THE GOVERNANCE OF CLINICAL TRIAL DATA

Frank Fagan*

ABSTRACT

Clinical trial results are among the most authoritative inputs in biomedical research and are increasingly reused as training, calibration, and evaluation data for Artificial Intelligence (AI) systems. Yet a large share of completed trials, especially small academic studies, pilot projects, and early-phase experiments, never disclose their findings. These omissions distort meta-analyses, hide negative results, and silently propagate upstream bias into (AI) models. Existing enforcement tools, including Food and Drug Administration (FDA) civil penalties and National Institutes of Health (NIH) grant sanctions, work tolerably well for high-stakes commercial trials when used, but fail across the long-tail, where investigators lack the resources or incentives to upload results once grant funds expire.

This Article reframes clinical-trial transparency as a training-data governance problem and proposes a complementary solution: modest fixed cost-offsets for compliant uploads, funded through a two-tier access model. Public summaries on ClinicalTrials.gov would remain free, while a licensed API would provide harmonized, machine-readable datasets to high-volume users such as insurers, pharmaceutical consortia, and AI developers. Aligning the distribution of costs and benefits makes transparency feasible for resource-constrained investigators and sustainable for the analytic institutions that depend on complete, unbiased data. The result is a light-touch institutional architecture that preserves fragile trial evidence and strengthens the reliability of both biomedical research and AI health systems.

* Professor of Law, South Texas College of Law Houston. Email: ffagan@stcl.edu. Comments welcome.

INTRODUCTION

AI in health care is advancing at a remarkable pace, magnifying the importance of every upstream evidentiary input. From drug discovery pipelines to diagnostic decision support and clinical risk stratification, machine learning systems promise to reshape medicine. But these systems are only as reliable as the data they ingest. Clinical trial results, long considered the gold standard of biomedical evidence, and increasingly used as training and calibration data in AI health systems, are a crucial input for AI-driven health tools. Yet clinical trial transparency remains incomplete and deeply biased. Between one-third and one-half of completed trials never disclose results, and those that are reported disproportionately emphasize positive findings.[1] The result is a skewed evidence base, distorted meta-analyses, and silent erasure of negative or inconclusive results.

This compliance gap is not new. For decades, researchers and policymakers have lamented the selective reporting of clinical trial outcomes.[2] Congress responded by enacting the FDA Amendments Act of 2007 (FDAAA), which requires registration and results reporting for many clinical trials.[3] The NIH followed with its own 2017 policy extending disclosure obligations to all NIH-funded trials.[4] These mandates carry

1. *See generally* Colby J. Vorland et al., *Publication of Results of Registered Trials with Published Study Protocols, 2011-2022*, 7 JAMA NETWORK OPEN (2024), https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813519#249418864 [https://perma.cc/DA7S-GCZV] (finding that 36% of registered clinical trials with published study protocols did not report their main results); Nicholas J. DeVito, Seb Bacon & Ben Goldacre, *Compliance with Legal Requirement to Report Clinical Trial Results on ClinicalTrial.gov: A Cohort Study*, 395 THE LANCET 361, 361 (2020), https://www.thelancet.com/article/S0140-6736(19)33220-9/fulltext [https://perma.cc/2PDN-8Z9W] (finding that only 40.9% of trials reported within one year and that only 63.8% of trials had reported at any time); Mayookha Mitra-Majumdar & Aaron S. Kesselheim, *Reporting Bias in Clinical Trials: Progress Toward Transparency and Next Steps*, 19 PLOS MED. (2022), https://pmc.ncbi.nlm.nih.gov/articles/PMC8769309/ [https://perma.cc/BFW2-LBXE] (citing Erick H. Turner et al., *Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy: Updated Comparisons and Meta-Analyses of Newer Versus Older Trials*, 19 PLOS MED. (2022), https://pubmed.ncbi.nlm.nih.gov/35045113/ [https://perma.cc/Q642-Y6W4]) (noting that despite an uptick in reporting following the passage of the FDAAA, bias toward publishing positive and statistically significant results still persists at high levels).

2. *See, e.g.*, Sally Hopewell et. al., *Publication Bias in Clinical Trials Due to Statistical Significance or Direction of Trial Results*, 2009 COCHRANE DATABASE SYSTEMATIC REVS., (2009) (noting persistent underreporting of negative trials in popular medical research repositories since 1950).

3. 42 C.F.R. § 11.22 (2025) (describing the various device and drug clinical trials that must be reported).

4. NAT'L INSTS. OF HEALTH, NOT-OD-16-149, NIH POLICY ON THE DISSEMINATION OF NIH-

significant penalties: the FDA can impose daily fines of up to $10,000 for failure to report when required, while the NIH can withhold grant funding from noncompliant investigators.[5] In practice, however, agencies rarely apply these sanctions.[6] Enforcement actions are practically nonexistent, and studies consistently show that compliance with disclosure rules has plateaued at unsatisfactory levels.[7]

The persistence of noncompliance reflects two distinct dynamics. At one end of the spectrum are high-value, late-phase "blockbuster" trials for drugs with billions of dollars at stake. In these cases, the reputational and strategic incentives to conceal negative results are so powerful that only robust enforcement in the form of substantial fines, grant clawbacks, or loss of valuable exclusivity periods, can alter behavior. At the other end lies the "long-tail" of clinical research: small academic trials, pilot studies, investigator-initiated projects, and early-phase experiments that fail to show efficacy. Here, noncompliance is less a matter of strategy than of inertia. Once grant funds are exhausted and staff disperse, the costs of preparing, formatting, and uploading results, even if they amount to just thousands of dollars, become prohibitive.[8] For the long-tail of clinical research, the

---

FUNDED CLINICAL TRIAL INFORMATION (2016) (establishing the "expectation that all NIH-funded awardees and investigators conducting clinical trials, funded in whole or in part by the NIH, will ensure that their NIH-funded clinical trials are registered at, and that summary results information is submitted to, ClinicalTrials.gov for public posting."). The policy went into effect January 18, 2017. *See id.*

5. 21 U.S.C § 801(2). The NIH policy notes that noncompliance may lead to actions described in 42 C.F.R. 11.66, which notes, "If it is not verified that the required registration and results clinical trial information for each applicable clinical trial for which a grantee is the responsible party has been submitted, any remaining funding for a grant or funding for a future grant to such grantee will not be released." *See* 42 C.F.R. § 11.66 (2026).

6. *See generally* Joan Stephenson, *In a First, FDA Warns Company to Remedy Failure to Post Clinical Trial Results*, 2 JAMA HEALTH FORUM (2021) (noting that despite enactment of the FAAA in 2007, the FDA did not issue a single compliance order until 2021 and that such compliance orders remain rare).

7. As of August 2024, the FDA has issued only five notices of noncompliance. *Clinic Succeeds in Pushing FDA on Clinical Trial Transparency*, COLUM. L. SCH., https://www.law.columbia.edu/news/archive/clinic-succeeds-pushing-fda-clinical-trial-transparency [https://perma.cc/D7HJ-NFSY] (last visited Sep. 18, 2025). As noted above, only one-half to two-thirds of clinical trials are reported. DeVito, Bacon & Goldacre, *supra*, note 1.

8. Monique L. Anderson et al., *Compliance with Results Reporting at ClinicalTrials.gov*, 372 NEW ENG. J. MED. 1031, 1037 (2015) (reporting that preparation of result summaries takes 4–60 hours). Johns Hopkins' Institutional Review Board notes that ClinicalTrials.gov estimates that 40 hours are needed for entering results. Institutional Review Board, *Registration of Clinical Trials on ClinicalTrials.gov*, JOHNS HOPKINS MED. (Dec. 2019), https://www.hopkinsmedicine.org/institutional-review-board/guidelines-policies/guidelines/clinical-trials [https://perma.cc/CSZ5-FJXM]. Freelance medical writers currently charge approximately $60–80 per hour. *See Top Freelance Medical Writers*

reputational stakes of publicizing unsuccessful trials are minor, but the administrative burdens are more cumbersome.

Today, the consequences of missing trial results extend beyond traditional evidence synthesis. In an earlier era, this opacity mainly undermined the integrity of meta-analyses and clinical practice guidelines.[9] Now, in the age of AI, the stakes are higher still. When biomedical data serve as training inputs for machine learning systems, publication bias and selective disclosure do not merely distort single studies, they propagate across models and persist in ways that are difficult to detect or correct.[10] AI systems that are trained on incomplete evidence will embed skewed priors that propagate across models and persist in ways that are difficult to detect or correct. This exaggerates benefits and downplays risks, and potentially generates systemic consequences for patients and future research paths.[11] If AI is to deliver on its promise in health care, its foundation must consist of adequately comprehensive and unbiased data. Yet current disclosure regimes fall far short of accomplishing this goal.

This Article suggests that new governance mechanisms are needed to address the compliance gap. Enforcement remains essential for blockbuster trials, where reputational and financial incentives to conceal are strongest. But for the long-tail of trials where cost and inertia are the dominant barriers, enforcement alone cannot solve the problem—a complementary tool is needed. This Article suggests the use of a contingent licensing system that pays modest cost-offsets, on the order of $3,000 to $5,000 per compliant upload, to make disclosure cost-neutral. Rather than treating reporting as an unfunded regulatory burden, this approach frames disclosure as a compensated contribution to a research commons.

The licensing model avoids the optics of "paying people to obey the law" by situating payments within a broader governance framework. Trial

---

for Hire, KOLBATREE, https://www.kolabtree.com/services/medical-writing [https://perma.cc/272U-9AXL] (last visited Sep. 19, 2025). For a forty-hour project, just the writing portion of the report costs $2,400–$3,200 at going rates. *See id.* Johns Hopkins Medicine estimates an additional $2,211 in administrative close-out fees. STANDARD COSTS AND FEES FOR SPONSORED CLINICAL TRIALS, JOHNS HOPKINS MED, 12 (Aug. 2023), https://www.hopkinsmedicine.org/-/media/research/documents/offices-policies/crss-standard-costs-and-fees-fy2024-v2-04-august-2023.pdf [https://perma.cc/D24W-TTFT].

9. Hopewell et al., *supra* note 2 (noting the persistent bias since 1950 and its impact on trustworthiness).

10. James L. Cross et al., *Bias in Medical AI: Implications for Clinical Decision-Making*, 3 PLOS DIGIT. HEALTH, (2024) (noting that bias compounds through the medical AI lifecycle).

11. *See, e.g.*, *id.* (noting that bias can diminish an AI health system's clinical utility).

sponsors and investigators would receive modest cost-offsets for uploading complete datasets. Downstream users such as insurers, pharmaceutical consortia, and AI developers, would help fund the system by paying for structured, harmonized, Application Programming Interface-enabled access (API) to aggregated trial data. Public summaries would remain free and accessible, as they are today, but advanced bulk access and harmonized datasets would generate revenue to sustain the pool. In effect, the system would redistribute resources from those who benefit most from trial data to those who bear the costs of disclosure, while simultaneously reinforcing transparency norms.

Concrete examples illustrate how modest offsets change the compliance calculus. Consider an R1 university principal investigator who has just completed a multi-million-dollar trial funded by the NIH. The trial fails, the grant expires, and staff scatter. Uploading results to ClinicalTrials.gov requires several days of additional Research Assistant (RA) time, but no funds remain. A $3,000 cost-offset would allow the PI to pay staff to complete the upload, preventing the data from vanishing. Or consider a small biotech firm testing a repurposed therapy for a rare disease. The trial proves inconclusive, investor interest wanes, and disclosure offers no commercial benefit. Yet a modest payment to offset compliance costs could bring those results into the commons, where they might prove valuable for future rare-disease research. Junior faculty running pilot studies and genomic researchers with expensive de-identification burdens face similar challenges. In each case, disclosure does not happen because it is costly and unrewarded, not because it is reputationally dangerous. Targeted cost-offsets can change the calculus.[12]

The proposal also connects to broader debates in law and policy. Licensing pools are familiar in intellectual property, from compulsory music licensing to collective rights organizations.[13] While exclusivity extensions and orphan drug incentives are long-standing statutory carrots in

12. Offsets may shift an investigator's preference from disfavoring reporting to mere indifference. In many cases, a modest affirmative benefit may also be necessary, such as a small bonus for timely reporting, institutional recognition incorporated into evaluation and promotion files, or compliance-based scoring boosts on future grant applications.

13. *See, e.g.*, Jane C. Ginsburg, *Fair Use for Free, or Permitted-But-Paid?*, 29 BERKELEY TECH. L.J. 1383, 1432–33 (2014) (cataloging various domains of compulsory licensing and collective rights regimes).

pharmaceutical regulation,[14] licensing pools have emerged in parallel as voluntary mechanisms to broaden access.[15] In the AI context, licensing is central to disputes over training data, fair use, and compulsory access to socially valuable but privately held content.[16] By framing clinical trial data disclosure through the lens of licensing, this Article links health law's transparency problem to AI governance debates, showing how a balanced regime of enforcement and incentives can sustain reliable inputs for machine learning.

Part I situates the compliance gap, explaining where and why data disappears. Part II presents the licensing model. Part III shows how the training data pool can be sustained through API-enabled cost-sharing. Part IV situates clinical trial data within the broader policy of training data governance.

I. EXISTING ENFORCEMENT TOOLS AND THEIR LIMITS

Enforcement is the backbone of modern disclosure law. In domains ranging from securities regulation to environmental reporting, Congress often embeds penalties to counteract strategic opacity and align private incentives with public needs. Clinical trial transparency follows this familiar pattern: both the FDAAA and the NIH trials policy rely on mandatory reporting requirements backed by financial sanctions. Yet in practice, these tools perform poorly. Their shortcomings are not accidental or episodic; they stem from structural economic features of the biomedical research environment. This Section develops that argument. It explains why, even assuming perfect enforcement authority, the present system cannot resolve the disclosure deficit and why the architecture of penalties fails to align with the incentive landscape of clinical research.

---

14. On the impact of exclusivity extensions, see, for example, Meyke A. Kang, *Incentivizing Lower Drug Prices Through Patent Extension*, 103 N.C. L. REV. 1245, 1245 (2025); on the impact of orphan drug incentives, see, for example, Robert Rogoyski, *The Orphan Drug Act and the Myth of the Exclusivity Incentive*, 7 COLUM. SCI. & TECH. L. REV. 1, 2–4 (2006) (describing the standard argument).

15. *See generally* Alberto Galasso & Mark Schankerman, *Licensing Life-Saving Drugs for Developing Countries: Evidence from the Medicines Patent Pool* (Nat'l Bureau of Econ. Rsch., Working Paper No. 28545, 2021) (documenting that inclusion of a patent in the Medicines Patent Pool produces an immediate and large increase in licensing by generics).

16. *See, e.g.*, Frank Fagan, *Training Data Governance*, N.Y.U. J. INTELL. PROP. & ENT. L. (forthcoming 2026) [hereinafter *Training Data Governance*]; Frank Fagan, *When Fair Use Fails*, Found. for Am. Innovation, Sep. 2025, at 1.

A. The Architecture of Enforcement

The statutory enforcement regime for clinical trial reporting rests on three pillars: (1) the FDA's daily monetary penalties for non-reporting, (2) the NIH's authority to suspend or withhold grant funding, and (3) reputational consequences that flow from regulatory noncompliance.[17] On paper, these tools are substantial. The FDA may impose civil monetary penalties that escalate daily for each violation. The NIH may classify noncompliant investigators as out of good standing, affecting their eligibility for future funding. Universities and research hospitals may discipline investigators who expose the institution to grant-risk through noncompliance. Together, these authorities create what Congress and the NIH envisioned as a credible threat.

From a law-and-economics perspective, such a regime operates by raising the cost of nondisclosure above the cost of compliance.[18] If the penalty is higher than the private gains from withholding or delaying results, rational actors should disclose. In a perfectly enforced system with homogeneous incentives, the logic is straightforward. But clinical research is not a homogeneous environment. Trials differ dramatically in scale, funding structure, commercial stakes, timelines, staffing, and in terms of the distribution of costs and benefits.[19] Because the statutory enforcement regime applies uniformly across this heterogeneous landscape, the result is predictable: the tools work tolerably well when incentives already point toward compliance, but collapse when incentives run in the opposite direction. Two distinctions are critical for understanding these limits: (1) the difference between high-stakes and low-stakes trials, and (2) the difference between strategic nondisclosure and cost-driven nondisclosure. Enforcement interacts very differently with each.

Enforcement performs best when nondisclosure is strategic. In settings where large commercial actors undertake expensive late-stage trials, the principal barrier to disclosure is not cost but strategic preference. Firms may prefer opacity when results undermine market expectations or weaken competitive standing. In such environments, penalties can operate as

17. *See supra* notes 4–6.
18. *See* A. Mitchell Polinsky & Steven Shavell, *The Economic Theory of Public Enforcement of Law* 38 J. ECON. LITERATURE 45, 46 (2000).
19. *See infra* §§ I.A–B.

meaningful deterrents because they are directed at actors who can absorb compliance costs and who are sensitive to regulatory pushback.[20] The underlying economics resemble insider-trading or securities-fraud enforcement: a concentrated set of large, sophisticated firms with clear legal obligations, ongoing interactions with regulators, substantial internal compliance capacity, and material exposure if they fail to comply. Enforcement is most effective where: (1) the responsible firm is readily identifiable and subject to oversight, (2) the relevant data lie within the firm's possession and control, (3) the consequences of noncompliance are predictable, and (4) the expected sanction outweighs the expected benefit of opacity. If the FDA imposed penalties consistently in the blockbuster setting, one would expect compliance to rise because commercial sponsors and their legal departments would treat reporting akin to any other regulatory requirement with real liability exposure. The point is not that blockbuster trials never generate nondisclosure; rather, it is that the on-the-books enforcement regime theoretically fits this domain. Where nondisclosure is chosen to maximize strategic advantage, penalties have the potential to alter the calculus even if, presently, they are rarely applied.

A stronger version of this claim becomes clear when one examines the settings in which enforcement actually does map onto the underlying incentives. This occurs most reliably in high-value late phase trials, where commercial and regulatory stakes converge.[21] These trials are typically run by large sponsors with dedicated compliance teams, centralized control over the data, and internal legal departments that closely track liability exposure.[22] Because these trials sit at the brink of FDA approval and future revenue, the expected surplus from strategic opacity and the expected cost of regulatory sanction are both concentrated in the same entity. That alignment matters. It ensures that a penalty, if imposed, lands on the actor that orchestrates and benefits from nondisclosure, and ensures that the sponsor has the organizational capacity to respond. In these late-stage environments, the economics resemble other settings where enforcement has real bite: a small number of sophisticated repeat players, predictable

20. *See* Polinksy & Shavell, *supra* note 18.

21. *Cf.* Zakariyya Mughal et al., *Sponsor-Level Compliance with ClinicalTrials.gov Reporting Requirements: A Comprehensive Analysis*, J. ACAD. PUB. HEALTH (Sep. 10, 2025) (reporting a higher compliance rate for industry versus academic sponsors).

22. *Id.*

obligations, and stakes large enough to make sanctions salient. The enforcement structure for blockbuster data disclosure remains conceptually sound, despite its infrequent use.

B. Enforcement in the Long-Tail

The conceptual logic reverses in the long-tail of clinical research. Here, nondisclosure does not arise from strategic concealment but from structural frictions embedded in the research production process. Enforcement tools falter—not because they are weakly drafted or sporadically imposed, but because they are misaligned with the underlying incentive structure.

i. Capacity-Constrained Compliance

Cost-based nondisclosure arises when various tasks required for compliance, such as data cleaning, table generation, module formatting, and quality control, must occur after the trial's funding has ended. At that stage, the marginal cost of compliance is high, whereas the marginal benefit to the investigator is minimal. When that investigator lacks funds to hire staff or devote uncompensated time to the upload, the legal threat of penalties does not supply the missing resources.[23]

In economic terms, enforcement cannot transform an infeasible action into a feasible one. Penalties are only effective when the regulated actor has the capacity to comply but chooses not to. When capacity has evaporated, penalties generate anxiety, not compliance. The investigator's decision problem is not "should I comply or risk sanctions?" but "I cannot comply given budget constraints; what is my exposure?" Under these conditions, increased enforcement may produce perverse effects by deterring investigators from initiating trials when downstream reporting obligations are perceived as administratively risky.

ii. Incentive-Constrained Compliance

Disclosure in the long-tail generates limited professional reward. It does not produce publication credit, usually does not strengthen grant prospects,

23. *See* Polinsky & Shavell, *supra* note 18, at 50 (noting that there will be underdeterrence if the non-compliant party's wealth is less than the fine).

and rarely enhances institutional prestige. The benefits are public. The costs are private. Penalties do nothing to reverse this distribution. They do not create positive incentives; they merely add negative ones. Consider the scenario from a principal-agent perspective. The NIH and FDA act as principals who want disclosure, but the agents (i.e., the thousands of individual investigators) do not internalize disclosure's value. Penalties can theoretically force alignment, but only if the agent can afford to comply. When the problem is under-incentivization and high marginal compliance costs, penalties do not induce effort; they merely raise the stakes of failure.

iii. Brittle Enforcement in a Decentralized Environment

Biomedical research is diffuse. There are thousands of university-based investigators, rotating project staff, hospital-based clinical research units, independent institutional review boards, contract research organizations, and small biotech firms.[24] Enforcement regimes struggle in such decentralized environments for at least three reasons:

- *Monitoring is costly*. Identifying noncompliant actors requires systematic scanning and verification of thousands of trial records.
- *Attribution is difficult*. Responsibility for disclosure can be ambiguous when teams disperse.
- *Targeting is imperfect.* Penalties are difficult to align because institutions and individual investigators bear responsibility at different stages of the trial.

The FDAAA regime treats clinical research as if it were functionally similar to securities markets. It expects a centralized and highly monitored environment that is amenable to targeted enforcement. In reality, the biomedical research environment resembles a federated system of thousands of small producers whose coordination depends heavily on localized administrative capacity. Although trials must be registered, registration does not make monitoring easy. Determining which trials are subject to results reporting, whether reporting deadlines have passed, and who bears responsibility for the upload remains a resource-intensive

24. *See* Mughal et al., *supra* note 21 (noting the diffuse landscape).

verification problem.

iv. Stunted Compliance Infrastructure

Sanctioning a university for its investigators' noncompliance generates a public goods problem. The cost of building compliance infrastructure must be borne centrally, but the benefits accrue primarily for public databases and downstream users such as AI developers, insurers, and researchers. Public institutions, therefore, have even weaker incentives to invest.[25]

v. Backward-Looking Penalties

Disclosure obligations arise only after a trial concludes, often years after the initial funding decision. Thus, enforcement functions on a severely delayed timeline. A private investigator who failed to anticipate the administrative demands of the upload stage cannot meaningfully change course once the grant is exhausted. Traditional regulatory logic that demands the possibility of future penalties to discipline present planning breaks down when planning requires future budgets that investigators do not control. In incentive terms, enforcement regimes assume that actors can internalize future sanctions ex-ante and allocate resources accordingly. But most academic trials operate under grant budgets that cannot be adjusted retroactively. The inability to reallocate funds undermines the forward-looking effect of penalties.

Beyond these structural incentive failures, the enforcement regime faces political-economy constraints inherent in the biomedical ecosystem. The FDA and NIH are relational regulators. Their effectiveness depends on cooperation with universities, hospitals, and scientific communities. Aggressive, high-profile enforcement against academic investigators risks political backlash and harms the agencies' relationships with the same institutions whose cooperation they need for other regulatory functions.[26]

25. *See id.* (noting that industry sponsors allocate greater resources to compliance administration because of repeat interactions).

26. *See, e.g.*, John T. Scholz, *Cooperative Regulatory Enforcement and the Politics of Administrative Effectiveness*, 85 AM. POL. SCI. REV. 115, 115 (1991) (developing a model in which political interests control bureaucratic outputs, but cooperation nevertheless increases compliance under narrow conditions).

Thus, even if enforcement tools were effective in theory, they are seldom deployed. Agencies face the classic problem of regulatory forbearance: sanctions exist but are politically costly to impose.

The failure of the enforcement regime follows from the interaction of these economic, institutional, and political features. The reason is not that enforcement is unnecessary. Indeed, in high-value commercial settings, it is essential. But enforcement without a deep commitment can only target strategic behavior, not structural inability or deeply misaligned incentives. Penalties to long-tail investigators do not supply capacity, create incentives, or build infrastructure. They merely punish failure. And where failure is systematic, predictable, and rooted in rational responses to post-grant cost structures, enforcement becomes a mismatched tool rather than a corrective one. In short, the problem is not insufficient severity of sanctions but a misaligned architecture of incentives. A reporting system that relies solely on the threat of penalty will remain incomplete so long as thousands of investigators face the same predictable cost-based barriers. The next Part therefore turns to the complementary tools required to fill these gaps. The aim of these tools is to provide capacity, correct incentive distortions, and generate a sustainable institutional framework for disclosure.

II. CONTINGENT LICENSING FOR THE LONG-TAIL

A complete transparency regime requires more than deterrence; it requires institutions that make disclosure practical, rewarded, and routine. Enforcement delineates the boundaries of lawful conduct, but it cannot on its own supply the resources, incentives, or administrative continuity that long-tail investigators systematically lack. What is needed is a complementary mechanism that aligns the costs borne by data producers with the benefits enjoyed by downstream users, and that treats disclosure not as an unfunded mandate but as a contributory act within a larger research infrastructure. A contingent licensing framework provides that alignment by channeling modest, targeted payments to those who generate trial data and by financing those payments through the parties who derive value from structured access to the resulting datasets.

A. The Licensing Model

A contingent licensing system begins from a simple premise: if disclosure is a public good, then the actors who benefit most from high-quality datasets should help defray the costs of producing them. Under this approach, a centralized pool would issue modest, fixed payments, on the order of $3,000 to $5,000, for each complete, timely upload of results to ClinicalTrials.gov. The amount is calibrated not to reward strategic behavior, but to neutralize the marginal costs that routinely block long-tail investigators from disclosing. Uploading a trial's results typically requires several days of research assistant time to clean data, populate structured tables, resolve validation errors, and coordinate quality review. In many academic settings, these tasks occur after grant funds expire, leaving investigators without the budgetary slack needed to complete the upload. A small, predictable offset allows those tasks to be completed without requiring investigators to subsidize compliance through uncompensated labor.

Framing the mechanism as a licensing system is crucial. The system does not "pay people to obey the law." Instead, it treats disclosure as a compensated contribution to a regulated data commons. When an investigator uploads a dataset, they are not merely satisfying a regulatory obligation; they are granting a license to place structured, reusable information into a shared repository that supports downstream research, evidence synthesis, and AI model development. The pool's payment functions as the consideration for that license. This framing is familiar from other data-governance domains, ranging from music licensing to collective rights organizations, where contributors are compensated for adding value to a resource that will be accessed and reused by many others.

The model also creates clear expectations and reduces administrative friction. Payments attach only to compliant uploads that meet objective criteria: complete datasets, adherence to ClinicalTrials.gov formatting, and resolution of error flags. Because the requirements are transparent and uniform, investigators know ex ante what is required, and institutions can plan accordingly. The predictable structure encourages universities to build light-touch administrative workflows: for example, designating a central staff member to assist with uploads once a trial concludes, knowing that the offset will cover their time. Over time, modest offsets can generate precisely

the kind of micro-infrastructure that enforcement alone has not produced, that is, a stabilizing layer of administrative capacity within decentralized research environments.

The model further avoids distortions that would arise from discretionary or negotiated payments. Offsets are fixed, not variable, and apply equally across therapeutic areas, institutions, and trial phases. This universality reduces strategic gaming and reinforces the idea that disclosure is part of the fabric of research, not an exceptional burden requiring special justification. The modest size of the payment also avoids incentive distortions. It is far too small to motivate the initiation of unnecessary trials, yet sufficient to neutralize the marginal costs that routinely block disclosure.

Equally important, the contingent licensing framework creates a direct economic link between data producers and data users. Trial sponsors and investigators contribute structured results; insurers, pharmaceutical consortia, and AI developers receive the benefit of harmonized, API-enabled access. The payment from the pool is thus the connective mechanism that ensures both sides of the market remain aligned. Where enforcement-based approaches assume that penalties will redirect incentives, the licensing model creates the incentives directly, allowing long-tail investigators to complete reporting obligations without bearing uncompensated costs, and allowing downstream users to access higher-quality datasets that improve the reliability of their analytical tools.

In short, the licensing model reframes disclosure from a compliance problem into a participation problem. By modestly compensating the act of adding structured data to a shared infrastructure, it transforms reporting from an afterthought into an ordinary step in the research lifecycle. Licensing transforms this step into one that researchers can plan for, budget around, and complete without friction. While the mechanism cannot replace enforcement in high-stakes settings, it does provide the missing architecture for the long-tail of clinical research, where capacity constraints and weak incentives are the central barriers to transparency.

B. Funding and Feasibility

Financing the licensing pool requires two elements: a stable public baseline and a sustainable mechanism for private cost recovery. Both are well within reach. Even under the most conservative assumption, in which

the system pays offsets for every applicable clinical trial each year, the total annual cost would be approximately $22 million to $37 million, based on the U.S. Department of Health and Human Services' (HHS) estimate that roughly 7,400 studies qualify as "applicable clinical trials" under the FDAAA.[27] As the Final Rule explains, about 5,150 of these are trials conducted under an Investigational New Drug application or Investigational Device Exemption (IND/IDE), and roughly 2,250 fall outside those pathways.[28]

This regulatory distinction maps neatly onto the framework developed earlier in the Article. IND/IDE trials, even when sponsored by academic investigators, are embedded in the commercial drug and device development pipeline and thus serve as a practical proxy for the high-stakes, enforcement-responsive "blockbuster" domain described earlier. The remaining non-IND/IDE studies can serve as a proxy for the long-tail: the small academic trials, pilot studies, investigator-initiated projects, early-phase experiments, and resource-constrained interventions identified in the Introduction as the core locus of cost-driven nondisclosure. Because only this latter subset requires structural support to achieve compliance, the number of trials needing offsets is likely to be substantially smaller than the statutory maximum. A pool sized to fund 1,000 to 2,000 uploads per year would require on the order of $3 million to $10 million, a negligible share of the NIH's budget and well within the scale of existing federal investments in research infrastructure.

Private contributions complement this foundation because the parties who benefit most from structured access to trial results are not investigators but downstream users. AI developers rely on harmonized trial data to train and calibrate diagnostic models, to build adverse-event prediction systems, and to benchmark model outputs against clinically validated evidence. These developers already pay substantial amounts for curated datasets, often millions of dollars for processed Electronic Health Record corpora, and would be willing to pay licensing fees for a unified API that delivers structured, machine-readable trial results. For a laboratory building a risk-stratification model, for example, the value of a single harmonized endpoint dataset may far exceed the modest fee required to sustain the licensing pool.

27. *See* Clinical Trials Registration and Results Information Submission, 81 Fed. Reg. 64982, 65125 (Sep. 21, 2016) (to be codified at 42 C.F.R pt. 11) (estimating 7,400 applicable clinical trials).

28. *Id.*

Similarly, insurers conducting health-technology assessments routinely subscribe to commercial evidence platforms to support coverage decisions. A licensing arrangement that provides standardized, API-enabled access to trial outcomes would reduce duplicative contracts and integrate evidence streams that insurers already purchase in fragmented form. Pharmaceutical consortia also stand to gain. Many firms rely on pooled registries to benchmark pipelines, evaluate competitor activity, or identify trial designs suitable for repurposing; a harmonized trial-level API reduces the transaction costs and data-cleaning burdens that currently impede such efforts. In each case, a predictable licensing fee represents a small fraction of the value these actors derive from structured data access.

This hybrid model of public baseline funding, supplemented by private licensing revenue, yields a self-replenishing pool. Public funds ensure the pool's continuity and signal a federal commitment to transparency. Private licensing fees scale naturally with demand for high-quality, harmonized data. As AI developers adopt trial-level datasets for calibration, bias detection, and safety evaluations, their usage generates the revenue that finances future uploads. As insurers integrate trial APIs into cost-effectiveness models or formulary decisions, they contribute directly to sustaining the infrastructure from which they benefit. And as pharmaceutical consortia expand portfolio-level analytics, their access fees help support the upstream activities that make those analytics possible. Because the marginal cost of uploading a completed trial is relatively fixed, and primarily based upon research-assistant time for data cleaning, formatting, and validation, the pool creates a stable, predictable institutional environment in which disclosure becomes routine.

The modest size of the offset also avoids distorting upstream behavior. A $3,000 payment is too small to induce unnecessary trials or meaningful changes in research strategy, yet large enough to remove the cost barriers that prevent compliance. For many genomic or biobank studies, for instance, the most significant reporting expense is Institutional Review Board-mandated de-identification; a small offset can cover this cost entirely. Small device or digital therapeutics firms, which commonly operate with lean administrative staff, similarly benefit when reporting obligations are no longer unfunded mandates. By neutralizing the marginal cost of compliance without altering the incentive to initiate research, the licensing pool targets precisely the friction point that enforcement fails to

address.

Seen in this light, the licensing pool is less a subsidy than an institutional mechanism for aligning the distribution of costs and benefits. Data producers, such as investigators, academic units, and small sponsors, supply structured trial results. Data users, such as insurers, pharmaceutical consortia, and AI developers, derive ongoing value from aggregated, high-quality access. The pool's payments and licensing fees form the connective tissue that coordinates these actors within a sustainable, transparent system. Enforcement remains essential in the high-stakes domain of late-phase commercial trials, but for the thousands of low-resource studies that form the backbone of biomedical evidence, a hybrid funding and licensing structure offers the missing institutional architecture that makes disclosure feasible and predictable.

III. API-BASED SUSTAINABILITY

A. A Two-Tiered Model for Evidence Access

A sustainable transparency system requires not only a mechanism to neutralize the costs of disclosure, but also an architecture for distributing the resulting data in a way that preserves public access while enabling cost recovery from high-volume users. The first tier preserves what exists today: free public access to summary results through ClinicalTrials.gov. These summaries, which take the form of concise tables, adverse-event counts, and high-level outcome measures, provide broad social value at minimal cost and function effectively for general public use, replication checks, and routine scientific reference. But they do not supply the volume, structure, or machine-readability required for downstream analytical uses, particularly in AI development.

The second tier provides an additional layer of governance. Here, harmonized datasets are made available through an API that exposes structured endpoints across trials, therapeutic areas, and time. This tier is designed for high-volume users such as pharmaceutical portfolio analysts, insurers conducting health technology assessment, and AI developers building risk models, who already pay substantial amounts for proprietary, curated datasets of far lower quality. Unlike the public summaries, the API would not merely replicate existing trial tables; it would additionally supply

normalized variable names, standardized outcome measures, consistent adverse-event taxonomies, and machine-readable metadata. These value-added features convert individual uploads into a usable research infrastructure.

This architecture parallels the tiered access systems that have emerged in other areas of data governance. Collective rights organizations routinely maintain a low-cost baseline for public use while charging for high-volume or value-added services.[29] In earlier work on training data governance, I argued that tiered licensing that divides free access for baseline content, and paid access for content that is at risk of exiting the commons, offers a way to preserve openness while addressing the economic realities of dataset creation.[30] Clinical trial evidence fits the same pattern. Summary results form a low-cost public baseline; harmonized, API-enabled datasets provide the value-added infrastructure that sophisticated users require and are willing to support. The licensing pool becomes fiscally sustainable because its funding sources map directly onto the distribution of benefits: investigators receive support to disclose, and analytic institutions pay for the structure that makes the resulting data usable at scale.

B. Funding Flows and Institutional Alignment

The two-tier structure also clarifies how resources move through the system. The core transaction is straightforward. When an investigator uploads a complete dataset that satisfies objective formatting and validation requirements, the pool issues a modest, fixed payment. The dataset then becomes part of the registry's structured layer, where it can be queried, aggregated, and linked through an API. High-volume users access this structured layer through a licensed endpoint and pay a predictable fee for bulk or value-added use.

This flow aligns costs with benefits in a way the current system does not. Investigators and small sponsors bear nearly all of the cost of making data usable, but capture little of the downstream value. At the same time, the actors who derive the greatest returns from structured access, i.e., firms

29. *See* Stanley M. Besen, Sheila N. Kirby & Steven C. Salop, *An Economic Analysis of Copyright Collectives*, 78 VA. L. REV. 383, 383–84 (1992) (discussing purpose of centralized copyright collectives).

30. *See Training Data Governance*, *supra* note 16.

building evidence pipelines, insurers modeling clinical effectiveness, and AI developers constructing diagnostic or triage systems, use the public system intensively without contributing to its upkeep. A licensing-based revenue stream reverses this mismatch. Payments flow toward the point of production, neutralizing the marginal reporting costs that block long-tail disclosure, while users who extract concentrated private value fund the infrastructure that makes that extraction possible.

The model also improves institutional planning. Universities and small sponsors can anticipate that compliant uploads will generate a defined offset, allowing them to assign staff to data cleaning and validation without relying on residual grant funds or uncompensated labor. As noted earlier, this work can be centralized within a single person or a small team who become skilled in formatting and uploading results, producing scale efficiencies that decentralized efforts cannot. On the user side, insurers and AI developers gain a stable, well-documented API that replaces the ad hoc data-wrangling they currently perform across heterogeneous trial reports. Because licensing fees attach to a standardized endpoint rather than idiosyncratic datasets, firms can integrate the API directly into evidence pipelines, risk models, and postmarket surveillance tools and incorporate the associated costs into ordinary budgeting and long-term planning.

Importantly, the pool becomes self-replenishing rather than dependent on continuous appropriations. Public baseline funding ensures the pool's availability and signals federal support for a functional disclosure infrastructure. Private API-based licensing fees scale naturally with demand for structured data. As AI developers incorporate trial-level endpoints into model calibration and bias detection, their usage helps finance future uploads. As insurers integrate standardized outcomes into coverage decisions or formulary design, their access fees sustain the infrastructure that lowers their own information costs. And as pharmaceutical consortia increasingly rely on harmonized trial registries to benchmark pipelines and study designs, they contribute directly to the system that produces those efficiencies.

By building a feedback loop between data producers and data users, the funding flow turns what is currently an unfunded mandate into a self-supporting evidence system. Enforcement still plays a role for blockbuster trials, but for the long-tail, routine disclosure becomes financially feasible and institutionally predictable.

C. The International Dimension

The proposed architecture also interacts naturally with existing international transparency regimes. The European Union's (EU) Clinical Trials Regulation[31] imposes results-reporting requirements broadly similar to the FDAAA, including structured tables and standardized outcome fields. The World Health Organization's International Clinical Trials Registry Platform (ICTRP) aggregates registration data from more than a dozen national registries and already functions as a federated search layer. Neither system, however, provides harmonized, machine-readable results capable of supporting large-scale evidence synthesis or AI development. Their public interfaces mirror ClinicalTrials.gov's current model: adequate for general reference but not optimized for advanced analytical use.

A two-tier structure provides a pathway, rather than a mandate, toward greater interoperability. If ClinicalTrials.gov exposes normalized endpoints through an API, foreign regulators and registries can adopt comparable output formats at minimal cost. The goal is not to create a unified global registry but to reduce the translation costs created by heterogeneous reporting schemas. Registries that adopt compatible formats would allow multi-site analyses, cross-jurisdictional meta-analyses, and model validation exercises that currently require substantial manual cleaning. For AI developers, whose models increasingly rely on international datasets, even partial alignment lowers the cost of incorporating foreign trial results into training or calibration workflows.

The appeal is functional rather than normative. The EU's Clinical Trial Regulation, ICTRP, and similar systems already generate structured data; the missing component is a value-added layer that exposes that structure in a consistent, machine-readable form. An API-enabled tier in the United States would demonstrate the feasibility of such an approach, and a modest amount of voluntary coordination could extend those efficiencies across jurisdictions. Nothing in the licensing model depends on international adoption, but the potential gains from reduced duplication, lower cleaning costs, and improved cross-registry comparability are significant for any actor working with global evidence streams.

31. Regulation (EU) No. 536/2014, 2014 O.J. (L 158) 1.

IV. BROADER GOVERNANCE IMPLICATIONS

A. Clinical Trials as Training Data Governance

Artificial intelligence recasts clinical-trial disclosure as a training-data governance problem. Trial results form a small but uniquely authoritative subset of the biomedical evidence space: they are endpoint-based, adjudicated, and methodologically comparable across studies. When these inputs are missing, downstream distortions propagate through meta-analyses, safety assessments, and model calibration. The resulting errors are not localized; they embed into pipelines that amplify the influence of upstream data conditions.

This dynamic parallels a more general challenge in maintaining the supply of high-quality training data. As developed in related work, many socially valuable inputs to AI systems occupy an at-risk position: they offer high downstream value but low private return for their creators.[32] The result is predictable undersupply. Markets cannot negotiate thousands of individualized transfers with dispersed, low-resource producers, and the absence of predictable institutional support erodes exactly the inputs that produce the greatest social benefit.

Long-tail clinical trials occupy the same structural niche. Their evidentiary value exceeds their commercial or reputational payoff, and their producers operate with the same limited administrative slack as other at-risk contributors to knowledge ecosystems. The licensing pool proposed here performs the same governance function that fallback licensing performs in training-data markets: it preserves fragile, socially valuable inputs by aligning the costs borne by producers with the concentrated benefits captured by downstream users. The two-tier data architecture, with its free baseline summaries and structured paid API layer, mirrors the tiered-access frameworks that have successfully stabilized other digital commons.[33]

In this sense, clinical-trial transparency is not an isolated disclosure problem; it is an instance of a broader institutional design challenge: how to

32. *See Training Data Governance*, *supra* note 16 (characterizing at-risk creators as producers whose willingness to generate high-value training data erodes when their outputs are repurposed by LLMs without any corresponding support for the costs of continued production).

33. *See id.* (proposing a tiered licensing framework that compensates value-added content producers who are demonstrably at risk of exit, while withholding payment from baseline producers whose participation is not threatened).

sustain data inputs whose social value exceeds their private incentives. The contingent licensing mechanism gives clinical-trial data the same governance architecture that other high-value, low-incentive data classes require to remain stable in an AI-driven environment.

B. Normative Justifications

The normative case for contingent licensing rests on standard welfare principles rather than moral appeals. Disclosure produces positive externalities for patients, clinicians, regulators, and developers. Patients benefit from more accurate assessments of safety and efficacy; clinicians and regulators gain a more complete evidentiary base; and developers benefit from training, calibrating, and auditing AI systems on unbiased inputs. Yet the private incentives of long-tail investigators do not capture these gains. The result is predictable underproduction of disclosures relative to the social optimum.

The licensing pool corrects this imbalance without distorting the upstream research process. The offset is too small to influence trial initiation, and the licensing fees fall only on actors whose private benefit from structured access easily exceeds the marginal cost. The mechanism therefore satisfies the core criterion for a justified intervention: it addresses a market failure in a targeted manner while leaving functioning markets, such as high-stakes commercial trials, undisturbed.[34]

Equally important, the model reframes trial disclosure as a contribution to a shared research commons rather than as an unfunded regulatory burden. The public retains free access to summary results, while high-volume users support the structured layer that they uniquely rely upon. The pool thus creates a stable, repeatable institutional environment in which essential data can be produced, maintained, and reused across both biomedical and AI applications. In an era where AI systems magnify the consequences of missing or biased inputs, the public value of this stability is considerable.

34. *See generally id.* As mentioned earlier, the blockbuster segment is, in a meaningful sense, a functioning market: private incentives and compliance capacity are already strong, and nondisclosure arises from strategic considerations rather than structural barriers. In this domain, penalties, and not offsets, are the appropriate tool.

CONCLUSION

Clinical-trial disclosure has long been treated as a compliance problem, yet the persistence of missing results reflects something more structural: a research ecosystem in which the costs of transparency fall on the actors least able to bear them, while the benefits flow to institutions and analytical systems far downstream. Enforcement remains necessary in the high-stakes commercial domain, where strategic nondisclosure can distort markets and patient expectations, but deterrence alone cannot resolve the predictable, capacity-driven failures that characterize the long-tail of academic and early-phase research. As AI systems increasingly rely on clinical evidence for training, calibration, and safety evaluation, the consequences of these gaps become more significant and durable.

A contingent licensing model offers a pragmatic, institutionally modest way to close this gap. By neutralizing the marginal costs of disclosure and financing those offsets through the users who benefit most from structured access, the mechanism aligns incentives without altering the underlying research pipeline. Free public summaries remain, while harmonized and API-enabled data provide a sustainable funding stream and a stable evidentiary foundation for AI development. The result is a light-touch governance layer that preserves fragile inputs and strengthens the evidentiary infrastructure on which both biomedical research and AI health systems increasingly depend.