IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

# AI Policies & Data Classification for Clinical Biotech

> 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 · 2026-04-12 · ai policies · data classification · clinical biotech · ai governance · clinical trials · health data privacy · eu ai act · regulatory compliance

Original article: https://intuitionlabs.ai/articles/ai-policies-data-classification-clinical-biotech



### In this article

1. Executive Summary
2. Introduction
3. The Role of AI in Clinical-Stage Biotech
4. Regulatory and Policy Landscape
5. Data Classification Frameworks in Biotech
6. Case Studies and Examples
7. Data Analysis and Evidence-Based Arguments
8. Implications and Future Directions
9. Conclusion

# Executive Summary

Artificial Intelligence (AI) is rapidly transforming clinical-stage biotechnology, promising faster drug discovery and trial processes while raising profound data governance challenges. Clinical-stage biotech firms handle vast volumes of sensitive data – from patient medical records to proprietary experimental results – making robust **AI policies** and **data classification frameworks** essential to ensure compliance, privacy, and data integrity. This report examines the historical evolution and current landscape of AI governance and data classification in clinical biotech. We analyze global and regional regulatory policies (e.g. the EU AI Act, HIPAA/GDPR and FDA guidance), industry best practices, and technical frameworks for categorizing data sensitivity. We highlight how leading companies are implementing multi-dimensional classification schemes (balancing privacy, product quality, confidentiality and intellectual property concerns (www.immuta.com)) and deploying organizational AI governance structures to manage risk (for example, establishing dedicated AI oversight committees (www.mdpi.com) (www.frontiersin.org)). Case studies illustrate real-world responses: from federated learning consortia that share AI insights while protecting proprietary data (e.g. the multi-pharma MELLODDY project with 2.6+ billion confidential data points (pmc.ncbi.nlm.nih.gov)) to startups using AI to automate trial documentation and patient matching (www.statnews.com) (hexaware.com). We synthesize evidence – including industry surveys and policy analyses – showing that AI use spans the product lifecycle, while a 2024 survey of life sciences and healthcare executives identifies gaps in formal governance in its sample (Arnold & Porter survey overview). These findings support reviewing privacy and compliance controls. Finally, we discuss emerging trends (such as the role of generative AI, federated learning for privacy, and evolving AI regulations worldwide) and outline recommendations for biotech organizations: develop comprehensive AI policies, adopt risk-based data classification frameworks, and maintain interdisciplinary teams for ongoing oversight. All claims are supported by recent studies, official guidelines, and expert reports (www.mdpi.com) (www.axios.com) (pmc.ncbi.nlm.nih.gov).

# Introduction

Clinical-stage biotechnology sits at the intersection of cutting-edge science and highly regulated medicine. These companies conduct **clinical trials** to test novel therapies in humans, generating enormous volumes of data – from detailed patient health records, genomic sequences, and imaging, to trial operations and adverse event reports. At the same time, many of these biotechs are integrating AI into their workflows, applying machine learning to streamline R&D, optimize trials, and analyze complex biological data. This dual trend – exploding data and pervasive AI – creates an urgent need for robust **AI governance policies** and **data classification frameworks** tailored to biotech. On the one hand, AI offers unprecedented capabilities (e.g. automating patient stratification (www.statnews.com) or accelerating trial logistics (time.com)). On the other hand, it raises risks of privacy breaches, biased algorithms, and regulatory non-compliance if data are mishandled or poorly secured (www.mdpi.com) (www.frontiersin.org).

Historically, data management in biotech has long been governed by standards like Good Clinical Practice (GCP), 21 CFR Part 11 (FDA rules on electronic records), and patient privacy laws (HIPAA in the U.S., GDPR in Europe). However, the AI era introduces new dimensions. AI systems depend on large training datasets and can infer novel insights (and potentially sensitive patient inferences), meaning that organizations must carefully control what data an AI can access and how models are audited. Simultaneously, regulators worldwide are evolving to address AI-specific concerns. For example, the EU implemented the **AI Act** (2024) with risk-based rules for AI in healthcare, and bodies like the U.S. FDA are actively engaging the biotech community on AI use in drug development (www.mdpi.com).

Data classification – the process of labeling data by sensitivity and intended use – is equally crucial. A sound classification framework allows biotech firms to identify protected health information (PHI), intellectual property, and trial data needing different security controls. Multi-tiered schemes (similar to NIST or ISO models) are often adopted, with labels like “Public, Internal, Confidential, Restricted, Top-Secret” or clinically-specific tags. Establishing these categories enables automated policy enforcement (e.g., where PHI demands encryption and restricted access) and supports regulatory compliance (e.g., meeting HIPAA’s privacy rule by labeling PHI (www.ncbi.nlm.nih.gov) or GDPR’s special category data (www.ncbi.nlm.nih.gov)).

This report provides a comprehensive, evidence-based analysis of AI policy and data classification frameworks in clinical-stage biotech. We first review the historical and regulatory context, then delve into current practices and standards. We examine technical frameworks for data classification (both generic and biotech-specific), and how they align with policies across regions. We include multiple perspectives – regulatory, technical, and industry – and detailed case studies of companies and projects. Throughout, we cite authoritative sources: peer-reviewed literature, government regulations, industry surveys, and expert analyses. Our goal is to offer an in-depth guide for stakeholders in clinical biotech to understand the landscape and prepare for the future of AI governance.

# The Role of AI in Clinical-Stage Biotech

AI technologies have permeated every stage of biotech R&D. In **drug discovery** and early development, machine learning helps identify molecular candidates, predict toxicity, and repurpose existing drugs. However, clinical-stage biotech companies (those conducting human trials) often find the bottleneck is not drug design but trial execution, patient recruitment, and regulatory documentation. As one industry leader notes, while AI has revolutionized discovery, the number of new drug approvals remains flat at ~50/year (time.com), signaling that the critical gains now lie in optimizing trials. For example, Formation Bio says its in-house clinical operations include patient recruitment, site management, and study monitoring (Formation Bio operating model). Similarly, startups like Nucleai use AI on histopathology images integrated with clinical data to predict treatment response and match patients to trials (www.statnews.com), illustrating AI's ability to stratify patient cohorts more intelligently.

These AI applications, while promising, come with data challenges. Clinical trials generate **complex, unstructured data**: electronic health record (EHR) extracts, genomic/omics datasets, imaging studies, and free-text reports (www.biospace.com). Maintaining data integrity and compliance is already a pain point; for instance, a BioSpace analysis noted that in many trials, patient data still have to be manually transcribed between systems (EHR to trial database), slowing processes and risking errors (www.biospace.com). Introducing AI tools into this ecosystem requires very clear data governance. Organizations need to know *what* data are being fed into an AI (personal health data vs. de-identified summary, commercial IP vs. public domain), and *why* (for training a model, making a regulatory submission, etc.).

The stakes are high because clinical data are extremely sensitive – they often involve identifiable patient health information. In the U.S., HIPAA's Privacy Rule broadly defines "Protected Health Information" (PHI) as any directly or indirectly identifiable health info (www.ncbi.nlm.nih.gov). HIPAA applicability requires a role-specific assessment: a biotech sponsor receiving data for research is not automatically a covered entity or business associate. Covered entities include health plans, clearinghouses, and healthcare providers that conduct specified electronic transactions; business-associate status depends on the functions or services performed for a covered entity (HHS). Research disclosures do not themselves require a business-associate contract: assess the permitted research disclosure route separately, such as individual authorization, an applicable IRB or Privacy Board waiver, or a limited data set with a data use agreement (HHS research-disclosure guidance; HIPAA and research). In Europe, the GDPR treats health data as a **special category** requiring extra protection (www.ncbi.nlm.nih.gov). Data like genetic information, biomarkers, or clinical outcomes clearly fall in this category. Thus, any AI system using such data must comply not only with machine-learning best practices but also with health privacy laws. This complexity is compounded by the fact that AI models themselves can inadvertently leak sensitive information (especially large language models trained on clinical notes), making robust data classification and oversight essential.

Given these considerations, clinical-stage biotech typically adopts a multi-faceted view of data. They recognize **clinical trial data** as a "crown jewel" requiring end-to-end protection, while other data (aggregate analytics, published research) may be less sensitive. They classify data along dimensions like *privacy* (Is it PHI/PII?), *GxP criticality* (impact on product quality or patient safety), *confidentiality level* (who can see it), *intellectual property significance*, and *data integrity requirements* (www.immuta.com). This ensures that, for example, a

Phase III patient dataset is tagged “Strictly Confidential – GxP-High – PHI”, which triggers maximum safeguards. We explore these schemes in detail later. What matters is that the intersection of AI and biotech drives an imperative: companies must have clear **AI policies** (covering development, validation, deployment) that integrate with **data classification frameworks** to manage risk.

# Regulatory and Policy Landscape

Both **governments** and **industry bodies** are responding to AI in biotech with new and evolving policies. These policies address multiple concerns: patient privacy, algorithmic transparency, safety of AI-driven decisions, and data governance. Below we outline key policies and their relevance.

**U.S. Framework:** The U.S. takes a largely sectoral approach. Health data privacy is governed by **HIPAA (1996)** and later amendments, which mandate standards for electronic health records and PHI (www.ncbi.nlm.nih.gov) but do not specifically address AI. FDA’s January 2025 draft guidance, *Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products*, proposes a risk-based framework for assessing AI-model credibility for a defined context of use. It is draft guidance, not for implementation, and contains non-binding recommendations (FDA draft guidance). On 14 January 2026, FDA and EMA released ten joint guiding principles for good AI practice in drug development (FDA announcement). Earlier, FDA hosted a 2024 workshop (co-chaired by academia) on AI in drug/biologic development (www.mdpi.com). The workshop emphasized needs for transparency, data quality, and algorithmic fairness but noted gaps in bias and privacy protection (www.mdpi.com). The FDA has also issued guidance for AI in medical devices (currently AI in diagnostics) and is piloting programs for advanced manufacturing technologies (which could include AI). Meanwhile, on **AI-specific law**, the U.S. lacks a comprehensive statute, but agencies like NIST and OSTP have labored on voluntary guidelines: e.g. NIST’s *AI Risk Management Framework (AI RMF)* (2023) provides a structure for assessing AI risks, and FTC/SEC have hinted at enforcement on deceptive or non-robust AI practices (www.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). U.S. states (e.g. New York’s bias audits for hiring algorithms, though in 2024 these mostly affect non-biotech uses) also signal growing scrutiny. Importantly, the U.S. still relies heavily on existing privacy law (HIPAA, as well as consumer laws enforced by FTC) to cover AI uses of health data (www.ncbi.nlm.nih.gov).

**European Framework:** The **EU AI Act** entered into force on 1 August 2024 and became generally applicable on 2 August 2026, with phased exceptions. The Commission’s current timeline gives 2 December 2027 for Annex III high-risk rules and 2 August 2028 for Annex I regulated-product high-risk rules (European Commission). Clinical use alone does not establish high-risk status. Under Article 6(1), the product-related route requires both an AI system that is an Annex I product or its safety component and a product that requires third-party conformity assessment; other high-risk uses are addressed through Annex III (Article 6). In addition, Europe’s **GDPR** (effective 2018) requires “special category” data (health, genetic) to have explicit consent or other safeguards (www.ncbi.nlm.nih.gov). Under GDPR, automated decision-making in healthcare may give patients rights to human review. The EU also encourages ethics; UNESCO’s global AI ethics recommendations and WHO’s digital health guidelines stress patient rights and fairness. Assess research, administrative, and clinical-decision uses separately by intended purpose, device status, conformity-assessment route, and AI Act

classification. Article 2 excludes systems developed and put into service solely for scientific research and development, and certain pre-market research activities; these exclusions have limits, including for real-world testing (Article 2).

**Other Regions and International Principles:** Other countries are catching up. The OECD has published *AI Principles* (2019) and is drafting sector-specific guides, including one for health (early 2024) emphasizing trust and privacy. China's new Personal Information Protection Law (PIPL) treats health/genetic data as highly protected; combined with recent AI guidelines (e.g. restricting "deep synthesis" tech), Chinese biotechs must navigate both. The UK's proposed AI regulation (post-Brexit) similarly follows a principle-based model, and British data law still mirrors GDPR (UK GDPR). Industry organizations (e.g. the Asia-Pacific Economic Cooperation or WHO) have privacy and AI charters; the key theme is cross-border. For biotech, global trials (e.g. multi-country studies) mean harmonizing data classification across jurisdictions.

**Corporate Policies:** Within companies, formal **AI governance frameworks** are emerging. Leading firms (e.g. AstraZeneca) have established internal AI oversight. A study on corporate AI governance in biopharma highlights that companies start with ethical AI principles but must translate them into operations – setting up committees, defining material scope of "AI" (since every data tool isn't necessarily AI), aligning decentralized groups, and measuring governance impact (www.frontiersin.org). The key recommendation is to *leverage existing quality and compliance structures*: for example, include AI data governance in standard GxP documentation practices. Many firms also require AI projects to undergo risk assessment (aligning with existing risk management) and to incorporate privacy-by-design. Arnold & Porter's 2024 survey of 100 executives and department heads across life sciences and healthcare identified gaps in formal governance among the respondents. Its published overview describes governance as lagging implementation (survey overview). This sample supports reviewing corporate AI policies; it does not establish that companies without formal policies have no oversight.

Below is a table summarizing some key AI/data governance frameworks relevant to biotech:

| Policy/Framework | Scope & Region | Key Requirements |
|---|---|---|
| **EU AI Act (2024)** | EU-wide, subject to the Act's scope | Risk-based rules; assess the intended use and applicable classification before determining obligations. The Commission distinguishes unacceptable, high, transparency, and minimal or no risk (European Commission). |
| **HIPAA / HITECH (1996/2009, USA)** | USA, for entities and data within scope | Research uses and disclosures of PHI require a permitted route, such as individual authorization, a qualifying waiver, or a limited data set with a data use agreement (HHS research guidance). Receiving data for research does not itself require a business-associate contract (HHS FAQ). |
| **GDPR (2018, EU)** | EU and jurisdictio ns | Treats health/genomic data as special category needing explicit consent or legal basis (www.ncbi.nlm.nih.gov). Grants data subjects rights (e.g. transparency, erasure). Automated decision-making rights (can contest AI decisions). |

| Policy/Framework | Scope & Region | Key Requirements |
|---|---|---|
| **FDA AI Draft Guidance & Principles** | USA (FDA); joint FDA–EMA principles | FDA's January 2025 draft guidance proposes a risk-based framework for assessing AI-model credibility for a defined context of use. It is not for implementation and contains non-binding recommendations (FDA draft guidance). On 14 January 2026, FDA and EMA released ten guiding principles for good AI practice in drug development (FDA announcement). The principles provide a foundation for good practice; they are not a regulation. |
| **NIST AI RMF (2023, USA)** | Voluntary (US) | Framework for managing AI risks throughout lifecycle; emphasizes classification of system impact and security requirements. Not legally binding, but influential as best practice. |
| **OECD AI Principles (2019, global)** | OECD member countries | Non-binding principles on trustworthy AI: inclusive growth, human rights, accountability, safety. Encourages proportionate regulation and interoperability. |
| **Company AI Governance (e.g. AstraZeneca)** | Internal | Example frameworks call for building AI governance on existing quality systems, using clear terminology and focusing on actionable risk controls (www.frontiersin.org); often create cross-functional AI committees and approval processes. |

*Table: Overview of selected AI and data governance frameworks relevant to clinical biotech (sources cited above).*

# Data Classification Frameworks in Biotech

A **data classification framework** provides labels or categories for data assets based on sensitivity, legal/regulatory requirements, and business-critical value. These categories inform how data must be handled (e.g., encryption, access controls, sharing rules). Clinical biotech data are especially varied, and firms typically adopt multi-dimensional classification schemes (www.immuta.com) (pmc.ncbi.nlm.nih.gov). Common dimensions include:

- **Privacy/Sensitivity:** Identifies personally identifiable information (PII) and protected health information (PHI). Under HIPAA, PHI is defined as any individually identifiable health info (www.ncbi.nlm.nih.gov). The EU GDPR covers "health" and "genetic" data as special categories (www.ncbi.nlm.nih.gov). In practice, biotech firms tag columns or files containing patient names, IDs, or sensitive traits (e.g. HIV status) as highly sensitive, requiring de-identification or strict access.
- **Regulatory Criticality (GxP Level):** Determines whether data are subject to Good Manufacturing/Clinical Practices. For example, manufacturing batch records or primary trial endpoints are GxP-critical (GxP-High) because inaccuracies could harm patient safety or regulatory filings. Non-critical data (e.g. preliminary research notes) might be GxP-Medium/Low.
- **Confidentiality:** How widely the data can be shared. "Public" data (e.g., published research literature) have no restrictions. "Confidential" might be internal use only. "Strictly Confidential" is limited to a very small group (e.g. specific R&D team), often triggering heavy encryption and logging. A table from industry illustrates these tags (www.immuta.com):

| Dimension | Description | Example Tag |
|---|---|---|
| **Privacy** | PHI/PII status; GDPR special category data. | `PHI`, `PII` |
| **GxP Criticality** | Impact on patient safety or regulatory approvals. | `GxP-High` |
| **Confidentiality** | Sharing scope (internal vs. public). | `Strictly Confidential`, `Public` |
| **Crown Jewel (IP)** | Strategic value to company (trade secrets, core IP). | `Yes`/`No` |
| **Integrity** | Required assurance level (accuracy, audit). | `Vital`, `Standard` |

*Table: Example multi-dimensional data classification tags in pharma/R&D (www.immuta.com).* (Each dataset or table column might carry multiple tags. For instance, a Phase III patient dataset could be “Strictly Confidential; GxP-High; PHI”.)

Underpinning these practices are national and international standards. In the U.S., frameworks like **NIST SP 800-53** (security controls) implicitly rely on classifying data by impact (low/medium/high). The NIST *Data Classification Practices* project notes that effective security depends on “organizations knowing what data they have, what its characteristics are, and what security/privacy requirements it needs to meet” (csrc.nist.gov). In healthcare, these characteristics often center on the *sensitivity* of patient vs. aggregated data, and on data’s criticality to product quality.

Notably, specialized healthcare classification guides have emerged. For instance, a 2025 industry case study discussed building deep-learning taxonomies to classify clinical trial documents by content (hexaware.com). The Egyptian government similarly formulated a health data classification methodology: it categorizes data into layers (e.g. public, internal, confidential, restricted, or top-secret) based on sensitivity and disclosure risk (pmc.ncbi.nlm.nih.gov). That model mirrors many corporate schemes: data labeled *restricted or top-secret* trigger maximum security controls, while *public* or *internal* data do not. The Egyptian analysis further emphasizes that classification is the **first step in data governance**, underlying all subsequent policies (pmc.ncbi.nlm.nih.gov).

Data classification in biotech must also consider emerging data types. Genomic and multi-omics datasets may carry uniquely identifying signals, so they often get the “most sensitive” label. Real-world images (e.g. radiology) and even structured metadata (dose levels, lab values) need scrutiny. For AI, an additional layer is *model ecosystem data*: training data provenance and synthetic data generation. For high-risk AI systems within Article 10’s scope, training, validation, and testing datasets are subject to data-governance duties appropriate to the intended purpose. These include examining specified potentially harmful biases, taking appropriate measures to detect, prevent, and mitigate them, and using sufficiently representative datasets; for systems that do not train AI models, the relevant duties apply to testing datasets (Article 10).

Overall, a robust classification framework is **instrumental for AI policies**. It feeds into access controls (ensuring only authorized algorithms or users can read sensitive data), audit trails, and privacy-preserving technologies. For example, often PHI-tagged data might only be used in de-identified form for model building, or accessed via secure enclaves. The biotech industry increasingly uses automated tools to enforce

classifications: e.g. AI/ML systems that scan new datasets, tag columns (e.g. “date_of_birth” → PHI), and assign labels reactive to context (www.immuta.com). Large organizations also integrate these tags with role-based access (if `GxP-High`, only validated QA personnel can alter the dataset).

# Case Studies and Examples

To illustrate how AI policies and data classification play out, we examine several real-world examples:

- **MELLODDY Consortium (Pharma Federated Learning):** Ten pharmaceutical companies (including AstraZeneca, Bayer, Janssen, Novartis, Merck, Amgen, and Boehringer Ingelheim) collaborated through MELLODDY on models trained using their private data without exchanging raw data (MELLODDY launch announcement). Using a privacy-audited federated platform, they pooled over **2.6 billion confidential activity data points** (covering 21+ million molecules) while each partner only received aggregate model improvements (pmc.ncbi.nlm.nih.gov). The results showed measurable gains in predictive accuracy for each firm’s QSAR models, demonstrating that *collaborative AI can boost R&D while protecting IP*. The study reports federated learning on confidential small-molecule assay data using a platform audited for privacy and security (MELLODDY study). A hypothetical clinical-biotech adaptation could add organizational confidentiality, GxP, and patient-data labels after assessing the actual datasets; those tags are not documented MELLODDY deployment features. Clinical applications would require their own privacy and governance assessment.
- **FDA AI Workshop (2024):** Regulatory insight can also shape policy. In August 2024, the FDA and Clinical Trials Transformation Initiative convened experts to discuss AI in drug development (www.mdpi.com). Although not a case per se, the workshop’s outputs highlight priorities. Presenters emphasized transparency of algorithms, ongoing data quality monitoring, and auditability. They also acknowledged gaps: for example, while sponsors focus on *model transparency*, they often under-address *data privacy*. The workshop report suggests that biotech firms should codify policies requiring bias testing, provenance tracking, and alignment with patient consent. This event reflects how regulators are gathering clinical-stage AI use cases to inform future FDA guidances.
- **Formation Bio (AI-augmented trials):** Formation Bio says its proprietary AI-native platform supports faster drug development (Formation Bio). This is a company description, not an independently verified estimate of time saved. As an illustrative governance exercise, a biotech team using patient-level data for trial operations should assess identifiability, permitted use, and access controls before deployment. Any example of separating identifiable records from aggregate analytics is a suggested design, rather than a verified description of Formation Bio’s implementation.
- **Clinical Trial Data Automation (Hexaware Case):** Hexaware’s public case overview describes a client in contract research services, healthcare data science, and technology solutions. Hexaware reports building taxonomies and deep-learning models to classify and digitize trial documents with minimal manual intervention (Hexaware case overview). As an illustrative extension, a biotech team could link document classifications to access and handling policies; this is a suggested design, rather than a verified feature of the reported deployment.

- **Biotech Leaders' Perspectives:** Industry voices also reveal evolving policies. An Axios report quotes Bayer executives: Simon Rosof (Bayer) noted that AI has "streamlined" gene-driven disease screening, while Brian Cantwell warned about patients self-diagnosing using tools like ChatGPT without oversight (www.axios.com). These remarks illustrate a policy milieu: companies are deploying AI for internal R&D gains but recognize new ethical pressures (e.g. AI health advice to patients). Arnold & Porter's 2024 survey also identified governance challenges in its sample of life sciences and healthcare executives (survey press release). This gap is a wake-up call: as AI use spreads (e.g. data mining large EHR or genomic repositories for trial insights), firms risk regulatory breaches unless they establish clear data classification (to avoid improperly using PHI) and governance protocols (to audit AI outputs).
- **Small Biopharma Innovation:** Opinion pieces highlight how smaller biotechs are agile in applying AI. Iddo Peleg (CEO of an AI for trials startup) notes that **automated EHR integration** is replacing manual data entry, despite big pharma's slow adoption (www.biospace.com). Companies like YonaLink now offer solutions to stream patient data directly into trial databases, which inherently requires classifying each field's sensitivity and encrypting the pipeline. In summary, these industry examples – from global consortia to scrappy startups – illustrate the same theme: success with AI in clinical biotech hinges on pairing powerful analytics with stringent data classification and governance.

# Data Analysis and Evidence-Based Arguments

Empirical data and expert analyses highlight both the promise and challenges of AI/data governance in biotech:

- **AI Adoption vs. Governance Gap:** Arnold & Porter's 2024 survey covered 100 senior executives and department heads across life sciences and healthcare. In that sample, 75% began implementing AI within the preceding two years; among companies currently integrating AI, 86% planned full deployment within the following two years (survey overview; press release). The report's governance section states that 53% of companies that had already implemented AI had policies and SOPs in place, and 51% had completed regular audits and assessments (report, p. 21). These findings indicate gaps in formal governance in this sample; they do not show that nearly half of all biotech firms lack oversight.
- **Data Privacy and Classification Importance:** Studies of privacy law (like Konnoth's 2024 analysis) underline the importance of classifying data by its inherent sensitivity, not just context (www.ncbi.nlm.nih.gov). For instance, GDPR's classification of "health data" pivots on content: any information about a person's physical/mental health is special-category (www.ncbi.nlm.nih.gov). In the U.S., HIPAA requires de-identification by removing 18 specified identifiers (www.ncbi.nlm.nih.gov) or an expert determination. These frameworks impose duties within their respective scopes; an organization's classification labels are an operational way to support compliance, rather than a blanket legal requirement to tag every item. HIPAA distinguishes PHI from qualifying de-identified information (HHS).

HL7 also notes that tagging individual resources is not always necessary to implement a policy (FHIR security labels). Biotech teams should assess identifiability, permitted use, and applicable safeguards rather than automatically applying the strictest controls to all health-related information.

- **Performance Gains from Federated Learning:** Quantitative outcomes from MELLODDY reinforce the data-classification approach. After training across 2.6 billion data points, each company reported significant improvements in their models' predictive metrics (pmc.ncbi.nlm.nih.gov). Importantly, this was achieved *without any partner relinquishing proprietary data*, validating federated learning as a privacy-preserving model. The evidence here supports the strategy: classify data as "cannot be moved", use federated learning instead, and still gain. Thus, for clinical-stage biotech with high IP sensitivity, this cooperation model provides a blueprint for deriving AI benefit under strict data classification.
- **Regulatory Workshops and Policy Formation:** The FDA's 2024 workshop was, in effect, an empirical needs assessment. Feedback from industry experts indicated "transparency, data quality, management, and algorithmic fairness" as key pillars (www.mdpi.com). This consensus among participants (supported by workshop notes) argues for policies that enforce these elements. For example, one may conclude that any AI policy in biotech should **require transparency (audit logs, explainability)** and **strict data quality controls** (aligned with how data are classified and curated). The lack of workshop focus on bias and privacy was also noted (www.mdpi.com), implying these areas need more attention in future guidance. This evidence informs our recommendations: policies must explicitly mandate data governance (classification, privacy filters) and fairness checks before AI deployment.
- **Ethical and Global Considerations:** Papers and handbooks emphasize global consistency on AI in health. For instance, an OECD brief ("AI in Health: Huge Potential, Huge Risks" 2024) stresses balancing innovation against data privacy and societal concerns (www.oecd-ilibrary.org). It highlights that not adopting AI has risks comparable to misusing it. This supports a data-driven approach: rather than halting AI, governments should formulate concrete policies (like classification frameworks) to safely harness it (www.oecd-ilibrary.org). The OECD's stance, backed by multi-country agreements, provides evidence that the international community is leaning toward risk-managed AI integration, reinforcing the need for biotech firms to align with such frameworks.

In sum, the data and expert consensus show a fast-growing AI ecosystem in biotech, coupled with a patchy governance landscape. This underscores the imperative arguments of our report: biotech companies must urgently develop AI policies that **prioritize data classification, risk management, and compliance** to fully capitalize on AI while safeguarding privacy and public trust (www.axios.com) (www.frontiersin.org).

# Implications and Future Directions

Looking ahead, the convergence of AI and biotech will have profound implications for both innovation and regulation. Several trends are especially important:

- **Increasing Regulatory Scrutiny:** The EU AI Act entered into force on 1 August 2024 and became generally applicable on 2 August 2026. The Commission's current timeline sets Annex III high-risk application for 2 December 2027 and Annex I regulated-product high-risk application for 2 August 2028

(European Commission). Firms developing systems within the applicable high-risk scope should plan for *data governance requirements* – for example, documenting dataset composition and demonstrating bias testing (www.hoganlovells.com). We may see regulators requiring auditable "AI submissions" akin to drug dossiers, where data lineage and model validation reports become standard. The U.S. may follow with similar codified expectations or updated FDA guidances. Emerging ethical guidelines (e.g. WHO's strategy for digital health) will also shape expectations, likely pushing for certifications or third-party audits of AI systems. FDA and EMA released ten joint guiding principles for good AI practice in drug development on 14 January 2026 (FDA announcement). These principles provide a shared foundation for good practice; firms should continue monitoring jurisdiction-specific requirements.

- **Evolution of Data Classification Schemes:** Data classification will evolve beyond static labels. Dynamic classification – where data can change sensitivity based on context or usage – may emerge. Aggregation alone should not trigger a public-release label. Before downgrading protections, document whether the output meets the applicable de-identification or anonymization standard and whether release is permitted; under HIPAA, qualifying de-identified information is no longer PHI (HHS). AI systems themselves could assist classification: tools that automatically recognize PHI in EHR notes (using NLP) and tag it. HL7 FHIR already defines security labels that attach security metadata to resources or bundles; access-control decisions also depend on policy, consent, provenance, and other context. Future developments may extend these labels or increase their adoption (FHIR security labels). We may also see international efforts (like of OECD) to unify classification terminology for AI (the OECD has proposed an AI system classification framework (www.frontiersin.org)) which biotech could adapt. Synthetic generation does not by itself establish that an output is anonymous or non-PHI. Before reducing protections or releasing synthetic clinical data, document an identifiability assessment under the applicable standard; HIPAA de-identification uses Expert Determination or Safe Harbor (HHS). As a governance practice, record provenance, assessed identifiability, and release permission as separate attributes, and reassess privacy risk over time (FCA/ICO discussion).
- **Expansion of Federated and Secure AI:** The success of projects like MELLODDY suggests federated learning will become commonplace in pre-competitive R&D collaborations. Future consortia might pool clinical trial outcome data (meta-analyses) under privacy guarantees to improve predictive models. Alongside, other approaches like secure multi-party computation and homomorphic encryption will mature, enabling even stricter data controls. This means biotech companies will also need policy frameworks that recognize these technologies – specifying, for example, how to classify and audit AI models trained under such schemes.
- **Organizational Change and Education:** On the corporate side, building AI policies is as much a people challenge as a technical one. The AstraZeneca/AOI study recommends continuous education and cross-functional teams (www.frontiersin.org). We predict more biotech companies forming dedicated AI governance boards or AI Officers, analogous to Data Protection Officers under GDPR. Training programs on AI ethics and data management for scientists and managers will spread. In the near future, proficiency in AI compliance may become a standard skill for biotech leadership. The implications for workforce and culture are significant: organizations will need to integrate AI risk metrics into their existing compliance KPIs.

- **Ethical and Societal Considerations:** Finally, AI in biotech carries societal implications. Beyond legal compliance, companies will face public scrutiny over issues like algorithmic bias (e.g. if an AI-guided trial disproportionately excludes minorities). Transparent communication about data use and safeguards will be critical to maintain trust. There may be calls for inclusive data sampling in training models (a classification dimension could even explicitly track demographic attributes to ensure diversity). Looking further out, as generative AI capabilities enter life sciences (e.g. designing novel molecules or simulating trials), regulatory frameworks will need to address data used in these models. For example, if an AI suggests a new drug target, how do we classify the underlying biological data and model outputs? Policies will likely co-evolve with technology.

# Conclusion

AI holds transformative potential for clinical-stage biotech, from smarter trial design to accelerated discoveries. However, realizing these benefits requires **meticulous governance** of data and algorithms. This report has shown that both regulatory bodies and industry are coalescing around risk-based AI policies and sophisticated data classification frameworks. Key takeaways include:

- **Data classification is foundational.** Clinical biotech must label and segregate data by sensitivity (PHI vs de-identified, GxP level, IP critical) so that AI systems use each dataset appropriately (www.immuta.com) (pmc.ncbi.nlm.nih.gov). Multi-axis schemes (privacy, confidentiality, integrity, etc.) align well with pharma's regulatory needs (www.immuta.com).
- **Global policies are evolving.** In the EU, the AI Act imposes new data governance obligations on high-risk AI, layered atop GDPR and medical device law (www.hoganlovells.com). In the U.S., HIPAA and emerging FDA guidelines emphasize data quality and fairness, though much relies on companies to self-regulate (www.mdpi.com) (www.ncbi.nlm.nih.gov). Biotech firms must understand these intersecting regimes and shape their internal policies accordingly.
- **Organizations must operationalize governance.** Best practices (drawn from case studies) highlight the need for cross-disciplinary AI governance committees, clear SOPs, and ongoing audits (www.frontiersin.org) (www.axios.com). Firms should embed classification tags into data pipelines and automate enforcement (e.g. automatically encrypting all "PHI" fields). Education and a culture of risk-awareness are crucial.
- **Collaboration is key.** Examples like MELLODDY show that federated AI allows knowledge sharing without breaking data classification rules. Partnerships among companies, regulators, and technology providers can address common governance challenges.
- **Future vigilance is required.** As AI tech advances (including generative models and secure AI), policies and frameworks will need continuous updates. Biotech stakeholders must stay engaged with new standards (e.g. updates to NIST, ISO, or clinical data standards) and participate in policy dialogues.

In summary, "AI policies and data classification frameworks for clinical-stage biotech" combine technical, legal, and ethical dimensions. By integrating robust classification of sensitive data with clear AI governance policies, organizations can foster innovation in drug development while safeguarding patient rights and data integrity.

The analysis above provides a deep, evidence-based guide for biotech executives, regulators, and practitioners navigating this complex landscape.

**References:**

- Pharma-industry data classification best practices (www.immuta.com)
- Regulatory perspectives on AI in drug development (www.mdpi.com)
- GDPR health data categories and focus on data nature (www.ncbi.nlm.nih.gov)
- NIST on data-centric security management (csrc.nist.gov)
- Egyptian health data classification framework (pmc.ncbi.nlm.nih.gov)
- Survey of AI adoption in life sciences (www.axios.com)
- Formation Bio's description of AI-assisted trial execution (Formation Bio technology)
- AstraZeneca AI governance lessons (www.frontiersin.org)
- Recent OECD/WHO discussions on AI in health (www.oecd-ilibrary.org) (www.hoganlovells.com)
- MELLODDY federated learning outcomes (pmc.ncbi.nlm.nih.gov)
- FDA/CTTI AI workshop key points (www.mdpi.com)
- Hexaware AI document classification case (hexaware.com)
- STATnews on Nucleai patient matching AI (www.statnews.com)
- Axios biotech AI summit notes (www.axios.com)

## Source links

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

1. https://www.immuta.com/blog/how-to-build-a-data-classification-framework-for-pharma
2. https://www.mdpi.com/1424-8247/18/1/47?type=check_update&version=1
3. https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.1068361/full
4. https://pmc.ncbi.nlm.nih.gov/articles/PMC11005050
5. https://www.statnews.com/2024/04/03/ai-biopharma-companies-clinical-trials
6. https://hexaware.com/case-study/intelligent-document-classification-of-clinical-trial-records-for-a-us-based-healthcare-data-company
7. https://www.arnoldporter.com/en/perspectives/topics/the-convergence-of-life-sciences-and-artificial-intelligence
8. https://www.axios.com/2024/11/14/life-sciences-ai-concerns
9. https://time.com/7372610/ai-drug-clinical-trials
10. https://www.ncbi.nlm.nih.gov/books/NBK613196
11. https://www.formation.bio/our-model
12. https://www.biospace.com/opinion-small-biopharma-is-bringing-ai-efficiency-to-clinical-trials
13. https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html
14. https://www.hhs.gov/hipaa/for-professionals/faq/is-a-business-associate-contract-required-for-a-covered-entity-to-information-to-a-researcher/index.html
15. https://www.hhs.gov/hipaa/for-professionals/special-topics/research/index.html

16. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological
17. https://content.govdelivery.com/accounts/USFDA/bulletins/4045c84
18. https://pmc.ncbi.nlm.nih.gov/articles/PMC12356831
19. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
20. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-6
21. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-2
22. https://pmc.ncbi.nlm.nih.gov/articles/PMC12611462
23. https://csrc.nist.gov/pubs/pd/2021/07/22/data-classification-practices-datacentric-security/final
24. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-10
25. https://www.melloddy.eu/new-research-consortium
26. https://www.formation.bio
27. https://www.axios.com/2025/11/21/axios-bfd-biotech-ai-bayer
28. https://www.arnoldporter.com/en/perspectives/news/2024/11/ai-is-transforming-life-sciences-but-raising-risk-concerns-new-benchmark-report-finds
29. https://www.arnoldporter.com/-/media/files/perspectives/publications/2024/11/report_the-convergence-of-life-sciences-and-artificial-intelligence_november-2024.pdf?hash=143DAAB41B6103F0C0992A8693856871&rev=bd08472ad0a641ba883d96278f84bc70
30. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
31. https://hl7.org/fhir/security-labels.html
32. https://www.oecd-ilibrary.org/en/publications/ai-in-health_2f709270-en.html
33. https://www.hoganlovells.com/en/publications/implications-of-the-eu-ai-act-on-medtech-companies_1
34. https://www.fca.org.uk/publications/research-articles/exploring-synthetic-data-validation-privacy-utility-fidelity
35. https://www.formation.bio/technology

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.