cognizant

CX CRM Advisory | Healthcare

Experience-led utilization management: Human + digital with Agentforce

Utilization Management is breaking at every handoff.
The fix is not more autonomy. It is a human + AI blended workforce.

Authors

Neetika
Mehulkumar Pandya
Rutuja Narkhede
Gaurav Gauba

Mentors

Andrew Bhosale
Samir Tamhane
Sougata Majumdar

Introduction

Utilization Management (UM) was built to confirm care is necessary and compliant. Instead, it has become healthcare's biggest source of friction. A physician and their staff spend 13 hours every week[1] on prior authorization work, and more than 80% of the denials they appeal are overturned[2], yet fewer than one in nine denials is ever challenged. At the same time, UM places a significant administrative burden on providers, who complete an average of 39 prior authorizations per week[3]. These challenges translate into unwarranted delays in care.

AI has proven it can cut administrative burden and speed up UM, but with 60% of physicians[4] linking it to more denials and trust in autonomous AI dropping by 16 % in a year[5], it cannot yet be trusted to act without human oversight.

This point of view, drawn from three advisory deep-dives, sets out a practical response: a blended human + AI workforce, powered by Salesforce Agentforce, with human judgment kept in the loop at every high-stakes decision.

1.1 The UM journey and where it breaks

Intake

Case, eligibility and clinical details are captured at first contact.

Prior authorization

Medical necessity is checked against payer criteria before care begins.

Concurrent

Ongoing treatment is reviewed against criteria while care is underway.

Retrospective

Completed claims are audited against clinical documentation.

Appeals

Disputed decisions are contested, reconsidered and resolved.

Every handoff adds manual work, waiting and doubt, and the same pattern repeats at each of the four transitions.

1.2 Why this matters now

Regulation has set a clock: most provisions of the CMS-0057-F Interoperability and Prior Authorization Final Rule take effect Jan. 1, 2026, and its prior-authorization API requirements follow on Jan. 1, 2027 (Centers for Medicare and Medicaid Services). The market is moving: US spend on UM solutions is projected to nearly double, from $613M in 2024 to $1.2B by 2030 (Grand View Research). And patience is exhausted: about seven in 10 insured adults now call prior authorization a burden (KFF polling).

Our understanding

1.3 What the evidence shows

93%

of physicians say prior authorization delays necessary patient care.

AMA Physician Survey, 2024[6]

82%

report that PA delays cause patients to abandon treatment.

AMA Physician Survey, 2024[6]

43% to 27%

decline in trust in fully autonomous AI within one year.

Capgemini Research Institute, 2025[5]

61%

of physicians fear health plans' use of AI is increasing denials.

AMA Physician Survey, 2024[6]

1.4 Human judgment at the center of AI-driven UM

Automating an opaque process only produces faster denials, and the high overturn rate shows that speed without trust simply moves the failure downstream. The question is not where AI can act but where human judgment must, and how to support that judgment at the moments it matters most.

Our understanding

UM does not need more autonomy; it needs deliberately engineered trust. The winning model is a blended Human + AI workforce in which Agentforce keeps routine work moving and assembles the evidence, while clinicians retain judgment and sign-off at every high-stakes hand-off.

Business solution

2.1 Use case 1: Dynamic policy adaptation

The problem: decision criteria age faster than they are refreshed

Payer medical policy changes continuously, in formats built for people rather than systems. The criteria libraries that drive live decisions refresh on slower, periodic cycles, so a share of every month's denials is issued against rules that are already out of date. This capability runs behind every decision and closes that gap before it becomes a wrongful denial.

10+ Policy updates/year[7]

Commercial payers typically update policies 10+ times annually, while Medicare issues regular and ad hoc revisions.

58,000+ policy sources[7]

Policy monitoring solutions track 58,000+ web pages across 475 insurers to detect coverage and authorization changes

What changes: Criteria stay current at decision time

- The agent ingests real-time payer policy feeds and regulatory bulletins, API-accessible rule libraries or even scrape payer portals (when necessary) to automatically apply the most current criteria.
- It updates UM decision engines so that criteria reflect the latest payer policy, closing the gap between a policy change and its application in live decisions.
- It applies today's criteria at decision time rather than last quarter's snapshot, and it flags every changed rule for review.

Where judgment stays human

- AI proposes rule updates and flags ambiguous or conflicting policy language for review; it never activates a change unsupervised.
- Compliance and clinical policy analysts review and approve every change before it goes live.
- Medical directors keep sign-off on high-impact changes and adjudicate edge cases, since misapplied policy directly affects patient access and compliance exposure.

What makes this difficult to deliver

- No two payers publish policy in the same way, so normalizing policy content is a substantial workstream in its own right.
- Every AI-parsed rule must be validated for accuracy before it goes live, with full version control and an audit trail.
- Policy provenance, version control, and regulatory traceability must be maintained across every update.

How it runs on the platform

SALESFORCE · Health Cloud + Agentforce

Data Cloud ingests payer feeds; Agentforce drafts rule updates from parsed policy language; Business Rules Engine and Health Cloud UM auto-apply criteria; Flow deploys versioned rules with an auditable change history.

MICROSOFT · Azure AI + Copilot

Azure AI Document Intelligence parses policy documents; Copilot Studio and Power Automate route proposed changes for analyst approval; Microsoft Purview logs the audit trail for regulators.

Value at stake and how to measure success

$515M

Enables annual savings across the healthcare ecosystem, benefiting payers and medical providers.

CAQH Index 2024[8]

14 minutes saved

Average time saved per authorization using electronic workflows.

CAQH Index, 2024[8]

Success measures

- Track the latency from policy change to live criteria.
- Measure the share of criteria parsed and approved.
- Monitor overturns tied to stale or misapplied rules.

Business solution

2.2 Use case 2: Appeal management

The problem: most winnable appeals are never filed

More than 80%[9] of appealed denials are overturned, yet barely one in nine denials is ever challenged, because assembling the evidence is slow, manual and clinician-intensive. This capability picks up the moment a denial lands and turns appeal drafting into an evidence-grounded, audit-ready workflow, so that every decision is defensible before it is contested.

88.5% Never appealed [9]

Only 11.5% of denied Medicare Advantage prior authorization requests were appealed in 2024, despite 80.7% of appealed denials being overturned.

$50M [10]

revenue loss reported by 35% of hospitals due to denied claims, according to the American Hospital Association report.

What changes: Evidence is assembled before submission

- The agent parses medical records, prior submissions and denial letters, matching clinical evidence against payer guidelines and medical policy.
- It drafts an evidence-grounded, citation-backed appeal letter in which every claim is traceable to a specific source.
- It builds the evidence file once, at the time of the original decision, so the appeal no longer starts from scratch.

Where judgment stays human

- AI assembles and drafts the first-pass appeal, with supporting evidence highlighted against the exact guideline it matches.
- Clinical and UM staff review, edit and validate the medical necessity rationale before submission.
- Clinicians own the narrative and the signature; accountability never leaves the human, since appeal outcomes carry legal and patient-safety weight AI alone should not own.

What makes this difficult to deliver

- Clinical evidence is often trapped in unstructured notes, faxes and scanned files that must be digitized reliably.
- There is zero tolerance for hallucinated or mis-cited evidence, so provenance is required on every line of the rationale.
- Teams must manage payer formats and tight deadlines while earning clinician trust in AI-assisted content.

How it runs on the platform

Salesforce · Health Cloud + Agentforce

Intelligent Document Automation pulls case history and drafts appeals inside Health Cloud case management. Unified data enables appeal reviewers to access the clinical evidence, member information and authorization details needed to support an appeal.

Microsoft · Azure AI + Copilot

Microsoft 365 Copilot with Azure OpenAI Service generates appeal narratives from case data. It enables the UM team to collaboratively review a comprehensive summary, thus reducing time spent in understanding a member's prior history and providing clinician sign-off.

Value at stake and how to measure success

95%

of appealed MA skilled nursing facility (SNF) admission denials were overturned.

HHS OIG, 2026 [11]

15%

improvement in denial overturn rates achieved through AI-enabled clinical appeal letters.

MIT Technology Review, 2025[12]

Success measures

- Track appeal cycle time from denial to submission.
- Measure the share of appeal claims with linked sources.
- Monitor first-level overturn and clinician edit rates.

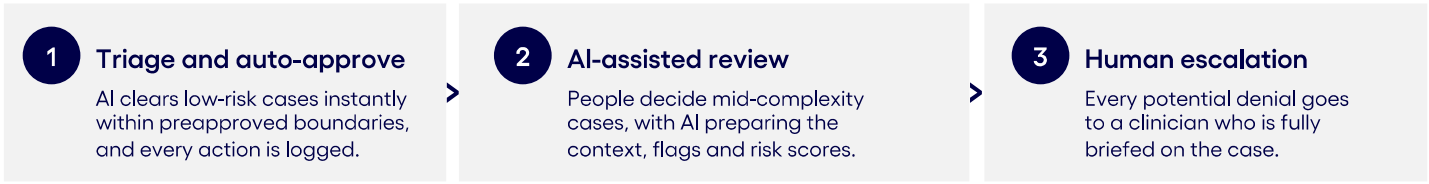
Our point of view: The trust-tiered UM model

Trust is not a tax on the value case. It is the value case itself. The next wave of UM value comes from giving people better information at the moments their judgment matters most.

3.1 The trust-tiered UM model

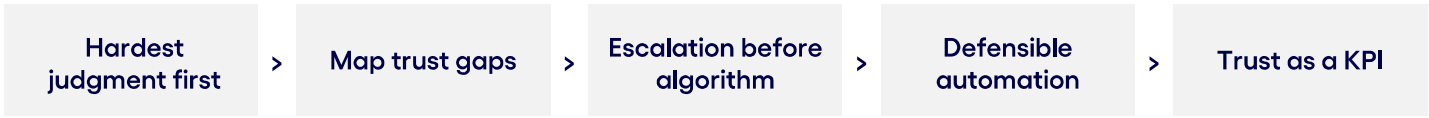
Our point of view is that the operating model, not the algorithm, is where trust is won. We call this operating model the trust-tiered UM model. Work splits across three tiers, so that automation earns its autonomy and people concentrate where the stakes are highest.

Exhibit 3 The trust-tiered UM model: Three tiers of human-in-the-loop control



Three design principles distinguish the model. It is governed, not gated: AI acts within preapproved boundaries at every tier, so speed never outruns accountability. It places judgment where it counts: human review concentrates on the cases with the highest clinical and compliance stakes. And it is auditable by design: every AI action, escalation and override is logged, so the model earns regulatory and member trust over time. Executive sentiment already sits here, since 90% of executives view human involvement in AI workflows as positive or cost-neutral, which makes the blended model the design rather than a compromise (Capgemini Research Institute, 2025)[5].

3.2 From model to motion



Adopting the model does not require a big-bang program. It requires sequencing. Leaders should start where judgment is hardest, such as policy adaptation and appeal defense over further prior authorization automation. They should map where trust breaks down across the five UM stages using patient and provider journey data, and design the escalation path before selecting a model, so that governance, thresholds and audit trails shape the build rather than follow it. They should automate only what they can defend, with provenance built in from day one and FHIR-first interoperability treated as a launch condition. And they should instrument trust as a KPI, tracking override rates and physician confidence alongside cycle time and cost, because UM's credibility problem is as large as its efficiency problem.

3.3 Operationalizing the model: The HAC decision framework

To move the trust-tiered model from principle to practice, we can utilize Cognizant's proprietary HAC (Human-Agent Collaboration) audit framework to bring rigor to agentification decisions. It scores each task against multiple decision parameters. These scores roll up into a single HAC score, which maps every task directly onto three tiers: Agentified, human-in-the-loop or 100% human.

For UM leaders, this turns tiering from a judgment call into a defensible, repeatable decision. It sequences where to deploy agents first, surfaces systemic gaps and gives compliance teams a documented, auditable rationale for every human checkpoint retained.

50-75%

of manual PA tasks automated with AI, boosting efficiency and allowing clinicians to focus on complex cases.

Mckinsey [14]

10 hours

per week that healthcare teams estimate agents could hand back to them,

Salesforce healthcare survey, 2025 [13]

Conclusion and references

4.1 Conclusion

Three independent deep dives converge on the same diagnosis: utilization management's crisis is one of trust rather than throughput. Adding more autonomy to an already opaque process only produces faster denials. The winning model, which we call the trust-tiered UM model, is a blended human + AI workforce in which Agentforce keeps policy criteria current and drafts every appeal with citation-backed evidence, while clinicians and medical directors retain judgment, signoff and accountability at every high-stakes handoff. Speed is a by-product of this design, never its goal. Organizations that move now will meet the 2026 and 2027 regulatory deadlines with an operating model that patients, providers and regulators can trust.

Utilization management does not need more autonomy. It needs deliberately engineered trust, with humans owning judgment at every high-stakes handoffs.

4.2 References

1. AMA, Fixing Prior Auth: Nearly 40% of Prior Authorizations a Week Away. https://www.ama-assn.org/practice-management/prior-authorization/fixing-prior-auth-nearly-40-prior-authorizations-week-way

2. KFF, Prior Authorization Topic Tracker, 2024–2025. https://www.kff.org/tag/prior-authorization/page/2/

3. AJMC, "AMA Survey Highlights Growing Burden of Prior Authorization on Physicians, Patients. https://www.ajmc.com/view/ama-survey-highlights-growing-burden-of-prior-authorization-on-physicians-patients

4. AMA, Press Release "Physicians Concerned AI Increases Prior Authorization Denials. https://www.ama-assn.org/press-center/ama-press-releases/physicians-concerned-ai-increases-prior-authorization-denials

5. Capgemini Research Institute, Rise of Agentic AI, 2025. https://www.capgemini.com/ch-en/insights/research-library/ai-agents/

6. AMA, 2024 Prior Authorization Physician Survey (published February 2025). Cited for: 93% report care delays; 82% report treatment abandonment; 13 hours per week on PA; 61% concerned AI is increasing denials; 40% employ staff dedicated to PA. ama-assn.org

7. MD Clarity, Payer Policy Changes: How Providers Catch Updates on Time to Avoid Denials. mdclarity.com

8. CAQH 2024 Index Report Key Takeaways FINAL.pdf

9. AHA News, Analysis: MA insurers made 53 million prior authorization determinations in 2024 activity. aha.org

10. HFMA, Health Systems Start to Fight Back Against AI-Powered Robots Driving Denial Rates Higher. https://www.hfma.org/revenue-cycle/denials-management/health-systems-start-to-fight-back-against-ai-powered-robots-driving-denial-rates-higher/

11. Medicare Advantage Organizations Overturned Nearly All Appealed Prior Authorization Denials for Skilled Nursing Facility Admission, Raising Concerns About Initial Denials | Office of Inspector General | Government Oversight | U.S. Department of Health and Human Services. oig.hhs.gov

12. MIT Technology Review, "From Pilot to Scale: Making Agentic AI Work in Health Care," 2025. https://www.technologyreview.com/2025/08/28/1122623/from-pilot-to-scale-making-agentic-ai-work-in-health-care/

13. Salesforce, Agentforce for Health announcement and healthcare workforce survey, 2025. Cited for: self-reported estimates of about 30% lower physician administrative burden and about 10 hours saved per week. salesforce.com

14. McKinsey, AI ushers in next-gen prior authorization in healthcare | McKinsey | mckinsey.com