



CHAI
COALITION FOR HEALTH AI

Use Case Best Practice Guide

*Prior Authorization:
AI-Supported Criteria
Matching*

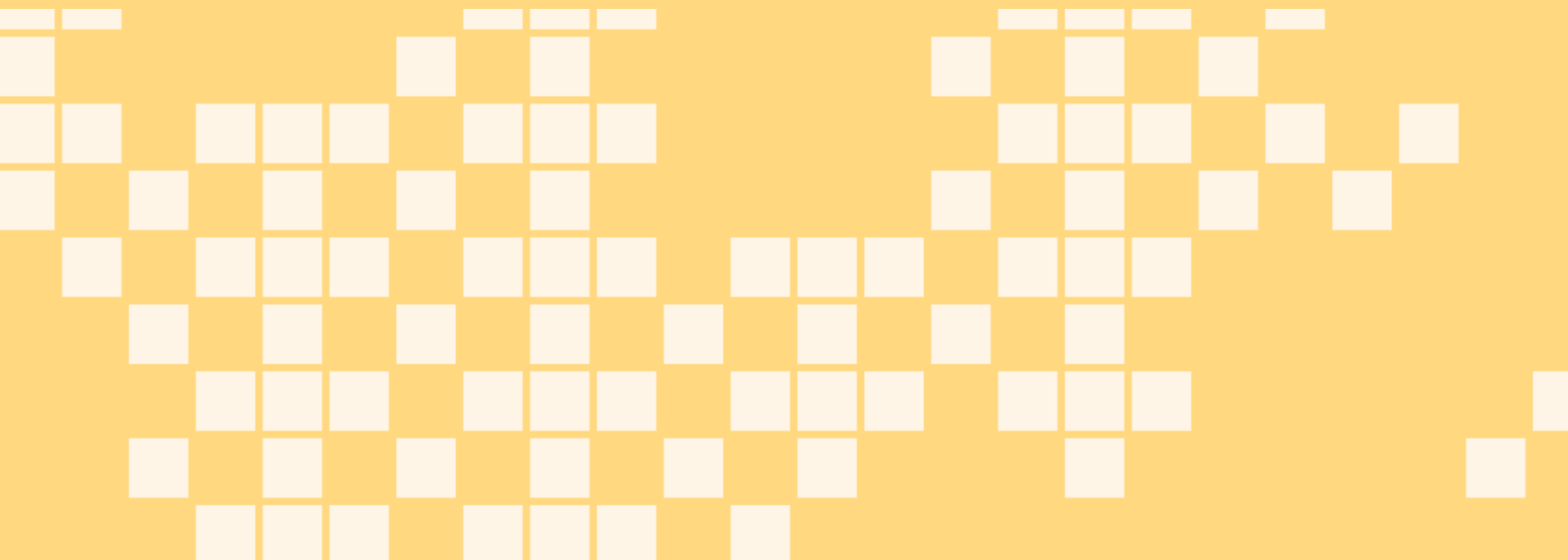




TABLE OF CONTENTS

03	What Does This Guide Include?
03	Use Case Description
04	Who is This Guide For?
04	Listening In: A Summary of Challenges & Insights
06	Best Practices for Developers
08	Best Practices for Implementers
09	Best Practices for Developers & Implementers
10	Appendix 1: Consensus Process
11	Appendix 2: Thank you & Contributors

Prior Authorization: AI-Supported Criteria Matching

For Developers & Implementers

What Does This Guide Include?

Use case specific best practice guides provide high level industry and consensus defined insights and recommendations, for the application of responsible AI principles to a specific use case. This guide focuses on an AI-supported criteria matching solution in prior authorization. The guide is organized by role (developer/implementer) and responsible AI principle areas where applicable (see figure below).



Useful, Usable & Effective

AI must solve specific problems, provide clear benefits, be easy to use, and perform reliably over time.



Fairness & Bias Management

AI systems should treat individuals and groups consistently, minimizing unjustifiable differences in outcomes caused by issues in data, design, deployment, or use.



Safe & Reliable

AI systems must not harm patients, requiring thorough testing, risk assessments, and continuous monitoring.



Transparent & Accountable

Stakeholders must understand how an AI system works, its limitations, and who is responsible for its impact.



Secure & Private

AI systems must protect patient data with strong security measures to prevent breaches and ensure confidentiality.

Use Case Description

To arrive at the best practice statements, work group members grounded in a specific use case. The criteria matching component of an AI-supported prior authorization (PA) system automates the process of assessing whether a healthcare service, procedure, or medication meets the payer's medical necessity guidelines. This step is critical in reducing administrative burdens on providers, improving turnaround times, and enhancing consistency in decision-making. AI facilitates the automatic extraction and

alignment of clinical documentation with established coverage criteria, minimizing manual effort and reducing the potential for errors or variability in approvals. AI-supported solutions cannot be used to automatically deny prior authorization. AI is used as a tool to support, not replace, clinical decision making.

Primary End Users

- Payers & Health Plans: Organizations ensuring AI-supported PA aligns with coverage policies
- Utilization Review & Medical Management Teams: Clinicians and administrative staff making final determinations on flagged cases
- Healthcare Providers: Physicians, nurses, and care teams submitting PA requests and reviewing AI recommendations

Scope Clarification & Note

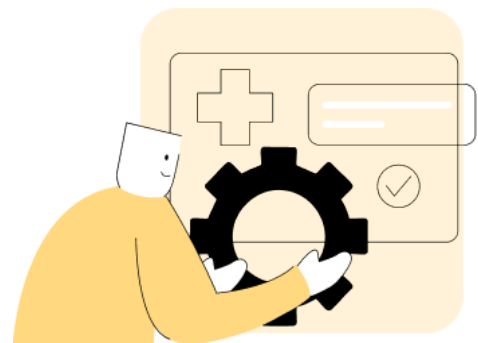


The AI-supported criteria matching component is limited to extracting and aligning documentation with coverage criteria. It does not determine approval or denial outcomes. Organizational routing rules (e.g., auto-approval pathways, triage logic, or manual review escalation) are separate, human-governed processes that may incorporate AI outputs but are not controlled by the AI model itself. All approval or denial decisions remain the responsibility of the implementing organization.

Who is This Guide For?



Developers: individuals involved in the software development process, including requirements gathering, design, coding, testing, and maintenance of software applications (derived from IEEE, 12207:2017)



Implementers: individual(s) responsible for the procurement, deployment, and/or overall realization of a system or component in accordance with a specified design (derived from IEEE 829 and IEEE 730)

Listening In: A Summary of Challenges & Insights

Below is a summary of some of the challenges and insights that emerged from the work group conversations.

Challenge #1: It is important to ensure fairness in a non-binary, contextual PA environment

- Fairness is difficult to assess in prior authorization because approvals and denials are not inherently positive or negative outcomes, making traditional parity metrics potentially misleading.

Challenge #2: Model performance may not generalize across payer populations, creating inequities when systems trained on commercial data are applied to Medicaid or Medicare beneficiaries.

- Vendor performance claims often lack specifics about data representativeness, evaluation methods, and generalizability to new populations.
- Vendors often train models on commercial insurance data that may not generalize to Medicaid or Medicare populations. Mismatches can reduce performance and introduce inequities.

Challenge #3: Ensuring effective human oversight is challenging because users may not fully understand the AI's limitations, leading to over-reliance on incomplete or incorrect outputs.

- Continuous monitoring cannot be fully automated in current practice because user signals (e.g., rejections, skips) may not reliably reflect correctness, requiring combined automated and manual review to maintain system accuracy and safety.
- Some tools rely on manual reviewers to catch what the AI missed, but users may not understand that model outputs can be incomplete or incorrect. Effective oversight depends on user understanding of the AI's fallibility.

Challenge #4: Ensure AI recommendations remain current, accurate, and supportive—never misleading or used for denials.

- AI systems in prior authorization cannot be used to render denials, and the model's role is limited to supporting users by surfacing relevant evidence or indicating where evidence is missing.
- Models are often trained on outdated payer policies. Without mechanisms for regular updates, tools may recommend incorrect actions or misalign with current coverage criteria.
- Generative AI may synthesize or pull information from incorrect or irrelevant parts of the patient record, leading to false conclusions about criteria matching.

Insight #1: Real-world data variability and incompleteness limit the reliability, accuracy, and generalizability of AI-supported criteria matching.

- Real-world clinical documentation is highly variable, often incomplete, and difficult to standardize across clients, making it challenging to build representative ground truth datasets.
- AI systems must handle missing or sparse documentation carefully, because payers often cannot control the quality of provider-submitted information and naïvely penalizing the AI or the patient can lead to inappropriate conclusions.

- Even highly trained annotators frequently disagree on whether criteria are met or not, revealing that the task itself has inherent ambiguity and setting realistic limits on attainable accuracy.

Insight #2: Model reliability depends on continuous updates, strong guardrails, and hybrid monitoring because AI systems can drift, misalign with policy, or produce unsafe outputs without proper oversight.

- Without strong prompt engineering and guardrails, AI outputs can be unpredictable or misaligned with intended use cases.
- AI systems in prior authorization cannot be used to render denials, and the model's role is limited to supporting users by surfacing relevant evidence or indicating where evidence is missing.

Insight #3: Transparency and traceability—such as verbatim evidence and clear citations—are essential for building trust and enabling users to validate AI outputs.

- Reliability and trust are strengthened when AI outputs use only verbatim, one-to-one excerpts from the clinical notes, preventing hallucinations and ensuring traceability.
- Trust increases when users can trace AI outputs to the source material. Citations also help identify hallucinations and support user validation.
- Some tools rely on manual reviewers to catch what the AI missed, but users may not understand that model outputs can be incomplete or incorrect. Effective oversight depends on user understanding of the AI's fallibility.
- Vendor performance claims often lack specifics about data representativeness, evaluation methods, and generalizability to new populations.

Insight #4: Effective use of AI in prior authorization depends on user understanding, workflow integration, and the ability to manage over-reliance or misinterpretation of AI outputs.

- Integration with existing systems and the ability to pull relevant documentation from multiple sources—including external records—is essential for effective criteria matching and reduces errors from incomplete information.
- Many users interacting with Prior Authorization AI tools—especially on the provider side—lack clinical training or foundational AI understanding, yet express high confidence in the tool. This creates a risk of misuse, over-reliance, and misinterpretation of AI-generated outputs.
- There's a continuum of prior authorization use cases—from low-risk eligibility checks to high-risk clinical necessity determinations. Not all require the same level of governance or trust-building.



Best Practices for Developers



Usefulness, Usability, & Efficacy

- Use all available, authorized data sources to construct a more complete clinical picture, even when full EHR interoperability is not possible. Prioritize flexible data ingestion pathways that capture key documentation (e.g., therapy notes, historical conditions) needed for accurate criteria matching.



Fairness & Bias Management

- Establish a disciplined, multi-annotator labeling and consensus-resolution process that continuously refines annotation guidelines and ensures the ground truth reflects real-world documentation complexity rather than idealized cases.
- Design AI behavior that defaults to “evidence not found” rather than inferring unmet criteria. For implementers, ensure such scenarios always route to human review without ever being used as implicit grounds for denial.



Safety & Reliability

- Establish processes to routinely update models or input rulesets as payer guidelines evolve.
- Combine automated anomaly detection with periodic human audits.



Transparency

- Implement strict guardrails that constrain the AI to extract and display only direct text retrieved from the patient's documentation, ensuring reviewers can easily verify the origin of every evidence citation.
- Enable the AI solution to output source citations for each criteria-matching statement.
- Provide clickable links to the original clinical documentation to support manual review and error detection.



Best Practices for Implementers



Usefulness, Usability, & Efficacy

- Incorporate brief onboarding, embedded in-tool guidance, and plain-language explanations that clearly define what the AI can and cannot do. Use role-specific prompts and verification reminders to calibrate user trust and prevent over-reliance.
- Involve domain experts in testing to verify that relevant, context-appropriate inputs are being used.



Fairness & Bias Management

- Supplement vendor validation with local testing or pilot evaluations to assess real-world performance and fairness.



Safety & Reliability

- Establish quality assurance protocols (e.g., inter-rater reliability checks) to ensure users are explicitly trained to independently verify model-suggested outputs.
- Adopt a tiered implementation and governance strategy. Start with lower-risk, automatable tasks (e.g., eligibility verification) and expand to more complex clinical criteria once trust and reliability have been demonstrated.
- Configure workflows and policies such that AI outputs are explicitly treated as decision-support—never as justification for denial—and ensure every case without clearly surfaced evidence is routed to a qualified human reviewer.
- Maintain staffing and processes for ongoing manual review of flagged cases to ensure safe system performance over time.

Note



AI outputs should be treated as evidence surfacing—not decision recommendations—and must not be used alone to drive automated approvals or denials. See CMS [FAQs](#) and [Final Rule](#)



Transparency

- Before procurement, request documentation of training data sources, population characteristics, and testing methodology. E.g., request vendor completes an AI model/solution card
- Clearly communicate to human reviewers that model outputs may include false positives or negatives.



Fairness & Bias Management

- For Implementers, during procurement and pre-implementation, request a breakdown of demographics from the vendor and test model performance on your local data before

deployment. For Developers, provide visibility into the training dataset (demographic breakdown, etc.) upon request from the implementing organization.

- Prioritize evaluating accuracy-based fairness metrics across demographic groups—such as precision and recall—rather than outcome parity, and ensure fairness assessments are grounded in clinical context rather than simplistic approval-rate comparisons.



Safety & Reliability

- Implement continuous monitoring and feedback loops to track AI accuracy and safety over time, particularly in high-risk clinical areas (e.g., orthopedics).



Transparency

- Benchmark AI performance against expert agreement rates, communicate these natural limits to end users, and avoid positioning the AI as achieving absolute objectivity when human reviewers themselves vary.

Appendix 1: Consensus Method

Best practice statements are collected from work group presentations and discussions. To ensure alignment across stakeholders, CHAI uses a three-phase consensus process for Best Practice Statements (BPS) generated through work group activities:

Phase 1: Initial Consensus Check

- **Purpose:** Gauge initial agreement on each draft BPS.
- **Voting Options:**
- *Include / Include Contextually / Exclude / Abstain*
- **Decision Rules:**
 - If $\geq 2/3$ vote “Include” → **Consensus achieved** (no further action).
 - If $< 2/3$ “Include”, but $\geq 2/3$ combined “Include” + “Include Contextually” → **Flagged for Phase 2.**
 - If $\geq 25\%$ vote “Exclude” → **Flagged for Phase 3.**
 - If $\geq 2/3$ vote “Exclude” → **Automatically excluded**

Phase 2: Revote with Revisions

- **Purpose:** Re-evaluate BPS that did not reach consensus in Phase 1 (but had < 25% “Exclude”).
- **Format:** Original and revised BPS shown side-by-side (based on Phase 1 feedback).
- **Voting Options:** *Include / Exclude / Abstain*, with an optional comment field.
- **Outcome:** Results used to determine final inclusion or exclusion.

Phase 3: Live Discussion and Vote

- **Purpose:** Address BPS with $\geq 25\%$ “Exclude” in Phase 1 (strong disagreement).
- **Steps:**
 - Facilitated group discussion of flagged BPS.
 - Live revote during the meeting + optional 1-week offline voting.
- **Voting Options:** *Include / Exclude / Abstain*
- **Outcome:** Final decision made based on discussion and revote results.

Appendix 2: Thank You and Contributors

We want to start by thanking every individual who showed interest, participated, listened, and came along with us in the early stages of our work. CHAI is at its core, a convener, and a member-driven non-profit. We are so grateful to be on this journey with you towards responsible AI in health for all. Your experiences, your feedback, your contributions, all make us who we are and help bring us to where we need to be.

For those who want to be credited directly by name, please reach out to us at program-management@chai.org to request contribution credit for the Prior Authorization Work Group. Below is a list of organizations who had at least one individual who showed interest and/or participated in our Prior Authorization Work Group.

If you want to learn more about our work groups (current and future), or have feedback on CHAI work groups, products, or services, please contact our Director of Responsible AI: merage@chai.org and Program Manager: anthony@chai.org.

Workgroup Leads

Name	Organization
Samta Shukla	BCBS Minnesota
Justin Brock	CVS Health
Chris Burnett	CVS Health
Christine Palermo	Encore Health
Jeremy Friese	Humata Health
Jan Sbarbaro	Lyric.ai
Akshay Sharma	Lyric.ai
Larry McEntire	MCG Health
Travis Bias	Solventum
Ajay Perumbeti	Solventum
Abby Steele	UnitedHealth Group
Tracee Coleman	UnitedHealth Group
David Rivkin	UnitedHealth Group
Fawad Butt	Penguin Ai

Participating Organizations

- Individual Contributor, Dorcas Yao, MD
- Airia
- ALIGNMT AI
- Ascension
- BCBS Minnesota
- Centene
- Children's Hospital of Philadelphia (CHOP)
- Claritev
- CVS Health
- Duke

- Encore Health
- Honor Health, Craig Norquist, MD
- Humata Health
- Hyro
- Innovaccer
- Interoice
- Johns Hopkins Health Plans
- Kaiser Permanente
- Key Software
- Lyric.ai
- Mayo Clinic, Nasibeh Zanjirani Farahani
- MCG Health
- Mercy
- MHK
- New York-Presbyterian
- Optum
- Penguin Ai
- Providence
- RAAPID
- Reggie Health
- Solventum
- Stanford
- Surescripts
- UnitedHealth Group
- WellSky
- Zelis