



Advancing Healthcare Administrative Automation: A Multi-Agent Framework for Prior Authorization and FHIR Interoperability

Yao Fehlis*

Independent Researcher, KUNGFU.AI, Austin, United States of America

Correspondence: Yao Fehlis, Independent Researcher, KUNGFU.AI, Austin, United States of America, E-mail: aprilcui11@gmail.com; DOI: <https://doi.org/10.56147/jbhs.3.4.165>

Citation: Fehlis Y (2026) Advancing Healthcare Administrative Automation: A Multi-Agent Framework for Prior Authorization and FHIR Interoperability. J Biol & Heal Sci 3: 165.

Abstract

Prior Authorization (PA) remains one of the most administratively burdensome processes in United States healthcare, delaying medically necessary care and generating billions of dollars in annual administrative costs. The CMS Interoperability and Prior Authorization Final Rule (CMS-0057-F), published in January 2024, requires impacted payers to expose standardized HL7 HL7® FHIR® application programming interfaces by January 1, 2027 and enforce shorter decision timelines beginning January 1, 2026: 72 hours for urgent requests and seven calendar days for standard requests. This regulatory shift creates a strong opportunity for agentic artificial intelligence systems capable of autonomously executing multi-step administrative workflows, invoking external tools, synthesizing heterogeneous clinical documents and generating structured, auditable outputs. We present a general-purpose agentic framework for healthcare administrative automation built around HL7® FHIR®-native data exchange, the Model Context Protocol (MCP) for secure tool orchestration and deterministic structured outputs designed for regulatory traceability. The framework comprises three cooperating agents: an Intake Agent for patient and coverage extraction, a Clinical Review Agent for multi-format document synthesis and a Policy Evaluation Agent for evidence-to-criterion mapping with confidence scoring. Each agent is grounded by typed FHIR tooling and governed by strict output schemas. We describe design principles for accuracy, latency, auditability and HIPAA compliance; map system components to CMS-0057-F requirements; and demonstrate the framework on a representative prior authorization case for lumbar spinal fusion. The results show how the multi-agent system assembles clinical evidence, evaluates policy criteria with calibrated confidence and produces structured outputs aligned with CMS-0057-F transparency requirements.

Keywords: Agentic AI; Large language models; Prior authorization; Healthcare interoperability; FHIR; CMS0057-F; Clinical decision support; Model context protocol

Received date: August 28, 2026; **Accepted date:** September 06, 2026; **Published date:** September 18, 2026

Introduction

Healthcare in the United States is caught between two converging pressures. On one side is an administrative apparatus of extraordinary complexity: Prior authorization alone requires an average of 16.4 hours of physician and staff time per week per practice, involves denial rates that vary from under 4% to over 13% across major payers and contributes to measurable delays in patient care, including

in oncology and surgery contexts where timeliness is clinically critical [1,2]. On the other side is an emerging generation of artificial intelligence systems capable of autonomous, multi-step reasoning over heterogeneous data sources systems that can read clinical notes, parse structured FHIR bundles, extract text from PDF attachments, cross-reference payer policy criteria and produce structured, auditable summaries at a speed and consistency that human reviewers cannot match.



The catalyst bridging these two realities is regulation. Published on January 17, 2024, the CMS Interoperability and Prior Authorization Final Rule (CMS-0057-F) establishes hard deadlines for federally impacted payers Medicare Advantage organizations, state Medicaid and CHIP programs and Qualified Health Plan issuers on federally-facilitated exchanges to implement HL7® FHIR® R4compliant APIs for patient access, provider access, payer-to-payer exchange and prior authorization support by January 1, 2027 [3]. Simultaneously, it requires that standard PA decisions be issued within seven calendar days and urgent decisions within 72 hours beginning January 1, 2026. These requirements represent a fundamental shift: Not just better data exchange, but measurably faster administrative throughput.

Achieving that throughput with existing human-centric workflows is, for most payers, operationally infeasible at scale. The average PA case involves 3-7 clinical documents, multiple format types (PDF, CCDA XML, FHIR bundles, questionnaire responses) and policy evaluation against 4-12 clinical criteria that vary by procedure, plan and clinical context. Manual review is inherently bottlenecked by reviewer availability, shift constraints and the cognitive overhead of synthesizing across heterogeneous sources. The compliance window is tight.

This paper argues that agentic AI LLM-based systems that can plan multi-step workflows, invoke typed external tools and produce structured reasoning traces is the correct architectural response to this regulatory mandate. Unlike prior generations of healthcare AI (static classification models, keyword-based clinical NLP, rules engines), agentic systems can adapt their data-gathering strategy to case-specific evidence availability, process any combination of document types encountered in a real-world PA case and explain their reasoning in terms directly mapped to payer policy criteria and regulatory requirements.

We make the following contributions:

- We present a general-purpose agentic framework for healthcare administrative automation built around three cooperating LLM-based agents with typed FHIR tooling and MCP-mediated payer system integration.
- We define a healthcare agent design taxonomy characterizing the distinct properties required of agents operating in regulated clinical-administrative environments, distinguishing them from general-purpose agentic systems.
- We provide a systematic mapping of agentic system components to CMS-0057-F compliance requirements, clarifying which regulatory obligations are amenable to agentic automation and
- which must remain under human accountability.
- We provide an illustrative end-to-end demonstration of the framework applied to a representative prior

authorization case, tracing evidence assembly, criterion evaluation and structured output generation across all three cooperating agents.

- We discuss safety, auditability and HIPAA alignment considerations that distinguish responsible deployment of agentic AI in healthcare administrative workflows from general-purpose automation.

The remainder of the paper is organized as follows. Section 2 provides background on PA administrative burden, CMS-0057-F, FHIR standards and LLM-based agentic systems. Section 3 surveys related work. Section 4 characterizes the agentic AI paradigm shift in healthcare administration. Section 5 describes the proposed framework architecture. Section 6 presents an illustrative case study. Section 7 maps the framework to CMS-0057-F requirements. Section 8 discusses limitations and broader implications. Section 9 concludes.

Background

The prior authorization crisis in US health-care

Prior authorization is a payer-imposed requirement that providers obtain approval before delivering certain procedures, medications or services to covered patients. While intended as a cost-control and appropriateness-verification mechanism, PA has become a significant source of administrative burden and a documented contributor to care delays. A 2022 American Medical Association survey found that 93% of physicians reported that PA delays care and 34% reported that PA-related delays had led to a serious adverse event for a patient [1].

The administrative cost of PA is substantial. The Council for Affordable Quality Healthcare (CAQH) estimates that a single electronic PA transaction costs the healthcare system approximately \$10.97 on the provider side when handled manually, vs. \$5.79 electronically a gap that persists because most PA workflows remain predominantly manual or semi-manual even where electronic channels exist [4]. With nearly 50 million prior authorization requests submitted to Medicare Advantage plans alone in 2023 and many more across commercial and Medicaid plans, the aggregate administrative waste is measured in the tens of billions of dollars annually [2]. From a data complexity standpoint, PA cases routinely involve:

- Structured FHIR resources (claim, coverage, patient, practitioner, condition, observation, procedure, medication request)
- Unstructured clinical notes and letters of medical necessity (PDF)
- Semi-structured clinical documents (HL7 CDA/CCDA: problem lists, medication histories, procedure notes, diagnostic results)



- Questionnaire responses capturing provider attestations
- Payer-specific medical policies defining criterion sets for each procedure code.

Assembling and evaluating this evidence against policy criteria is inherently a multi-step, multimodal reasoning task precisely the class of task for which agentic AI systems offer the greatest advantage over both human reviewers and narrower automation approaches.

CMS-0057-F: Regulatory mandate for interoperability

The Centers for Medicare & Medicaid Services (CMS) finalized the Interoperability and Prior Authorization rule (CMS-0057-F) on January 17, 2024. The rule applies to Medicare Advantage organizations, Medicaid and CHIP fee-for-service programs, Medicaid and CHIP managed care plans and Qualified Health Plan issuers on federally-facilitated exchanges (collectively, impacted payers) [3].

Core API requirements

CMS-0057-F mandates implementation of five HL7® FHIR® R4 APIs, all required in production by January 1, 2027:

- **Patient access API:** Adds prior authorization data to the existing Patient Access API, including authorization status, relevant clinical data and reasons for denial. Members and authorized third-party apps access claims, encounters and PA information *via* FHIR.
- **Provider access API:** Enables in-network treating providers to retrieve member data for treatment purposes *via* FHIR R4 endpoints, including both individual and bulk access, contingent on accurate provider-patient attribution lists.
- **Provider directory API:** A publicly accessible, unauthenticated FHIR directory of contracted providers. Payers must reflect provider roster changes within 30 calendar days of the change.
- **Payer-to-payer API:** Transfers a member's data when coverage changes to maintain continuity of care, covering five years of clinical history and active (non-denied) prior authorization data.
- **Prior authorization API:** Requires impacted payers to support (a) checking whether PA is required for a given item or service, (b) surfacing documentation requirements, (c) electronic PA submission and (d) electronic PA decisions. CMS recommends alignment with the HL7 Da Vinci Prior Authorization Support (PAS) Implementation Guide, though alternative FHIR R4-compliant implementations are permitted.

Decision timeline requirements

Effective January 1, 2026, Medicare Advantage

organizations, Medicaid and CHIP programs and Medicaid and CHIP managed care plans must issue PA decisions within:

- 72 hours for urgent (expedited) requests
- 7 calendar days for standard requests.

These requirements do not apply to QHP issuers on federally-facilitated exchanges. The timelines are substantially shorter than current practice in many plans and effectively require automation of significant portions of the PA intake, triage and initial clinical review workflow to be sustainable at scale.

Transparency and reporting

Impacted payers are additionally required to report PA metrics publicly, including approval rates, average decision times and denial reasons broken down by request type (standard vs. expedited) and by items and services subject to PA. This transparency requirement creates regulatory exposure for payers that cannot demonstrate efficient, evidence-grounded decision-making further incentivizing automation of PA workflows with auditable reasoning chains.

Implementation timeline

- January 1, 2026: Operational requirements (decision timelines, denial notification standards)
- January 1, 2027: API implementation requirements (all five APIs in production)

FHIR standards and the Da Vinci ecosystem

HL7® FHIR® R4 (Release 4) is the current stable version of the HL7 international standard for clinical data exchange. FHIR resources model clinical and administrative entities as JSON or XML documents with defined schemas and RESTful API conventions. The HL7 Da Vinci Project has produced a suite of implementation guides (IGs) that define standardized FHIR workflows for the PA lifecycle: the Prior Authorization Support (PAS) IG for PA submission and response; the Coverage Requirements Discovery (CRD) IG for surfacing documentation and authorization requirements at the point of care; and the Documentation Templates and Rules (DTR) IG for distributing procedure-specific documentation questionnaires from payers to providers. Together these IGs define the data substrate on which an agentic PA system operates. Agents that consume and reason over standard FHIR bundles are inherently portable across payers, EHR systems and care settings a critical property for scalable deployment.

Agentic AI and LLM tool use

The term agentic AI describes AI systems that can autonomously plan and execute multi-step workflows, invoke external tools or APIs, observe the results of those



invocations and update their subsequent plan based on retrieved information all in service of a high-level objective [5,6].

The enabling mechanism for agentic behavior in LLMs is tool use (also called function calling): the ability of a language model to emit structured tool invocation requests as part of its generation, which the host system executes and feeds back into the model's context. Modern frontier models including GPT-4o, Claude 3.x/4.x and Gemini 1.5/2.x support rich tool-use interfaces with typed parameter schemas, enabling reliable invocation of external functions, APIs and data sources.

Healthcare presents a particularly demanding setting for agentic deployment. The relevant literature identifies several properties that distinguish healthcare agent requirements from general purpose agentic tasks [7-9]:

- **Regulatory accountability:** Every agent decision must be traceable to specific evidence, with documented reasoning that can be reviewed by human clinicians and regulators.
- **Structured output fidelity:** Downstream systems (EHRs, payer platforms, reporting pipelines) require structured JSON or FHIR resources, not natural language narrative.
- **Multi-format document comprehension:** Clinical evidence arrives as PDFs, CCDA XML, FHIR bundles, questionnaires and free-text notes sometimes all in a single case.
- **Calibrated uncertainty:** Agents must indicate when evidence is absent, ambiguous or contradictory not merely when it is present.
- **Deterministic safety:** Clinical and administrative decisions require low-temperature generation and schema-constrained outputs to prevent hallucination of clinical evidence.

Recent benchmarking work demonstrates that current frontier LLMs are capable of structured clinical reasoning and tool-use in PA-relevant tasks, but most published work focuses on diagnostic reasoning rather than the administrative-clinical interface that PA occupies [10]. This paper addresses that gap.

Related Work

LLM-based clinical decision support

Substantial recent work has demonstrated the clinical reasoning capabilities of frontier LLMs on diagnostic tasks clinical note summarization and medical question answering [11-13]. However, most of this work focuses on clinical reasoning rather than the administrative-clinical interface that PA occupies a distinction that matters because PA requires evidence retrieval, policy application and structured output as much as clinical knowledge.

Agentic AI in healthcare

Recent surveys and position papers characterize the emerging landscape of agentic healthcare AI [7,8]. Anand et al. propose a HIPAA-compliant agentic framework with dynamic policy enforcement [14]. Alkhalaf et al. describe an MCP-based multilingual healthcare agent [15]. Lee et al. present H-AdminSim, a multi-agent simulator for hospital administrative workflows with FHIR integration, demonstrating the value of purpose-built evaluation infrastructure for LLM-based administrative automation [16]. In a sister domain, prior work on multi-agent systems for drug discovery laboratory automation and cycle-time analysis provides precedent for the cooperative multi-agent designs that we extend to the prior authorization setting [17-19]. Our work complements this body of literature by providing the first systematic mapping of agentic AI components to CMS-0057-F regulatory requirements, grounding the framework in the specific FHIR APIs and decision timelines mandated by the rule.

Prior authorization automation

Existing PA automation products rely primarily on rules engines and optical character recognition. Pandey et al. propose a multi-agent LLM system for medical necessity justification that decomposes PA checklist evaluation across specialized agents, demonstrating strong criterion-level accuracy [20]. Our framework extends this direction with a FHIR-native data layer, MCP tool orchestration and explicit CMS-0057-F regulatory alignment properties absent from their system.

Bedi et al. introduce Health Admin Bench, a benchmark of 135 healthcare administration tasks across simulated GUI environments; their finding that the best computer-use agent achieves only 36.3% end-to-end task success on PA workflows underscores the difficulty of the problem and motivates our API-native approach, which avoids the brittleness of GUI-based automation entirely [21]. Commercial systems such as Cohere Health and Olive AI have deployed production PA automation, but their architectures and evaluation methodologies are not publicly documented.

FHIR-based interoperability and MCP

The Da Vinci project has produced implementation guides that provide the data foundation for FHIR-native PA workflows [22]. Ehtesham et al. present an open-source MCP-FHIR framework integrating LLMs with FHIR R4 EHR data *via* the Model Context Protocol for clinical decision support, establishing the feasibility of the FHIR+MCP integration pattern [23]. Our framework applies this pattern to a multi-agent cooperative architecture for PA, adding heterogeneous clinical document handling (Questionnaire Response, PDF, HL7 CDA/CCDA) and alignment to the Da Vinci PAS, DTR and CRD implementation guides under CMS-0057-F.



The Agentic AI Paradigm Shift in Healthcare Administration

From static automation to adaptive orchestration

Healthcare administrative automation has historically relied on three technological layers: (1) rules engines that encode deterministic decision logic (*e.g.*, CPT code eligibility tables, coverage criteria lookup), (2) Optical Character Recognition (OCR) and template-based extraction for document processing and (3) Robotic Process Automation (RPA) for UI-driven workflow execution. Each layer addresses a narrow slice of the PA problem and fails at the boundaries: rules engines cannot handle novel criteria, OCR fails on variable clinical document structure and RPA breaks with any UI change. LLM-based agentic systems address what these layers cannot. Rules engines, OCR pipelines and RPA remain useful for well-defined deterministic tasks eligibility lookups, document digitization, legacy system integration but they fail at the reasoning layer that PA requires: synthesizing heterogeneous evidence, evaluating criteria against ambiguous clinical language and adapting when expected data is absent. A single agent with access to typed tools can:

- Decide which documents to retrieve based on case

context

- Choose how to process each document type dynamically
- Extract and synthesize evidence across multiple sources simultaneously
- Evaluate criteria using natural language reasoning grounded in retrieved data
- Adapt its strategy when expected evidence is absent (*e.g.*, request additional documentation rather than defaulting to denial).

This adaptive orchestration is not merely a performance improvement it represents a qualitative change in what the system can handle at the reasoning layer. A rules engine requires explicit programming for every new procedure code, criterion set and document type. An agentic system requires a well-defined tool interface and a clear task specification; the reasoning generalizes.

Key properties of healthcare agents

We characterize healthcare administrative agents along seven dimensions that distinguish them from general-purpose agents (**Table 1**):

Table 1: Distinguishing properties of healthcare administrative agents vs. general-purpose agentic systems.

Property	General agents	Healthcare admin agents
Output format	Flexible/narrative	Strictly typed FHIR/JSON
Uncertainty handling	Optional	Mandatory (confidence scoring)
Evidence citation	Rarely required	Required per criterion
Regulatory traceability	Not required	Mandatory (CMS, HIPAA)
Failure mode	Graceful degradation	Hard stop+human escalation
Data standards	Ad hoc	FHIR R4, X12, ICD/CPT/HCPCS
Hallucination risk tolerance	Moderate	Near-zero (clinical safety)

These properties impose concrete design requirements: low generation temperature, schema constrained output, mandatory evidence sourcing, graceful escalation paths and strict separation between data retrieval (deterministic) and reasoning (generative).

The model context protocol as interoperability layer

The Model Context Protocol (MCP), proposed by Anthropic and increasingly adopted across the industry, provides a standardized interface for LLMs to invoke external tools *via* a well-defined JSON-RPC protocol [24]. MCP is particularly well-suited to healthcare interoperability contexts because:

- **Tool name spacing:** MCP servers expose named tools with typed parameter schemas, preventing ambiguous invocations.

- **Deployment separation:** The MCP server (owned and operated by the payer or EHR vendor) can be entirely distinct from the AI service (owned by the technology provider), preserving data governance boundaries without requiring full API standardization.

- **Auditability:** MCP tool calls and responses are logged at the transport layer, creating an immutable trace of every data access event critical for HIPAA compliance and CMS-0057-F reporting obligations.

- **Composability:** Multiple MCP servers can be composed into a single agent context, allowing a PA agent to simultaneously access payer UM systems, EHR APIs, clinical knowledge bases and drug formularies without tightly coupling those integrations in application code.

In our framework, payer systems expose their FHIR resources *via* an MCP server, while the AI orchestration layer accesses those resources through the MCP protocol.



This cleanly separates data ownership from intelligence, a boundary that is increasingly important as healthcare AI regulation evolves.

Framework: A FHIR-Native Agentic System for PA Automation

System architecture overview

Figure 1 illustrates the high-level architecture of the proposed framework. The system is organized into four layers:

- **Data layer:** Payer-owned FHIR R4 repositories, UM systems, policy knowledge bases and attachment storage. Accessed exclusively through the MCP server interface.
- **Tool layer:** MCP tools exposing typed FHIR operations (case fetch, attachment retrieval, policy criteria lookup) and local function tools for document processing (PDF extraction, CCDA parsing, FHIR bundle traversal, questionnaire answer extraction).
- **Agent layer:** Three specialized LLM-based agents (Intake, Clinical Review, Policy Evaluation) each operating with a dedicated system prompt, tool subset and output schema.
- **Interface layer:** REST API endpoints for EHR integration, payer workflow platforms and human reviewer UIs. Structured JSON outputs feed downstream FHIR resources and reporting pipelines.

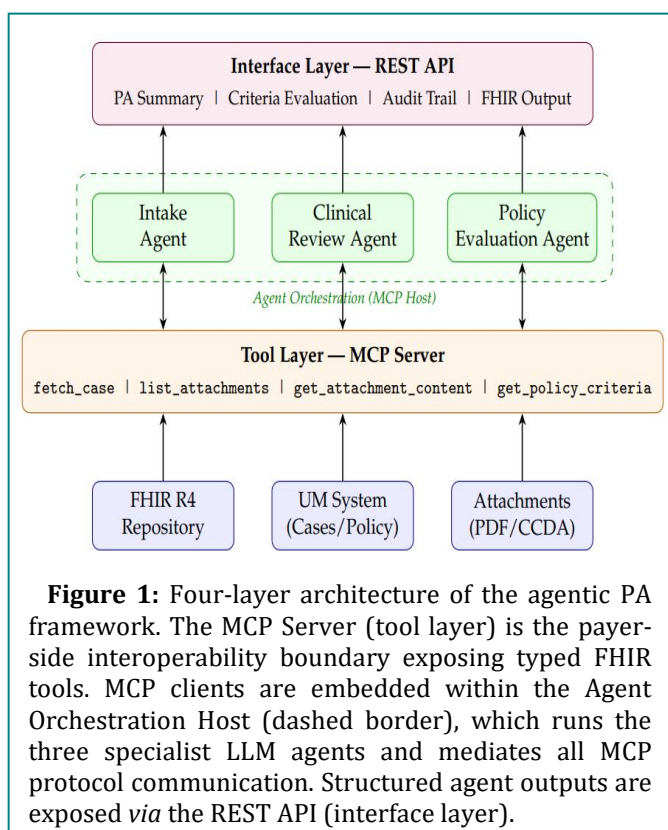


Figure 1: Four-layer architecture of the agentic PA framework. The MCP Server (tool layer) is the payer-side interoperability boundary exposing typed FHIR tools. MCP clients are embedded within the Agent Orchestration Host (dashed border), which runs the three specialist LLM agents and mediates all MCP protocol communication. Structured agent outputs are exposed *via* the REST API (interface layer).

Agent taxonomy

Intake agent

The Intake Agent extracts patient identity, coverage and cost-sharing information from digital insurance card data. It is deliberately narrow operating on structured FHIR input with no clinical reasoning which makes it fast and reliable. Its output is shared with downstream agents without re-fetching, replacing the manual eligibility verification step that consumes meaningful front-desk staff time per encounter [4].

Clinical review agent

The Clinical Review Agent is the most computationally complex component of the framework. It retrieves the complete PA case from the payer system, inventories all attached clinical documents and dynamically selects the appropriate processing approach for each document type extracting text from PDFs, parsing structured sections from HL7 CDA/CCDA documents, reading clinical resources from FHIR bundles and extracting structured answers from Questionnaire Response resources. It synthesizes all extracted evidence into a coherent clinical narrative for downstream policy evaluation. Cases that exceed processing bounds are flagged for human reviewer attention rather than resolved autonomously.

Policy evaluation agent

The policy evaluation agent receives the clinical narrative produced by the Clinical Review Agent and the payer's criterion set for the requested procedure and evaluates each criterion against the assembled evidence. Each per-criterion record contains the criterion identifier and name, a status drawn from {met, not_met, partial, insufficient_evidence}, a confidence level (high, medium or low), a quoted evidence excerpt and a source citation traceable to a specific attachment and section. *insufficient_evidence* distinguishes absence of evidence from evidence of absence a meaningful clinical-administrative distinction. This structure avoids the binary approve/deny framing that would require medical licensure and instead provides a structured evidence summary for human review.

This design deliberately preserves human accountability for final PA decisions while automating the evidence assembly and criterion mapping that currently dominates reviewer time.

Tool design principles

We identify four principles governing tool design in healthcare agentic systems:

- **Principle 1:** Typed schemas over free-form arguments

Every tool accepts strongly-typed parameters (case ID, attachment index, CPT code) rather than free-text queries. This prevents the agent from constructing ill-formed tool



calls and enables the MCP server to validate inputs against a fixed schema.

➤ **Principle 2:** Deterministic retrieval, generative reasoning

Tool implementations are deterministic: given the same inputs, they return the same outputs. Generative reasoning is confined to the LLM agent, which reasons over deterministic data. This separation makes the system debuggable and auditable.

➤ **Principle 3:** Result truncation with priority ordering

Long clinical documents are truncated before being returned to the agent context. Truncation is structured to preserve high-signal content (diagnosis sections, procedure notes, assessment and plan) over low-signal content (administrative headers, boilerplate consent language).

➤ **Principle 4:** Failure modes that escalate gracefully

Tool call failures (network errors, malformed documents, missing case IDs) return structured error objects that the agent can reason about triggering document inventory re-check, requesting human clarification or generating a “missing evidence” flag on the affected criterion.

Output schema and auditability

The complete framework output for a PA case is a structured summary comprising six independently consumable zones: a request header (patient, provider, requested service), coverage context, a clinical narrative, per-criterion evaluations with evidence citations, a missing-items list for evidence gaps and an attachment inventory for audit purposes. Schema-constrained output is enforced at the system prompt level and validated before forwarding to downstream systems, with schema failures escalated to human review rather than silently passed through.

Illustrative Case Study: Lumbar Fusion Prior Authorization

To demonstrate how the framework orchestrates multi-agent evidence assembly in a realistic clinical-administrative context, we trace a single PA case for lumbar spinal fusion (CPT 22612) through the three cooperating agents. This is an illustrative walkthrough, not an empirical benchmark; it shows how the framework resolves the multi-format evidence problem described in Section 2.1. **Figure 2** summarizes the end-to-end flow.

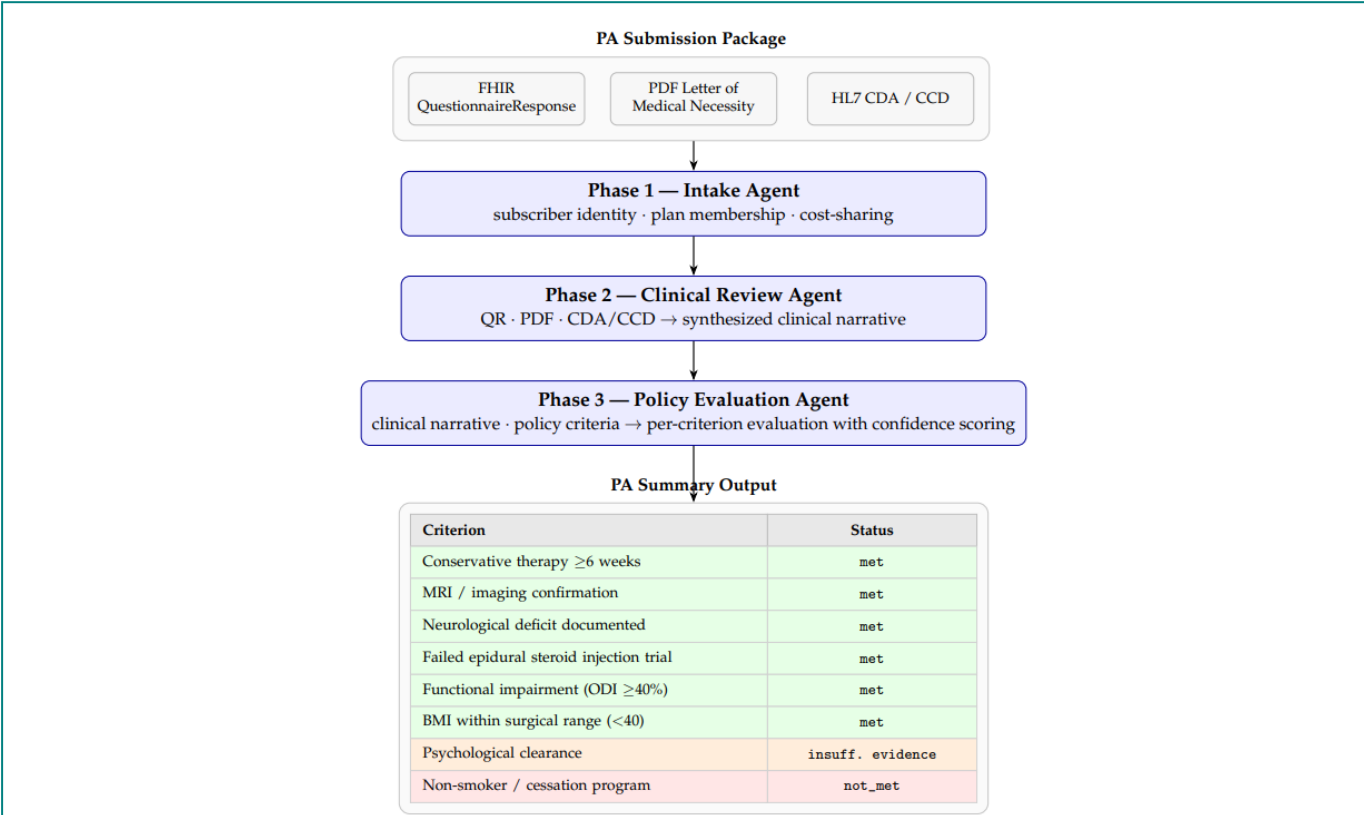


Figure 2: End-to-end case study flow for lumbar spinal fusion PA (CPT 22612). The submission package is processed sequentially through the three agent phases. The PA Summary Output table lists all eight policy criteria with color-coded status: green (met) orange (insufficient_evidence), red (not_met). Six criteria are satisfied; the two flagged items direct human reviewer attention to the specific evidence gaps.



Case context

A 58-year-old patient with chronic lumbar stenosis presents for prior authorization of single-level posterolateral lumbar fusion at L4–L5. The treating orthopaedic surgeon submits the PA request *via* the provider EHR, which routes to the payer UM system through the Da Vinci PAS API. The submitted package contains:

- A FHIR QuestionnaireResponse (14 clinical questions completed by the ordering surgeon)
- A PDF letter of medical necessity (3 pages, free-text narrative)
- An HL7 CDA Continuity of Care Document (CCD) including problem list, procedure history, laboratory results and medication history.

The payer's medical policy for CPT 22612 requires evaluation against a set of mandatory and discretionary clinical criteria.

Phase 1-intake agent

The Intake Agent resolves the patient's subscriber identity, plan membership, effective coverage period and applicable cost-sharing tier in a single call to `parse_coverage_bundle`. The resulting `IntakeSummary` is attached to the case context and made available to both downstream agents without re-fetching. This replaces a manual eligibility verification step that, per industry benchmarks, consumes 5–10 minutes of administrative staff time per case [4].

Phase 2-clinical review agent

The Clinical Review Agent receives the PA case identifier and autonomously orchestrates the following tool call sequence:

- `fetch_case("PA-2026-7742")` returns the FHIR Bundle containing the Claim resource, Coverage reference and DocumentReference pointers to all three submitted attachments.
- `list_attachments("PA-2026-7742")` confirms three documents: A FHIR Questionnaire Response, a PDF clinical note and an HL7 CDA document.
- `extract_questionnaire_answers(bundle)` returns structured Q&A pairs—including conservative therapy duration ("12 weeks of physical therapy and two epidural steroid injections"), functional impairment score ("Oswestry Disability Index: 56%"), BMI ("28.4") and smoking status.
- `extract_pdf_text(pdf_base64)` returns the full text of the letter of medical necessity for evidence indexing.
- `parse_ccda(ccda_base64)` returns structured clinical sections: Problems (L4–L5 stenosis, lumbar

radiculopathy), Procedures (two epidural steroid injections, September 2025 and January 2026), Results (MRI 2026-02-14 confirming bilateral foraminal stenosis with L5 nerve root impingement) and Medications.

The agent synthesizes these heterogeneous sources into a coherent Clinical Summary narrative documenting the patient's 18-month conservative therapy trajectory, radiographic findings, functional impairment and comorbidities relevant to surgical risk.

Phase 3-policy evaluation agent

The Policy Evaluation Agent receives the ClinicalSummary and calls `get_policy_criteria` with CPT code 22612 to retrieve the applicable criteria. It evaluates each against the assembled evidence:

- **Conservative therapy ≥ 6 weeks (met, high confidence):** QR documents 12 weeks of physical therapy; CCDA procedure history confirms two ESI attempts satisfying modality requirements.
- **MRI confirmation of stenosis (met, high confidence):** CCDA results section documents L4–L5 bilateral foraminal stenosis on MRI 2026-02-14.
- **Neurological deficit (met, high confidence):** QR and PDF letter document dermatomal sensory loss and motor weakness grade 3/5 in the left lower extremity.
- **Failed ESI trial (met, high confidence):** CCDA confirms two ESI procedures; QR reports less than 50% sustained relief.
- **Functional impairment (ODI $\geq 40\%$) (met, high confidence):** QR reports ODI 56%.
- **Psychological clearance (insufficient_evidence):** No attachment contains a psychiatric assessment. The agent flags this as a missing-items entry rather than `not_met`.
- **BMI < 40 (met, high confidence):** QR reports BMI 28.4.
- **Non-smoker/cessation plan (not_met, medium confidence):** QR confirms active smoking; no cessation plan is present in any attachment.

The complete output is a structured PA Summary with all six zones populated. The UM nurse reviewer receives this artifact and can focus attention on the two gaps missing psychological clearance and the tobacco use criterion rather than re-reading all three source documents from scratch.

Framework properties illustrated

This walkthrough illustrates four properties of the framework that architecture diagrams alone cannot convey:

- **Adaptive evidence routing:** The agent dynamically



- selected different parsers for each attachment type (JSON questionnaire, PDF text, CCDA XML) without pre-programmed case-routing logic.
- **Cross-source synthesis:** The conservative therapy criterion was evaluated by combining evidence from the QR (duration) and the CCDA (modality confirmation) a task requiring synthesis across two distinct document types and time periods.
 - **Calibrated gap detection:** The agent correctly assigned insufficient_evidence for psychological clearance, distinguishing absence of evidence from evidence of absence an important clinical-

administrative distinction.

- **Structured auditability:** Every criterion evaluation carries an explicit source field traceable to a specific attachment index and section, satisfying CMS-0057-F transparency requirements without post-processing.

Regulatory Alignment with CMS-0057-F

Table 2 maps each major CMS-0057-F requirement to the corresponding framework component that addresses it.

Table 2: Mapping of CMS-0057-F requirements to framework components.

CMS-0057-F requirement	Framework component	Implementation mechanism
Prior authorization API (check PA required; surface documentation needs; electronic submission and decisions)	MCP server+FHIR tool layer	Payer exposes FHIR R4 endpoints <i>via</i> MCP; agent checks coverage rules, surfaces documentation requirements and submits/retrieves PA bundles aligned with the Da Vinci PAS IG (recommended, not mandated)
Provider directory API (public provider roster; changes reflected within 30 days)	FHIR tool layer	Intake Agent queries payer-hosted provider directory to validate treating-provider attribution before PA submission
72h/7-day decision timelines	Clinical review+policy evaluation agents	Agentic orchestration replaces serial human review steps; tool call iteration limits bound processing time; complex cases are flagged for priority human review
Provider Access API (treating provider data access)	Intake agent+fhir tool layer	Provider queries return patient coverage and PA status <i>via</i> structured FHIR bundle
Denial transparency (reason codes, evidence citations)	Policy evaluation agent output schema	not_met criteria include explicit evidence and source fields mapped to denial reason codes
Public reporting of PA metrics	Audit trail (zone 6)	All tool calls, timestamps and criterion evaluations logged for aggregation and regulatory reporting
Payer-to-payer API (member transition data exchange)	FHIR tool layer	Payer FHIR R4 history bundles can be ingested and parsed by Clinical Review Agent during member transitions

HIPAA compliance considerations

Healthcare agentic systems operate on Protected Health Information (PHI) and must comply with the HIPAA Privacy and Security Rules. Our framework addresses HIPAA requirements through:

- **Minimum necessary access:** The Intake Agent accesses only coverage bundles; the Clinical Review Agent accesses only attachments for the specific case under review. Tool schemas enforce case-scoped access (case ID required on all calls).
- **Audit logging:** Every MCP tool call including the requesting agent, timestamp, case ID and response size is logged to an immutable audit trail.
- **No PHI in LLM training:** Framework deployments use inference-only API access under zero data-retention agreements offered by major LLM providers, ensuring that PHI is not used for model training.
- **Data residency:** The MCP server and all PHI storage remain under payer control; only deidentified reasoning traces pass through the AI service layer.

- **Human in the loop:** Final PA decisions are made by licensed reviewers using the agent’s structured summary as a decision-support artifact, not as a binding determination.

Explainability and regulatory auditability

CMS-0057-F’s transparency requirements extend beyond decision timelines to the reasoning behind denials. The framework’s per-criterion output schema provides a natural explainability layer: For each not_met or insufficient_evidence criterion, the output captures exactly what evidence was sought, what was found (or not found) and from which source. This maps directly onto CMS’s requirement that payers provide specific clinical reasons for PA denials, rather than generic codes.

The tool call trace (order of tool invocations, arguments and response summaries) constitutes a process-level audit trail documenting how the evidence assembly was conducted. Payers can retain this trace to demonstrate that their PA review process is evidence-driven and reproducible.



Discussion

Limitations

The framework has not been evaluated on production PA cases; prospective evaluation with a live payer partner is a necessary next step. LLM-based agents can synthesize evidence in ways that are clinically plausible but not always traceable to a specific passage a non-trivial failure mode in a regulated context that retrieval-augmented generation and post-hoc citation verification can mitigate. The case study covers musculoskeletal surgery; generalization to oncology, behavioral health and cardiology PA categories requires additional work, as each introduces distinct document types and criterion structures. Finally, the framework relies on frontier model tool-use capabilities that require version pinning and regression testing before production deployment.

Implications and future directions

CMS-0057-F is one of several converging regulatory pressures alongside state-level PA reform and the ONC HTI-1 Final Rule creating the standardized FHIR data infrastructure that agentic AI requires. The architectural pattern described here (FHIR-native data layer, MCP-mediated orchestration, specialist agents, human review at the final decision boundary) is likely to generalize to adjacent workflows: Claims adjudication, utilization management and care coordination. Future work includes a supervisor agent for cross-agent workflow state management, streaming criterion evaluation for real-time reviewer feedback, adaptive policy retrieval for novel procedure codes and prospective evaluation on de-identified production cases.

Conclusion

CMS-0057-F is a performance mandate as much as a data exchange mandate: 72-hour urgent decision timelines, FHIR API requirements and public reporting obligations create a compliance environment that manual and rules-based workflows cannot sustainably meet at scale. Agentic AI with autonomous multi-step evidence retrieval, calibrated policy evaluation and structured auditable output is architecturally well-suited to this challenge.

We have presented a FHIR-native three-agent framework for PA automation grounded in MCP tool orchestration, schema-constrained output and human accountability at the final decision boundary. An illustrative lumbar fusion case demonstrates adaptive evidence routing, cross-source synthesis and calibrated gap detection aligned with CMS-0057-F transparency requirements. We have mapped each framework component to specific regulatory requirements, clarifying where agentic automation can support compliance and where human accountability must remain. The regulatory timelines are now measured in months; the framework

offers a path toward compliance that preserves both auditability and clinical accountability.

References

1. American Medical Association (2022) Prior authorization physician survey.
2. Kaiser Family Foundation (2023) Medicare advantage prior authorization requests and denial rates, 2022-2023.
3. Centers for Medicare & Medicaid Services (2024) Interoperability and prior authorization final rule (CMS-0057-F). U.S. Department of Health and Human Services.
4. CAQH (2023) CAQH index: Closing the gap. Council for Affordable Quality Healthcare.
5. Wang L, Ma C, Feng X, et al. (2024) A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18: 186345. [Crossref] [Google Scholar]
6. OpenAI (2024) Function calling and tool use in GPT-4 models.
7. Nie X, et al. (2025) A survey of LLM-based agents in medicine: How far are we from Baymax? [Crossref] [Google Scholar]
8. Yuan H, et al. (2025) Agentic large language models for healthcare: Current progress and future opportunities. *Medicine Advances* 3: e70000. [Crossref] [Google Scholar]
9. Rao A, Kim J, Bhatt D, et al. (2024) AI agents in clinical medicine: A systematic review. *JMIR Medical Informatics*. [Crossref] [Google Scholar] [Indexed]
10. Hadi MU, et al. (2025) Agentic AI in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents. [Crossref] [Google Scholar]
11. Singhal K, Azizi S, Tu T, et al. (2023) Large language models encode clinical knowledge. *Nature* 620: 172-180. [Crossref] [Google Scholar] [Indexed]
12. Achiam J, Adler S, Agarwal S, et al. (2023) GPT-4 technical report. [Google Scholar]
13. van Veen D, Van Uden C, Blankemeier L, et al. (2024) Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* 30: 1134-1142. [Crossref] [Google Scholar] [Indexed]
14. Anand V, et al. (2025) Towards a HIPAA-compliant agentic AI system in healthcare.
15. Alkhalaf M, et al. (2025) Agentic-AI healthcare: Multilingual, privacy-first framework with MCP agents. [Google Scholar]
16. Lee J, Son MH, Choi EH (2026) H-AdminSim: A multi-agent simulator for realistic hospital administrative workflows with FHIR integration. *Proceedings of the Conference on Health, Inference and Learning (CHIL) 2026*. [Google Scholar]
17. Fehlis Y, Crain C, Jensen A, et al. (2025) Accelerating drug discovery through agentic AI: A multi-agent approach to laboratory automation in the DMTA cycle. [Google Scholar]
18. Fehlis Y, Crain C, Jensen A, Watson M, Juhasz J, et al. (2025) Technical implementation of tippy: Multi-agent architecture and system design for drug discovery laboratory automation.
19. Fehlis Y (2025) Uncovering bottlenecks and optimizing scientific lab workflows with cycle time reduction agents.



20. Pandey H, Amod A, Shivang (2024) Advancing healthcare automation: Multi-agent system for medical necessity justification. Proceedings of the BioNLP Workshop, ACL 2024. [Crossref] [Google Scholar]
21. Bedi S, Welch R, Steinberg E, Wornow M, et al. (2026) Health admin bench: Evaluating computer-use agents on healthcare administration tasks. [Google Scholar]
22. HL7 Da Vinci Project (2023) Da Vinci Prior Authorization Support (PAS) implementation guide, STU 2.0.1.
23. Ehtesham A, Singh A, Kumar S (2025) Enhancing clinical decision support and EHR insights through LLMs and the Model Context Protocol: An open-source MCP-FHIR framework. [Crossref] [Google Scholar]
24. Anthropic (2024) Model context protocol specification.