

Prior Authorization Decision Support: AI Product PRD

Date: October 16, 2026 | **Author:** Sankar Kumar Palaniappan | **Status:** Draft | **Version:** 1.0

Scope note: the working prototype (React + Vite, scripted data, six cases) is the MVP described here. Production-scale items are labeled as target state. All figures are targets or hypotheses to validate, not measured results.

1. Problem (AI-Specific)

- **Problem:** Utilization management (UM) nurses decide whether a requested inpatient admission meets level-of-care criteria. The evidence sits in 14+ page faxed packets (cover sheet, H&P, cardiology clearance, labs, surgeon orders). Finding each fact, checking it against layered coverage policy, and documenting the rationale is slow, and decisions run against regulatory clocks (CMS: 7 days standard, 72 hours expedited).
- **Why AI:** Inputs are unstructured and inconsistent across providers, and the same fact appears in different words ("multi-day stay" vs. "2-midnight admission"). Rules alone cannot read the packets. A model can extract and map facts, but only if every claim is traceable and uncertain cases reach a human.
- **Status quo:** Manual read-through and copy-paste into a checklist. Inconsistent depth of review, no uniform citation trail, and clarification requests to providers are written ad hoc.

2. User

- **Primary persona:** UM nurse reviewer. **Secondary:** medical director (handles escalated and denial decisions). **Tertiary:** auditor or compliance reviewer reading the trail afterward.
- **Job to be done:** Reach a defensible determination (approve, request clarification, escalate) quickly, with evidence to back it.
- **AI interaction pattern:** Suggestion-and-approve. The AI extracts, maps, and drafts; a human makes every determination. No auto-approval or auto-denial.
- **Trust requirements:** High. Every checkmark must be verifiable in one click against the source page and sentence. Denials are made only by the medical director.

3. Core Metric

- **North Star:** Time from packet receipt to a defensible determination, with zero uncited findings.
- **Baseline:** Not yet measured. First task in the pilot is to time the current manual process on a sample of cases.
- **Target (hypothesis):** At least 40% reduction in nurse review time on straight-through cases, with 100% of findings carrying a source citation.

- **Leading indicators:** Citation click-through rate; share of cases routed to each review tier; clarification round-trip count; reviewer override rate on AI criterion status.

4. MVP Features

Feature	Priority	Description
Packet extraction	P0	Pull procedure, comorbidities, clearance, and expected length of stay from the packet, each tagged with document and page
Hierarchical criteria checklist	P0	Four inpatient level-of-care criteria, each marked Met / Ambiguous / Missing / Not Met, mapped to the governing tier (NCD, LCD, payer policy, MCG)
Verifiable citations	P0	Clicking a citation opens the real packet PDF at the page with the cited sentence highlighted; public CMS policy links out to the CMS Coverage Database
Risk-scoped review tiers	P0	Banner explaining in plain language: lighter-touch review, standard review (one gap), or escalation (two or more gaps, or any Not Met)
Determination actions	P0	Approve, request provider clarification, escalate to medical director, deny (medical director only), each with an AI-drafted, human-readable summary shown before confirming
Audit trail	P0	Timestamped record of every action and the exact AI summary shown, exportable
Ask about this document	P1	In-page chat over the packet with citation links back to source
Parse-failure path	P1	Illegible packets are not guessed at; the case routes to manual review with a provider resubmission request
Case queue and CMS clocks	P1	Queue by status (needs action, escalated, closed) with standard and expedited deadline indicators

Feature	Priority	Description
Reports view	P2	Aggregate view of cases by tier and outcome

5. Constraints

- **Technical:** Prototype is front-end only with no backend. Production needs OCR for faxed scans, PDF text-layer mapping for highlighting, and page-level latency suitable for interactive use.
- **Data:** Prototype uses synthetic, scripted cases only. Real use requires PHI handling and policy content licensing (MCG and InterQual are proprietary).
- **Regulatory:** HIPAA; CMS decision timeframes; payer and state UM rules. The tool supports clinical judgment and is not a determination engine. The cited LCD in the demo is illustrative.
- **Organizational:** Single builder, portfolio-scale scope. No claim of clinical validation.

6. Grounding Strategy

- **Source of truth:** The submitted packet for patient facts; a versioned criteria library for policy (NCD, then LCD, then payer policy, then MCG, in precedence order).
- **Retrieval approach (target):** Extraction constrained to structured fields, each requiring a page and quoted span; criteria evaluation against a fixed, versioned rule set rather than free-form recall of policy. Prototype scripts this behavior.
- **Freshness:** Criteria library versioned with effective dates; each determination records the version used.
- **Citation and traceability:** Mandatory. A finding with no verifiable quote cannot be marked Met.

7. Hallucination Guardrails

- **Detection:** Automated check that every quoted span exists verbatim on the cited page; any finding failing this is downgraded to Missing.
- **Prevention:** Structured output with required `quote` and `source` fields; the model is instructed that "not found" is a valid answer.
- **Fallback:** Ambiguous criteria are never resolved by the AI. They generate a targeted provider question (for example, "confirm expected length of stay in midnights"). Conflicting documents are flagged as conflicts, not reconciled. Unreadable packets route to manual review with no extraction attempted.
- **User-facing signals:** Four-state status per criterion, review-tier banner, and a visible reason for each non-Met status.

8. Cost Budget

- **Cost per case (target):** Under \$0.50 in inference for a 14-page packet, versus nurse time measured in tens of minutes.

- **Model tier:** Mid-tier model for extraction and criteria mapping; escalate to a larger model only for conflicting or low-confidence documents.
- **Volume assumptions:** To be sized with the pilot site; assume low thousands of cases per month for budgeting.
- **Cost controls:** Cache extraction per packet version, cap pages per call, route simple cases to the cheaper tier, and re-run only changed sections after a provider response.

9. Eval Strategy

- **Offline:** Golden set of packets with labeled facts and expected criterion status. The prototype's six cases are the seed set: clean approval, ambiguous length of stay, multi-ambiguous with conflicting notes, corrupted packet, resolved-after-clarification, and expedited clarification. Expand with real de-identified packets.
- **Online:** Shadow mode first (AI output not shown to the nurse), then assisted mode with override tracking and sampled second-reviewer audit.
- **Metrics:** Extraction accuracy, citation validity (quote exists on cited page), criterion-status agreement with expert labels, false-Met rate (the critical error), and latency.
- **Cadence:** Every prompt, model, or criteria-library change; monthly drift check on production samples.

10. Production Readiness

- **Monitoring:** Citation-validity rate, override rate, parse-failure rate, tier distribution, cost per case, and time-to-determination against CMS clocks.
- **Rollback:** Prompt, model, and criteria-library versions pinned per release; one-step revert to the prior version.
- **Abuse and cost prevention:** Authenticated roles (nurse, medical director, auditor), per-user rate limits, and page-count caps.
- **Launch checklist:** False-Met rate below the agreed threshold on the golden set; security and privacy review; medical director sign-off on the escalation rules; on-call coverage for clock-critical cases.

11. Engineering Context

- **Architecture:** Browser app with case queue, case workflow (extraction, criteria, review, determination), packet viewer, and audit trail. Target adds an extraction service (OCR, structured extraction), a criteria service (versioned rules), and a secured audit store.
- **Integrations (target):** Fax or portal intake, EHR and payer systems of record, identity provider.
- **Team and ownership:** PM and builder (this project); clinical SME and compliance reviewer needed for a pilot.
- **Dependencies:** De-identified real packets, licensed criteria content, security review.

12. Product Intelligence

- **Usage analytics:** Time per stage, citation clicks, clarification and escalation rates, and where reviewers disagree with the AI.
- **Feedback loop:** Reviewer overrides become labeled examples for the golden set; recurring provider omissions inform intake guidance.
- **Personalization:** None in v1. Decisions must be uniform across reviewers.
- **Market signal:** Payers face growing regulatory pressure on prior authorization turnaround and transparency, which favors tools that show their evidence.

13. Design & UX

- **Interaction model:** Guided flow with a persistent case header, side-by-side checklist and source evidence, and a dark and light theme.
- **Latency handling:** Staged loading states during extraction; the checklist renders as results arrive.
- **Error and uncertainty states:** Dedicated Ambiguous, Missing, and Not Met treatments with the reason shown; a full failure state for unreadable packets.
- **Edit and override:** The nurse can disagree with any criterion status; every action opens a confirmation showing the AI summary that will be stored in the audit trail.

14. Technical Implementation Framework

The prototype is a React 19 and Vite single-page app, not Next.js. A production build would follow this shape:

- **Frontend:** React with Tailwind CSS v4; pdfjs-dist for in-browser rendering and sentence highlighting; Mermaid for workflow diagrams.
- **API layer (target):** Server routes for case retrieval, extraction jobs, and audit writes; streaming for extraction progress.
- **Model integration (target):** Structured-output calls with schema validation and the citation check from section 7.
- **State and data:** Client-side runtime state in the prototype; a database with immutable audit records in production.
- **Deployment:** GitHub Pages via Actions for the demo at sankar.work/Pre-Authorization; hosted app with secrets management and PHI controls for production.

15. Feature-Wise Build Plan

Feature	Phase	Owner	Target Date	Dependencies
Scripted end-to-end demo (six cases)	Phase 1 (done)	PM / builder	Oct 2026	None

Feature	Phase	Owner	Target Date	Dependencies
Live extraction with citation validation	Phase 2	Engineering	Q1 2027	Model access, de-identified packets
Golden-set evals and shadow mode	Phase 3	PM + clinical SME	Q1 2027	Labeled data
Backend, roles, and durable audit trail	Phase 4	Engineering	Q2 2027	Security review
Pilot with one UM team	Phase 5	PM + clinical lead	Q2 2027	Phases 2 to 4, compliance approval

Open Questions

1. What is the acceptable false-Met rate, and who signs off on it?
2. Which policy sources can be licensed or embedded, and under what terms?
3. How should conflicting documents be shown: flag only, or propose a resolution?
4. Which review-tier thresholds should be configurable per payer?
5. What does a defensible audit export need to contain for regulators?

Assumptions Made

- Format: AI Product PRD, because model grounding, hallucination risk, and evals shape the product.
- The prototype is the MVP; all dates after October 2026 are planning targets, not commitments.
- Baselines and targets are hypotheses, since no real-world measurements exist yet.
- Four criteria and the review-tier thresholds are taken from the current implementation.
- The first user is a payer-side UM nurse; provider-side prior authorization submission is out of scope.
- The demo uses synthetic data and is not a clinical tool.