The Prior Auth API Economy: How CMS-0057-F, CMS-0062-P, Da Vinci FHIR Rails, State Gold Carding Laws, AI Guardrails, and the AHIP/BCBSA 257M Commitment Turn UM Into a Programmable Transaction

APR 22, 2026

Abstract

- Thesis: PA is shifting from labor-arbitrage cost containment to API-driven infrastructure; value is migrating from BPO and in-house staff to software, data, and middleware

- Regulatory stack: CMS-0057-F (Jan 2024 final, Jan 2027 compliance, four FHIR APIs), CMS-0062-P (Apr 2026 proposed, Oct 2027 proposed compliance, drug PA), ONC HTI-1 (EHR certification), AHIP/BCBSA June 23, 2025 commitments covering 257M Americans, state laws in WA, NE, TX, AR, MD, CA, VA, IN, AK, MI, IA, IL, VT, CO

- Scale of dysfunction: 39 PAs/week per physician, 13 hrs/week of physician+staff time, 40% of practices have dedicated PA staff, 29% of docs report a serious adverse event from PA delay, only 35% of PAs processed electronically (CAQH), only 9% of surveyed orgs could support an ePA API by 2027 (CAQH), nearly half of PA requests still submitted by fax or phone per BCBSA Jan 2026, $515M annual CAQH savings estimate, $15B CMS 10-yr savings estimate

- Critical regulatory unlock: CMS granted enforcement discretion in 2024 allowing an all-FHIR PA workflow in place of the legacy X12 278, effectively waiving the HIPAA X12 mandate for PA and opening the door to pure-FHIR infrastructure

- Four CMS-0057-F APIs: Patient Access, Provider Access, Payer-to-Payer (the mechanism for 90-day continuity carryover, shipping up to 5 years of claims, encounters, USCDI data, and PA history between plans on patient opt-in), Prior Authorization

- Technical stack: HL7 Da Vinci PA suite (CRD for discovery, DTR for documentation, PAS for submission/adjudication), FHIR-to-X12 bridge available but no longer

required, CDS Hooks cards in EHR workflow

- AI boundaries: state laws (MD HB0820, CA SB 1120, NE LB 77, CA SB 363) require individualized basis, mandate human decision-making, require quarterly audits, impose up to $1M per case fines for high app
(UnitedHealth) and PXDX (Cigna) as caution

- Opportunity map: provider-side orchestrat
a service"), PA portability/Payer-to-Payer lay
aware clinical decision support, patient-faci
specialty benefit manager modernization, an
PA dataset

- Named players: provider-side (Cohere, Rhy
Itiliti Health); payer infra (Smile Digital Hea
specialty benefit managers (Evolent, eviCore,
modernization targets

- Honest risks: payer foot-dragging via thin I
transition, gold card complexity varying acro
clearinghouses and EHRs, utilization rebou

Table of Contents

How Bad Is It Actually

Before getting into the opportunity, it helps to sit with how genuinely broken the current state is, because the scale of the dysfunction is what makes the rebuild economically obvious. The AMA's 2024 Prior Authorization Physician Survey, which pulled responses from 1,000 practicing docs, reads like a clinical fever chart. Physicians are knocking out an average of 39 PA requests per week, burning roughly 13 hours of combined physician and staff time to do it. Forty percent of practices have dedicated PA staff who do literally nothing else. About one in three physicians say their requests get denied often or always, and 75 percent say the denial volume has climbed over the past five years.

Then there's the part that should make any policymaker queasy. Ninety-three percent of physicians report that PA delays necessary care, 82 percent say it leads to treatment abandonment, and 29 percent, which is more than one in four, report that PA has caused a serious adverse event for a patient under their care. Hospitalizations in 23

percent of cases. Life-threatening events in 18 percent. Permanent disability or death in 8 percent. For a process whose entire justification is cost containment, the morbidity tail is unusually loud.

The money side is just as ugly. The CAQH 2024 Index pegs the savings from full ePA adoption at $515 million annually, with 14 minutes of staff time shaved per authorization compared to the current mixed bag of portals, faxes, and the occasional API. CMS has projected $15 billion in savings over ten years from its interoperability rule alone. Those numbers are the market size. They explain why every incumbent clearinghouse, EHR vendor, UM platform, and health tech seed company is suddenly interested in a transaction type they mostly treated as infrastructure plumbing five years ago.

The Adoption Gap Nobody Wants to Talk About

The most damning numbers in the CAQH 2024 Index are not the savings figures. They are the adoption figures. Only 35 percent of prior authorizations are currently processed electronically at all. That is not "not fully automated." That is "roughly two-thirds of all PA transactions in American healthcare still move on paper, fax, portal, or phone in 2024." Every other major administrative transaction in healthcare (eligibility, claims, remittance) has adoption rates above 90 percent. PA is the last unreconstructed workflow in the whole X12 catalog.

The second number is worse. Only 9 percent of surveyed organizations reported that they could currently support an ePA API of the kind CMS-0057-F mandates by January 1, 2027. Nine percent. That is not a compliance gap, that is a compliance chasm, and it is the single most important statistic in the entire PA economy. It says the industry has roughly eight months of actual build time from the publication of this essay to stand up infrastructure that 91 percent of surveyed orgs cannot yet support. Either the rule gets watered down or delayed, which is possible but not assumed, or the next 14 months are going to be a historic scramble for FHIR-native middleware, systems integrators, and anyone who can plausibly claim to ship production-grade FHIR endpoints at payer scale. Market-sizing exercises that ignore the 9 percent number are missing the scale of the urgency.

BCBSA in a January 2026 article confirmed the provider-side picture bluntly: nearly half of PA requests are still submitted by fax or phone. Half. In 2026. At payer plans that are publicly committed to 80 percent real-time ePA by January 2027. The gap between the aspirational commitment and the operational reality is roughly the size of the opportunity.

The Federal Rails and the Four APIs of CMS-0057-F

The piece that actually forces the rebuild is CMS-0057-F, the Interoperability and Prior Authorization Final Rule that CMS dropped in January 2024. It applies to Medicare Advantage organizations, state Medicaid and CHIP fee-for-service programs, Medicaid managed care plans, CHIP managed care entities, and Qualified Health Plan issuers on the federally-facilitated exchanges. The operative word is breadth. These plans collectively cover a huge slice of the insured population, and by January 1, 2027 they all have to implement a Prior Authorization API built on HL7 FHIR standards. In addition, payers must send PA decisions within seven days for standard requests and 72 hours for urgent ones, and provide a specific reason for any denial.

What tends to get lost in casual coverage is that CMS-0057-F actually requires four distinct FHIR APIs, not one. The Prior Authorization API gets the headlines because it is the operational showpiece, but it is only one layer of the stack. The other three are just as important to the economics of the new PA economy, and each of them underpins a different business model category.

The Patient Access API requires payers to make claims, encounter data, formulary information, and (by 2027) PA information available to members through third-party apps on request. This is the rail that lets patients see their own PA status, approvals, denials, and the reasons for those decisions in real time through consumer-grade apps. The Provider Access API requires payers to share claims, encounters, and PA data with in-network providers for attributed members, which gives clinicians longitudinal visibility into prior care and prior PA history when they are making a decision. The Payer-to-Payer API is the most operationally consequential for the portability business model: when a member opts in, the previous payer must ship up to five years of claims, encounter data, USCDI-aligned clinical data, and PA history to the new payer. That is the technical mechanism by which the 90-day continuity of care carryover actually happens. And the Prior Authorization API is the real-time decisioning surface.

These four APIs together constitute the FHIR rail layer. Every business model in the new PA economy sits on top of one or more of them. Treating “the PA API” as the whole story is like treating the checkout step as the whole story of e-commerce. The sleeper provision in CMS-0057-F, separate from the APIs themselves, is the public reporting mandate. Payers have to annually report approval rates, denial rates, average decision timeframes, and the percentage of requests approved only after appeal. This is the first time in American health insurance history that PA behavior becomes a

standardized, comparable, public dataset. The implications for analytics, benchmarking, and market access products are enormous and mostly unpriced.

ONC's HTI-1 final rule closes the loop on the EHR side by updating the Health IT Certification Program to require certified EHRs to support FHIR-based data exchange. You can have all the payer APIs in the world, and if the EHR cannot consume them cleanly, the value never reaches the clinician. HTI-1 fixes that on the certification side, though actual vendor implementation quality is a separate and more cynical conversation.

The Quiet Unlock: CMS Waives X12 for PA

This is the piece of the regulatory package that does not get nearly enough attention outside the standards community, and it is arguably the single most important enabler of the new infrastructure category. Under HIPAA, electronic PA transactions have historically been required to use the X12 278 standard. That requirement is what kept the clearinghouse ecosystem in the middle of every PA, because clearinghouses are the plumbing layer for X12 EDI. Any company trying to build a pure-FHIR PA product was either violating HIPAA or had to maintain a FHIR-to-X12 bridge for the electronic transaction of record.

In 2024, alongside CMS-0057-F, CMS granted enforcement discretion allowing payers and providers to use an all-FHIR PA workflow in place of the X12 278. In effect, this waives the HIPAA mandate for PA specifically. The regulatory permission slip to skip the clearinghouse entirely is now issued. Any pure-FHIR infrastructure company that stands up a Da Vinci-compliant PAS pipeline can now legally operate as the full transaction rail for PA without needing to round-trip through X12 or a traditional clearinghouse. That is not an incremental improvement. That is the moment a whole category of legacy intermediaries becomes bypassable by regulation rather than by ambition.

This is why the Da Vinci PAS guide defines a FHIR-to-278 translation as optional for backwards compatibility rather than as a required bridge. Companies can run pure FHIR end to end. The clearinghouses know this, which is why their FHIR investments are accelerating. But by the time most of them ship FHIR-native product, several of their largest transaction categories will have already migrated.

CMS-0062-P and the Drug Rule

In April 2026 the Trump administration followed up with CMS-0062-P, a proposed rule that extends the same framework to prescription drugs. Impacted payers would

have to support electronic PA for medications, adopt updated FHIR and NCPDP SCRIPT standards for drug transactions, and report their interoperability API endpoints and usage metrics back to CMS. The deadlines on the drug side are tighter. Medicaid plans would get 24 hours to respond to drug PA requests. ACA marketplace plans would get 72 hours for standard and 24 for expedited. The comment period runs until June 15, 2026, with proposed compliance in October 2027. Assuming it survives comment substantively intact, the drug rule locks in what the medical rule established: PA is now a federally regulated, API-delivered, time-bound transaction, and pharmacy is next.

The June 2025 AHIP Announcement and Why 257M Is the Real Number

The commercial sector, which covers most working-age Americans, is largely outside the direct scope of CMS-0057-F. The industry knows that if it does not move voluntarily, Congress or a future administration eventually will. On June 23, 2025, AHIP and the Blue Cross Blue Shield Association announced a joint voluntary commitment, built in partnership with HHS and CMS, designed to streamline PA across the commercial sector.

The headline number is the one that matters: the participating plans collectively cover roughly 257 million Americans. That includes UnitedHealthcare, Cigna, Aetna, Elevance Health, Humana, Centene, Kaiser Permanente, and the full Blue Cross Blue Shield federation. That roster is effectively every payer in the country that matters at scale. When a voluntary commitment covers 257 million lives, it stops being a PR gesture and becomes a de facto industry standard. It is also a flare to regulators and Congress that further rulemaking is unnecessary, which is the political function of such commitments generally.

The announcement laid out six specific commitments with staggered 2026 and 2027 effective dates. Honoring existing approvals for at least 90 days when a patient switches plans. Reducing the volume of PA requirements by identified service categories. Providing clear communication and transparency on denials and appeal rights. Expanding real-time responses across plans, culminating in 80 percent of electronic PA requests being addressed in real time by January 1, 2027. Ensuring continuity of care during transitions. And committing to specific measurement and public reporting of performance.

The April 2026 progress report was the first real read on whether the commitment was being executed. Participating plans reported eliminating 11 percent of prior authorizations across various medical services, which translated to 6.5 million fewer

requests for patients. In Medicare Advantage the reduction was over 15 percent. BCBSA CEO Kim Keck re-committed the Blues specifically to the 80 percent real-time ePA target by January 1, 2027. Real time is the key phrase. Real time at the point of care is not a workflow improvement. It is an infrastructure requirement. It rules out batch, portal scraping, and any architecture that cannot guarantee sub-second response for the common cases. Hitting 80 percent real time on ePA forces FHIR-native backends, and a 257-million-life cohort of commercial payers is now collectively on the hook for that build by January 2027.

States Are Writing the SLA

While the federal and industry layers are setting the direction, state legislatures are the ones actually prescribing the operational constraints. The December 2025 NAIC Prior Authorization White Paper grouped state activity into four buckets: turnaround time requirements, gold carding, continuity of care protections, and AI guardrails. Each of them turns a discretionary administrative tool into a regulated SLA product, which is the exact moment when a workflow becomes a software market.

Washington State's ESSHB 1357, passed in 2023, is the most technically prescriptive. It required health plans to build and maintain a FHIR-based PA API by January 1, 2025, with 72 hours for urgent and 7 days for standard turnaround. Miss the deadline and the request is automatically approved. Auto-approval on SLA miss is a bright line that effectively converts the SLA from aspirational into a liability-bearing legal obligation, because now a missed deadline has a measurable dollar cost. Nebraska's LB 77 in 2025 did something similar, setting 72 hours for urgent and 7 days for non-urgent with auto-approval on miss, tightening urgent down to 48 hours starting in 2028.

Other states have piled on. Indiana's SB 0480 requires 48 hours for urgent and 5 business days for non-urgent, and mandates electronic receipt of PA requests. Alaska's HB 0144 requires 72 hours for standard and 24 hours for expedited. Michigan imposes 72 hours for urgent and 7 days for standard, with automatic approval on a missed deadline, using the same liability mechanism as Washington. Iowa's HF 303 sets 48 hours for urgent and 10 days for non-urgent, and requires annual reporting to the state insurance commissioner. Illinois and Vermont have adopted 90-day continuity windows for plan switches, tracking the AHIP/BCBSA commitment. Colorado has passed rules allowing multi-year authorization validity for stable chronic regimens, which is operationally significant for specialty drug access. And California's SB 363 goes in a different direction entirely by imposing denial-rate disclosure requirements and fines of up to $1 million per case where more than 50

percent of appeals are overturned. That last provision is effectively a statutory penalty for overly aggressive algorithmic denial, and it is going to be cited in every investor deck for an AI audit company for the next three years.

Gold Carding as a Data Problem in Disguise

Texas kicked off the gold carding trend with HB 3459 in 2021, which exempts physicians from PA for specific services if their approval rate over the previous 12 months is 90 percent or higher. The 2025 amendment, HB 3812, extended the evaluation window from six months to a full year and forced insurers to submit annual gold-card data reports to the Texas Department of Insurance. Arkansas followed with HB 1301 in 2025, applying gold carding at the group practice level.

Gold carding is usually sold as a patient-access mechanism, which it is, but operationally it is a data problem. To run a compliant gold card program at scale, a payer has to track provider-level approval rates by service category across rolling 12-month windows, with enough granularity to identify when a provider crosses the threshold and enough auditability to defend the calculation to a state regulator. That is the definition of a structured longitudinal dataset. Once a payer has that, the same infrastructure can be resold back to providers as a benchmarking tool, to employer groups as a network-performance metric, and to life sciences as a provider-segmentation input. Gold carding accidentally creates a provider-performance database as a regulatory byproduct.

The 90-Day Carryover Is a Payer-to-Payer API Problem

State laws are increasingly protecting patients from care disruptions when they switch plans. The AHIP and BCBSA commitments include a 90-day transition period where an existing PA must be honored for benefit-equivalent, in-network services. Virginia's HB 736 from 2026 goes further, requiring that initial PAs remain in effect for at least six months and continued requests for at least 12 months. Nebraska's LB 77 provides that approved PAs are valid for one year in most cases and can follow a patient for 60 days after a plan switch. Illinois and Vermont both now have statutory 90-day continuity windows. Colorado allows multi-year authorization validity for stable chronic regimens.

The mechanism that makes all of this work is the Payer-to-Payer API in CMS-0057-F. The statutory text of the carryover commitments presumes that an existing PA and its

clinical documentation can physically move between payers in a timely, reliable, machine-readable way. That does not happen on its own. The Payer-to-Payer API is the rail: when a member opts in, the old payer ships up to five years of claims, encounters, USCDI clinical data, and PA history to the new payer in FHIR. Without that data flow, the 90-day honoring commitment reduces to the new payer taking the patient's word for what was previously approved, which no payer will do at scale.

Think about what this actually requires in code. A PA approval is a state object. It has attributes: service, duration, clinical justification, approving clinician, expiration date, and associated documentation. It has to be portable across competing payers. It has to survive plan switches, network changes, and PBM transitions without losing integrity. Building the rails to securely transfer that state object between payers that are actively fighting for the same members is not a feature. It is a neutral intermediary layer, and no incumbent wants to be the one who builds it for the other side. That is a greenfield opportunity disguised as a compliance burden, with the Payer-to-Payer API as the underlying transport.

AI Guardrails and Where the Lawsuits Are Pointing

For anyone building AI for utilization management, the most important state-level trend is the wave of laws explicitly regulating algorithmic decision-making in PA. At least five states have enacted specific guardrails. Maryland's HB0820, effective October 1, 2025, closely tracks California's SB 1120 from 2024. Both require carriers, PBMs, and private review agents to ensure that any AI tool used in utilization management bases its coverage decisions on the enrollee's individual medical and clinical history, not on group or demographic statistics. The Maryland law mandates at least quarterly reviews of any AI used in UM, requires carriers to report metrics on AI use in adverse decisions, and makes AI tools available for audit by the insurance commissioner. Violations can trigger misdemeanor charges, monetary penalties, and revocation of certificates. Nebraska's LB 77 goes a step further and prohibits AI as the sole basis of a denial, requiring that adverse decisions be made by a physician. California's SB 363 adds denial-rate disclosure and the million-dollar-per-case fine structure for high appeal-overturn rates.

These laws did not emerge in a vacuum. UnitedHealthcare's nH Predict algorithm, used to set appropriate lengths of post-acute care for Medicare Advantage patients, has been the subject of class action litigation and congressional attention. Denial rates allegedly climbed from 10.9 percent in 2020 to 22.7 percent in 2022, a period that coincided with UnitedHealth's acquisition of naviHealth, which built the tool.

Plaintiffs claim the algorithm has a 90 percent error rate and routinely overrides physician recommendations. A federal judge in April 2026 ordered UnitedHealth to produce AI claim denial documents in discovery, which is going to be interesting to watch.

Cigna's PXDX system, exposed in a 2023 ProPublica investigation, allegedly allowed medical directors to deny hundreds of thousands of claims per month without reviewing individual patient files by auto-flagging claims that did not match a list of pre-approved condition-procedure pairs. Reporting suggested a single medical director could push through more than 50 denials in seconds. Whether you view these tools as efficient triage or algorithmic rubber-stamping, the regulatory consensus has crystallized: AI can assist, but cannot be the decider. Any architecture that assumes otherwise is betting against the direction of every relevant legislature and most relevant plaintiffs' firms.

The Da Vinci Stack in Plain English

The technical foundation for all of this is the HL7 Da Vinci Project, a private-sector standards initiative that brought payers, providers, and EHR vendors together to build FHIR-based implementation guides for value-based care workflows. CMS formally endorsed the Da Vinci PA guides as the basis for CMS-0057-F compliance, which means the Da Vinci PA suite is now effectively regulation by reference.

The ePA stack has three pieces. The first is Coverage Requirements Discovery, or CRD. When a clinician initiates an order or referral inside the EHR, a CRD call fires in the background to the patient's payer and asks, in effect, does this service require PA for this member on this plan. The payer replies with a CDS Hooks card that drops into the clinician's workflow and tells them before the order is locked in. Catching the PA requirement at the point of decision is the single highest-leverage intervention in the whole workflow, because it is the moment at which alternatives, documentation, or a different service choice are still possible.

The second piece is Documentation Templates and Rules, or DTR. Once PA is confirmed as needed, the DTR API pulls the payer's specific clinical documentation requirements and business rules and presents them to the provider as a SMART on FHIR app or a structured questionnaire. DTR can also pre-populate answers by pulling from the patient's existing clinical data in the EHR, which is where AI documentation synthesis has its biggest near-term footprint.

The third piece is Prior Authorization Support, or PAS. This is the submission layer. Once DTR has assembled the packet, PAS bundles the clinical data with the PA

request using FHIR resources, submits it to the payer, and returns either a real-time decision or a tracking ID for pending cases. Historically the PAS guide included a FHIR-to-X12 278 mapping for backwards compatibility, but with the 2024 CMS enforcement discretion that mapping is optional rather than required. The response comes back as a structured FHIR resource, which means the EHR can auto-update the record and the workflow without human re-entry. When it works end to end, the whole thing looks less like prior authorization and more like a payment authorization on a credit card network, which is broadly the right mental model.

Where the Plumbing Actually Breaks

The architecture is clean on paper. Real life is messier. Even with the CMS waiver on X12, the X12 278 transaction still dominates current production UM environments and most clearinghouses are still overwhelmingly X12-based. The transitional reality is multi-hop architectures, latency budgets blown on translation, and error surfaces at every mapping boundary. Every translation is a place where bugs live.

The AMA numbers on the provider side are ugly. Only 23 percent of physicians say their EHR offers ePA for prescriptions, and 30 percent say the PA requirement information in their EHR is rarely or never accurate. Which means even where payers do stand up FHIR endpoints, the provider-side tooling to consume them is frequently absent, broken, or stale. Layer on BCBSA's January 2026 admission that nearly half of PA requests are still submitted by fax or phone, and the picture is that fax machines and portal scraping are persistent because the alternative has not reliably arrived at the clinician's desktop. The 2024 CAQH Index's 35 percent electronic processing figure says the same thing from a different angle. And the 9 percent of surveyed orgs that can support a 2027-compliant ePA API says it from a third.

The clearinghouse layer is the other structural drag. Clearinghouses sit between providers and payers, routing transactions and translating formats. Most of the majors are optimizing their X12 infrastructure rather than rebuilding FHIR-native. That creates a real risk that the new FHIR rails get routed through legacy clearinghouse plumbing that neutralizes the whole latency premise of real-time adjudication. The CMS enforcement discretion theoretically lets anyone skip this layer entirely for PA. The question is who actually builds a credible alternative at payer scale fast enough to matter.

What AI Can and Cannot Do Here

AI is both the most interesting and the most dangerous layer in the new PA economy. On the provider side, AI genuinely shines in orchestration and synthesis. Give it a

patient's unstructured clinical notes, a payer's DTR rule set, and a target service, and modern models can reliably identify the criteria, map documentation to them, and auto-populate the questionnaire. The clinician is still in the loop; the AI is doing the tedious evidence-assembly that a human has been doing with a highlighter and a prior fax for 20 years. That work is low-controversy, high-value, and straightforward to productize.

On the payer side, AI can route incoming requests intelligently, flag missing documentation, identify clear auto-approvals based on objective guideline adherence, and surface patterns that warrant clinician review. CMS itself has begun incorporating AI into review workflows, with an explicit requirement that licensed clinicians sign off on final decisions.

Where AI breaks is where vendors try to collapse clinical judgment into a model. The nH Predict and PXDX cases are not abstract: they are the templates for the next decade of litigation. The AMA survey found 61 percent of physicians are worried AI will increase denial rates, and there is no regulatory trajectory anywhere in the country in which AI gets to be the sole decider on an adverse determination. The durable business model is AI as triage and documentation, clinicians as deciders, with clean audit trails that prove which role each played. Anyone architecting the opposite will eventually be a discovery exhibit.

Provider-Side Orchestration as a Business

On the provider side, the immediate pain is unambiguous. Practices are spending $20 to $30 per PA transaction, and 40 percent of them have dedicated PA staff. The market for end-to-end orchestration is huge, underserved, and not yet consolidated.

The shape of the winning product is by now pretty clear. Sit on top of the EHR via SMART on FHIR or a direct integration. Use AI to ingest clinical notes and map them against payer-specific rules, using CRD and DTR where available and falling back to portal scraping or direct payer connectivity where they are not. Auto-package and submit the PAS request. Track state in real time. Close the loop back into the EHR with a structured result. The moat is coverage breadth. Any orchestration tool that only handles FHIR-native payers is a toy, because roughly half of PA volume is still fax or phone per BCBSA. The one that handles all payers and all service types, drugs included now that CMS-0062-P is bringing pharmacy PA into the electronic framework, is the one that wins the practice.

The competitive field on provider-side orchestration is already busy. Cohere Health, Rhyme, and Infinitus are among the better-known names. Myndshft, Infinx, Anterior, and Itiliti Health are each executing on meaningful slices of the workflow, with different specialty depths and different integration footprints. The category is not yet consolidated and specialty-specific orchestration (oncology, advanced imaging, MSK, high-cost drugs) still has whitespace where the generic horizontal tools tend to underperform.

The interesting second-order revenue model here is gold carding. An orchestration platform that consistently gets its providers to 90 percent approval rates across the service categories that matter is effectively shrinking the total addressable volume of PAs for those providers, which is an unusual and potentially disruptive product promise. It also makes the state-by-state complexity of gold card programs a moat, because a national tool that can track and administer gold card status across Texas, Arkansas, and any future state that enacts a program is meaningfully harder to build than it looks.

Payer-Side FHIR Middleware

Payers have a hard deadline and a hard problem. CMS-0057-F requires four FHIR APIs by January 1, 2027. Most legacy UM stacks were built on mainframes and X12 EDI and physically cannot hit sub-second response times. A 2026 Becker's Payer survey found 28 percent of payers estimated spending $1 million to $5 million just on API implementation, which probably undercounts the true total cost of ownership once clinical policy digitization, integration testing, and ongoing SLA monitoring are priced in. The CAQH 9 percent readiness figure puts a number on the gap: the overwhelming majority of payers cannot actually ship to the rule as currently written.

That gap is the market for white-labeled FHIR middleware, or, as the pitch decks are now calling it, Prior Auth as a Service. The product ingests inbound PAS requests, runs them against digitized clinical guidelines (InterQual, MCG, internal policies), auto-approves the clear cases, and intelligently routes the gray ones to medical directors. The payer keeps control of clinical policy. The vendor handles the FHIR compliance, the latency engineering, and the public-reporting data layer. This is fundamentally a distributed systems problem, not a clinical one, which is why cloud-native engineering teams will likely outcompete legacy health IT vendors on this layer. The AHIP and BCBSA 80 percent real-time commitment is functionally a performance spec for the middleware category.

The infrastructure layer already has credible players, though none that has yet consolidated the market. Smile Digital Health (formerly Smile CDR) and Firely are the

most prominent pure-play FHIR infrastructure vendors, with deep HL7 engineering benches and broad payer deployments. HealthEdge and CareEvolution are each building middleware-adjacent products aimed at the same 2027 deadline. None of them is an obvious default yet, which is unusual for a market with a hard deadline 14 months out.

The Portability Layer Nobody Owns Yet

The 90-day carryover from AHIP and BCBSA, combined with state rules like Nebraska's 60-day portability, Virginia's minimum duration requirements, and the Illinois and Vermont 90-day windows, creates a structural need for a neutral PA portability layer. The Payer-to-Payer API is the plumbing, but the plumbing needs an operator. When a member switches plans, the new payer has to receive the existing approvals and the clinical documentation that justified them in a machine-readable, standardized format, and the mechanism has to work across every combination of incumbent and recipient payer. The two payers are competitors. They have no existing data-sharing relationship. Neither wants to be the one to build the exchange that honors the other's decisions at scale.

That is the definition of a neutral intermediary opportunity. The analogy is not perfect, but credit bureaus are instructive: a shared infrastructure that holds a history that any party in the ecosystem needs but none wants to operate individually, with a per-transaction fee model and a SaaS layer for access. A FHIR-native PA portability service that brokers Payer-to-Payer API calls, holds PA state objects, enforces data provenance, and exposes them to payers and providers on request would solve a compliance headache for the entire industry and would benefit from a strong regulatory flywheel as more states mandate continuity of care. The fact that nobody has obviously won this space yet, 14 months out from the 2027 deadlines, is mildly remarkable.

Compliance Tooling for the Algorithmic Audit Era

The Maryland and California AI laws, the parallel moves in Nebraska, and California's SB 363 fine structure create a compliance surface that incumbent UM vendors are not structurally equipped to handle. Maryland alone requires quarterly audits of any AI used in UM, written policies and procedures for AI use, metrics reporting on AI involvement in adverse decisions, and the ability to make AI tools available for commissioner inspection. California layers on denial-rate disclosure and up to $1 million per case where appeals are overturned at rates above 50 percent. Any payer

that has deployed AI into its UM stack now has a running compliance obligation that needs tooling.

There is a nascent market here for independent AI audit and compliance platforms. The product tests payer algorithms against clinical guidelines, monitors denial patterns for demographic disparity (the "unfair discrimination" standard in Maryland's law), generates the reporting required by state insurance commissioners, and maintains the documentation needed to defend an algorithm under regulatory review. ONC's HTI-1 algorithm transparency requirements, which require certified health IT developers to disclose information on predictive algorithms and assess them for fairness, appropriateness, validity, effectiveness, and safety, extend this demand to the vendor side.

The analog is SOC 2 or HITRUST, but for clinical algorithms. Payers will eventually want to be able to point at an external audit report the way they currently point at a SOC 2 Type 2. Someone is going to build that playbook and sell it to every UM vendor in the country.

PA-Aware Clinical Decision Support

One business model category that is hiding in plain sight inside the Da Vinci stack is PA-aware clinical decision support. The CRD API, by design, tells a clinician at order entry whether a service will require PA for a specific member on a specific plan. Flip that around and it is the raw material for a different kind of CDS: a tool that shows the clinician, in real time, which clinically equivalent alternative would be auto-approved, which would trigger PA, and which would likely be denied, before the order is committed.

This is a subtle reversal of the usual CDS flow. Traditional CDS surfaces clinical appropriateness. PA-aware CDS surfaces approvability as a first-class property of the order, alongside appropriateness. For service categories where multiple clinically equivalent options exist (imaging modalities, drug classes, step therapy alternatives, DME configurations), this tooling can eliminate a huge share of PA volume at the point of order entry by quietly steering clinicians toward the approvable option. The payer likes it because utilization conforms to guidelines without a denial fight. The provider likes it because the administrative burden disappears. The patient likes it because there is no delay. The CRD output is the input. Building this well requires deep integration with the EHR, strong payer rule coverage, and a UX that clinicians actually tolerate, which is a high bar, but the economic alignment across all three parties is unusually strong.

Patient-Facing PA Transparency and Appeal Tools

The Patient Access API in CMS-0057-F extends to PA status by 2027. That is a quietly important development for the patient financial experience category. Patients will be able to see, in real time through consumer-grade apps, which of their PAs are approved, which are denied, what the reasons for denial are, and where they stand in the appeal process. Historically that information has lived in payer portals that patients never log into, in letters that arrive after the fact, and in occasional calls to member services.

Once that data is accessible through a consumer-facing API, a new product category becomes possible: patient-facing PA transparency and appeal assistance tooling. Think of it as a mashup of patient financial experience software and legal tech. The product ingests a patient's PA history through Patient Access, explains denials in plain language, automatically generates appeal letters with the clinical justification populated from the patient's own clinical data, and tracks the appeal through to resolution. For high-cost services where PA denials are most consequential (specialty drugs, cancer care, rare disease therapies, complex surgical procedures), this is a genuinely novel category. It is also a natural adjunct to existing patient advocacy, price transparency, and financial navigation products. The CMS rule is the enabler; the market does not yet have a clear leader.

Specialty Benefit Manager Modernization

One area that is conspicuously absent from most PA economy analyses is the specialty benefit manager segment. Companies like Evolent, eviCore, Carelon, and Lumeris operate carved-out PA workflows for specific benefit categories (oncology, cardiology, radiology, musculoskeletal, post-acute) under delegated utilization management arrangements with payers. These pipelines are structurally fragmented, often built on bespoke legacy technology inherited from a decade of acquisitions, and not generally FHIR-native.

CMS-0057-F treats delegated UM vendors the same as the underlying payer for compliance purposes, which means the specialty benefit managers have the same January 2027 deadline as their payer clients. That creates two distinct business opportunities. One is a modernization play: building FHIR-native UM infrastructure specifically for the specialty benefit manager segment, which is operationally different from a horizontal payer middleware product because the rules, the specialties, and the clinical guidelines are narrower and deeper. The other is a consolidation play: a well-

capitalized specialty benefit manager that modernizes its stack and then rolls up weaker competitors could meaningfully reshape the segment. Either way, this is a category that has been quietly overlooked relative to its PA transaction volume, and it deserves more attention than it gets.

The Dataset Is the Real Prize

Of all the layers in the PA economy, the one most chronically underpriced by entrepreneurs is the data layer. Once PA fully digitizes on FHIR rails, the exhaust is extraordinary. The CMS-0057-F public reporting requirements will produce, for the first time, a standardized, comparable, longitudinal dataset on payer PA behavior across Medicare Advantage, Medicaid, and ACA marketplace plans. CMS-0062-P extends the same framework to drugs. California's SB 363 denial-rate disclosure adds state-level granularity. Over time this dataset captures payer decisioning patterns by service category and geography, denial rates by provider and specialty, provider performance against guidelines (the basis of gold carding), drug access patterns by formulary tier, appeal overturn rates at the level California fines on, and the correlation between PA denial rates and downstream clinical outcomes.

That dataset is monetizable in a dozen directions. Life sciences companies pay serious money to track market access for new drugs. Providers will pay to benchmark and optimize their gold card status. Payers will pay to benchmark their own UM efficiency against peers. Employers will pay to evaluate health plan performance on behalf of their beneficiaries. Even the plaintiffs' bar will pay for pattern analysis, particularly in jurisdictions with California-style fine structures. The regulatory framework here is not just creating transparency. It is creating a structured public dataset that becomes the foundation for a new analytics category. The companies that aggregate, clean, and enrich this dataset first will have durable advantages that are hard to unwind once public reporting is established.

Incumbents, New Entrants, and Where the Whitespace Is

The current competitive map sorts roughly into five buckets. Legacy clearinghouses (Change Healthcare now under Optum, Availity, Waystar) own the X12 278 transaction layer and are slow on FHIR, partly because FHIR cannibalizes some of their existing economics and CMS just waived the regulatory requirement that kept them in the middle. EHR vendors (Epic, Oracle Health) are building CRD, DTR, and PAS into their platforms, but they are constrained by enterprise sales cycles and the

sheer complexity of supporting thousands of payer-specific configurations inside a single product.

Specialty PA vendors on the provider side (Cohere Health, Rhyme, Infinitus, Myndshft, Infinx, Anterior, Itiliti Health) are executing well on AI orchestration but are not in the infrastructure layer. Payer infrastructure vendors (Smile Digital Health, Firely, HealthEdge, CareEvolution) are in the middleware race but have not consolidated the market. UM platform incumbents and specialty benefit managers (eviCore, Carelon, Evolent, Lumeris) hold payer relationships but tend to run on legacy stacks and have the most to lose from genuine disintermediation.

The whitespace is almost entirely in the infrastructure plane and the adjacent data and compliance layers. FHIR middleware for payer-side compliance. A neutral portability layer brokering Payer-to-Payer API transactions. Algorithmic audit and compliance tooling. PA-aware CDS. Patient-facing PA transparency and appeal tooling. Specialty benefit manager modernization. A standardized data products layer sitting on top of the CMS public reporting data. These are all platform businesses with genuine network effects, deep switching costs once the integrations are built, and a regulatory tailwind that is effectively a customer acquisition subsidy.

The other underappreciated angle is the drug side. CMS-0062-P specifically targets pharmacy PA, which is where patient harm from delay is often most acute (specialty drugs, oncology, rare disease therapies) and where the existing tooling is arguably even further behind the medical side. Surescripts dominates ePrescribing, but the ePA flow for drugs has been historically messy. A clean FHIR and NCPDP SCRIPT-native drug PA orchestration layer, tuned for specialty pharmacy workflows, is a highly defensible wedge if the rule finalizes on something like the proposed timeline.

Honest Risks and What Could Go Sideways

Any serious entrepreneur or investor in this space needs to hold five risks clearly in mind, because each of them has meaningful probability and each kills specific theses.

The first is payer foot-dragging. The history of payer compliance with federal interoperability rules is one of minimum viable effort. The plausible 2027 outcome is that a large share of payers ship thin FHIR veneers bolted on top of legacy UM stacks, technically satisfying the rule while delivering almost none of the real-time, sub-second experience the rule was designed to produce. Products that presume deep, high-quality FHIR implementation on day one will underperform. Products that assume a multi-year maturation curve for payer endpoints will be better positioned.

The second is dual-stack necessity. During the transition, which will last years, any provider-side product that only handles FHIR-native payers is non-viable, because half of PA volume is still fax or phone per BCBSA and 65 percent is still non-electronic per CAQH. The winning products will run FHIR where available and degrade gracefully to portal automation, fax, and phone where FHIR is absent. Architecting for dual-stack is a meaningful engineering investment that pure-FHIR purists will skip, to their commercial detriment.

The third is gold card complexity. Every state that enacts a gold card program writes its own rules: different thresholds, different evaluation windows, different service categories, different reporting obligations. A national tool that cannot handle the full compliance matrix across every gold-card state will get outcompeted by state-specific tools in its weakest jurisdictions. This is a classic compliance-as-moat dynamic, but only if the product genuinely gets it right.

The fourth is consolidation pressure. Clearinghouses and EHR vendors have strong incentives to buy FHIR middleware, audit tooling, and orchestration startups rather than build them. This is a positive outcome for founders but a material risk for investors who are underwriting independent platform trajectories. Strategic acquirers will pay for distribution and regulatory clocks; the timing of those deals will shape returns more than any individual product win.

The fifth is utilization rebound. This one is the most philosophically uncomfortable. If PA becomes frictionless (real-time, auto-approved for most requests, low administrative burden), payers may rationally expand the scope of services subject to PA rather than shrink it, because the cost of administering PA on a new service category drops to near zero. The end state may be more services subject to PA, not fewer, with the volume flowing through the new API rails. That is still a better patient experience than the status quo, but it is worth being honest: PA is being rebuilt, not eliminated. The rails are the market, and the rails are getting more capacity, not less. Anyone whose thesis depends on PA shrinking in aggregate is probably going to be disappointed.

Closing Thoughts

The Prior Authorization API Economy is not speculative. It is under active construction. Federal rules are laying the FHIR rails across four APIs. State laws are writing the SLA parameters and drawing the AI guardrails across at least a dozen jurisdictions. A voluntary commitment covering 257 million Americans is pulling the commercial sector onto the same standards without waiting for Congress. CMS has already waived the X12 mandate for PA, which is the quiet legal unlock that makes

pure-FHIR infrastructure viable. And the documented dysfunction of the status quo (65 percent non-electronic, 9 percent of orgs actually ready, half by fax or phone, $15 billion of projected savings, 13 hours per physician per week, adverse events affecting nearly one in three physicians' patient panels) makes the economic case for rebuild unambiguous.

The simple framing for entrepreneurs and investors is this. PA is becoming a programmable transaction, and programmable transactions always generate platforms. The companies that build the orchestration, the middleware, the portability, the audit tooling, the PA-aware CDS, the patient-facing transparency tools, the specialty benefit manager modernization, and the data products on top of the new FHIR rails will capture the value that currently evaporates in fax machines, peer-to-peer phone tag, and administrative overhead. Regulatory pressure is not the headwind in this market. It is the forcing function creating the market. Programmable medical necessity is a new category of health technology infrastructure, and 2027 is the year the rails get pressure-tested. The practical question is not whether to build here but which layer to build at: rails, trains, or stations. Each has a different capital profile, a different moat, and a different buyer, and each is wide open.

Subscribe to Thoughts on Healthcare Markets and Technology

By Special Interest Media · Hundreds of paid subscribers

We cover broad topics at the intersection of health tech, health policy, health entrepreneurship and health investing.

1 Like • 1 Restack

← Previous

Discussion about this post

Comments Restacks

Write a comment...