IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Pharma AI Pilots: Fixing Data Foundations for Scale

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-03-03 · pharma ai · ai scaling · data foundation · data governance · fair data · data fragmentation · life sciences ai · ai pilots

Original article: https://intuitionlabs.ai/articles/pharma-ai-pilots-data-foundation



In this article

1. Executive Summary
2. Introduction and Background
3. The Promise of AI in Pharma
4. Why Pharma AI Pilots Fail
5. 1. Data Quality, Readiness, and Fragmentation
6. 2. Infrastructure and Integration Gaps
7. 3. Organizational and Cultural Barriers
8. 4. Regulatory and Validation Challenges
9. Data Foundation: The Key to Scaling AI
10. Data Architecture and Integration
11. Data Quality and Compliance Frameworks
12. People, Process, and Governance
13. Case Studies and Evidence
14. Building a Robust Data Foundation
15. 1. Establish Unified Data Platforms
16. 2. Deploy Knowledge Graphs and Semantic Layers
17. 3. Rigorous Data Quality and Compliance Processes
18. 4. Organizational Change and Training
19. Data, AI and Business Value: Analysis and Evidence
20. Key Survey Findings on Data and AI (Table)
21. Failure Modes vs. Solutions (Table)
22. Case Studies of AI Scaling (Real-World Examples)
23. A. Pharma Launch Delayed by Data Inconsistency
24. C. Government AI Pilot (FDA Elsa)
25. D. Novartis “Supercharged R&D” Initiative
26. E. GSK Partnerships for AI-Driven Discovery
27. Implications and Future Directions
28. Transforming Pharma with a Solid Data Foundation
29. Near-Term Actions and Investments
30. Long-Term Directions
31. Conclusion

Executive Summary

AI offers opportunities for pharmaceutical research and operations, but deployment planning must account for technical limits. RAND notes that some AI projects fail because the task is too difficult for the technology to solve (RAND). Key obstacles include fragmented, poor-quality data; siloed legacy systems; inadequate data governance; and cultural resistance. For example, a Pistoia Alliance poll reported that only 36% of respondents were plugging internal documents into models (Pistoia Alliance).

To move from isolated proofs-of-concept to enterprise-scale AI, pharma must **solve the data foundation**. Industry leaders now emphasize FAIR data practices, unified knowledge graphs, and robust governance as the groundwork for trustworthy AI (www.fiercepharma.com) (events.reutersevents.com). In Pistoia Alliance's December 2025 conference poll, **27%** of respondents either did not know what scientific content their organization's AI or LLM systems used or relied only on titles and abstracts (Pistoia Alliance). The proposed remedy is a "traceable, audit-ready data pedigree" built with FAIR metadata and knowledge-graph infrastructures (www.fiercepharma.com). In practice this means creating enterprise-wide **unified data layers** that eliminate fragmentation (events.reutersevents.com), validating and harmonizing clinical and operational data, and enforcing clear ownership and quality standards across departments (events.reutersevents.com).

This report surveys the landscape of "AI at Scale" in pharma (2026), documenting why so many pilots fail and outlining how to build a strong data foundation. We review historical context and current state-of-practice, present multiple perspectives from industry (survey data, expert analyses, and case studies), and provide evidence-based recommendations. Key findings include:

- **Data Quality and Fragmentation.** Pharma data is often siloed in hundreds of systems (ELNs, LIMS, clinical databases, CRM, etc.) with divergent formats (sakaradigital.com) (nttdata-solutions.com). Fragmentation, duplicate records, missing metadata, and unvalidated sources severely limit AI utility. Surveys show roughly **50%** of pharma leaders cite "AI-ready data" as the top obstacle (www.bio-itworld.com), and data readiness should be assessed for the intended use (FDA principles).
- **Governance and Trust.** Without data governance, all AI suffers. Inconsistent HCP profiles, misaligned terminologies, and stale records erode end-user trust (pharmavanguard.com) (www.axios.com). One pharma study notes that field users may "stop believing" AI outputs the moment they encounter an obvious data error (pharmavanguard.com). Conversely, teams with a single **"source of truth"** see far more adoption. Successful organizations invest first in master data management and traceability – e.g. fully **auditable data lineages** and FAIR metadata – so that every model can be validated against real, current data (www.fiercepharma.com) (events.reutersevents.com).
- **Infrastructure and Integration.** Legacy IT architectures hinder scale. Many companies never designed PLM, CRM, or manufacturing systems for large-scale data-intensive AI. As one analysis notes, "few enterprises originally architected IT for the data intensity, performance sensitivity and dynamic scaling" needed by modern AI (www.techradar.com). Signs of infrastructure bottlenecks include data scientists spending more time wrangling fragmented datasets than training models (www.techradar.com). The fix involves cloud-native data lakes or "AI factory" platforms and alignment between IT and business sites (www.techradar.com).
- **Organizational and Cultural Gaps.** Pharma companies often "love the tech, not the outcome." There is frequently a mismatch between R&D, clinical, manufacturing, and commercial teams, each with its own data "truth" and KPIs. In practice, having decentralized data stewards and a governance board is essential. Training and cross-functional collaboration (e.g. " AI literacy" programs) are needed so that data producers (scientists) and data consumers (analysts) speak the same language (www.bio-itworld.com) (www.janeasystems.com).Companies that treat AI like a capital project – with defined ROI targets and KPIs – consistently outperform those that chase buzzwords (www.techradar.com) (sakaradigital.com).

- **Regulatory and Timing Constraints.** The pharmaceutical industry's high regulation intensifies delays. For drug-development AI, the January 2026 good-practice principles call for validation, risk mitigation, and oversight proportionate to the context of use and model risk (FDA principles). Long data-processing times also blunt AI's impact: one survey noted some teams spend *100–200 days* just cleaning data or negotiating access contracts before modeling (pharmavanguard.com). By the time insights emerge, the clinical or market window may have passed.

Recommendations center on **"fixing the data foundation first"** (www.techradar.com) (sakaradigital.com). Building centralized but interoperable data repositories, adopting standard ontologies, automating pipeline validation, and leveraging AI-friendly architectures (knowledge graphs, connected lakes/fabrics) are critical. We document frameworks like the Veeva "connected data and software" approach and the NTT Data "pilot-to-production" methodology. We also compare success stories – from big pharma partnering with AI-genomics firms to academic AI tools for drug repurposing – highlighting how the underlying data strategy enabled (or in many cases, defeated) them. The report closes with future directions, noting that as pharma embraces agentic AI and personalized therapies, the quality of data infrastructure will be the limiting factor in realizing AI's full promise.

Introduction and Background

The Promise of AI in Pharma

Artificial Intelligence (AI) has been hyped as a game-changer across industries, and nowhere is the promise greater than in pharmaceuticals. The drug development process is notoriously long (often 10–15 years) and expensive (billions per drug), with high attrition rates at each stage (apnews.com). AI and machine learning techniques offer potential to **accelerate drug discovery**, improve the design of clinical trials, optimize manufacturing, and personalize patient care. For example, deep learning models (like AlphaFold) now predict protein structures in seconds, and generative models can suggest novel chemical compounds. In clinical operations, AI can assist in patient recruitment, safety monitoring, and real-world evidence analysis. In commercialization, machine-learning analytics can refine segmentation of doctors (HCPs/HCOs) and forecast prescription demand.

Pharma companies have invested heavily: Industry analysts estimate **>$3.7 billion** annually in AI initiatives (sakaradigital.com). Virtually every major pharmaceutical or biotechnology firm has a number of pilot projects and strategic partnerships with tech startups or academia. For instance, GSK announced in early 2026 a flurry of **multi-year AI collaborations** (genomics data partnership with Helix, $50M to biotech Noetik) to realize a "multimodal" drug discovery approach linking genetic to clinical data (www.pharmaceutical-technology.com). Insitro (an AI-driven biotech) has signed deals with Lilly and BMS to use ML on chemical and biological datasets for new therapeutics (apnews.com). Moreover, FDA says Elsa's models do not train on regulated-industry submissions, while the tool allows employees to access internal documents in a high-security GovCloud environment (FDA).

Moving from a proof of concept to routine use requires evaluation of the complete system. For drug-development AI, the January 2026 good-practice principles recommend risk-based performance assessment that includes human–AI interactions, fit-for-use data, and metrics appropriate to the intended context of use (FDA principles).

This report examines the core problem: **Pharma AI at Scale – Why Pilots Fail and How to Fix the Data Foundation**. We begin with a brief history of AI in pharmaceuticals and define the current landscape of "AI pilots" vs "at scale" deployments. We then analyze the root causes of pilot failure from multiple angles (data, infrastructure, organization, and regulation), with extensive evidence from surveys, industry reports, and case examples. The middle of the report is devoted to the *data foundation*: We detail what a robust data architecture and governance framework looks like in pharma, and how it addresses those failure points. We interweave concrete case studies (e.g. a major pharma's delays due to poor HCP data (pharmavanguard.com), an AI-driven repurposing tool for rare diseases (www.axios.com), and best-practice frameworks like Veeva's Global Data Cloud (www.veeva.com)). Finally, we discuss future implications: how fixing data foundations will enable next-generation AI (agentic AI, digital twins, precision medicine) and what pharma organizations must do to seize this opportunity. The conclusions provide a synthesis of actionable steps backed by the evidence.

Why Pharma AI Pilots Fail

Although AI is emerging across many pharma domains, studies and anecdotes strongly indicate that early pilots often **fizzle before delivering business impact**. In this section, we dissect the main reasons – with an emphasis on data-related issues – that cause AI projects to stall or fail in pharmaceutical and life-science settings. The analysis draws on industry surveys, technical articles, and subject-matter expertise. We find that the root causes fall into four overlapping categories: **(1) Data Quality and Fragmentation, (2) Infrastructure and Integration, (3) Organizational and Cultural Barriers, and (4) Regulatory/Validation Challenges**.

Each category is discussed below, with specific examples and data from pharma-relevant sources.

1. Data Quality, Readiness, and Fragmentation

Fragmented Data Silos: Pharma companies accumulate data in hundreds of specialized systems – clinical trial databases, laboratory notebooks (LIMS/ELN), manufacturing ERP/MES, commercial CRM, medical affairs systems, etc. – often acquired or built independently. These systems rarely share common schemas or integration. As one analysis bluntly states: "Pharmaceutical data is distributed across hundreds of systems… not designed for the kind of cross-functional data integration AI requires" (sakaradigital.com). Without unification, AI models cannot easily draw on all relevant information. For example, an AI targeting oncology trials would need to link genomics, patient history, lab results, and trial outcomes, but these may reside in separate databases with different IDs. A recent industry agenda noted that bridging **structured and unstructured sources across commercial, medical, and access silos** is essential to "reveal complete patient profiles" (events.reutersevents.com).

Surveys confirm that pharma data is highly **unstructured and fragmented**. In a Pistoia Alliance poll (2025), roughly *50%* of respondents cited “AI-ready data” as a top obstacle (www.bio-itworld.com), and 96% of respondents in Veeva’s commercial-biopharma survey believed their data was not well-structured or AI-ready for scaling (Veeva). Fragmentation has tangible costs: one Veeva study highlighted that inconsistent HCP/HCO data caused a **two-month** product launch delay and roughly **15% lower prescription volume** in the launch phase (pharmavanguard.com) (www.veeva.com). In practice, data silos cause teams to duplicate effort (each department keeps its own “gold copy”), leading to delays in access and analysis. Data residency issues (on-premises vs cloud) and ownership conflicts (e.g. a CRO holding trial data, a hospital owning patient records) further complicate sharing.

Poor Data Quality: Even when data is collected, quality is often an issue. Errors, missing values, duplicates, and inconsistent coding plague life-sciences datasets. As one technical guide advises, rigorous “data cleaning” is usually needed: fixing outliers, imputing missing values, aligning timestamp formats, and reconciling measurements with the true standard (nttdata-solutions.com). In pharma this can mean reconciling different assay units or ensuring lab results match the official source of truth. If the raw data is wrong or incomplete, any AI model built on it will produce unreliable outputs. As Axios reported in 2024, developers of an AI drug-repurposing tool cautioned that *“the tool’s success is only as good as the medical knowledge it uses”* (www.axios.com). In other words, domain knowledge embedded in the data must be comprehensive and accurate.

The lack of metadata and lineage is another dimension of poor quality. In Pistoia Alliance’s December 2025 conference poll, 27% of respondents either did not know what scientific content their organization’s AI or LLM systems used or relied only on titles and abstracts (Pistoia Alliance). This combined category should not be interpreted as a provenance-failure rate. Teams should document the source material used by each model and how it was prepared.

Data Timeliness and Currency: Pharma R&D and commercialization timelines are fast-paced. Data that is months or years out-of-date can make AI useless. Pharma functions operate under tight schedules, so that insights arriving “after the window has closed” are worthless (pharmavanguard.com). Data readiness times can be very long, underscoring inefficiency: one report noted that teams may spend **100 days** just cleaning HCP data and **200 days** navigating data access contracts before analysis can even begin (pharmavanguard.com). By the time an AI report is delivered, the market or clinical decision points may have already passed, nullifying any potential benefit. Thus, *not only quality but speed* of data preparation is a critical factor.

Data Governance and Trust: Closely related to quality is governance. Ambiguity about data ownership, inconsistent definitions, and lack of standards all undermine trust in AI. For example, multiple internal systems may use different codes or definitions for the same medical condition. Unless these are harmonized, AI models will “learn” from inconsistent signals. A conference session on Pharma Data & Tech Europe 2026 emphasized standardizing segment/condition definitions across Marketing, Medical, and Access to avoid “mixed messages” (events.reutersevents.com). In the absence of governance, duplicate or conflicting records abound. A commercial AI pilot may output an HCP recommendation that is obviously wrong because the underlying specialties or territories data was stale. One observer notes that **“even a technically excellent prediction is**

useless if users don't trust it", and that trust in pharma quickly fractures when "the data basis is inconsistent: duplicate HCPs, misaligned specialties, stale affiliation updates" (pharmavanguard.com). In other words, AI amplifies any underlying data errors, so weak governance dooms adoption.

In summary, data issues in pharma include fragmented sources, poorly formatted inputs, unclear governance, and stale or delayed datasets. Data readiness deserves attention alongside technical feasibility: RAND's interviewees also described failures caused by fundamental limitations in what AI can achieve (RAND).

2. Infrastructure and Integration Gaps

Technical infrastructure – both hardware and software – is the next key bottleneck. Many pharma companies' IT architectures were built for traditional applications, not for high-throughput AI workloads. One TechRadar article observes that most enterprises' legacy IT "was never designed for the data intensity, performance sensitivity and dynamic scaling" of modern AI (www.techradar.com). In such legacy environments, data may be physically siloed on-premises, tape libraries, or outdated databases that are hard to query. As a result, bringing AI into production often hits computing and integration walls.

A common pattern is the **"architecture first"** lesson: companies cannot hope to chew off enterprise data for AI without rethinking their infrastructure. Leading analysts advise that pilot success in AI is meaningless if the underlying systems remain disjoint. As one expert put it, "many enterprises still run on siloed content repositories, legacy systems, and fragmented integrations", meaning an AI cannot see half the data it needs (www.techradar.com). Data scientists may have built a prototype on a small extract, but at scale the agentic AI cannot retrieve records across legacy lab systems and the ERP system simultaneously. In effect, infrastructure becomes a "structural barrier to progress" (www.techradar.com).

Concrete indicators of infrastructure weakness include:

- **Slow Data Access:** When data is spread across old servers, retrieval times can be long. If engineers "spend more time dealing with slow retrieval" and broken pipelines than training models, the pipeline is broken (www.techradar.com).
- **Lack of Scalability:** Many systems lack cloud or elastic scaling. They cannot handle concurrent access by AI agents, large batch analytics, or real-time serving.
- **Fragmented Data Pipelines:** Even if companies have data lakes, they are sometimes batch-based and not integrated. TechRadar notes that if data scientists fix pipelines manually for each pilot, "the organization has hit a major infrastructure bottleneck" (www.techradar.com).

In practice, solving these gaps often entails re-architecting for the AI era. This might mean migrating lab and trial data into a **cloud-native data platform**, or building an *"AI factory"* that standardizes data flow end-to-end (www.janeasystems.com) (www.techradar.com). The AI & Big Data Expo (London 2025) emphasized *decentralized AI platforms* that empower business units rather than isolating them (www.janeasystems.com). Aligning IT (digital) with Operational Technology (lab/manufacturing systems) is also crucial. For manufacturing AI, use validation and oversight proportionate to the intended use and model risk (FDA principles). Pharma faces the same need to bridge automation (e.g. bioreactor controls) and analytics.

Until this integration is solved, AI prototypes remain tethered to the lab. Successful scaling demands unified, high-throughput pipelines not only for batch model training but also for embedding AI into real-time workflows—whether that is a clinical supply chain or an automated quality control line. Without this underlying infrastructure, pilots effectively fall into **pilot purgatory**, no matter how promising the algorithms are. As one expert observed, “AI agents need to be woven into the enterprise fabric, connected to the right data and workflows” or “autonomy quickly becomes chaos” (www.techradar.com).

3. Organizational and Cultural Barriers

Even with good data and a robust platform, many AI pilots stall for organizational reasons. In pharma, cross-functional coordination is notoriously difficult. Different departments often have competing priorities (e.g. R&D cares about candidate molecules, sales cares about HCP behavior) and insufficient shared accountability for “digital” initiatives. This leads to several failure modes:

- **Lack of Business Alignment:** Projects often begin as R&D or data science initiatives without clear business sponsorship. Statistics from TechRadar highlight that many pilots are “anchored in [no] clear business plan with measurable ROI” (www.techradar.com). Without aligning AI goals to metrics like increased throughput or lower development cost, pilots look like fascinating toys that never pay off. Industry advice is to treat AI projects like capital investments: define ROI targets up front and track them (www.techradar.com).
- **Siloed Teams:** Pharma typically separates functions (Clinical, Manufacturing, Medical/Commercial, Quality). These silos often operate in isolation, each maintaining their own IT systems and data practices. It is common for IT and OT (or Lab and IT teams) to “operate in isolation” (www.techradar.com), so nobody owns end-to-end integration. Clear governance structures and RACI matrices are often absent. For example, in data governance the question “who is the master of truth for HCP addresses?” is sometimes unanswered. This fragmentation means no team is incentivized to fix the data across departments, so inconsistent data circulates.
- **Data Culture and Expertise:** Pharma scientists and clinicians have highly specialized knowledge but may lack AI/data literacy. There can be skepticism about AI from domain experts who fear “garbage in, garbage out.” As one life-sciences data leader put it, data analysts are often frustrated: “We share only summary reports with leadership; now AI needs the raw data.” Many researchers are accustomed to publishing curated analyses, not sharing raw tables for machine learning. Surveys show a deep cultural challenge: scientists often “resist relinquishing control of their data, preferring to present conclusions in reports rather than providing raw data for broader reanalysis” (www.bio-itworld.com). Shifting culture to open data (even internally) requires strong leadership.
- **Trust and Change Management:** Once an AI model is deployed, success depends on user trust and adoption. Pharma field teams (sales reps, medical liaisons) are notoriously skeptical of automated suggestions. Early failures can cause “the first casualty” of trust; as soon as a user spots an error – even if due to bad data – they may ignore the system entirely (pharmavanguard.com). The flipped

dynamic (users ignoring AI) kills momentum. A positive example is Novartis' strategy of embedding AI specialists with biologists and chemists to build "super-intelligent" teams (www.novartis.com), but most firms lack this cross-pollination.

- **Organizational Readiness:** Many companies lack dedicated data science or AI Ops teams, or fail to upskill staff. Notably, Novartis instituted enterprise-wide training (Microsoft Copilot and its own "Data Science Academy") to build digital fluency (www.imd.org). Those that do not are left with isolated data scientists who cannot operationalize their models across the business.

In short, while these causes are not strictly "data problems", they interact strongly with data issues. Disjointed leadership, poor governance, and cultural resistance amplify the impact of fragmented data and weak infrastructure. Conversely, companies that succeed in scaling AI typically fix both simultaneously: they integrate governance and business processes as they clean and centralize data. We will see in Case Studies how firms that invested in enterprise data platforms also created cross-functional data councils and drives for adoption.

4. Regulatory and Validation Challenges

Pharmaceutical development is one of the most heavily regulated activities in industry. Any changes to processes (even software) often require validation and documentation under Good Practice (GxP) regulations. AI for pharma thus faces unique hurdles:

- **Rigorous Validation:** FDA's January 2025 draft guidance addresses AI used to generate information or data supporting regulatory decisions about drug safety, effectiveness, or quality. It contains nonbinding recommendations, rather than a universal validation rule or deployment-cost multiplier (FDA draft guidance).
- **Data Privacy and Usage Restrictions:** Patient data and proprietary trial data are sensitive. Privacy rules like HIPAA, GDPR, and contract limitations restrict how data can be used or shared. While federated learning and encryption can help, they add complexity to data pipelines. Regulatory bodies also worry about how AI uses patient data. For example, FDA insisted that its internal AI tool "Elsa" **is not being trained on proprietary manufacturer data** (www.axios.com) to avoid compliance breaches. In pharma companies, similar concerns must be addressed – building AI without violating data agreements.
- **Speed vs Caution Paradox:** Paradoxically, the very quest for speed may trigger caution. Regulators and executives are rightfully risk-averse: a mistaken AI decision could lead to patient harm or a regulatory recall. The consequence is often that **deployments are over-cautious or stalled**. Safety-critical use cases (e.g. GC/MS analysis, trial patient matching) encounter multi-level sign-offs and pilot inertia, adding to the delay. Broad surveys of healthcare executives have noted that the "AI pilot era" has shifted mere experimentation toward a longer-term prudence in 2026 (hitconsultant.net) (Note: the pivot from "pilot" hype implies institutions are scrambling to manage risk, not just hype).

In summary, pharma's regulatory environment demands that any AI solution is demonstrably reliable and compliant. This means robust audit trails for the data, strict version control of models, and often human-in-the-loop processes. It raises the bar for what "success" looks like. The good-practice principles call for traceable documentation of data provenance, processing steps, and analytical decisions, alongside governance that protects sensitive data (FDA principles).

Data Foundation: The Key to Scaling AI

Given the failure modes above, the prescription is increasingly clear: **AI won't work at scale without fixing the data foundation first** (www.techradar.com) (sakaradigital.com). In practice this means treating data as a strategic asset, not a byproduct. In this section we delve into what a robust data foundation entails for pharma: the architecture, governance, and practices needed to support AI reliably. We draw on best practices from the industry, conference frameworks, and technology approaches (like knowledge graphs and data pipelines) that have been promoted specifically for pharmaceutical applications.

Data Architecture and Integration

At the core, pharma companies should build **unified data layers** that eliminate the fragmentation discussed above (events.reutersevents.com). This does not mean arbitrarily dumping all data into one database – that can violate privacy and scale poorly – but rather creating strategic integration points. Common approaches include:

- **Enterprise Data Lakes and Warehouses:** Consolidating diverse datasets (raw and processed) in a central repository (often in the cloud). Data from clinical systems, laboratory results, EMR, marketing CRMs, etc., are ingested (after cleaning) into the lake. In one session at Pharma Data & Tech Europe 2026, experts explained that "creating unified data layers that eliminate fragmentation across systems" is the first step to enterprise AI value (events.reutersevents.com). In practice, vendors like Veeva and AWS offer Life Sciences data lakes that come with pre-built connectors to common sources.
- **Master Data Management (MDM) and Standardization:** Persistent, "golden records" for key entities must be established. For example, every clinical site, study protocol, drug compound, HCP, or patient should have a unique identifier and agreed-upon attributes. MDM tools or custom reference tables are needed to de-duplicate and harmonize these across systems. During a Veeva data report, life-sciences leaders emphasized that *trust, speed, and consistency* are the core data issues (www.veeva.com), and building a single source-of-truth repository is a proven solution. Boehringer Ingelheim, Bayer, and Astellas are cited as companies actively "standardizing on a global data model" to achieve this consistency (www.veeva.com).
- **Data Lakes vs Data Fabric vs Data Mesh:** Beyond a single repository, modern strategies include **data fabrics or meshes**, which keep data at its source but make it interoperable via metadata and APIs. Data mesh, for example, empowers each domain team (R&D, Manufacturing, Commercial) to manage its data as a "product" but publishes standardized schemas to a catalog. The key is **interoperability**. Every architecture must allow an AI pipeline to query and join across domains. As the Pharma

conference notes, integration platforms should “bridge structured and unstructured data sources” across silos (events.reutersevents.com). This might involve a messaging bus, or a unified ontology that connects table columns.

- **Knowledge Graphs:** One powerful emergent pattern is the use of knowledge graphs, which the industry is keen on. A webinar described knowledge graphs as the “operational infrastructure” to make fragmented pharma data AI-ready (www.fiercepharma.com). In a knowledge graph, entities (drugs, genes, proteins, patients, clinical conditions, operational units, etc.) become nodes, with edges for relationships. This allows the system to answer complex queries (e.g. “which trials have similar patient cohorts and target paths?) and to tie AI answers to evidence. The same FiercePharma event promoted a “GraphRAG” approach: linking large-language-generation outputs to this graph so that every answer can be traced to structured evidence (www.fiercepharma.com). In effect, a knowledge graph imposes an enterprise-wide schema dynamically and can greatly reduce data misinterpretation.
- **Clinical and Real-World Data Repositories:** For trials and patient data, pharma should curate specialized registries and RWD (real-world data) lakes. Integrating EMR data is especially tricky. Sessions at data conferences advise converting unstructured EMR notes into structured formats (NLP + curation) and then using **standardized protocols** for queries (events.reutersevents.com). Access pipelines (with anonymization/minimization) are set up so that BI tools and AI can tap into them under governance. This requires IRB/consent alignment and robust audit logs on usage (events.reutersevents.com).

Overall, the architecture goal is to enable **“enterprise-wide analytics”**. Using the AI & Big Data Expo framework, data systems should empower cross-department teams, not isolate them (www.janeasystems.com). Every AI model in production must draw from this shared foundation rather than isolated sandboxes. Critically, building these foundations is *hard work*–one deep-dive roadmap lists it as the first three of ten steps for scaling pharma AI (nttdata-solutions.com).

Data Quality and Compliance Frameworks

A stable platform must be coupled with strong governance and quality controls. These include:

- **Metadata and Lineage (FAIR Data):** Following the FAIR principles (Findable, Accessible, Interoperable, Reusable) is often cited as best practice (www.fiercepharma.com). Each dataset or column in the lake should have metadata: source application, last update timestamp, data steward, quality score, and transformation history. Governance tools can automate lineage tracking: any time an AI pipeline consumes data, that usage is logged. These provenance records are crucial for regulatory audits. In the webinar on trustworthiness, speakers stressed creating an “audit-ready data pedigree” via FAIR metadata (www.fiercepharma.com).
- **Data Quality Rules and Monitoring:** Implement automated checks for common data issues (e.g. unexpected nulls, duplicate IDs, outliers beyond scientific plausibility). For example, a pharma-grade pipeline might include validation rules that blood analyte values must lie within human ranges, or that time stamps increase monotonically. Such rules can be encoded in data processing (e.g. using tools like Great Expectations) so that any breach stops the pipeline and alerts data stewards. Continuous

monitoring dashboards can track data drift and completeness over time. According to NTT Data's scaling guide, the "Optimize and validate the dataset" step is essential: scripts or tools should correct obvious errors and align data modalities before modeling (nttdata-solutions.com).

- **Data Ownership and Stewardship:** Define clear data owners for each domain. Governance frameworks should assign responsibility (e.g. clinical leads for patient data, IT for systems integration, medical affairs for HCP data) and give them metrics to maintain. At Pharma 2026 conferences, speakers repeatedly emphasize needing "aligned objectives and coordinated investment by applying proven models of integrated leadership" to break down silos (events.reutersevents.com). In practice, many companies form *Data Councils* with delegates from each function, who approve data standards and oversee quality KPIs.
- **Standards and Ontologies:** Use industry data standards wherever possible. In therapeutic areas, CDISC standards for clinical trial data (SDTM, ADaM) can ensure common structure. Ontologies like SNOMED or UMLS can unify medical terms. Internally, pharma companies may develop corporate ontologies for company-specific concepts (e.g. compound numbering schemes). The agenda for Pharma Data Europe 2026 lists "harmonized definitions and models" for consistent insights (events.reutersevents.com), underlining that semantic consistency is key to merging data from multiple sources.
- **Data Privacy and Security Controls:** As data is centralized, strict controls are implemented. Role-based access, encryption-at-rest/in-transit, and data masking for sensitive fields are part of the foundation. For example, one conference slide suggests "controlling access by role and purpose… and auditing every query" for patient data (events.reutersevents.com). This ensures compliance with regulations while still enabling AI use. It also builds trust – data owners will share more if they know secure templates and logs govern access.
- **Data Catalogs and Self-Service:** To make data **accessible**, a modern foundation often includes data discovery tools. Analysts should be able to search a data catalog (with metadata and usage docs) to find data sets. This democratizes AI so that departmental users can discover and request data themselves, rather than relying on IT tickets. Again, such catalogs usually leverage FAIR principles to ensure Findability.

In sum, a Pharma data foundation is not just a technical schema, but a **governance ecosystem** that enforces consistent and trustworthy data across the enterprise. Many of the shortfalls in pilot projects can be directly traced to gaps here: e.g. 67% of pharma leaders admitted they abandoned AI projects "due to foundational data issues" (pharmavanguard.com). By contrast, organizations that build rigorous QA processes and data pedigrees create the baseline needed for AI success.

People, Process, and Governance

A foundation includes people and processes as much as technology. Leading approaches stress that AI readiness rests on three pillars: **People, Systems, and Governance** (www.janeasystems.com). In practice, this means:

- **Cross-Functional Teams:** Embed AI projects in teams that include data scientists **and** domain experts (clinicians, chemists, operations staff). For example, Novartis reports that their AI innovation teams physically co-locate AI engineers with biologists, chemists, and MSLs to “challenge each other” and build mutual understanding (www.novartis.com). Such integration helps in correctly labeling data and interpreting anomalies.
- **Skill Building and Literacy:** World-class pharma AI organizations invest in training. Novartis created a “Data Science Academy” and rolled out tools like Microsoft Copilot to raise baseline data skills among its workforce (www.imd.org). Upskilling is essential so that many staff, not just IT, are comfortable with AI concepts and can contribute to data initiatives.
- **Governance Structures:** Formalize policies and ethics. This includes committees or roles (Chief Data Officer, AI Ethics Officer, etc.) and written standards for data usage. For instance, one big pharma described strict operating models where all AI projects require end-to-end oversight, version control, and “defined KPIs aligned with operational goals” before they commence (www.techradar.com). Data governance boards typically establish the company’s master data definitions and approve new sources for inclusion in the data platform.
- **MLOps and Continuous Deployment:** Modern AI at scale requires MLOps practices (CI/CD for models). This is often neglected in pilots. A robust data foundation supports this by providing automated pipelines that feed data into models and automatically evaluate model outputs on new data. Teams adopt tools that track model performance over time (detecting drift) and manage retraining triggers. Without stable data inputs and automated pipelines, it is impossible to regularly update models on fresh data.
- **Accountability and Metrics:** Ultimately, behavior changes when incentives and metrics are aligned. Successful programs tie data and AI outcomes to executive KPIs (faster trial times, improved yield, patient satisfaction, revenue lift) (www.techradar.com). When leaders explicitly require teams to report standardized AI success metrics, it forces attention on scaling and not just demonstrating a cool technology.

The bottom line of these people/process changes is that technology deployment is not enough; the foundation is social as well as technical. Companies that regarded AI as a business transformation (with data as the foundation) – rather than an ad-hoc IT project – have demonstrably broken the cycle of failure.

Case Studies and Evidence

To illustrate these points, we examine several real-world examples and analyses from late 2025 through early 2026. These cover both failures and successes in pharma AI, demonstrating how data foundations made (or could have made) the difference.

Veeva Commercial Data Report (2025)

A key source of industry statistics is the Veeva "State of Data, Analytics and AI in Commercial Biopharma" report (2025). Surveying 116 senior commercial analytics leaders, it found a **stark AI paradox**: strong ambition but weak outcomes (www.veeva.com) (pharmavanguard.com). Among its findings:

- **Ambitious Use Cases:** Companies are pursuing advanced AI/ML in commercial (72% using AI to analyze HCP engagements, 62% for customer segmentation, 60% for 360° views of HCPs) (www.veeva.com). This shows high commitment: these are not trivial tasks but core of sales/marketing strategy.
- **Pilot Breakdown:** Despite the ambition, "the vast majority [of AI's] value potential at scale remains uncaptured" (www.veeva.com). The survey highlights that most pilots do not expand. Notably, **67% of organizations have already abandoned AI projects due to foundational data issues** (pharmavanguard.com). Many projects loop back to "isolated pilot projects that fail for the same unaddressed data reasons" (www.veeva.com).
- **Specific Stats:** As noted earlier, 89% of respondents said they "couldn't scale more than half" of their AI initiatives; 96% believed their data was not well-structured or AI-ready for scaling (Veeva); 67% have given up on projects for data-related reasons (pharmavanguard.com). In context, one highlighted case had poor HCP/HCO data lead to *two-month market delays and 15% fewer early prescriptions* (pharmavanguard.com) (www.veeva.com). These are concrete business impacts.

Veeva's report goes on to identify the core problems as *"trust, speed, and consistency"* of data (www.veeva.com). It concludes that for the 89% of companies that failed to scale, prioritizing data foundation is key (www.veeva.com). Specifically, they advocate *"connecting pre-harmonized data and software"* so that analytics can operate on clean, consolidated information (www.veeva.com). This aligns precisely with our thesis: only by unifying data (e.g. via Veeva's own **Data Cloud**, which pre-integrates HCP/HCO profiles with CRM) can pharma harness AI.

Tech Industry Analyses

While not pharma-specific, several broad-tech articles reinforce the pattern. For instance, a TechRadar feature on agentic AI (Jan 2026) interviewed an enterprise tech leader who lamented: "Many early adopters of AI tools are struggling. Pilot projects stumble, costs escalate, and results fail to match expectations" due to "moving too fast without the strategy, infrastructure, and data foundations required" (www.techradar.com). This expert cites industry reports that *"80% to 90% of all enterprise data is unstructured"* (www.techradar.com), underscoring the data challenge even outside pharma.

Another TechRadar analysis (Nov 2025) of manufacturing AI found **nearly 90%** of AI pilots stall before scaling (www.techradar.com). Crucially, it observed that "most pilots fail not because the algorithms don't work, but because the data beneath them is fragmented, poor quality or locked away in silos" (www.techradar.com). This directly parallels pharma: the cause is not unique to factories. This manufacturing piece also noted the issue of alignment between IT and operational teams (www.techradar.com) – exactly the problem pharma faces among R&D, clinical operations, and IT departments. Thus, even absent pharma specifics, the stats and diagnoses from adjacent industries affirm that data fragmentation and silos are endemic to scaling AI.

The TechRadar April 2026 report on agentic AI goes further: in enterprises, “the most common reason [AI] projects stall is a **lack of data maturity**” (www.techradar.com). Specifically it cites many businesses still operating with “fragmented data, duplicated sources and unclear ownership” (www.techradar.com). These conditions “make it difficult to verify whether the information is current, accurate or even still relevant” when AI uses it (www.techradar.com). The report bluntly concludes that strengthening the data foundations is often the **first deployment step** for any AI that operates autonomously in workflows (www.techradar.com). In other words, the advice across the board is not to double-down on bigger models, but to first fix the underlying data infrastructure.

Pharma Pilot Stuck in Purgatory (Sakara Digital, 2026)

A detailed blog by Sakara Digital focuses on pharma AI deployment. For primary evidence, Veeva’s 2025 commercial-biopharma report describes difficulty moving initiatives from pilots to enterprise-wide scale (Veeva). Sakara identifies three unique challenges in pharma: the extreme **regulatory burden** (validation requirements can double project effort) (sakaradigital.com), **data fragmentation** across hundreds of independent systems (sakaradigital.com), and **risk aversion** given patient-safety stakes (sakaradigital.com). These amplify any data or infrastructure gaps. The article emphasizes that pilots are “forgiving” (data can be hand-curated, edge cases excluded), but production demands *no* hand-holding – the entire pipeline must be robust (sakaradigital.com). Sakara’s conclusion is a blueprint: to fix pharma AI, organizations must attack these root causes systematically. For deployment planning, assess predictive performance in the intended context of use rather than assigning an unsupported universal causal percentage to data and process problems (FDA principles).

AI Repurposing Tool (Harvard/Nature Medicine, 2024)

On the research side, a recent Nature Medicine publication (covered in Axios Sept 2024) developed an AI tool (TxGNN) that screens ~17,000 diseases against existing drugs for repurposing (www.axios.com). The researchers explicitly cautioned users: the AI’s success “is only as good as the *medical knowledge it uses* to derive conclusions” (www.axios.com). This limitation – heavy reliance on the underlying knowledge graph – highlights a key data foundation point: if the knowledge base is incomplete or biased, AI output will be flawed. Thus, even for cutting-edge generative models, the human-curated medical and pharmacological data layer is the foundation. This example underscores that in pharma R&D, “beautiful credo [algorithms] need solid data behind them” – a theme echoed in many expert comments (e.g. the Pharma Vanguard article noted that “AI is not magic – it magnifies the cracks in your data” (pharmavanguard.com)).

FDA AI Tool (Elsa)

The FDA’s internal rollout of “Elsa” (June 2025) provides another cautionary note. Elsa is a generative-AI assistant for reviewers, but the FDA had to ensure it doesn’t compromise proprietary data. They explicitly stated Elsa “isn’t being trained with proprietary data submitted by drug and device manufacturers” (www.axios.com). The tool was built in AWS GovCloud to handle sensitive data. For pharma companies, this is lesson in itself: any domestic AI must adhere to strict data boundaries. It also shows how an organization with

nearly unlimited resources still focused on **trust and compliance** in their data foundation. As one expert put it, the integration of AI within an organization must also answer *"which models, which inputs"* to remain secure (www.axios.com).

Scaling Frameworks: Pharma 2026 Conferences

Finally, industry gatherings highlight collective wisdom on data foundation. At Reuters Events' Pharma 2026 conferences, leaders shared scaling frameworks built "from the ground up" (events.reutersevents.com). These stressed enterprise-wide visibility (shared KPIs and governance), harmonized definitions (consistent segment/condition codes) (events.reutersevents.com), and **certified shared datasets** to "accelerate launches" (events.reutersevents.com). One session explicitly entitled *"Set the foundation: Solve pharma's data quality crisis"* presented bullet points on unifying data layers, enforcing governance standards, and integrating structured/unstructured sources (events.reutersevents.com). In multiple presentations, the refrain was the same: **without trusted, high-quality data, AI is a dead end**. The fact that these topics dominate official Pharma 2026 agendas is itself evidence of the consensus that data must be fixed.

Implementation Roadmap (NTT Data, 2020)

In a 2020 white paper by NTT Data (still widely cited), the authors outline the steps to move from proof-of-concept to enterprise AI in pharma. It begins with **exploratory data analysis**: mapping all relevant sources (LIMS, MES, batch records, etc.) and assessing quality (nttdata-solutions.com). The next step is meticulous **data cleaning and validation**: scripts to fix errors, impute missingness, and ensure that data aligns with "source of truth" (nttdata-solutions.com). These foundational steps – often neglected in hurried pilots – are presented as mandatory. Later steps in their roadmap cover model building and MLOps, but only after the foundation is laid. This industry-backed guide reinforces our theme: a clear data strategy is the prerequisite for any scaled implemention of AI in pharma.

These case studies illustrate data and process challenges. They do not establish a universal cause of pilot failure: RAND's interviewees also identified failures caused by fundamental limitations in what AI can achieve (RAND). For example, companies that created a **"trusted data foundation"** (complete records, governance policies, integrated platforms) report that field teams begin to adopt AI (trusting recommendations, scaling pilots) (www.veeva.com). By contrast, the majority who skip these steps get stuck. We saw multiple sources—from pharma surveys to vendor roadmaps—emphasize the same remedies: knowledge graphs, FAIR pipelines, data cataloging, and domain-aligned governance. The next section synthesizes these lessons into concrete strategies for building that foundation.

Building a Robust Data Foundation

Use the architecture, quality controls, and governance described in "Data Foundation: The Key to Scaling AI" as the baseline. The steps below focus on implementation priorities.

1. Establish Unified Data Platforms

- **Enterprise Data Lake/Cloud:** Select a platform using the options in “Data Architecture and Integration.” For the Veeva example, see its commercial data report (www.veeva.com). Assign owners for ingestion pipelines and update schedules, using the unified-data-layer goal in the Pharma Data & Tech Europe agenda as a planning reference (events.reutersevents.com).
- **Data Warehouse / Data Mart for Key Domains:** Identify which structured reporting datasets need a governed warehouse or data mart. Assign owners and document which version is authoritative for each reporting use.
- **Federated Access (Optional):** For the federated approach outlined in “Data Architecture and Integration,” decide which datasets should remain at source and assign responsibility for query access.

2. Deploy Knowledge Graphs and Semantic Layers

Use the knowledge-graph approach described in “Data Architecture and Integration” to scope an initial implementation:

- **Unification of Ontologies:** Agree on identifiers and vocabulary mappings for the entities needed by the first use case.
- **Query Across Domains:** Define one cross-domain query and the relationships needed to answer it before expanding the graph.
- **Explainability and Auditing:** Graph-based AI (often combined with LLMs) can provide traces for each answer. The FiercePharma “GraphRAG” approach ties an LLM’s output to specific nodes/edges in the graph (www.fiercepharma.com), giving auditability. In high-stakes pharma, this explainability is critical.
- **Curation and Growth:** Assign reviewers for proposed relationships and set a schedule for expanding the graph after the initial use case.

For the “operational infrastructure” framing discussed earlier, see the knowledge-graph webinar (www.fiercepharma.com). Use that discussion to inform the scope and staffing of the initial graph project.

3. Rigorous Data Quality and Compliance Processes

- **Automated Validation Pipelines:** Turn the checks listed in “Data Quality and Compliance Frameworks” into pipeline acceptance criteria. Follow the NTT roadmap’s recommendation to clean data before model development (nttdata-solutions.com), and assign reviewers for logged anomalies.
- **Monitoring and Alerts:** Select the quality metrics described above, set alert thresholds, and assign a responder for each dataset.
- **Governance Frameworks:** Apply the ownership, lineage, and access policies in “Data Quality and Compliance Frameworks.” For versioned definitions and certified dashboards (events.reutersevents.com), assign approval responsibilities and document the release process.

- **Continuous Improvement:** View data foundation work as ongoing. Every AI project should produce remediation backlogs: each time a data issue blocks a model, it should be fixed in the core platform. Track progress with project-specific measures of source coverage and traceability. Veeva recommends a trusted single source of truth to give analytics teams' models credibility (Veeva).

4. Organizational Change and Training

Technical fixes must be paired with people changes:

- **Data Stewards and Champions:** Put the domain ownership model described above into practice: name each steward and document how pipeline failures and source-system changes are escalated.
- **Education Programs:** Turn the literacy goals in "People, Process, and Governance" into role-specific training on the chosen platform, data standards, and limits of AI outputs.
- **Agile Methodology for Data:** Treat data improvements like a product. Use agile sprints and Kanban boards for data tasks (e.g. "clean X dataset", "load new batch records"). This ensures progress is visible and prioritized with other IT backlog. In some models, analytics teams and IT form squads focused on high-value use cases (e.g., one squad for "supply chain AI", one for "trial design AI"). Each squad helps define the data needs end-to-end.
- **Cross-Functional Governance:** Give the council described in "Data Quality and Compliance Frameworks" a review schedule and authority to prioritize data remediation and allocate resources.
- **Vendor Partnerships:** In many cases, partnerships with specialized vendors can jumpstart the foundation. For example, pharma companies often license curated healthcare provider/hospital databases (as Veeva's OpenData) to get a reliable starting point. They also use SaaS platforms (like Snowflake, Dataiku, or domain-specific AI platforms) to leverage best practices. Finnish pharma, for example, partnered with AWS to build a ML-ready data lake, benefitting from AWS's data management tools without reinventing the wheel.

Track these responsibilities alongside the technical rollout so that stewardship, training, and platform work progress together.

Data, AI and Business Value: Analysis and Evidence

We now return to some of the metrics and evidence that connect a strong data foundation to business outcomes. This includes quantitative survey results, performance indicators, and the contents of pilot vs. production studies. Where possible, we summarize findings in tables to highlight comparisons.

Key Survey Findings on Data and AI (Table)

The following table collates key statistics from industry reports and surveys in pharma and related sectors, illustrating the scale of the problem and its alignment across sources:

Source / Survey	Year	Finding
Veeva — commercial-biopharma survey (Veeva)	2025	Among respondents, 89% were unable to scale more than half of their initiatives; 96% believed their data was not well-structured or AI-ready for scaling; 67% reported having to abandon an AI initiative due to bad data. For the separate launch-delay case, see "Veeva Commercial Data Report (2025)."
Pistoia Alliance — US conference poll (Pistoia Alliance)	2025	27% of respondents either did not know what scientific content their organization's AI or LLM systems used or relied only on titles and abstracts; 36% were plugging internal documents into models.
TechRadar (Manufacturing) (www.techradar.com)	2025	~90% of AI pilots across industries stall before scaling. Majority of failures due to *fragmented, poor-quality data in silos*; IT/OT teams' isolation further hamper transition from lab to factory.
TechRadar (Agentic/Enterprise AI) (www.techradar.com)	2025–26	60–90% of enterprise AI projects at risk of failure by 2026. Industry experts note *"messy data"* and governance as root causes. Gartner predicts 60% of orgs will miss AI value by 2027 due to incohesive governance.
RAND — AI project failure research (RAND)	2024	Interviewees described project failures caused by fundamental limitations in what AI can achieve.
Qlik — cross-industry agentic AI study (Qlik)	2025	18% reported fully deploying agentic AI; 46% said scale was three to five years away. Deployment status does not establish a project-failure rate.

These findings converge on a narrative: *nearly all surveyed executives acknowledge serious data issues*, and most AI initiatives do not scale without addressing them. The table highlights that different authoritative sources – consulting firms, industry analytics, and trade media – all identify **data fragmentation, poor quality, and lack of governance** as the root causes behind project failures.

Failure Modes vs. Solutions (Table)

Below is a summary table mapping common failure causes in pharma AI pilots to the corresponding fixes in the data foundation. Each row provides concrete examples or strategies:

Failure Mode	Consequence	Data Foundation Fix
Fragmented Data Silos (multiple incompatible systems) (sakaradigital.com)	AI models lack complete context, produce partial/misleading outputs. E.g. Clinical and sales data never joined, so AI can't correlate trial success with market uptake.	Build **unified data layers** (data lake or fabric) that ingest from all systems. Use ontologies/PK mappings to merge keys (e.g. patient IDs). Implement data catalogs so teams know what data exists (events.reutersevents.com) (nttdata-solutions.com). Ensure enterprise data warehouse for critical domains.

Failure Mode	Consequence	Data Foundation Fix
Poor Data Quality (missing, incorrect values) (nttdata-solutions.com)	Models trained on bad data yield unreliable predictions. E.g. Anomaly detection flags due to mislabeled lab units. Loss of trust.	Enforce **data cleaning pipelines**: automated scripts to fix obvious errors, impute or flag missing values, standardize units. Require source reconciliation (compare with gold sources) in preprocessing (nttdata-solutions.com). Monitor data drift & quality metrics, with alerts for anomalies.
Lack of Metadata/Lineage	Impossible to audit or trace data provenance; compliance risk. E.g. Audit finds AI decision but chain of custody for input data is unclear.	Adopt **FAIR data stewardship**: attach metadata, versioning, and lineage to every dataset. Maintain data dictionaries. Use data lineage tools to trace every model input back to source system (www.fiercepharma.com). Aim for "traceable, audit-ready data pedigree" (www.fiercepharma.com).
Siloed Teams & Ownership	Data gaps go unaddressed. E.g. R&D fixes data for its pilot, but Marketing never gets updates.	Establish **cross-functional governance**: data owners for each domain collaborate. Use centralized data governance boards. Align KPIs and incentives across functions for data sharing (www.techradar.com) (www.janeasystems.com). Projects treated as enterprise (C-suite) priorities.
Legacy Infrastructure	Slow, unreliable data pipelines. E.g. Queries take hours, blocking iteration.	Invest in **cloud-native, high-speed infrastructure**. Modernize ETL (streaming, APIs). Consolidate old servers and databases into scalable platforms (www.techradar.com). Ensure real-time or near-real-time data access for urgent insights.
Regulatory & Compliance Delays	Validation requirements halt deployment (months/years).	Build validation and compliance into pipeline from start. Keep immutable logs. Use secure, compliant platforms (FedRAMP/GxP cloud). Emulate FDA's Elsa – segregate sensitive data appropriately (www.axios.com). Pre-certify data sets (e.g. mappable to CDISC) to streamline audits.
Cultural Resistance	Stakeholders distrust or ignore AI output. E.g. Sales rep rejects AI suggestion citing "data is wrong".	Increase user trust by **improving transparency** (show model confidence, cite data sources). Involve end-users early for feedback. Upskill personnel on new data tools (www.imd.org). Create early wins by focusing on high-value, low-risk cases to build credibility.
Scaling to Production	Pilots require manual fixes and never become stable.	Implement **MLOps/ DataOps**: automated pipelines that retrain models when new data arrives. Integrate AI into workflows (e.g. CRM, LIMS) so it becomes part of the process. Use feedback loops: capture user corrections to refine data/models.

This table demonstrates that for nearly every downside of a stalled AI project, there is a corresponding data-centric strategy. For instance, if fragmented systems are to blame, the solution is to unify data sources (knowledge graphs, data lake) (events.reutersevents.com). If slow execution is a problem, the fix is better infrastructure and engine (GPU clusters, optimized pipelines) (www.techradar.com). By systematically pairing failure modes with these fixes, organizations can read this as a checklist of actions (and all are supported by references above).

Note that implementation of these fixes requires deliberate investment and time. But case evidence suggests it pays off: companies that have built such foundations report moving teams from *"data wrangling to insight generation"* (www.veeva.com).

Case Studies of AI Scaling (Real-World Examples)

The examples below draw practical takeaways from the earlier case discussions and add a discovery-partnership example.

A. Pharma Launch Delayed by Data Inconsistency

For the launch-delay example, see “Veeva Commercial Data Report (2025)” above and Veeva’s account (Veeva). The proposed response below focuses on what to do before expanding a related deployment.

Recommended response: Reconcile HCP records and establish a governed master dataset before expanding a related AI deployment. This is a proposed response, not a documented recovery sequence for the reported case.

This case underscores that even advanced predictive models depend on mundane data hygiene – a conclusion consistent with Veeva’s call for a “trusted single source of truth” (www.veeva.com) and the Pharmavanguard warning that AI “magnifies the cracks in your data” (pharmavanguard.com).

C. Government AI Pilot (FDA Elsa)

For the Elsa rollout discussed above, see “FDA AI Tool (Elsa)” and the FDA’s announcement (FDA).

Takeaway: Distinguish model-training restrictions from document access when specifying an internal AI tool’s data controls.

D. Novartis “Supercharged R&D” Initiative

For the Novartis example discussed in “People, Process, and Governance,” see its “Supercharging R&D” story (www.novartis.com) and the workforce-development discussion (www.imd.org). The augmentation perspective is also described in the Novartis story (www.novartis.com).

Relevance: Pair the workforce-development measures discussed in “People, Process, and Governance” with the technical rollout.

E. GSK Partnerships for AI-Driven Discovery

At JP Morgan Healthcare 2026, GSK’s leadership emphasized that AI was only a tool within a broader **“multimodal approach”** (www.pharmaceutical-technology.com). GSK’s deals (e.g. Helix genomics, Noetik AI oncology) were accompanied by commitments to integrate those partners’ data into GSK’s platform. For example, GSK is developing internal knowledge graphs and combining Helix’s genomic cohorts with their clinical compound library. This is an example of proactively **expanding the data foundation** when acquiring AI capability. GSK’s CSO explicitly noted that tackling “candidate attrition at Phase II” requires linking data from *genetics through phenotype to clinical readouts* (www.pharmaceutical-technology.com).

This case shows that frontier companies view data partnerships as part of AI strategy. The raw talent of Noetik's algorithms would be wasted without access to GSK's datasets, and vice versa. GSK's successful announcements were therefore not just about the algorithms, but about committing the data assets and integration needed to run those algorithms at scale.

Implications and Future Directions

Transforming Pharma with a Solid Data Foundation

The evidence is overwhelming: companies that invest in their data foundation enable meaningful AI transformation, while those that do not get left behind. For the pharma industry at large, this means the data architecture decisions made today will have strategic consequences tomorrow. Some forward-looking implications:

- **End of "Pilot Era":** Industry analysts have predicted that by 2026 the era of aimless pilots will end, and companies will either deliver operational AI or mothball the concept entirely (hitconsultant.net). As AI matures, organizations must evolve or risk obsolescence. Those who cement robust data foundations will move beyond experiments.
- **Operational Excellence and Speed:** With reliable data, AI can be embedded in real-time workflows. For example, **digital twins** of manufacturing processes become feasible only if the underlying data (sensor feeds, batch logs) are clean and accessible. Pharma 4.0 initiatives – such as adaptive control systems for bioreactors – rely on the same foundations.
- **Regulatory Evolution:** The January 2026 good-practice principles emphasize context-specific, proportionate validation and oversight (FDA principles).
- **From Hype to Value:** If done right, AI at scale could genuinely boost R&D productivity (e.g. by higher success rates in trials), improve patient outcomes (through personalized medicine models), and optimize costs (through predictive maintenance of equipment). McKinsey cautions that generative AI needs a proper data architecture to deliver results (McKinsey).
- **Emergence of Trusted AI (Explainability):** A robust data foundation is also the basis for **explainable AI**, a growing requirement in healthcare. When an AI model can cite exact trial data or molecular assays for each recommendation (as modern XAI tools attempt), it builds clinician trust. This will be especially important as the industry moves toward AI in precision medicine: genomic or diagnostic-driven treatment decisions cannot remain black boxes.
- **Competitive Advantage of Early Movers:** Large pharma and biotech that build advanced data infrastructures will gain a competitive edge over smaller companies. Just like having a larger chemical library, having a richer data asset will amplify future AI development. These leaders might also monetize their data – e.g. by contributing to consortia or licensing to AI partners – generating new revenue streams.

- **Iterative Improvement:** A solid data foundation enables continuous learning. As models generate new predictions, outcomes feed back into the system, refining both data and algorithms. This virtuous cycle requires the stable pipelines and governance we've described. Without them, each new project starts from scratch; with them, the platform improves over time.
- **Crisis Resilience:** The COVID-19 pandemic showed the value of agile data systems in pharma (e.g. for rapid vaccine rollout). In future health emergencies, companies with clean, integrated data can reposition and react faster (e.g. repurposing drugs via AI, as in the Harvard example (www.axios.com)). By contrast, companies still wrestling with data chaos will be too slow to adapt.

Near-Term Actions and Investments

Based on our analysis, recommended near-term steps for pharma CEOs and CIOs include:

1. **Conduct a Data Audit:** As a first practical step, organizations should inventory all AI-related data assets. Following NTT's guidance (nttdata-solutions.com), map out systems (LIMS, CRM, trial DBs, etc.), evaluate quality, and prioritize gaps. This audit should feed a remediation backlog with deadlines and owners.
2. **Create an AI Data Charter:** Formalize policies and responsibilities in writing. This charter might specify data standards (formats, ontologies), quality thresholds, and ethical use guidelines. As FiercePharma advised, move from "unclear policies" to a clear, FAIR-oriented framework (www.fiercepharma.com).
3. **Assemble Cross-Functional Teams:** For at least one high-impact project, set up a dedicated team with data engineers, data scientists, domain experts, and IT. Give them end-to-end ownership and required resources to break silos. Novartis-style integrated squads are a useful model.
4. **Invest in Technology Enablers:** Deploy or upgrade to modern data platforms. Options include commercial Life Sciences data clouds (Veeva, AWS for Health), analytics platforms (Databricks, Snowflake), and graph/database tools. Evaluate vendors on pharma-specific features like HIPAA compliance, clinical/reference integration, and support for GxP. Also invest in MLOps tools (Kubeflow, MLFlow) to handle model lifecycle.
5. **Prioritize a "Data Fix" Pilot:** Rather than coding novel algorithms immediately, run a pilot project specifically on data engineering. For example, pick one use-case (say, predictive maintenance in manufacturing) and focus first on ingesting and cleaning all relevant data over 3 months. Measure the ROI of data cleansing to build the business case.
6. **Measure and Report Foundation Health:** Establish metrics for the data foundation itself (perhaps data uptime, error rates, number of integrated sources). Report these at exec levels as you would any business KPI. Over time, improvements in these metrics should correlate with more successful AI deployments.

Long-Term Directions

Looking ahead, the relationship between pharma data foundations and AI will evolve significantly by 2030:

- **Foundation Models and Pharma-Specific LLMs:** The industry is moving toward building large AI models trained on biomedical corpora or multi-modal (chemistry + biology) data. Examples include efforts to adapt GPT-style architectures to drug discovery ("BioGPT"). These models can handle global knowledge, but their usefulness in pharma hinges on fine-tuning with high-quality proprietary data. Hence, the better the data foundation (clean, rich internal datasets), the more effective such models will become.
- **AI Agents and Automation:** Agentic AI systems (that act autonomously across workflows) will enter pharma more widely. Google announced Agent2Agent (A2A), an open protocol that complements Anthropic's Model Context Protocol (MCP) for providing tools and context to agents (Google). These agents will require extremely solid data foundations because they will make tasks like laboratory scheduling or supply chain adjustments. The TechRadar prediction that ~40% of agentic projects will be canceled if foundations are weak (www.techradar.com) underscores the risk.
- **Federated and Collaborative Data Ecosystems:** Pharma increasingly realizes that single-company data is limited. We may see consortia (e.g. Pistoia-style) that share cross-company, anonymized datasets via federated learning. Foundations will be key to participate: companies must ensure their in-house data is trustworthy before connecting to others.
- **Patient-Centric Data Integration:** The dream of personalized medicine (AI suggesting best treatment for an individual) depends on linking diverse patient data sources (genomics, wearables, EHR, claims). Pharma companies are pushing into this space (often via partnerships with providers). Data governance will extend to patient consent and cross-institution trust frameworks. Preparing for this will further stress data foundations.
- **Regulatory Evolution:** January 2026 good-practice principles already set out a foundation for developing AI good practice in drug development (FDA principles). These principles should not be presented as a guarantee of faster approval.
- **Economic Impact:** BCG reports that companies hope AI can improve organizational efficiency and preserve margins (BCG).

Conclusion

The journey from AI pilot to enterprise-scale transformation in pharma hinges on the often-overlooked data foundation. Our comprehensive review – citing surveys of executives, analysis by analysts, and real case examples – leads to a clear conclusion: *data readiness is one part of deployment planning*. RAND also identifies failures caused by fundamental limits in what AI can achieve (RAND).

Key actionable insights are:

- **Prioritize Data Quality and Governance.** Treat data as the first deliverable, not an afterthought. Invest resources to clean, integrate, and catalog data across R&D, manufacturing, and commercial functions. Use modern tools (data lakes, knowledge graphs) and enforce standards (MDM, FAIR).

- **Build Multi-Disciplinary Teams.** Ensure business, IT, and analytics staff collaborate throughout. Educate leaders to make data readiness a metric of AI ROI. Provide training and incentives to promote data sharing and continuous improvement.
- **Modernize Infrastructure.** Replace brittle legacy systems with scalable platforms designed for AI workloads. Adopt cloud and automation to ensure fast, elastic data pipelines.
- **Govern and Validate Continuously.** Implement robust metadata, audit trails, and compliance checks from the outset. Adapt the rigorous validation mindset of GxP to AI deployments.
- **Learn and Iterate.** Use pilot projects not only to test models, but to expose data gaps. Each failed pilot is a lesson in what to fix. Track metrics on data quality and use them to guide investments.

Pharma is finally at a tipping point: AI technology has matured, but its impact will be limited by how well the industry addresses these foundational issues. Companies that heed the warnings – investing in trusted data lakes, integrated systems, and strong data governance – will unlock the promised efficiencies and innovations. Those that don't risk relegating AI to "innovation theater" forever.

In short, **the cure for pilot failure is a healthy data foundation**. By establishing clear data provenance, unified knowledge structures, and strong oversight, pharma can transform scattered pilots into powerful, value-driving AI at scale. The evidence is clear, the solutions are known – it's now up to industry leaders to act on them.

Sources: All statistics and assertions are supported by the cited industry reports, surveys, and analyses (www.fiercepharma.com) (www.bio-itworld.com) (pharmavanguard.com) (www.techradar.com) (events.reutersevents.com) (www.axios.com) (www.veeva.com) (www.pharmaceutical-technology.com) (apnews.com) (sakaradigital.com) (www.axios.com) (nttdata-solutions.com), among others. These include reputable trade publications, conference proceedings, whitepapers, and news coverage of pharma and AI trends. Each claim above is traceable to specific excerpts from those sources.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.rand.org/pubs/research_reports/RRA2680-1.html
2. https://pistoiaalliance.org/news/1-in-4-life-sciences-professionals-dont-know-what-data-their-ai-models-use
3. https://www.fiercepharma.com/premium/webinar/1379583
4. https://events.reutersevents.com/pharma/data-and-tech-europe/agenda
5. https://sakaradigital.com/blog/pharma-ai-pilots-fail-production-3
6. https://nttdata-solutions.com/bnl/blog/machine-learning-for-pharmaceuticals-blog-series-part-3
7. https://www.bio-itworld.com/news/2026/01/20/ai-in-pharma--progress-made--but-data-challenges-persist
8. https://www.fda.gov/media/189581/download
9. https://pharmavanguard.com/opinion/why-pharmas-ai-pilots-stall-and-how-medical-affairs-commercial-can-actually-win
10. https://www.axios.com/2024/09/25/old-drugs-repurposed-rare-diseases
11. https://www.techradar.com/pro/five-signs-your-infrastructure-is-stalling-your-ai-strategy

12. https://www.techradar.com/pro/why-agentic-ai-pilots-stall-and-how-to-fix-them
13. https://www.janeasystems.com/blog/ai-big-data-expo-2025
14. https://www.techradar.com/pro/why-ai-pilots-fail-and-how-manufacturers-can-break-the-cycle
15. https://www.techradar.com/pro/why-most-agentic-ai-projects-fail-and-how-to-avoid-being-one-of-them
16. https://apnews.com/article/004c0ce0442b72c37bfec6e032796808
17. https://www.pharmaceutical-technology.com/news/jpm26-gsk-banks-on-recent-ai-deals-to-offset-patent-expiries
18. https://www.fda.gov/news-events/press-announcements/fda-launches-agency-wide-ai-tool-optimize-performance-american-people
19. https://www.veeva.com/resources/the-state-of-data-analytics-and-ai-in-commercial-biopharma
20. https://www.veeva.com/wp-content/uploads/2025/10/The-State-of-Data-and-Analytics-Commercial-Biopharma-eBook-NA.pdf
21. https://www.novartis.com/stories/supercharging-rd-how-ai-helps-us-bring-medicines-patients-faster
22. https://www.imd.org/entity-profile/novartis-ai-maturity-2025
23. https://www.fda.gov/media/184830/download
24. https://www.axios.com/2025/06/02/fda-launches-ai-tool
25. https://hitconsultant.net/2025/12/16/2026-executive-predictions-why-the-ai-pilot-era-is-officially-over
26. https://www.axios.com/2025/05/12/fda-ai-drugs-company-data-questions
27. https://www.techradar.com/pro/why-more-than-half-of-ai-projects-could-fail-in-2026
28. https://www.qlik.com/us/news/company/press-room/press-releases/qlik-2025-agentic-ai-study-budgets-surge-but-data-readiness-delays-scale
29. https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality
30. https://docs.databricks.com/aws/en/introduction
31. https://docs.snowflake.com/en/user-guide/intro-key-concepts
32. https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability
33. https://www.bcg.com/press/8january2026-biopharmas-innovation-engine-powers-ahead-business-model-faces-strain

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.