responsible for defining program-level scope, arbitrating escalated issues, and approving the enterprise-wide data governance strategy.

  Alongside the EGC, the **Data Governance Council (DGC)** provides cross-functional leadership, translating business needs into governance directives, arbitrating data-related conflicts, and reviewing organizational performance

[9] Plotkin, D. (2020). *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press.

[10] Basti, G., & Vitiello, U. (2023). Structuring Data Governance for AI: A Multilevel Accountability Model. *Journal of Governance Studies*, 11(2).

on data-related KPIs.[11]

- **Operational Responsibility**: At the operational level, **Business Data Owners** represent their respective business areas and are held accountable for data governance metrics established by the DGC. They are responsible for enforcing internal controls and ensuring familiarity with data governance roles, principles, and tools.

  Supporting them are the **Business Data Stewards**, who serve as the point of contact for data within their domains. These stewards translate business needs into data requirements, define data elements, maintain business rules, and escalate data issues as needed.

  Operational execution is further enabled by the **Data Governance Organization (DGO)**, which provides centralized, non-IT support to the DGC, including structural documentation, prioritization of data domains, and synthesis of analytical dashboards.[12]

- **Technical and Compliance Enablement**: On the technical and compliance front, multiple specialized roles contribute to horizontal and vertical governance alignment. **Compliance officers, privacy leads, and regulatory experts** contextualize governance frameworks to regional or domain-specific requirements (e.g., GxP, GDPR, HIPAA).

  IT Data Stewards and the broader IT organization deliver the technological scaffolding for governance implementation, managing data flows, databases, and tool integrations while conducting impact analyses on data changes.

[11] Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

[12] Plotkin, D. (2020). *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press.

Finally, **Data Specialists and Data Producers/Consumers** serve at the base of the accountability chain. They ensure the accuracy, integrity, and adherence to data creation and maintenance procedures—thereby closing the loop between strategic policy and operational execution.[13]

This **vertical** stratification reinforces traceability and auditability throughout the data lifecycle, while ensuring that decisions are matched with appropriate authority and expertise.[14]

[13]Deloitte. (2021). *Data Governance: Driving Value While Mitigating Risk.*

[14]Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

Table 4.1: Summary of Polycentric Governance

Dimension	**Description**
Polycentric Governance	
Definition	Distributed governance model where each data domain operates as a semi-autonomous unit, aligned by shared principles and centralized platforms.[15]
Governance Structure	Horizontal distribution of authority across domains (e.g., R&D, Supply Chain), enabling flexibility and local responsiveness while maintaining enterprise coherence.[16]
Key Features	• Multi-level authority with global platform integration (e.g., SAP MDG)[17] • Inclusive stewardship with diverse stakeholder input[18] • Adaptability via deliberation and co-design[19] • Use of "bilinguals" bridging domain and data expertise[20]
Benefits	Improves regulatory alignment, flexibility, and stakeholder engagement across AI and data-intensive environments.

Table 4.3: Summary of Layered Accountability

Dimension	**Description**
Layered Accountability	
Definition	Stratified model that aligns responsibilities across strategic, operational, and compliance levels, based on role and risk.[21]
Governance Structure	Vertical distribution of accountability, supported by councils, stewards, and operational roles, ensuring traceability and authority delegation.[22]
Key Features	• Strategic: Executive Governance Committee, Data Governance Council • Operational: Business Data Owners, Business Stewards, DGO • Compliance/Technical: Privacy, IT Stewards, Data Producers/Consumers[23]
Benefits	Reinforces auditability, regulatory readiness, and decision transparency across the data lifecycle.[24]

Building a Resilient and Scalable Governance Framework

By **combining polycentric governance** (distributed authority) **with layered accountability** (stratified responsibility), enterprises can build a model that overcomes fragmentation without compromising control.[25]

This structure acknowledges that authority and accountability are not the same, and that separating them avoids two common pitfalls:

1. *Over-centralization*, which creates bottlenecks and stifles innovation.

2. *Diffuse responsibility*, where accountability is so scattered that no one is answerable for ethical breaches, data inconsistencies, or model failure.[26]

In pharmaceutical enterprises—where data must flow securely across borders and be reused across AI applications—this dual model offers a blueprint for scaling governance with integrity, agility, and compliance.[27]

4.2 Making Governance Valuable, Visible, and Investable

Despite increasing recognition of data as a strategic asset, data governance initiatives within pharmaceutical organizations often suffer from underinvestment and organizational inertia.[28]

[25]Brous, P., Janssen, M., & Herder, P. (2020). The Dual Effects of Data Governance: A Systems Theory Perspective on Smart City Data Governance. *Information Systems Frontiers*, 22, 1109–1127.

[26]Otto, B. (2011). Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. *Communications of the Association for Information Systems*, 29(3).

[27]McKinsey & Company. (2022). *Transforming Pharma with Data and AI.*

[28]McKinsey & Company. (2021). *The State of AI in 2021.*

As detailed in Chapter 3, a major reason for this is the difficulty of demonstrating the immediate, tangible value of governance.

While the long-term benefits—such as regulatory compliance, audit readiness, and reduced data risk—are widely acknowledged, they are frequently perceived as indirect or difficult to quantify.[29]

This perception leads to a classic free-rider dynamic, particularly acute in AI-intensive pharmaceutical companies. Here, governance plays a critical role not only in ensuring data quality but also in maintaining the traceability, reliability, and regulatory defensibility of AI models.[30]

Nonetheless, investments in foundational governance capabilities—such as metadata management, version control, lineage documentation, or automated access control—are often deprioritized in favor of short-term efficiencies.[31]

To shift data governance from perceived overhead to strategic enabler, organizations must **embed governance into the core of their operational and analytical ecosystems**.

This begins with establishing a clear value proposition for governance that resonates across functions.

In pharmaceutical contexts, this includes quantifiable improvements in cycle time for regulatory submissions, fewer errors in product labeling, reduced compliance rework, and faster onboarding of new suppliers.[32]

Governance outcomes should be linked to enterprise KPIs, not only in **data terms**

[29]Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data Governance: A Conceptual Framework, Structured Review, and Research Agenda. *International Journal of Information Management*, 49, 424–438.

[30]Deloitte. (2021). *Data Governance: Driving Value While Mitigating Risk.*

[31]Khatri, V., & Brown, C. V. (2010). Designing Data Governance. *Communications of the ACM*, 53(1), 148–152.

[32]IQVIA. (2022). *Digital Transformation in Life Sciences: Realizing the Value of Data.*

(e.g., % of clean master records, audit trail completeness), but also in **business terms** (e.g., time-to-submission, safety signal response time).[33]

Furthermore, governance activities must be made *visible*.

One way to do this is through governance **dashboards** or data governance **scorecards** that track participation, data quality, and compliance milestones by business unit.

These tools should be available not just to IT or compliance teams, but to executive leadership and domain leads as part of standard operational reporting. For example, embedding governance metrics in product lifecycle reviews or risk management frameworks reinforces its importance and ensures it remains a cross-functional priority.[34]

One of the most effective ways to transform data governance from a behind-the-scenes activity into an enterprise-wide capability is to make its operations and impacts visible.

Governance Dashboards

Governance dashboards are **interactive**, often platform-integrated tools that display real-time or near-real-time indicators of governance activity and health.[35]

Typical metrics include:

- Data quality scores (e.g., completeness, conformity, accuracy) by domain or dataset

[33]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

[34]Informatica. (2023). *Data Governance Maturity Model and Best Practices.*

[35]Collibra. (2023). *The Business Value of Data Intelligence.*

- Policy compliance rates (e.g., percentage of assets with ownership assigned, classified per GDPR)
- Lineage coverage (e.g., percentage of critical datasets with active data lineage graphs)
- Access activity and anomalies (e.g., unusual access events, expired user credentials)
- Governance workflow metrics (e.g., average time to approve data change requests)

In pharma, these dashboards are often **embedded within enterprise MDM** (e.g., SAP MDG), **cataloging** (e.g., Informatica Axon, Collibra), or **compliance platforms**.

For example, a clinical trial data dashboard might show real-time metadata validation rates, protocol completion gaps, or inspection readiness based on document traceability.[36]

Dashboards provide operational teams, data owners, and even external auditors with a live window into governance performance — improving both control and confidence in AI, regulatory, or analytical use cases.

Governance Scorecards

Where dashboards are dynamic, governance scorecards are **structured reports** issued periodically (e.g., monthly or quarterly) that summarize governance KPIs and trends.[37]

[36]Veeva. (2023). *Veeva Vault Clinical Data Management Suite Overview.*
[37]Gartner. (2022). *Measuring the Value of Data Governance Programs.*

They are especially useful for:

- *Benchmarking governance* maturity across departments or business units
- *Tracking domain participation* in governance processes (e.g., policy reviews, data stewardship training)
- *Highlighting areas of risk or inaction*, such as datasets without assigned owners, or unvalidated models in production

These reports are typically reviewed by data governance councils, domain leads, and executive sponsors to prioritize remediation, budget allocation, and platform investment.

Data Product Thinking for Strategic Governance

For governance to become strategic, metrics must be integrated into business operations.[38]

This means:

- Including governance KPIs in product lifecycle reviews (e.g., ensuring all new compounds have validated metadata before submission to regulatory authorities)
- Making stewardship and compliance metrics part of risk registers and internal audit frameworks
- Including governance health in executive dashboards, often side-by-side with operational KPIs like cycle time, approval rates, or system uptime

[38]Ladley, J. (2019). *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program* (2nd ed.). Morgan Kaufmann.

For example, in a pharma manufacturing context, data quality breakdowns in supplier or batch records could delay regulatory filings. Linking such events to governance health metrics helps build the business case for improved stewardship, training, and platform funding.[39]

The issue of uneven contribution, where some domains invest more than others while benefits are broadly shared, can be addressed through **data product thinking**.

Under this model, datasets (e.g., clinical trial master data, pharmacovigilance records, supplier metadata) are treated as products with designated owners, usage SLAs, and internal "customers".[40]

Governance becomes an integral **part of each data product's lifecycle**, with ownership linked to outcomes, funding, and accountability.

This incentivizes departments to maintain high-quality, well-documented datasets and to collaborate across boundaries, since their data products are reused across AI and regulatory workflows.

Traditional data governance models in pharmaceutical organizations often reinforce a compliance-first culture, in which data is collected and validated as a regulatory requirement rather than a reusable asset.

Dehghani's Data Mesh challenges this mindset by introducing a paradigm shift: **data should be treated as a product**, designed, maintained, and consumed intentionally, with defined ownership and measurable value.[41]

This concept is especially relevant for pharma, where data must support repeatable AI pipelines, cross-functional R&D collaboration, and end-to-end regulatory traceability.

[39]IQVIA. (2022). *Digital Transformation in Life Sciences: Realizing the Value of Data.*
[40]Dehghani, Z. (2022). *Data Mesh: Delivering Data-Driven Value at Scale.* O'Reilly Media.
[41]Dehghani, Z. (2022).

Dehghani outlines eight essential characteristics of data products.

In a governance context, the most relevant for pharmaceutical enterprises include:

1. **Discoverable**

    - All data products should be registered in a global catalog with metadata, schema definitions, and descriptions of business context (e.g., trial phase, regulatory classification).
    - Governance requirement: enforce mandatory metadata registration and make discoverability a prerequisite for data sharing.

2. **Addressable and Secure**

    - Each data product must have a stable, versioned endpoint accessible through APIs or query interfaces, with identity-based access control.
    - Governance requirement: implement role-based access policies tied to domain-level compliance rules (e.g., GDPR for patient records, GxP for manufacturing data).[42]

3. **Trustworthy and Accurate**

    - Product teams must define and meet service level objectives (SLOs) for freshness, accuracy, completeness, and schema conformance.
    - Governance requirement: establish federated quality scorecards, validation pipelines, and automated alerts for SLA breaches.

4. **Self-Describing and Well-Understood**

[42]Informatica. (2023). *Data Governance and Privacy Best Practices.*

- Every data product should publish a data contract, including documentation, sample queries, lineage, and business logic.
- Governance requirement: use metadata layers and ontologies (e.g., IDMP, MedDRA, SNOMED CT) to enforce shared semantics across domains.[43]

5. **Interoperable and Composable**

   - Data products must follow agreed-upon standards so they can be joined or reused. For pharma, this means harmonizing patient IDs, molecule codes, or adverse event categories across geographies and use cases.
   - Governance requirement: define enterprise taxonomies and ensure semantic alignment via data modeling reviews.[44]

Integrating data product thinking into governance provides a pragmatic foundation for responsible, scalable data use.

In AI-powered pharmaceutical enterprises, it enables teams to deliver traceable, compliant, and usable data at scale—aligning local autonomy with enterprise-wide integrity. Through this model, governance becomes not a bottleneck, but a catalyst for AI adoption and trustworthy data-driven innovation.

From Investment to Impact

However, these efforts must be supported by **sustained and intentional investment** to move from isolated governance initiatives to a scalable, enterprise-wide capability.[45]

[43]EMA. (2022). *Identification of Medicinal Products (IDMP): EU Implementation Guide.*
[44]Gartner. (2022). *Master Data Management and Governance Best Practices.*
[45]Deloitte. (2021). *Accelerating Data Governance Adoption for Digital Transformation.*

Research from Deloitte and Informatica emphasizes that without **long-term strategic funding** and **executive sponsorship**, even the most promising governance models struggle to achieve measurable impact or cultural adoption.[46]

Organizations must commit to developing governance as a permanent business function, not a temporary compliance fix.

A foundational element of this investment is the creation of **structured training programs** that build data literacy across both technical and non-technical staff.[47]

These initiatives help domain experts understand their governance responsibilities and empower stewards to actively manage data assets according to defined quality, privacy, and regulatory standards.

As Tableau and DAMA International highlight, governance success is directly correlated with how well stakeholders are educated and engaged.[48]

In parallel, effective change management strategies are essential to reinforce new behaviors and norms. This includes **governance onboarding** for new hires, **playbooks** for data product teams, and governance **communities of practice (CoPs**) where stewards, data owners, and analysts can collaborate.

According to Semarchy and Lumenalta, visibility into the value of governance, through dashboards and well-communicated success stories, can significantly increase participation and long-term resilience.[49]

Equally important is the institutionalization of **cross-functional governance roles** such as data product owners, business-aligned stewards, and governance leads embedded in operational units.

[46]Informatica. (2022). *CDAO Survey: The State of Data Governance and Responsible Data Use.*
[47]Tableau. (2021). *The Role of Data Literacy in Data Governance Success.*
[48]DAMA International. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2nd ed.). Technics Publications.
[49]Semarchy. (2022). *Why Data Governance Programs Fail—and How to Fix Them.*

These roles should be formally resourced, trained, and integrated into incentive and performance structures. As emphasized by Informatica and McKinsey, decentralized governance cannot function without clearly accountable roles supported by consistent tooling and authority.[50]

In regulated pharmaceutical environments, sustained investment must also support validation, auditability, and compliance readiness.

This includes maintaining traceable records in systems like SAP Master Data Governance (MDG) and Veeva Vault, automating regulatory submissions, and embedding requirements such as GxP, GDPR, and IDMP directly into operational data practices.[51]

Regulatory bodies like the FDA and EMA increasingly require organizations to demonstrate not just data quality, but also the presence of **structured governance processes** that underpin data integrity and decision-making.[52]

To support these requirements, **tooling and automation** play a crucial role.

When integrated with MLOps pipelines or real-world evidence (RWE) platforms, governance shifts from a reactive activity to a *proactive*, embedded capability—enforcing policies, tracking data lineage, and managing compliance metadata in real time.[53]

In this way, governance becomes part of the invisible infrastructure of pharmaceutical operations: continuous, transparent, and seamlessly woven into daily workflows without adding friction.

Ultimately, sustained investment in governance is **not a sunk cost**, it is a lever for unlocking digital transformation, accelerating AI integration, and building institu-

[50]McKinsey & Company. (2023). *Scaling AI With Effective Data Governance.*

[51]Veeva. (2023). *Vault Quality Suite: Regulatory and Data Integrity Compliance.*

[52]FDA. (2021). *Framework for the Use of Real-World Evidence to Support Regulatory Decision-Making.*

[53]OECD. (2021). *Recommendation on the Governance of Artificial Intelligence.*

tional trust.

As highlighted across multiple industry reports (Gartner, OECD, Deloitte), organizations that treat governance as a **strategic capability**, rather than an afterthought, realize greater long-term ROI through reduced compliance risk, faster data availability, and higher model confidence.[54]

In conclusion, overcoming the free-rider dynamic and the perception of governance as a cost requires a **dual transformation**: *cultural and structural*.

Governance must be positioned not as a compliance checkpoint, but as an enabler of safe, efficient, and scalable AI adoption. Only when governance is made valuable, visible, and investable can it become embedded in the enterprise fabric—fueling innovation while ensuring accountability and trust.

4.3 Designing for Trust: Transparency, Fairness, and Explainability

Extending Governance to Upstream AI Stages

Bias mitigation must begin at the point of data creation—well before model training or deployment. Following the principle of **"prevention over correction"**, data governance frameworks should enforce rigorous standards during the data collection, curation, and labeling phases of the AI pipeline.[55]

Governance teams should mandate bias impact assessments for any dataset used in

[54]Gartner. (2022). *Measuring the Business Value of Data Governance.*

[55]Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AAAI/ACM Conference on AI, Ethics, and Society*.

AI development, with a specific focus on documenting:

1. *Subgroup representation* (e.g., age, sex, ethnicity, comorbidities) and identifying gaps—such as underrepresentation of older adults, pregnant patients, or ethnic minorities in clinical trial datasets;[56]

2. *Data source provenance*, ensuring that datasets come from traceable and compliant sources (e.g., GxP-validated systems, verified EHR exports, real-world data repositories).[57]

To ensure these expectations are met, organizations should implement the following operational protocols:

- **Data Entry Validation Rules**: Define structured input constraints using master data frameworks (e.g., SAP MDG) or validation layers that flag missing or imbalanced values at ingestion. For instance, ensuring age values fall within valid ranges or that required demographic fields are not null before integration into training pipelines.[58]

- **Data Profiling and Source Audits**: Use automated tools (e.g., Talend, Informatica, or Great Expectations) to scan for representation gaps, frequency imbalances, or missing metadata.[59]

- **Annotation Governance**:

[56]Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

[57]EMA. (2022). *Guideline on Computerised Systems and Electronic Data in Clinical Trials*.

[58]SAP. (2023). *Master Data Governance for Health and Life Sciences*.

[59]Great Expectations. (2023). *Open Source Data Quality Framework*. https://greatexpectations.io/

  - Inter-Annotator Agreement (IAA): Require labeling teams to report on agreement metrics (e.g., Cohen's Kappa, Krippendorff's Alpha) and set a threshold below which labels must be re-reviewed.[60]
  - Annotation Protocols: Adopt structured guidelines such as the Amazon Mechanical Turk Quality Guidelines or use tooling platforms like Labelbox, Snorkel, or Prodigy that allow metadata logging (e.g., annotator ID, confidence score, review status).
  - Protected Attribute Auditing: Define explicit governance over sensitive fields like gender or race, including de-identification rules, the use of proxies (e.g., ZIP code as a proxy for ethnicity), and compliance checks with privacy laws (e.g., GDPR, HIPAA).[61]

- **Metadata Capture and Stewardship Assignment**: Align data assets with designated data stewards responsible for maintaining data lineage and fitness-for-use across domains.[62]

By embedding these practical protocols into the data intake and labeling stages, organizations can ensure that governance is not retrofitted after model development, but integrated early enough to shape the model's foundations.[63] This approach is not only technically sound but also essential for meeting regulatory and ethical expectations in high-impact domains such as pharmaceuticals and healthcare.

[60]Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.

[61]European Commission. (2016). *General Data Protection Regulation (GDPR)*; U.S. Department of Health & Human Services. (2022). *HIPAA Privacy Rule*.

[62]Informatica. (2023). *Data Stewardship and Metadata Management Best Practices*.

[63]Mitchell, M., et al. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Mandating Multi-Stage Fairness Audits

In alignment with the **principle of monitoring key data indicators** throughout the lifecycle, fairness audits must be institutionalized at all critical AI development stages: training, validation, and post-deployment.[64]

These audits are not one-time checks but **recurring governance activities** that ensure sustained equity across patient populations, clinical trial phases, and therapeutic domains.

To operationalize this, data governance frameworks should mandate:

- **Use of Standard Fairness Metrics**: Models must be evaluated using formal statistical definitions of fairness, such as:
  - Demographic Parity (equal positive prediction rate across groups)
  - Equal Opportunity (equal true positive rates)
  - Equalized Odds (equal false positive and true positive rates)

  These metrics should be calculated not only on training sets but also across test and live inference datasets, ensuring fairness holds across environments and deployment contexts.[65]

- **Deployment of Fairness Cards and Dashboards**: AI systems should be accompanied by standardized *"Fairness Cards"* or *audit dashboards*.[66]
  - Metric breakdowns by sensitive subgroup (e.g., age, sex, race, trial arm)

[64]Raji, I. D., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *ACM FAT*.

[65]Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In *NeurIPS*.

[66]Google AI. (2021). *Fairness Indicators and Model Cards*. https://ai.googleblog.com

  - Threshold comparisons against internal fairness standards or regulatory benchmarks (e.g., EMA/ICH equity guidelines)
  - Visualizations of error rates and risk differentials

- **Governance Review Boards for Contextual Alignment**: Each audit result must be reviewed by a cross-functional governance body, such as a Data Governance Council or AI Ethics Committee.
  - Interpret metric deviations in context (e.g., why a model might appear "biased" if trained on a rare-disease cohort)
  - Evaluate ethical implications where strict parity may not apply due to medical constraints
  - Approve or deny model deployment based on predefined risk tolerance criteria[67]

- **Documentation and Traceability**: Audit results, deliberations, and remediation actions must be:
  - Logged in metadata repositories or compliance platforms (e.g., Veeva Vault, SAP MDG)
  - Version-controlled and linked to the model's lifecycle metadata
  - Auditable by both internal QA functions and external regulators (e.g., FDA or EMA inspections)[68]

In pharmaceutical AI applications—where models guide trial inclusion, patient outreach, or treatment pathways—multi-stage fairness audits are not just a best prac-

[67]OECD. (2021). *OECD Framework for Classifying AI Systems and Risk*.
[68]U.S. FDA. (2021). *Good Machine Learning Practice for Medical Device Development: Guiding Principles*.

tice. They are a critical governance safeguard against the replication of systemic disparities.[69] Regular audit cycles, governed through transparent documentation and multi-disciplinary review, ensure that fairness is not incidental—but designed, measured, and maintained.

Documenting and Governing Trade-offs

One of the most persistent challenges in AI governance is the lack of transparency around trade-offs between fairness and performance.[70]

In many private-sector contexts—particularly in data-driven pharmaceutical enterprises—operational efficiency and predictive accuracy are often prioritized without explicit documentation of the ethical consequences.[71]

To address this, data governance frameworks should institutionalize a formal **Model Impact and Ethics Review (MIER) protocol**. This review must be required whenever fairness, representational balance, or subgroup equity is knowingly compromised in favor of utility.

The MIER process should include **cross-functional deliberation, executive or ethics board approval**, and **written justification** of the trade-off decision, thereby ensuring consistency with the organization's risk tolerance and ethical standards.[72]

All decisions should be versioned and stored in **centralized metadata reposito-**

[69]Leslie, D. (2019). *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*. The Alan Turing Institute.

[70]Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

[71]OECD. (2021). *State of Implementation of the OECD AI Principles: Insights from National AI Policies*.

[72]Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).

ries, where they can support traceability, facilitate future re-assessments, and enable internal or external audits.[73]

Revalidation Triggers and Contextual Fairness in Pharma

AI models, even those that initially meet fairness and accuracy standards, are not static assets—they are vulnerable to performance degradation over time due to evolving clinical data, population shifts, or label drift.[74]

In regulated environments such as pharmaceuticals, where predictive tools may influence patient inclusion criteria, treatment pathways, or adverse event detection, this variability poses both ethical and regulatory risks.

To mitigate these risks, data governance frameworks must include clearly defined, enforceable **revalidation and retraining triggers**.[75]

These triggers should be both *policy-based and automated*, activating re-assessment processes when pre-established thresholds are exceeded.

Typical scenarios include significant drops in model performance across specific patient subgroups, the incorporation of new data sources (e.g., real-world evidence or new clinical trial arms), or observable drift in feature distributions. To operationalize this, MLOps pipelines should incorporate monitoring tools capable of **tracking fairness metrics** and **subgroup-specific accuracy**, with alerts routed to governance teams for intervention.[76]

At the same time, fairness in pharmaceutical AI cannot be governed in abstrac-

[73]Leslie, D. (2020). *Understanding bias in AI for healthcare*. The Alan Turing Institute.

[74]Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, 28.

[75]U.S. FDA. (2021). *Good Machine Learning Practice for Medical Device Development: Guiding Principles*.

[76]Microsoft. (2022). *Responsible AI Maturity Model: Operationalizing Responsible AI*.

tion—it must reflect contextual equity, accounting for variations across therapeutic areas, trial phases, and demographic segments.

Governance protocols should mandate that **model performance** is monitored and reported across relevant strata, such as age, gender, disease severity, or comorbidity status.[77]

By combining lifecycle-based revalidation triggers with context-sensitive fairness evaluation, organizations can establish a governance architecture that is both **responsive and robust**.

This approach not only enhances trust and compliance, but also ensures that AI models continue to deliver equitable and reliable outcomes as clinical realities evolve.

4.4 Readiness to Govern: Infrastructure, Standards, and Scalability

As AI adoption accelerates in pharmaceutical organizations, many governance gaps stem not from policy absence, but from weak infrastructure, outdated tooling, and a lack of interoperability.

Data governance must therefore evolve into a scalable technical and organizational foundation, capable of supporting trustworthy AI at enterprise scale.

Foundations: Metadata, Validation, and Audit Readiness

Effective governance depends on robust metadata systems and lifecycle registries that go beyond fairness tracking to support data discoverability, lineage, ownership,

[77]European Medicines Agency (EMA). (2022). *Regulatory Science Strategy to 2025*.

and reusability across the enterprise.[78]

These systems form the connective tissue of scalable governance, enabling technical auditability and organizational accountability.[79]

Unlike version-controlled audit logs (discussed in 4.3), **metadata registries** support day-to-day data operations by allowing cross-functional teams to access structured metadata—such as data classification, source provenance, usage entitlements, and stewardship assignments.

Tools such as *Collibra, Informatica EDC, Alation, or Data Mesh Catalogs* (e.g., DataHub, OpenMetadata) offer integrated solutions for lineage tracing, glossary management, and federated data product registration.[80] In regulated contexts, these platforms can be configured to capture GxP-critical metadata fields, validation status, and data privacy labels in compliance with frameworks like IDMP, GDPR, and FDA audit readiness.[81]

Metadata governance should also be supported by **automated validation templates, data quality dashboards, and reporting pipelines**, ensuring that all clinical and AI-critical datasets are tagged, discoverable, and monitored through their full lifecycle.

[78]Informatica. (2021). *Metadata Management: The Foundation for Data Intelligence*.

[79]DAMA International. (2017). *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK2)*. Technics Publications.

[80]Gartner. (2022). *Market Guide for Data and Analytics Governance Platforms*.

[81]European Medicines Agency. (2023). *Data Standards and Metadata Management Requirements for Regulatory Submission*.

From Legacy to Scalable Architecture

Many pharmaceutical enterprises continue to operate within legacy IT environments that were not originally designed to support dynamic or AI-centric data governance.[82] These infrastructures, often built around rigid data warehouses or minimally governed data lakes, present structural limitations that hinder downstream enforcement of quality, compliance, and auditability. For instance, the absence of schema enforcement or integrated metadata tracking in traditional data lakes prevents traceability and complicates validation, even when formal governance policies exist.

Transitioning toward a scalable governance model necessitates **architectural modernization**. A core requirement is the implementation of **orchestration frameworks** capable of managing complex, multi-stage workflows across the AI lifecycle. Tools such as *Apache Airflow, Dagster*, or container orchestration systems like *Kubernetes* facilitate the modular execution of data ingestion, transformation, training, and deployment pipelines.[83]

These orchestration layers offer a foundation for embedding policy enforcement into operational workflows.

Scalability further requires leveraging **cloud-native architectures**, which provide the elasticity needed for real-time validation, automated compliance scans, and parallelized lineage propagation at enterprise scale.[84] Such architectures support dynamic provisioning of compute resources and facilitate integration with monitoring tools that ensure continuous alignment with governance thresholds.

Crucially, scalable governance in the pharmaceutical context must also ensure inter-

[82]McKinsey & Company. (2021). *Rewiring data governance to improve AI outcomes in life sciences*.

[83]Uber. (2020). *Introducing Michelangelo: Uber's Machine Learning Platform*.

[84]Google Cloud. (2022). *Data Governance for AI and Analytics on Cloud-Native Platforms*.

operability with **domain-specific standards**. Compliance with frameworks such as *CDISC* (for clinical trial data), *HL7/FHIR* (for healthcare interoperability), and *SNOMED CT* (for medical terminology standardization) enables pharmaceutical organizations to streamline regulatory submissions and support multi-stakeholder collaboration.[85]

The technical underpinnings of scalable governance must thus balance performance, compliance, and interoperability across evolving data ecosystems.

Integration with MLOps and Platform Governance

As artificial intelligence becomes increasingly embedded within both clinical and commercial operations, data governance can no longer function solely as a static policy layer.

Instead, it must be deeply **integrated** into the software and infrastructure that supports AI lifecycle management.[86]

This transformation calls for a shift from retrospective oversight to **platform-embedded governance logic**, where governance checkpoints are operationalized within the systems that orchestrate AI development and deployment.

Modern MLOps platforms—such as *MLflow, Kubeflow, and Azure Machine Learning*—provide native support for model versioning, reproducibility, and lineage tracking.[87] These platforms can be extended to enforce governance requirements by embedding validation gates at key decision points.

For example, models may be prohibited from progressing to deployment if they

[85]CDISC. (2023). *Standards in Clinical Research: Current Landscape and Best Practices*.

[86]Anderson, M., & McGinnis, B. (2022). *Embedding Governance into MLOps*. O'Reilly Media.

[87]Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, 28.

fail to meet predefined thresholds for fairness metrics, documentation completeness, or performance stability.[88] Governance becomes an integral part of the DevOps pipeline, monitored and enforced in real time rather than audited post hoc.

Moreover, enterprise governance requires **policy propagation** across systems. Access controls, audit flags, and data usage entitlements should remain consistent as data and models transition from experimental notebooks to production endpoints.

This level of interoperability is only possible through tightly coupled **CI/CD pipelines**, which also enable features such as automated rollback protocols, trigger-based revalidation, and real-time alerts for non-compliance.[89]

Through this approach, governance-by-design becomes not only a compliance mechanism but also a means of scaling accountability and trustworthiness across distributed AI systems.

Regulatory Agility and Enterprise Alignment

Given the stringent regulatory environment of the pharmaceutical sector, governance infrastructure must be designed for both **regulatory compliance and adaptability**.[90]

Agencies such as the FDA and EMA are increasingly focused on the digital traceability, reproducibility, and validation of AI-driven processes, particularly as models are deployed in contexts that affect patient safety or clinical decision-making.

To meet these evolving expectations, organizations must align their governance systems with regulatory principles such as GxP compliance, software-as-a-medical-

[88]Hummer, W., Feki, M., Gilbert, S., & Bianchini, D. (2021). DevOps for AI: Managing the lifecycle of machine learning applications. *ACM Computing Surveys*, 54(4), 1-35.

[89]Deloitte. (2022). *Operationalizing AI through MLOps and Model Governance*.

[90]European Medicines Agency (EMA). (2021). *Reflection paper on regulatory requirements for machine learning applications supporting human decisions*.

device (SaMD) validation, and audit-readiness.[91]

A key enabler of this alignment is the use of **audit logs and data flow maps** that conform to emerging regulatory validation frameworks. These tools should be capable of producing reproducible records of model behavior, decision provenance, and data lineage across the lifecycle of AI products.

Additionally, data pipelines must embed **regulatory metadata markers**, including consent provenance, usage constraints, and compliance status flags, to support proactive and automated governance.

Beyond regulatory readiness, **enterprise-wide alignment** across departments, regions, and subsidiaries is essential.[92]

Governance infrastructure must serve as a unified reference system that harmonizes practices while allowing for local execution and risk tailoring.

This entails the adoption of **shared data definitions, cross-functional stewardship roles, and a common framework** for policy interpretation.

Only through this combination of *central coordination and contextual flexibility* can pharmaceutical enterprises manage the scale, complexity, and risk inherent in AI-driven innovation.

[91]U.S. Food and Drug Administration (FDA). (2023). *Good Machine Learning Practice for Medical Device Development: Guiding Principles*.

[92]Informatica. (2022). *Data Governance in Regulated Industries: Enabling Global Coordination and Local Flexibility*.

4.5 Conclusion: Toward Scalable and Responsible Data Governance in AI-Driven Pharmaceutical Enterprises

This thesis has examined the evolving challenges and requirements of data governance in AI-enabled, data-intensive environments, with a particular focus on the pharmaceutical industry.

Through a structured analysis of governance gaps and a review of emerging frameworks, it has become evident that traditional governance models are no longer sufficient to support the complexity, risk, and strategic importance of modern data ecosystems.

In response, this work proposes an **integrated governance architecture** tailored to multinational pharmaceutical enterprises that operate under high regulatory scrutiny while pursuing AI-driven innovation.

A foundational element of this architecture is the implementation of **polycentric governance**, which distributes decision-making authority across autonomous yet coordinated data domains.

This structure recognizes the functional specificity of departments such as R&D, regulatory affairs, and pharmacovigilance, while promoting enterprise-wide alignment through shared principles, metadata frameworks, and compliance thresholds. Polycentric governance enables organizations to balance local flexibility with global coherence—a critical requirement in complex, regulated environments.

To complement this horizontal distribution of authority, the framework introduces **layered accountability**, a stratified model that aligns responsibilities with risk

exposure and organizational role. From executive committees that define strategic priorities, to operational stewards managing day-to-day data quality, this structure ensures that governance is both actionable and auditable across all levels.

Such clarity is particularly vital when AI decisions carry ethical implications, regulatory consequences, or patient safety risks.

However, structural models alone are insufficient unless governance becomes a visible, valuable, and investable enterprise function. Embedding **governance metrics** into dashboards, KPIs, and product lifecycle reviews elevates its relevance across business units.

The adoption of **data product thinking** further reinforces this approach by assigning ownership, service-level expectations, and accountability to datasets treated as reusable enterprise assets.

Sustained impact requires **long-term organizational investment**. This includes ongoing training programs, change management strategies, and the institutionalization of governance roles across functions.

In the pharmaceutical sector, this also means dedicating resources to audit-readiness, regulatory compliance, and documentation standards such as GxP, GDPR, and IDMP. Platforms like SAP MDG, Veeva Vault, and Collibra can be leveraged to automate validation, version control, and traceability, making governance both scalable and unobtrusive.

To ensure fairness, explainability, and accountability in AI applications, governance must be embedded across the full model lifecycle. This includes upstream bias detection, multi-stage fairness audits, and contextual subgroup monitoring.

Mechanisms like the **Model Impact and Ethics Review (MIER**) and retraining triggers based on model drift further strengthen the system's responsiveness and

integrity.

Finally, governance must be supported by a **modern, interoperable infrastructure** capable of scaling across cloud-native environments and integrating with MLOps platforms.

Toolchains should support metadata lineage, automated policy enforcement, and compliance traceability. Alignment with global data standards (e.g., HL7, CDISC, SNOMED CT) and regulatory expectations from FDA and EMA ensures that governance is not merely compliant but anticipatory.

In conclusion, this thesis argues that effective data governance is not a static policy framework but a **dynamic, multi-layered capability** that integrates organizational design, technological infrastructure, and ethical foresight.

For pharmaceutical enterprises operating at the intersection of scientific discovery, regulatory scrutiny, and AI innovation, such governance is not optional: it is essential.

Only by embedding governance into the operational DNA of the organization can companies ensure that their data is not only usable and compliant, but trustworthy, equitable, and ready for the future.

Chapter 5

Case Study – Strengthening Data Governance in a Global Pharmaceutical ERP Transformation

This chapter presents a case study of a global pharmaceutical company undertaking a **large-scale digital transformation initiative** centered on enterprise resource planning (**ERP**). For confidentiality reasons, the name of the company and specific identifying details have been omitted.

Specifically, the company is transitioning from SAP ECC to SAP S/4HANA and SAP Master Data Governance (MDG), with the explicit **aim of modernizing its master data infrastructure and embedding enterprise-wide data governance practices**.

This transformation is not merely technical, it reflects a strategic effort to stan-

dardize, control, and align data governance capabilities in support of AI adoption, regulatory compliance, and operational agility.

As the pharmaceutical industry becomes increasingly data-intensive, the quality and governance of **master data**—spanning *products, customers, suppliers, and regulatory entities*—has become a critical success factor.

In this context, the case study examines how the organization is designing and operationalizing a data governance model to support its SAP MDG implementation. The focus is on creating sustainable stewardship roles, defining metadata standards, managing cross-domain accountability, and enabling scalable data governance aligned with AI-readiness and GxP/GDPR compliance.

5.1 Overview of the Company's Use of AI and Data Infrastructure

The data governance transformation initiative reflects a strategic effort to build a scalable, future-proof **Master Data Operating Model** capable of supporting enterprise-wide data quality, compliance, and innovation.

The overarching goal of the transformation is to define a **holistic**, **four-dimensional** governance framework that aligns the organization, people, processes, and technologies involved in managing master data.

As depicted in the operating model blueprint, this governance framework seeks to establish a shared understanding of ownership, accountability, and control over data, treating master data as a corporate asset.

Holistic Governance Dimensions

1. **Organization:** The governance model defines a structured policy and process architecture that strikes a balance between standardization and business-specific flexibility. The organizational layer provides the backbone for **aligning governance with business needs** while ensuring scalability across functions and geographies.

2. **People:** Clearly assigned roles and responsibilities are at the core of the model. By **formalizing ownership and stewardship** across business units, the company aims to ensure accountability for data quality, while enhancing consistency and compliance across all data domains.

3. **Governance Process:** This dimension encompasses the **workflows and decision-making mechanisms** that guide policy creation, approval, and enforcement. It supports **cross-functional collaboration** among governance bodies and enables scalable execution of governance rules.

4. **Data & Technology:** The transformation is underpinned by the deployment of SAP MDG and SAP S/4HANA, along with other tools such as Veeva, OMP, and potentially Collibra. These technologies support **real-time validation, lineage tracking, and automated policy enforcement**.

Master Data Domains in Scope

A central element of the transformation program is the governance of *core master data*, which serves as the foundation for transactional integrity, regulatory compliance, and analytical accuracy across the organization.

Master data in MDG is subject to more centralized validation and control processes, while data in satellite systems—although relevant—remains out of governance scope during this implementation phase.

The categorization of master data domains by managing system is summarized below:

Table 5.1: Master Data Categorization by System

System	**Master Data Types**
SAP MDG (Governed)	Supplier, Customer, Finance, Core Material types
SAP S/4HANA (Transactional)	Internal Orders, Pricing Conditions, Project Work Breakdown Structures (WBS)
Other Core Systems	Material PLM, Planning, and Quality Management data
Satellite Systems (Excluded from MDG Scope)	Supplier Qualification and Risk (e.g., Ariba, Oro), Contracts, Procurement fields

The **core master data**—including supplier, customer, finance, and material records—will be fully governed within **SAP MDG**, enabling centralized validation, change control, and traceability across the enterprise. These domains are considered in-scope for the initial rollout and are tightly integrated with the Target Operating Model.

Conversely, **data maintained in S/4HANA**—such as pricing and project structures—plays a key transactional role but is **not directly governed through MDG workflows**.

Other supporting data sets (e.g., quality management attributes or planning hierarchies) reside in legacy or satellite systems and, while critical, fall outside the primary scope of governance in this phase.

This deliberate scoping allows the organization to focus on building robust gover-

nance foundations around its most valuable and sensitive data assets—those that affect finance, compliance, and supplier/customer interactions—while enabling future expansion into broader data domains.

Strategic Goals and Implementation Approach

The company's operating model transformation is guided by a phased methodology structured in three key cycles:

- **Cycle 1 – Setup:** Involves **assessing the current maturity** of master data governance across business lines, identifying existing pain points and growth opportunities.
- **Cycle 2 – Co-Design:** Focuses on **defining the Target Operating Model (TOM)** and strengthening data capabilities. This includes formalizing governance bodies, defining roles, setting interaction models, and establishing RACI matrices for key governance processes.
- **Cycle 3 – Roadmap:** Defines a **structured evolution plan** for implementing and scaling the TOM, including integration milestones for MDG go-live, partial deployment of S/4, and full alignment of both platforms.

The project ultimately aims to embed governance as an enabler of business value. By treating data as a reusable and strategic asset, the company can improve decision-making, reduce compliance risks, and accelerate AI and analytics adoption.

5.2 Governance Challenges and Opportunities Identified

The implementation of a scalable data governance framework in a multinational pharmaceutical enterprise presents inherent complexities, particularly in the context of a full-scale ERP transformation.

As part of its migration from SAP ECC to SAP S/4HANA and SAP MDG, the company undertook a comprehensive **"as-is" assessment** to examine its current governance landscape and maturity.

This diagnostic effort revealed a **fragmented environment** shaped by historical practices, uneven accountability structures, and differing levels of data governance maturity across functions and geographies.

The organization's model topology was evaluated across **three perspectives**:

- functional
- business line
- geographical

each of which exhibits unique governance challenges and degrees of readiness.

In particular, the **functional dimension** emerged as the focal point for designing the future-state operating model.

Functional areas such as *Global Supply Chain, Procurement, Commercial, Finance, and Global Functions – Data & Analytics (GF D&A)* exhibit significant variation in how they manage master data, ranging from mature federated models to fragmented, siloed approaches.

To navigate this complexity and design an effective governance model, it is essential to distinguish between core master data domains and functional dimensions:

Master Data Domains vs. Functional Dimensions

The **core master data domains**—namely *Finance, Customer, Supplier, and Material*—serve as the foundation for the enterprise's digital processes.

These domains are **not confined to a single business function**; rather, they flow across and support multiple functional dimensions. Each master data domain also comprises **distinct master data views**, reflecting the varied structuring and grouping of information based on the specific needs and processes of the contributing functions.

For example, customer master data is jointly managed by GSC (Global Supply Chain), Commercial, and Finance functions. Similarly, material master data traverses R&D, Manufacturing, Quality, and Procurement.

This **functional interdependence** creates both operational dependencies and governance ambiguity.

While master data domains require centralized standards, each functional dimension often implements its own processes, priorities, and controls. As a result, the company faces a **dual challenge**:

1. strengthening data ownership at the domain level
2. enabling cross-functional collaboration and governance alignment

Domain-Specific Governance Gaps

Finance Master Data. In the finance domain, roles such as Data Owner, Process Owner, and System Owner are formally assigned, and responsibilities are relatively well defined. However, several issues persist:

- **Lack of duplicate checks** during master data creation, leading to redundancies and compliance risks.
- **Weak version control** for hierarchy management, resulting in inconsistent historical tracking.
- **Inadequate segregation of duties**, with certain user groups able to alter master data without formal approval.
- **Poor visibility** into task status, SLAs, and ownership across lines of business (LoBs).

These gaps expose the organization to audit risks and erode confidence in financial reporting.
A centralized governance layer, coupled with clear RACI matrices and version-controlled audit trails, is needed to reinforce data quality and traceability.

Customer Master Data. Customer data governance suffers from unclear ownership, inconsistent processes, and fragmented stewardship across regions and teams. Pain points include:

- **No clear process or system ownership** for tax attributes or intercompany data.

- **Absence of access controls** on critical attributes (e.g., Import Duty, Country of Origin).
- **No retention policy** for obsolete customer records, leading to cluttered datasets.

The shared ownership model between Commercial, GSC, and Finance makes it difficult to enforce uniform rules.

To address this, the organization should define data attribute ownership at a granular level and formalize governance workflows for each phase of the customer lifecycle.

Supplier Master Data. Supplier data governance is marked by siloed operations, minimal local representation, and weak enforcement of existing policies. Common issues include:

- **Undefined ownership roles and non-standard practices** across regions.
- **Lack of centralized governance** for indirect vendors.
- **Sensitive supplier attributes** not consistently managed through proper TPF approval workflows.

While the Supplier Enablement Team ensures basic quality checks, the absence of strong governance roles results in data quality degradation and operational inefficiencies.

Stronger role clarification, data routines, and escalation workflows are required to stabilize supplier data governance.

Material Master Data. Material master data presents some of the most complex governance challenges due to its technical nature and cross-functional reach. Key issues include:

- **Fragmented maintenance** across systems, resulting in inconsistent definitions and classifications.
- **Missing audit trails and workflow visibility** for material creation and updates.
- **Poor alignment between system policies and actual business processes**, particularly in regulatory and planning contexts.

With up to 16 departments managing different types of material master data, a harmonized material hierarchy and a unified governance model are essential to reduce duplication and enforce standards.

5.3 Application of Best Practices: Data Governance Strategy and Bias Mitigation

As part of its ERP transformation journey, the company has adopted a **Hub-and-Spoke operating mode** to establish a unified, scalable, and accountable Master Data Governance (MDG) framework.

This model is designed to reflect how data is owned, managed, and used across business functions, supporting the complex interdependencies of core master data such as Customer, Supplier, Finance, and Material.

By integrating business and technology roles into a single governance framework, this structure provides the foundation for improving data quality and mitigating governance-related risks, including potential data bias.

The Hub-and-Spoke Operating Model

The Hub-and-Spoke model has been selected as the preferred governance architecture for its ability to **balance centralized coordination with decentralized execution**.

This structure is particularly well-suited to the complexities of master data management in a global pharmaceutical enterprise, where data must be governed across both functional and geographic boundaries.

- **The Hub** functions as the *central governance authority*.

  It is responsible for:

  1. defining cross-domain strategies
  2. establishing enterprise-wide data standards
  3. ensuring overall oversight

  The Hub also coordinates master data domains, that span multiple business functions and require alignment across ownership boundaries.

- **The Spokes** represent *individual business functions* (e.g., Finance, Global Supply Chain, Procurement, R&D, Commercial), within which master data domains are operationally managed.

  Each Spoke is:

1. accountable for the data linked to its specific processes
2. responsible for maintaining quality, ownership, and stewardship of that data.

- **Local Data Spokes** operate at the *plant or local operating company (LOC) level*.

  They provide frontline support for master data management activities, ensuring responsiveness to regional needs and adherence to global standards.

- **MDM Operations** is responsible for the execution of day-to-day data activities, including data entry, updates, and maintenance.

  This team supports the overall governance model by aligning data operations with the policies and standards defined by the Hub.

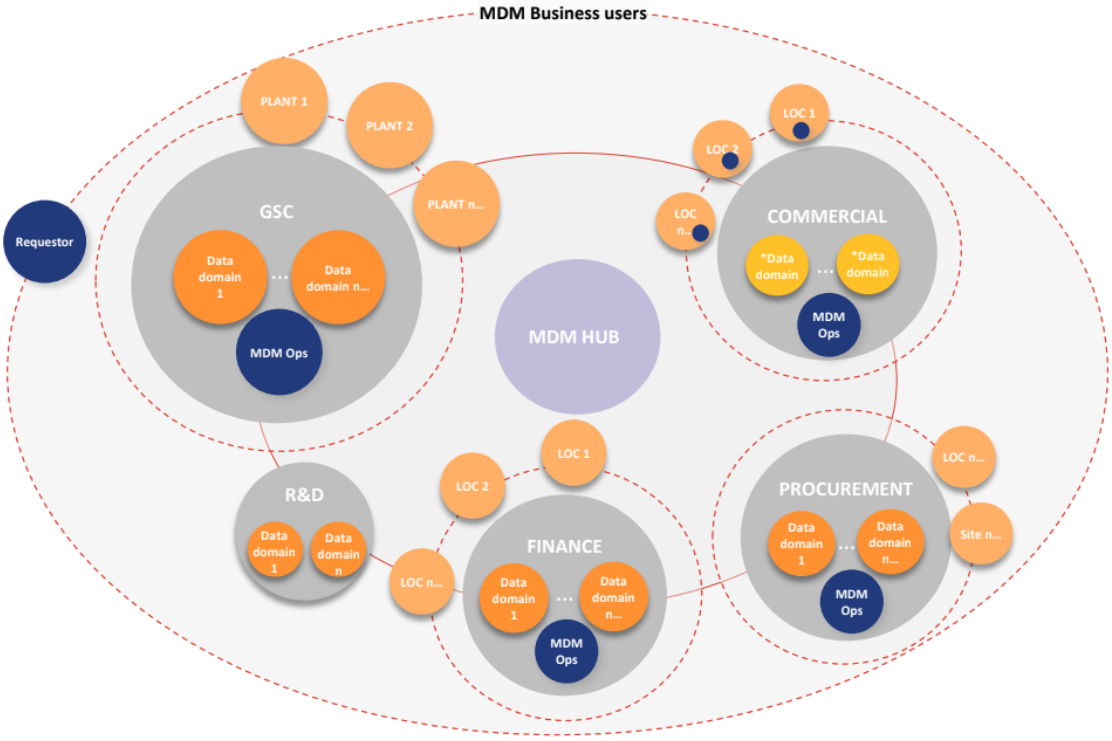

Figure 5.1: MDM Governance Hub-and-Spoke Operating Model

The Master Data Hub: Central Coordination

The Master Data Hub serves as the central organization responsible for harmonizing business and technology functions to govern master data consistently across the enterprise.

It acts as the operational engine of data governance, ensuring standardization, coordination, and oversight across all core master data domains and business functions.

Functionally, the Hub is tasked with driving enterprise-wide master data manage-

ment (MDM) efforts. It provides a unifying structure that enables strategic alignment, operational control, and quality assurance for critical data assets. **Its core responsibilities include:**

- Defining and communicating the enterprise vision for MDM, ensuring that all stakeholders understand the strategic importance of high-quality, governed data.
- Establishing global data standards, performance indicators (KPIs), and quality thresholds, particularly for data cleansing and validation processes.
- Managing enterprise-wide master data systems, notably SAP MDG and SAP S/4HANA, including tool configuration, usage policies, and integration workflows.
- Facilitating alignment and coordination among Spokes by enabling cross-functional discussions and decision-making processes.
- Providing centralized stewardship support for complex, cross-domain master data entities such as Customer and Material, which often traverse multiple organizational boundaries.

The Hub is structured around a set of clearly defined leadership and **coordination roles**, each with specific responsibilities:

- **MDM Sponsor(s)**: Senior executives who provide strategic oversight and executive sponsorship. Their role is to secure funding, champion the value of master data, and ensure business alignment with data initiatives. They act as key enablers in integrating MDM efforts into broader organizational performance and efficiency goals.

- **MDM Lead**: A senior business leader responsible for coordinating all MDM-related projects and initiatives across the organization. This role owns the business side of MDM tools, supports issue resolution (particularly for the "Big 4" domains: Customer, Supplier, Finance, and Material), and ensures that the MDM strategy is operationalized effectively across functions. The MDM Lead also acts as a bridge between business and IT stakeholders.

- **Domain MDM Leads**: These individuals are accountable for the orchestration and maintenance of CRUD (Create, Read, Update, Delete) processes within their assigned domain (e.g., Customer, Supplier, Finance, Material). Their mandate includes enforcing enterprise data standards and governance policies, coordinating with local and business data stewards, and driving data quality improvement efforts within their scope.

- **Tech MDM Lead**: This role translates business requirements into technical specifications and oversees the configuration, enhancement, and maintenance of MDM tools and repositories. The Tech Lead also plays a critical role in data harmonization, cleansing, and conversion activities, ensuring that the system landscape supports business goals.

- **Tech Data Steward**: Supporting the technical implementation of MDM policies, this role ensures the operational integrity of master data platforms. It is particularly relevant in "thin spoke" contexts, where local data stewardship is limited and requires centralized assistance.

The establishment of these roles—particularly the Domain MDM Leads—is a direct response to the high complexity and cross-functional nature of core master data. With more than 100 workflows across Customer, Supplier, and Finance domains and

over 40 distinct account groups, **centralized orchestration** becomes essential. It not only ensures process consistency and quality control but also plays a key role in bias mitigation by standardizing decisions, clarifying responsibilities, and reducing variability in data handling practices across functions.

The Spokes: Functional Ownership

The Spokes represent the **functional dimensions** of the organization—such as Finance, GSC, Procurement, R&D, and Commercial—**within which one or more core master data domains are operationally managed**.

Each Spoke assumes responsibility for the execution of governance activities tied to its relevant processes and plays a critical role in translating centrally defined standards into day-to-day operational practices.

Within a given Spoke, **multiple data domains may coexist**, each with designated ownership and stewardship responsibilities.

These domains are coordinated by an MDM Lead, who ensures consistent application of governance policies, fosters alignment with the central Hub, and facilitates collaboration across other Spokes.

In addition to the **central MDM Lead**, each Spoke also includes its own **Domain MDM Lead**, who is embedded within the functional area and directly manages the master data processes relevant to that domain.

This model enables domain-level accountability while preserving cross-functional coherence.

The **primary responsibilities** of Spokes include:

- Executing CRUD (Create, Read, Update, Delete) operations in accordance

with standards and workflows defined by the Hub.

- Monitoring and remediating data quality at the local and functional level.
- Escalating unresolved data issues and conflicts that may affect broader system integrity.

Key roles embedded within the Spokes structure include:

- **Domain MDM Lead**: Present within each Spoke, this role is responsible for orchestrating and maintaining master data within the specific domain (e.g., Customer, Supplier, Finance, Material). As detailed in the previous section, Domain MDM Leads ensure governance standards are enforced locally, coordinate with business and technical stewards, and align data operations with enterprise-wide requirements.
- **Data Owner**: Defines the data standards and business policies for a specific domain. The Data Owner is accountable for developing the data quality strategy and overseeing its implementation, ensuring that governance objectives are embedded in functional operations.

  In each Spoke, multiple Data Owners will be present, each responsible for a specific master data view, reflecting the functional segmentation of the domain.
- **Data Steward**: Supports the Data Owner by monitoring data quality, validating adherence to standards, and coordinating with technical stewards. This role acts as a translator between business and IT, turning technical quality insights into actionable business interventions. The Steward also assists the MDM Lead in defining, orchestrating, and maintaining domain-specific CRUD processes and coordinates local stewardship efforts.

In each Spoke, multiple Data Stewards will be present, each responsible for a specific master data view, reflecting the functional segmentation of the domain.

- **Local Data Steward**: Positioned at the plant or local operating company (LOC) level, this role applies centrally defined rules to local data realities. The Local Steward handles region-specific data management tasks and acts as the first point of contact for operational data quality issues.

To accommodate varying levels of data governance maturity across functions, Spokes are categorized into:

- **Thick Spokes**: Represent mature functions (e.g., Global Supply Chain) with established stewardship teams and well-defined governance practices.
- **Thin Spokes**: Represent less mature or resource-constrained areas (e.g., R&D) that rely more heavily on the Hub for governance guidance and support.

By aligning operational execution with centralized direction, the Spoke model enforces shared accountability, reduces the risk of governance fragmentation, and enhances data integrity across the enterprise.

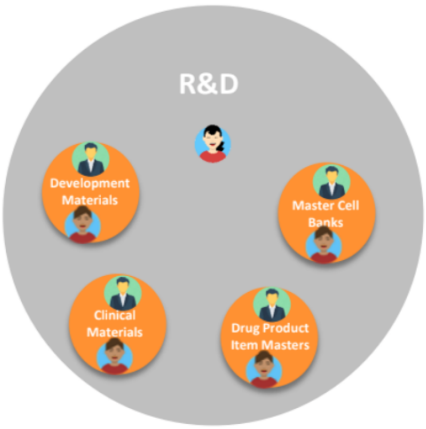

Figure 5.2: **R&D spoke** is defined as a thin spoke, meaning it has only partially mature data management operations and requires additional support. This is particularly relevant for specialized data sets such as Development Materials, Clinical Materials, Master Cell Banks, and Drug Product Item Masters.

Figure 5.3: **The GSC spoke** functions as a fully formed spoke with a high level of operational maturity. It includes a variety of sub-functions such as Logistics, Finance, Planning, Manufacturing, and Asset Management, all contributing to the Material Master Data ecosystem.

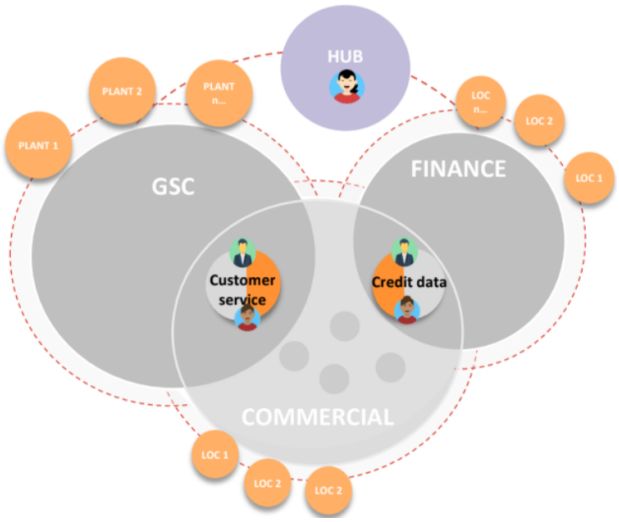

Figure 5.4: **The Commercial spoke** is characterized as a virtual overlay rather than a standalone spoke. It plays a critical role in governing Customer Master Data, which spans both the GSC and Finance domains.

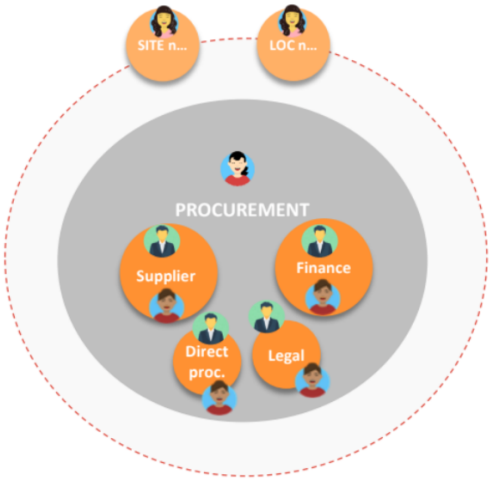

Figure 5.5: **The Procurement spoke** manages master data related to Suppliers, Finance, Direct Procurement, and Legal. It has established functional-level governance but continues to mature.

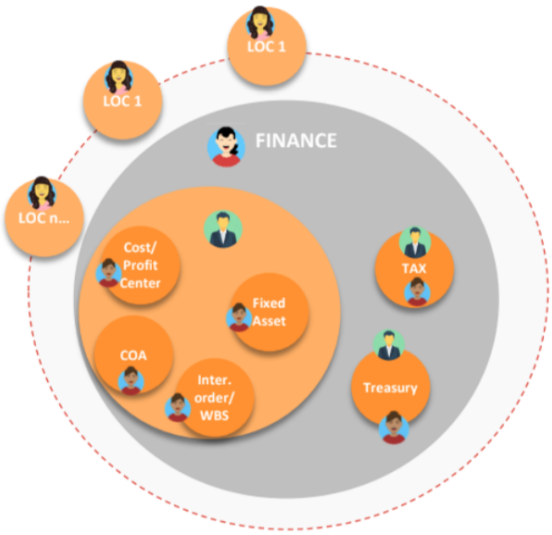

Figure 5.6: **The Finance spoke** manages a wide array of master data including Chart of Accounts (COA), Cost and Profit Centers, Fixed Assets, Internal Orders/WBS, Tax, and Treasury.

Transition Plan and Phased Implementation

Finally, to manage the complexity of the ERP and MDM transformation, the company follows a structured, three-phase rollout:

1. **Current State**: SAP ECC remains active, with local governance variations and limited standardization.
2. **Partial Deployment**: SAP MDG goes live in parallel with ECC, initiating the shift to the Hub-and-Spoke governance model.
3. **Target End State**: Full transition to SAP MDG and S/4HANA, with centralized governance, harmonized processes, and integrated tools.

This phased approach enables the organization to progressively embed governance into operations, reduce ambiguity, and strengthen control over master data quality and compliance.

This phased approach ensures stable adoption of governance processes, supports user readiness, and reduces the risks of inconsistency during the transition.

5.4 Conclusions and Strategic Implications for Data Governance in the Pharmaceutical Industry

The case study presented in this chapter offers a detailed account of how a global pharmaceutical company has approached the challenge of building a scalable, cross-functional Master Data Governance (MDG) framework during its transition from SAP ECC to SAP S/4HANA.

This transformation involved the adoption of a hub-and-spoke model, the introduction of SAP MDG, and the formalization of stewardship roles. These efforts reflect a significant commitment to establishing robust data governance structures.

However, when contrasted with the theoretical model outlined in Chapter 4, several areas of **alignment and divergence** become apparent.

This section analyzes those findings and offers actionable recommendations that can support more resilient, transparent, and ethically grounded governance in pharmaceutical enterprises.

Structural Alignment with Best Practices

A clear point of convergence lies in the adoption of a **federated governance structure** based on the hub-and-spoke paradigm. As discussed in Chapter 4, this approach reflects **polycentric governance** principles, where centralized coordination, handled by the hub, coexists with decentralized, domain-specific execution through the spokes.

The case study company has effectively operationalized this model: the hub focuses on enterprise-wide enablement (e.g., standard-setting, tool support, training), while spokes are accountable for CRUD operations, data quality, and stewardship at the domain level.

The **differentiation of roles** into MDM Leads, Data Owners, and Data Stewards also mirrors the thesis's emphasis on clarifying responsibilities between technical and business actors.

Particularly, the integration of "thick" and "thin" spokes demonstrates maturity-sensitive governance: more advanced functions like Finance or GSC act as au-

tonomous spokes, while emerging or less mature areas, such as R&D, rely more heavily on centralized guidance.

Furthermore, the case study's **phased deployment strategy**—with a foundational wave followed by an optimization wave—aligns with the incremental implementation model suggested in Chapter 4, which prioritizes capability **ramp-up and change management** to mitigate resistance and promote long-term adoption.

Key Gaps and Strategic Recommendations

Despite these structural strengths, several limitations emerge when examining the case study against the governance principles discussed in this thesis. These limitations are not unique to this organization but reflect broader challenges common across the pharmaceutical industry.

1. Missing Conflict Arbitration Mechanisms

Although the company has introduced MDM Leads who also act as cross-domain coordinators, the current model **lacks an explicitly defined process for arbitration and resolution of conflicts** between domains or functions.

This absence is particularly problematic in a regulated and siloed industry like pharma, where master data such as customer, material, or supplier records often span multiple functional units (e.g., Finance, Commercial, Procurement).

Institutionalize a *Cross-Domain Data Governance Council*—a formal governance layer with decision rights and escalation procedures, would ensure alignment across domains and mediate disputes over data ownership, stewardship, and lifecycle changes.

Such structures are endorsed by enterprise governance frameworks like DAMA-DMBOK2[1].

2. Limited Visibility into Data Quality and Cross-Domain KPIs

While data ownership roles are defined, the current governance framework does **not fully operationalize performance metrics** that measure either domain-specific health or the impact of cross-domain dependencies. As highlighted in Chapter 4, data accountability must be supported by concrete KPIs, audit trails, and usage analytics.

A *dual-layer KPI framework*, should be implemented to enable effective performance monitoring, distinguishing be- tween:

- **Domain Health KPIs**: accuracy rates, SLA adherence, steward backlog.
- **Cross-Domain KPIs**: conflict resolution times, object reuse rates, lifecycle propagation quality.

This transition supports a shift from compliance-centric governance to value generation, as advocated in McKinsey's data maturity models[2] and Informatica's trust scorecard methodology[3].

3. Governance Challenges During ERP Transition Phases

One of the most vulnerable moments for data governance occurs during the transition from legacy ERP systems (SAP ECC) to the target state architecture comprising

[1]DAMA International. (2017). *DAMA-DMBOK2: Data Management Body of Knowledge.*

[2]McKinsey & Company. (2022). *Data Transformation: Driving Business Value through Data Governance.*

[3]Informatica. (2023). *Data Governance Maturity Framework: From Compliance to Value Creation.*

SAP S/4HANA and SAP MDG. This coexistence phase—referred to as the "hybrid state"—is expected to span from mid-2026 to the end of 2027. During this period, SAP MDG will be live for selected key domains and regions, while SAP ECC will remain active in others, resulting in a dual-system landscape.

Although the company recognizes the inherent complexity of this transition, it currently lacks a structured framework to govern the overlapping lifecycle of master data objects across systems. This absence introduces risks such as role duplication, conflicting ownership, unsynchronized policies, and inconsistent accountability.

To mitigate these risks:

- **A structured Governance Playbook** for Transition should be developed. This playbook should define interim ownership and control models for legacy systems, introduce temporary "shadow steward" roles to maintain continuity in ECC, and establish synchronization protocols for harmonizing policies across platforms. Communication guidelines should also be included to align expectations and ensure stakeholder engagement throughout the transformation. Aligning this playbook with Kotter's 8-Step Change Management Framework[4] can ensure it supports both technical change and behavioral adoption.

- **Early activation of central governance roles** within MDG-covered domains. These roles—already defined as part of the long-term Target Operating Model (TOM)—should not be postponed until full S/4HANA deployment but instead embedded into operations during the hybrid phase.

  These roles are foundational to long-term governance maturity and must therefore be clearly defined, resourced, and operationalized during the hybrid mode. This includes:

[4]Kotter, J. P. (1996). *Leading Change*. Harvard Business Review Press.

  - establishment of a **RACI model** to clarify who validates, enriches, and approves each MDG object. Role readiness should be supported by training, onboarding procedures, and performance measurement frameworks (e.g., KPIs tied to data quality and lifecycle adherence).

- A **governance role matrix** should be maintained throughout the transition. This matrix ensures visibility into system-specific responsibilities, identifies overlaps or gaps, and tracks the shift of accountability from ECC to MDG. An example is shown below in Table 5.2.

Table 5.2: Example Governance Role Matrix During Hybrid ERP Phase

Object	**System**	**Domain Lead**	**Data Owner**	**Data Steward**
Material	MDG	Yes	Yes	Yes
Vendor	ECC	No	TBD	Local Only
Customer	MDG	Yes	Yes	Yes
Finance WBS	ECC	No	TBD	Local Only

This governance mapping serves three critical purposes:

1. it prevents redundancy by avoiding parallel ownership structures
2. it exposes governance gaps that may delay data readiness

3. it prepares teams for the eventual consolidation of processes in S/4HANA.

By establishing governance continuity through a phased implementation—supported by central roles, structured playbooks, and dynamic matrices—the organization can minimize transition risks, accelerate change adoption, and reinforce governance maturity in parallel with system transformation.

4. Absence of Programmatic Governance in AI-Driven Workflows

Although the focus is on master data, the growing use of such data in AI workflows (e.g., R&D, supply chain, pharmacovigilance) calls for embedded governance across the model lifecycle. Currently, the organization has not operationalized AI governance mechanisms within its MLOps environments.

Governance should be embedded into AI workflows by integrating:

- Quality checkpoints at model ingestion.
- Lineage tracking and audit trails.
- Automated revalidation routines and compliance gates.

These actions align with emerging regulatory guidelines such as the OECD AI Framework[5] and support ethical, explainable AI adoption.

Broader Implications for the Pharmaceutical Industry

This comparison reveals that building a hub-and-spoke model is only the first step. True governance maturity entails the integration of ethical oversight, AI lifecycle governance, and adaptive structures tailored to domain-specific maturity.

[5]OECD. (2021). *Framework for the Classification of AI Systems.*

- Move beyond structural design by embedding meta-governance layers to reinforce transparency, equity, and cross-domain accountability.
- Codify governance adaptability through formal maturity models and capability assessments.
- Provide targeted training for MDM actors on both operational and ethical dimensions of data stewardship.

The identified gaps do not suggest failure, but rather reflect the increasing complexity of enterprise data governance. By incorporating the insights and best practices outlined in this thesis, pharmaceutical companies can develop governance systems that are operationally effective, ethically grounded, and ready to support AI-driven transformation.

Ultimately, this thesis argues that effective data governance—when embedded across systems, roles, and lifecycles—is not merely a compliance exercise, but a strategic enabler. In a context where AI adoption accelerates and regulatory demands evolve, building robust, federated, and ethically aligned governance frameworks is essential. The hub-and-spoke model, enhanced by polycentric governance and layered accountability, provides a viable path forward.

Only by treating governance as a dynamic capability can organizations navigate the complexities of modern data ecosystems and realize the full potential of trustworthy AI.