82 percent. AI regulation in health insurance and the variable it refuses to measure.

From a Stanford study published in January 2026, and what it reveals about a system self-stabilized by unobservability.

One system may regulate by a threshold of algorithmic quality. Another may regulate by a threshold of administrative exhaustion. These two regulations are not of the same kind, and the American health insurance system has, in effect, ended up operating on the second mode.

A study by Mello, Trotsyuk, Djiberou Mahamadou and Char published in *Health Affairs* in January 2026, *The AI Arms Race in Health Insurance Utilization Review*, delivers a figure that regulatory doctrine will have to digest: 81 percent of prior authorization denials issued in Medicare Advantage plans and appealed by beneficiaries are overturned upon review. The 2024 data published by KFF from CMS reporting give 80.7 percent of appeals partially or fully overturned out of 4.1 million denials. The order of magnitude is stable: roughly four appeals out of five result in a partial or total reversal. This number is not a punctual defect. It is an operating point.

The figure, in context

Before any interpretation, one must establish the order of magnitude, and resist the temptation to aggregate measures that do not bear on the same cohorts.

In 2024, Medicare Advantage insurers issued 52.8 million prior authorization decisions. 4.1 million were denied (7.7 percent). Of those denials, 11.5 percent were appealed, and 80.7 percent of those appeals were partially or fully overturned (KFF, from CMS data, January 2026). This is the rigorous aggregate datum.

The complaint *Estate of Lokken v. UnitedHealth Group* alleges, in the post-acute care segment managed by the nH Predict model, an appeal rate of approximately 0.2 percent only. This figure is plaintiff allegation, not aggregate finding. It is consistent with the underlying mechanics: the more vulnerable the cohort (elderly patient, care with a narrow clinical window, competing mortality), the lower the appeal rate collapses. The two figures do not measure the same thing; they describe the same structure at two resolutions.

Out of 4.1 million denials, 3.63 million are never contested. How many of them would have been overturned if they had been? The system does not measure it, and the central argument of this article is that this unobservability is not a methodological accident. It is a structural property of the apparatus.

Mechanics and components of friction

For a beneficiary to move from denial to reversal, a documented chain of obstacles must be cleared. Seven days now, since the CMS-0057-F rule effective January 1, 2026 for impacted payers (Medicare Advantage, Medicaid managed care, CHIP, and QHP plans on federal marketplaces), for the first standard decision; 72 hours for expedited cases. The rule also imposes specific reasoning of denials. Several associated API obligations have distinct deadlines, notably 2027 for the operational implementation of the Prior Authorization APIs. Retrieval of the medical record, letter from the treating physician, appeal form, exchange with the appeals department. In case of denial of the first-level appeal: second-level, hearing before an Independent Review Entity, Medicare Appeals Council, judicial review in federal court.

For post-acute care (skilled nursing facility care, rehabilitation), the appeal timeframe often exceeds the relevant clinical window. The appeal arrives after the need has ceased to exist.

This friction breaks down into three components that are useful to distinguish, without prejudging their relative weight:

1. **Administrative friction**: procedural delays, forms, levels of appeal, deadlines, submission formats;
2. **Informational friction**: fragmentation of the clinical record, cost of retrieval, translation into the grammar of utilization review, certification of the care trajectory;
3. **Cognitive friction**: understanding of the right of appeal, capacity to mobilize the prescriber, beneficiary exhaustion, asymmetry of literacy in the face of the denial letter.

The relative share of these three components is not measured to date. Reducing one does not mechanically reduce the others: a perfectly orchestrated record does not erase the legal complexity of appeal, nor the exhaustion of an elderly beneficiary in post-acute care. The distinction matters because it conditions the scope of any regulatory intervention.

On the provider side, the picture is complementary. According to the AMA, physicians spend on average 12 hours per week on prior authorization requests; 94 percent report associated care delays; 78 percent indicate that patients abandon a treatment in the face of the cost of waiting. At the national level, providers spend around 19.7 billion dollars per year litigating denials.

The thesis, and a doctrinal refinement that strengthens it

AI regulation in American health insurance does not operate through guarantee of algorithmic quality. It operates through friction of use. The regulatory compliance of the system is compatible with a high rate of initial decisions unsupported after challenge, as long as the cost of appeal for the beneficiary remains prohibitive.

A doctrinal refinement is required here. The term *calibration* suggests a centralized intent, an actor explicitly tuning the cost of appeal to a target mobilization threshold. That assertion would be rhetorically stronger and empirically weaker. The available elements (the parametric *dial* reported by former EviCore employees, risk contracts indexed on spending reduction, the 3-to-1 commercial pitch) converge toward a distributed industry optimization, not toward a central command.

The defensible position is finer, and harder to attack: the system does not need to be designed for non-recourse in order to operate by it. An emergent optimization is harder to regulate than a centralized intent, precisely because it has no identifiable author. And a system without an author has no designated regulatory respondent either.

The system is therefore *structurally equivalent* to a calibration. It functions *as if* calibrated, without it being necessary to prove central intent. This nuance does not weaken the thesis. It displaces it: what is regulable is not intent, but the structural property.

Unobservability as a political variable

An observable error rate is regulable. A non-observable error rate is not.

The current system does not measure how many uncontested denials would have been reversible after challenge. It does not measure it *by construction*: non-contestation is precisely the event that prevents measurement. The regulatory consequence is radical.

A quality standard can apply only to what allows itself to be observed. It therefore applies only to contested denials, which is 11.5 percent of the total in the aggregate cohort, and within that, the segment biased by selection toward cases where contestation appeared viable. The uncontested segment, mathematically the larger one, is regulatorily inaccessible. This is not a hole in regulation. It is the blind spot the system maintains as the condition of its own stability.

This property makes the system self-stabilized. The higher the cost of challenge, the wider the uncontested segment. The wider that segment, the more the regulatorily measurable share concentrates on marginal cases. And it is on this marginal share that regulators work. Reality escapes the instrument applied to it.

Hence the doctrinal anchor: the regulatory defect is not that the model denies too much. The defect is that the system does not measure the cost required to transform a denial into a contestable object. The initial decision is observable. The appeal is observable. But the politically decisive zone, all the denials that never become appeals because the evidence is too costly to reconstitute, remains off-metric.

The question, then, is not only who decides. It is who holds the architecture of evidence at the moment of denial.

Ex post friction is not the whole apparatus

Ex post friction (the cost of appeal after a denial) is what doctrine usually describes. But it constitutes only one half of the actual system.

The other half is ex ante deterrence: the anticipated non-demand of care. What former executives of the prior authorization industry call the *sentinel effect* designates the physician who, having internalized the statistical cost of a request likely to be denied, stops formulating it. The beneficiary who, having internalized the history of denials in their care category, no longer asks the prescriber for the intervention. None of these non-demands appears in prior authorization statistics. They are structurally invisible because there exists no accounting category for care never requested.

The system is therefore not only a system of denials. It is a system that modifies the distribution of demand. Ex ante deterrence precedes ex post friction, and acts on a volume that no reporting captures. If one were to name the overall property, it would be: regulation by ex ante deterrence of demand, completed by regulation through ex post friction of appeal. The two mechanisms compose an apparatus whose operational efficiency is measured not by the denial rate, but by the ratio of initial requests to potential requests, a ratio that is never published, because it is never computed.

Unobservability therefore does not bear only on uncontested denials. It bears on requests never formulated. At each layer of the apparatus, the politically decisive segment is the one that leaves no trace.

The political economy of non-recourse, under constraints

The simplified yield equation, $Profit \approx f(denials \times (1 - appeal\_rate \times reversal\_rate))$, implicitly assumed that maximizing denials was always optimal for the payer. This is not the case. Several constraints weigh on the equilibrium:

Medical Loss Ratio: the ACA rule requires that a minimum share of premiums be returned in care (80 percent on the individual market, 85 percent in group). Massive denials reduce paid care and may push the ratio below the legal threshold, triggering mandatory rebates.

Administrative cost: review, compliance, legal defense. Each additional denial carries a processing cost.

Legal risk: class actions like Lokken, regulatory exposure, CMS sanctions.

Reputational risk and CMS Star Ratings: the latter conditions financial bonuses in MA and incorporates service indicators.

The relevant equation is therefore:

$$Profit \approx f(\ denials \times (1 - appeal\_rate \times reversal\_rate) - admin\_cost - legal\_risk - reputation\_cost\ )$$

under constraints: $MLR \geq legal\_threshold$; $litigation\_exposure \leq tolerance$; $Star\ Rating \geq threshold$.

What is optimized is not the gross volume of denials. It is the marginal equilibrium under constraints. Yet at this equilibrium, non-recourse remains a central stabilizing factor. An uncontested denial is, all other things equal, the economically most favorable denial: it maximizes the net yield of the individual operation, generates neither jurisprudence nor precedent, and does not expose the payer to reputational risk. Its aggregate contribution to the MLR is another matter. A denial reduces the numerator of medical expenses and may, in cumulative terms, push the ratio below the legal threshold, thereby triggering rebates; but this effect depends on the full portfolio and not on the individual denial. The MLR constraint weighs on the global strategy of the payer, not on the marginal value of each uncontested denial.

It is this equilibrium that reproduces itself, without any actor having to design it as such.

The pattern: Lokken, PxDx, EviCore

Three ongoing cases suffice to illustrate the motif. Their evidentiary status differs and must be kept visible.

1. ***Estate of Lokken v. UnitedHealth Group*** **(D. Minn., 0:23-cv-03514)**: active judicial proceeding. The complaint targets UnitedHealthcare's use of nH Predict, a tool developed by naviHealth (an Optum subsidiary acquired for 2.5 billion dollars in 2020), to manage post-acute care authorizations for Medicare Advantage beneficiaries. The complaint alleges an error rate of 90 percent and an appeal rate of 0.2 percent. An investigation by the Senate Permanent Subcommittee on Investigations (October 2024), a publicly established finding distinct from the plaintiffs' allegation, documented that UnitedHealth's denial rate in post-acute care rose from 8.7 percent to 22.7 percent between 2019 and 2022, after the deployment of nH Predict. On March 9, 2026, the District Court of Minnesota ordered extensive discovery; the response to the motion to compel is expected on April 29, 2026.

2. ***Kisting-Leung v. Cigna Corporation*** **(E.D. Cal., 2:23-cv-01477)**: an earlier but structurally parallel proceeding. The ProPublica investigation of March 2023, based on internal documents and interviews, established that the PxDx system (procedure-to-diagnosis) had rejected more than 300,000 requests in two months of 2022, at an average rate reported as 1.2 seconds per decision. Cigna disputes the qualifier *artificial intelligence* and argues that PxDx checks code conformity; the technical qualification changes without changing the structural effect.

3. **EviCore by Evernorth**, a Cigna subsidiary covering prior authorization for more than 100 million Americans, was the subject of a joint ProPublica / Capitol Forum investigation (October 2024). The elements are of differing status. The public commercial material of EviCore promises insurers a return on investment of 3 dollars saved on paid care for 1 dollar spent: a point established on primary source. Some contracts indexing compensation on spending reduction were documented through internal documents consulted by the investigators. The existence of an internal parametric threshold, nicknamed *the dial*, and the denial rate of approximately 20 percent in Arkansas fall under elements reported by investigation on testimonial sources and published data. They converge with the rest, but do not judicially prove it.

The motif is shared: an algorithmic apparatus optimized for a denial volume exceeding human review capacity, backed by an appeal mechanism whose friction guarantees that it will be marginally used.

Why current regulation does not touch the equilibrium, and risks aggravating friction

The regulatory movement is coherent, but its scope is marginal.

The reference is California. **SB 1120**, in effect on January 1, 2025, provides that a denial, delay or modification based on medical necessity cannot be decided on the sole basis of an AI system and must be reviewed by a qualified professional. Several other states (Arizona, Maryland, Nebraska, Texas) have adopted or discussed similar texts, with varying perimeters and effective dates; Indiana (HB 1271, March 4, 2026) addresses downcoding more specifically. At the federal level, CMS-0057-F frames deadlines and imposes specific reasoning of denials, as recalled above.

The regulatory lever that matters is not the quality of the initial decision. A human review downstream of an algorithmic denial does not change the equilibrium as long as the review is mobilized by only 11.5 percent of those denied, and as low as 0.2 percent in the most vulnerable cohorts.

Worse: certain regulations targeting decision quality may mechanically aggravate the friction of contestation. The requirement of specific reasoning of the denial has two opposing effects. On one hand, it increases the cost of issuing the denial for the payer, exerting downward pressure on volume. On the other hand, it increases the technical complexity of justification, and therefore the symmetric cost of refutation for those without documentary resources. If the net effect tips toward the second side, the regulation aggravates the friction it intends to correct. This ambivalence is not an empirical certainty; it is an incentive structure that no current reporting allows to measure.

The liberal objection, *the market corrects*, presupposes a market. Medicare Advantage beneficiaries operate within an enrollment apparatus constrained by calendar windows, high switching costs, and radical information asymmetry: the payer's algorithm is not inspectable by the insured. The market cannot correct what the user cannot observe.

The reverse objection, *state bills resolve the problem*, conflates an intervention on *decision quality* with an intervention on *the mechanics of appeal*. As long as evidentiary reconstruction remains prohibitive, the system's yield is insensitive to model improvement.

The regulator-operator paradox

On January 1, 2026, the Center for Medicare and Medicaid Innovation (CMMI) launched the WISeR Model (Wasteful and Inappropriate Service Reduction). The pilot tests the use of technologies, including AI, in utilization review for targeted services in *traditional* Medicare, across six states (Arizona, New Jersey, Ohio, Oklahoma, Texas, Washington) and six years. The scope covers seventeen service categories; the operation runs through third-party entities with human clinical review provided for adverse determinations. It is a controlled pilot, with limited perimeter.

But its direction of travel matters. The federal regulator becomes operator of an apparatus structurally equivalent to the one it regulates elsewhere. This dual posture, regulator of Medicare Advantage insurers and operator of an algorithmic utilization review on traditional Medicare, displaces the political problem. CMS can no longer present itself as an external observer of the apparatus it evaluates. It is a part of the system it regulates. This configuration does not prejudge the quality of the pilot; it bears on the institutional credibility

of any subsequent restriction.

Symmetrically, Executive Order 14365 (December 2025) charges federal agencies with challenging state rules deemed *too constraining* on AI and with promoting a *minimal* national standard. The risk of preemption weighs on the state legislative mosaic before it has even had time to produce its effects.

The concept

This regulatory mode must be named. For the case of AI in health insurance, the useful distinction is doctrinal.

Medical abandonment (*déshérence médicale*) designates the situation in which the patient is left without follow-up by default of a chain of care. The 82 percent, the 11.5 percent, the 0.2 percent describe an *administrative abandonment*: the situation in which the beneficiary is left without recourse by default of a chain of contestation. This is not the same phenomenon as medical abandonment, but it is the same structure: the system functions because the user cannot negotiate its exits. This administrative abandonment can in fact compound the medical abandonment we have already described.

Two adjacent concepts illuminate the position. *Perimeter bias*: regulation evaluates what it names, not what produces the effect. The quality of the AI is regulated; the quality of the friction is not. *Governability debt*: a system can be locally compliant and globally ungovernable. Each denial is individually defensible in law; the aggregate constitutes a stable political distortion.

What should be regulated?

The cost of evidentiary reconstruction, the informational component of friction.

Concretely: effective portability of the beneficiary's clinical record, real-time access to the data necessary for contestation, automatic pre-instruction of an appeal by the same system that produced the denial, transparency on the parametric thresholds of the denial model, non-aggregated publication of denial and reversal rates by service category and clinical cohort. The FHIR Prior Authorization API planned by CMS for 2027 goes in the right direction, on the condition that its scope cover contestation and not only submission.

None of these measures acts on the quality of the model. All act on the informational mechanics of evidentiary reconstruction.

A clarification is necessary here, so as not to overload the argument. Four-sided architectures, as conceptualized by our cohort of students in the MIT Chief Product Officer program around the CHIP (Comprehensive Healthcare Intelligence Platform) concept, articulating patients, providers, payers, and developers, demonstrate that an alternative orchestration of data is technically feasible. Patient, provider, payer, developer are the four fragmented holders of the evidence that the beneficiary today must reconstitute alone. CHIP is a feasibility proof, not a causal proof. The absence of such an architecture is not the sole cause of the current system; informational friction is probably a necessary condition for maintaining the equilibrium, but not a sufficient one. A perfect orchestration of data would not, on its own, suppress procedural and cognitive frictions. It would shift the mobilization threshold. Whether that shift, supposing it occurs, is enough to invert the regime of the system is an open question. But it can be posed only on condition that informational friction is attacked.

Limits of the thesis

Four limits must be kept visible.

1. First, the 80.7 percent reversal rate conditional on appeal is selection-biased: appealed cases are those where contestation appeared viable. One cannot, by simple extrapolation, conclude that the same proportion of unappealed denials would also be reversible. The structural argument does not depend on this extrapolation. It rests on the inverse fact: friction prevents measurement of the true reversal rate, and the unknowable is itself the political variable.

2. Second, a portion of initial denials is clinically justified: prior authorization also has a utility function, documented by MACPAC. The thesis does not imply abolition of the device; it implies measuring what it does *beyond* its explicit function.

3. Third, informational friction is necessary, not sufficient. An empirical decomposition of the three components (administrative, informational, cognitive) has not been conducted to date. The proposal to attack the cost of evidentiary reconstruction is a lever whose net effect on the appeal rate can be predicted only with an assumed range of uncertainty.

4. Fourth, the mechanics described here bear on the American system. European regimes (French Sécurité sociale, British NHS, Bismarckian systems) operate within different cost and incentive structures. The translatability of the concept of administrative abandonment to those regimes deserves separate work.

Closing

AI regulation in health insurance is technically feasible, legally framed, politically debated. It is also structurally backed by an apparatus of friction that is not recognized as regulatory, and that reproduces itself without any actor having to design it as such.

As long as this apparatus is not named, state bills will continue to concentrate on the quality of the algorithmic decision without touching the equilibrium of the system. Prior authorization vendors will continue to optimize their administrative approval rate, indifferent to the reversal rate on appeal. Beneficiaries will continue, at a minimum of 88.5 percent, to play the unpaid role of operating point. And the politically decisive segment, denials never contested and requests never formulated, will continue to escape any measurement, because its non-measurability is precisely what stabilizes the whole.

The 82 percent is not a statistic of failure. It is the numerical expression of a zone outside the metric, and of a system self-stabilized by the unobservability it maintains.

What remains to be known is who pays for it to reproduce itself. The arithmetic answer is known.

#ArtificialIntelligence #AIGovernance #DigitalHealth #Regulation #HealthInsurance

Primary sources:

• Mello M.M., Trotsyuk A.A., Djiberou Mahamadou A.J., Char D., *The AI Arms Race in Health Insurance Utilization Review*, Health Affairs 2026;45(1):6-13. DOI: 10.1377/hlthaff.2025.00897.

• Fuglesten Biniek J., Sroczynski N., Freed M., Neuman T., *Medicare Advantage Insurers Made Nearly 53 Million Prior Authorization Determinations in 2024*, KFF, January 2026.

• *Estate of Gene B. Lokken et al. v. UnitedHealth Group, Inc.*, 0:23-cv-03514 (D. Minn.); discovery order of March 9, 2026, *2026 WL 658883*.

• *Kisting-Leung v. Cigna Corporation*, 2:23-cv-01477 (E.D. Cal.).

• ProPublica & Capitol Forum, *Inside the Company Helping America's Biggest Health Insurers Deny Coverage*, October 2024 (EviCore).

• U.S. Senate Permanent Subcommittee on Investigations, *Refusal of Recovery*, October 2024.

• CMS Interoperability and Prior Authorization Final Rule, CMS-0057-F.

• CMMI, WISeR Model Operational Guide, January 1, 2026.

• California SB 1120 (effective January 1, 2025).

• Sunstein C.R., *Sludge: What Stops Us From Getting Things Done and What to Do About It*, MIT Press, 2021.