S-ACAD: Safe and Rapid Clinical Algorithm Development through Multi-AI Collaboration and Human Governance

Anonymous Author(s)
Affiliation
Address
email

Abstract

Healthcare leaders face a strategic dilemma: traditional expert-led content development ensures safety but is too slow for digital innovation, whereas AI automation offers speed but introduces unacceptable risks from hallucination. We present S-ACAD (Semi-Automatic Clinical Algorithm Development), a Human-in-the-Loop governance framework that positions continuous human oversight as the core management strategy for navigating AI-enabled uncertainty. In a prospective proof-of-concept study on pediatric febrile seizures, S-ACAD produced a parent-actionable algorithm in 245 minutes using multiple AI agents (Gemini 2.5 Pro, ChatGPT o3, Perplexity Pro, Consensus) for parallel data collection and Claude Opus 4 for critical review ('AI Sparring'). Two independent pediatric specialists confirmed zero critical safety errors. By comparison, a fully automated approach (F-ACAD) completed the task in 68 minutes but generated 17 issues including 9 high-priority safety concerns. These preliminary findings suggest that S-ACAD creates a viable pathway for 'Active Governance' of AI-assisted clinical content development, balancing operational efficiency with rigorous safety standards.

1 Introduction

Digital health organizations face a strategic paradox when deploying AI for clinical content development. Traditional expert-led methods protect clinical safety, but they can take weeks to months, creating a knowledge translation gap. AI can synthesize quickly, yet hallucinations can produce plausible but medically inaccurate advice, creating unacceptable liability for leaders accountable for patient safety [1]. Managing this speed–safety trade-off has become a core executive challenge.

This strategic dilemma is particularly acute in high-stakes clinical scenarios. Consider pediatric febrile seizures—the most common convulsive disorder in childhood, affecting 2–5% of children between ages 6 months and 5 years [2,3]. Although most febrile seizures are benign, witnessing one's child suddenly lose consciousness causes extreme anxiety for parents. In such situations, parents need immediate, actionable guidance—not clinical guidelines designed for healthcare professionals.

The traditional process of translating complex medical guidelines into simple, actionable algorithms is resource-intensive, typically requiring multidisciplinary teams working over several weeks to months [4]. Recent advances in large language models (LLMs) have demonstrated significant potential for accelerating medical knowledge synthesis [5]. However, the risk of 'hallucination'—where AI confidently generates plausible but medically inaccurate information—poses critical challenges. Recent studies document hallucination rates affecting diagnosis or management in AI-generated clinical content [6].

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.

This study describes the development process and formatively evaluates S-ACAD (Semi-Automatic Clinical Algorithm Development), an organizational governance framework that positions continuous human oversight as the core management strategy. S-ACAD leverages AI for rapid data collection and synthesis while freeing the expert to focus on high-stakes clinical validation. To explore feasibility and make speed–safety trade-offs explicit, we prospectively applied S-ACAD to pediatric febrile seizures and benchmarked it against a fully automated variant (F-ACAD).

2 Methods

This prospective, single-day case study was conducted on July 25, 2025, to execute and document the S-ACAD workflow for developing pediatric febrile seizure guidance. The entire process was designed to be transparent and reproducible, with all artifacts documented in the supplementary materials (Figure 1).

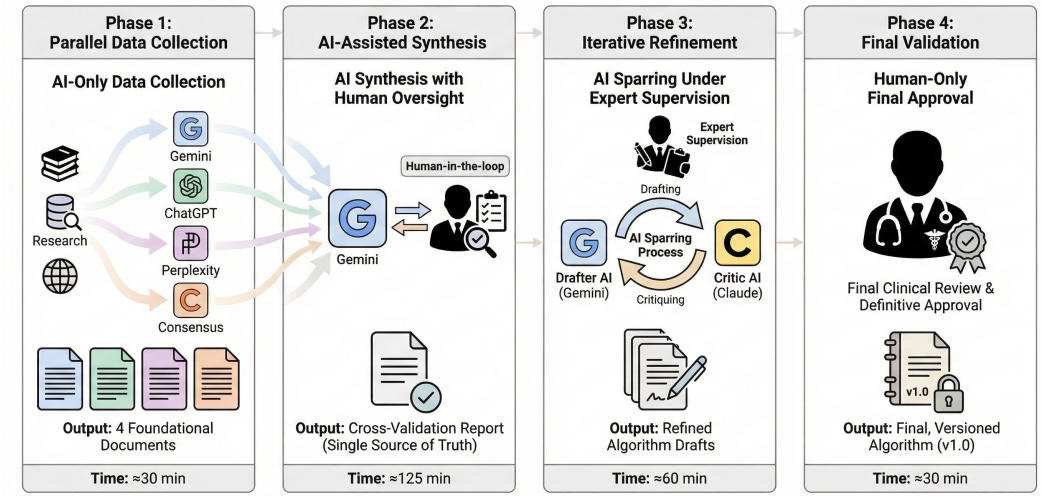

Figure 1: Overview of the S-ACAD workflow used to develop a parent-actionable pediatric febrile seizure response algorithm. The 4 phases combine parallel multi-LLM data collection, expert-validated synthesis (human-in-the-loop), AI-assisted iterative drafting ("AI Sparring"), and final clinician approval with version control. Phase durations shown are approximate single-run estimates and may vary by topic, team experience, and tooling.

2.1 AI models and configuration

Data was collected using the 'deep research' modes of Gemini 2.5 Pro (Google), ChatGPT o3 (OpenAI), Perplexity Pro (Perplexity AI), and Consensus. The critical review in Phase 3 was performed using Claude Opus 4 (Anthropic) in 'deep thinking' mode. All AI systems were accessed via subscription-based web interfaces with default configurations.

2.2 S-ACAD workflow

The S-ACAD workflow consists of four phases: (1) Parallel foundation data collection using multiple AIs—four AI services with different strengths were utilized simultaneously to build a comprehensive information base, compensating for potential biases of any single model; (2) Expert-led, AI-assisted cross-validation—the human expert conducted primary review of AI outputs, then used Gemini to generate a 'Cross-Validation Synthesis Report' identifying consensus, conflicts, and unique information; (3) AI–human collaborative algorithm generation—an initial draft was generated through cyclical refinement ("AI Generation → AI Critique → Human Revision"), with Claude serving as the 'AI Sparring' critic; (4) Final expert review—the human expert verified all decision pathways and approved the content as clinically safe.

2.3 Data collection and metrics

The following metrics were recorded in real-time: time spent on each phase, number of AI prompt iterations, number and type of human expert interventions, volume of information generated, and time spent resolving conflicts. Human interventions were categorized into: Clinical Judgment, Safety Review, UX Optimization, and Prompt Adjustment.

2.4 Independent expert review

The final algorithm was reviewed by two board-certified pediatric specialists—a pediatric emergency physician and a primary care pediatrician—using a structured survey instrument with 5-point Likert scales across four domains: Clinical Accuracy, Completeness, Safety, and Usability.

2.5 Comparative analysis with F-ACAD

Following S-ACAD completion, an identical task was performed using F-ACAD (Fully Autonomous Clinical Algorithm Development) via Genspark's multi-agent system. The system autonomously deployed 14 specialized AI agents without human intervention beyond initial task specification. The same clinical objective and scope were provided for fair comparison.

3 Results

The final developed algorithm follows a structured decision tree from risk screening to hospital visit determination (Figure 2).

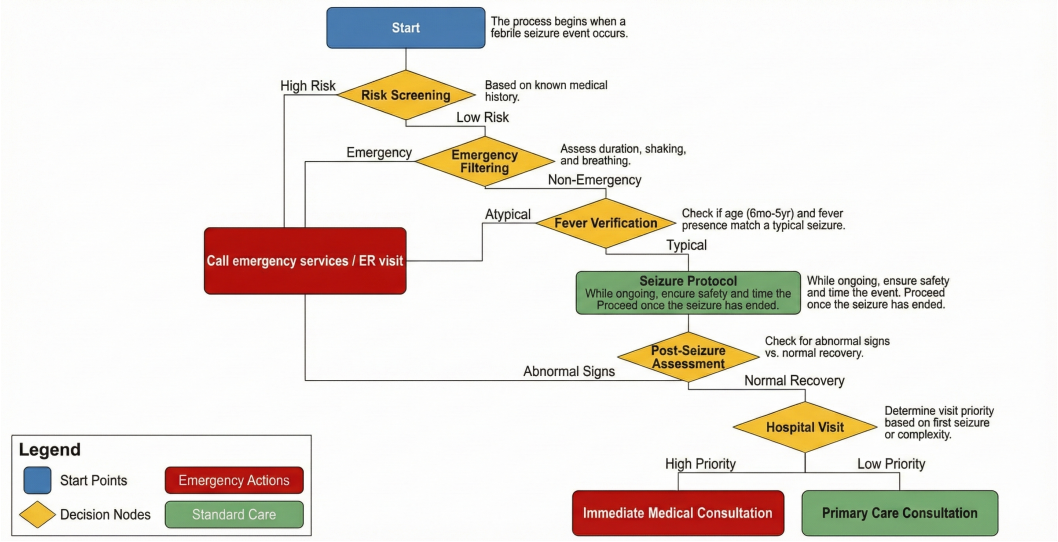

Figure 2: Simplified flowchart of the final pediatric febrile seizure response algorithm (v1.0) produced via the S-ACAD workflow. The diagram presents risk screening, emergency filtering, and post-seizure assessment steps, with color-coded nodes distinguishing emergency actions from standard-care guidance.

3.1 Workflow efficiency

The S-ACAD workflow was completed in approximately 245 minutes (4 hours 5 minutes), with 8 AI calls and 19 distinct human interventions. Phase 2 (Cross-Validation) accounted for the largest share of time ($\sim$125 minutes), highlighting that expert validation—rather than raw drafting—was the dominant workload. Phase 3 (Algorithm Generation) had the most human interventions (7 times), indicating intensive human–AI collaboration.

Table 1: Time and resource consumption by S-ACAD workflow phase

Phase	Time (min)	AI Calls	Human Interventions
Phase 1: Data Collection	∼35	4	4
Phase 2: Cross-Validation	∼125	1	5
Phase 3: Algorithm Generation	∼65	2	7
Phase 4: Final Review	∼20	1	3

3.2 AI sparring effectiveness

The Claude-based 'AI Sparring' phase identified 16 significant improvement suggestions. Of these, 14 (87.5%) were adopted after expert review: Logic Gaps (3, all adopted), UX Improvements (7, 5 adopted), Missing Scenarios (4, all adopted), and Safety Enhancements (2, all adopted).

3.3 Human intervention analysis

Analysis of the 19 human interventions showed that Clinical Judgment (42.1%) and Safety Review (26.3%) were most frequent, followed by UX Optimization (15.8%) and Prompt Adjustment (10.5%) (Figure 3).

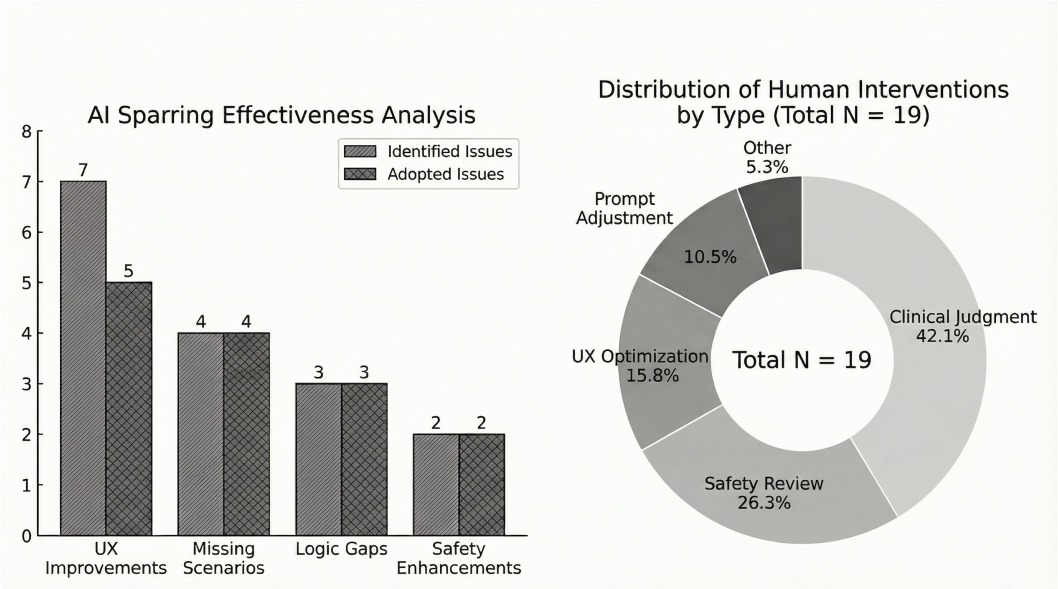

Figure 3: Human expert interventions recorded across all phases of the S-ACAD workflow (total N=19). Panel A summarizes issues identified during Phase 3's AI Sparring critique versus issues adopted after expert review, by category. Panel B shows the distribution of all intervention types across the workflow.

A critical intervention example: the AI initially allowed parents to directly administer emergency medication if seizures lasted over 5 minutes; the expert modified this to mandate calling 911 for real-time instructions before administration.

3.4 Expert review results

Both independent pediatric specialists confirmed zero medically inaccurate sections or critical safety errors requiring mandatory correction. Clinical validity ratings were 9.0/10 and 9.5/10. Both recommended the algorithm as 'Usable after minor revisions' for parent-facing deployment.

3.5 Comparative analysis: S-ACAD vs. F-ACAD

F-ACAD completed the task in 68 minutes—3.6 times faster than S-ACAD's 245 minutes. However, F-ACAD's AI critics identified 17 issues requiring correction, including 9 high-priority concerns related to emergency response clarity and standard-of-care alignment. These were precisely the areas where S-ACAD's human interventions proved crucial.

Table 2: Comparison of S-ACAD and F-ACAD approaches

Metric	S-ACAD	F-ACAD
Total Time	245 min	68 min
Human Interventions	19	0
Critical Issues Identified	0	17 (9 high-priority)
External Validation	Zero errors	Not performed

4 Discussion

4.1 Active governance as management strategy

S-ACAD addresses the challenge of AI-enabled uncertainty by redefining Human-in-the-Loop. Rather than serving merely as a passive final reviewer, the human expert engages in 'Active Governance'—directing AI, applying critical clinical judgment, and enforcing safety oversight throughout the development lifecycle. The 19 documented interventions demonstrate what this governance looks like in practice and map directly onto the safety failures observed under full automation.

4.2 Strategic implications

S-ACAD's efficiency gains arise from restructuring the development workflow. Traditional methods are slowed by sequential reviews; S-ACAD uses parallel AI processing and 'AI Sparring' to compress weeks of waiting time into minutes, significantly reducing development effort compared to traditional methods (Figure 4).

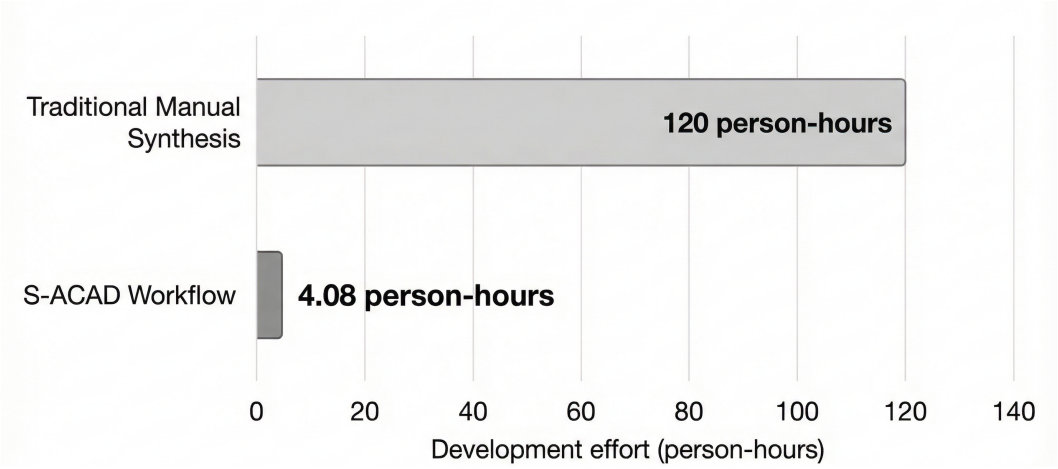

Figure 4: Comparison of algorithm development effort between traditional manual synthesis and the S-ACAD workflow. The traditional approach is presented as a representative effort estimate (≈120 person-hours), whereas the S-ACAD proof-of-concept required 4.08 person-hours.

Notably, expert validation (Phase 2) remained the rate-limiting step—this is the methodology's most critical safety feature, not a bottleneck to eliminate. The key managerial implication is the necessity of shifting expert roles from routine data collection to high-stakes validation and accountability. S-ACAD demonstrates that the optimal strategy is not a simple speed–safety trade-off but a redesigned workflow where human expertise is strategically allocated to governance and validation.

4.3 Limitations

This is an 'n-of-1' study: a single expert applied the workflow to a single clinical topic. The time profile ($\sim$4 hours) reflects a single-run operational measurement and cannot characterize variability across topics or settings. Significant potential for researcher bias exists, as the human expert was also the methodology's originator. The retrospective coding of interventions was performed without independent verification. Future work should include multi-expert, multi-topic validation with formal inter-rater reliability assessment.

5 Conclusion

Preliminary findings suggest that S-ACAD creates a viable pathway for an 'Active Governance' model of Human-in-the-Loop for AI-assisted clinical content development. By combining rapid AI drafting with continuous expert oversight, S-ACAD can reduce turnaround time without compromising safety. The methodology's core contribution is demonstrating how to harness AI's speed while maintaining human accountability for patient safety—a critical capability for healthcare leaders navigating AI-enabled uncertainty.

References

[1] Thirunavukarasu, A.J., et al. (2023) Large language models in medicine. *Nature Medicine*, **29**(8):1930–1940.

[2] Steering Committee on Quality Improvement and Management (2008) Febrile Seizures: Clinical Practice Guideline. *Pediatrics*, **121**(6):1281–1286.

[3] Leung, A.K., et al. (2018) Febrile seizures: an overview. *Drugs in Context*, 7:212536.

[4] Grimshaw, J.M., et al. (2012) Knowledge translation of research findings. *Implementation Science*, 7:50.

[5] Singhal, K., et al. (2023) Large language models encode clinical knowledge. *Nature*, **620**:172–180.

[6] Kumah-Crystal, Y.A., et al. (2024) Evaluating large language model accuracy in clinical note generation. *JAMIA Open*, **7**(1):ooae002.

A Appendix / Supplemental Material

The full supplementary materials (37 pages), including detailed prompts, logs, and evaluation forms, are attached below.

Multimedia Appendix 1: Detailed Prompts Used for Each AI Model

Multimedia Appendix 2: Full Text of the Cross-Validation Synthesis Report

Multimedia Appendix 3: Complete Flowchart of the Final Algorithm (v1.0)

Multimedia Appendix 4: Detailed Log of the AI Sparring and Iterative Development Process

Multimedia Appendix 5: Comparative Analysis of F-ACAD Methodology and Results

Multimedia Appendix 6: The Expert Evaluation Form

Multimedia Appendix 7: Anonymized Qualitative Feedback from External Experts

Multimedia Appendix 8: External Validation Raw Data (Quantitative Scores)

Multimedia Appendix 1. Detailed Prompts Used for Each AI Model

This appendix provides the detailed prompts used in Phase 1 (Parallel Data Collection) of the S-ACAD workflow, in both English and Korean, to maximize the capabilities of each AI model and ensure consistency in the outputs.

1. Prompts for Foundational LLMs (Gemini, ChatGPT)

Objective: To generate a comprehensive, encyclopedic 'Knowledge Base' on pediatric febrile seizures, serving as a baseline document for cross-validation. The focus is on factual description rather than rule-based instructions.

"You are a 'Medical Knowledge Researcher' tasked with building a medical knowledge database based on pediatric neurology textbooks and the latest clinical papers. Please create a comprehensive baseline document on 'Pediatric Febrile Seizures.' This document will be used as a 'Baseline Document' for cross-validation with other information and for expert review. The output should be in an encyclopedic, descriptive format detailing objective facts, NOT in an 'IF-THEN' rule-based format.

I. General Overview

1. Definition: Clinical definition of a febrile seizure.
2. Epidemiology: Peak age, incidence, prevalence.
3. Pathophysiology: Currently known mechanisms of occurrence.
4. Classification: Detailed criteria for simple vs. complex febrile seizures and the clinical significance of each item.

II. Clinical Aspects and Management

1. Risk Factors: Detailed description of factors influencing first occurrence and recurrence (e.g., family history, age, underlying conditions).
2. Management [Before Seizure]:
    - Importance of differentiating the cause of fever.
    - Current medical stance on the use of antipyretics.
3. Management [During Seizure]:
    - Step-by-step first-aid procedures at home.

147 o Absolute and relative criteria for calling emergency medical services (e.g., 911).

4. Management [After Seizure]:

    o Characteristics of the post-ictal state and key observation points.

    o Typical questions asked during a medical visit, and types/purposes of possible examinations.

III. Prognosis and Additional Considerations

1. Long-term Prognosis: Rate of transition to epilepsy, impact on neurodevelopment and cognitive function.

2. Recurrence: Recurrence rates and associated factors.

3. Differential Diagnosis: Other conditions that must be distinguished from febrile seizures (e.g., meningitis, afebrile seizures).

IV. [Additional/Enhanced Section] Special Situations and Various Scenarios

1. Considerations by Population:

    o Age: Specific characteristics and precautions for different age groups, such as infants (6-12 months) and children over 5 years.

    o Underlying Conditions: Differences in management for children with developmental delays, epilepsy, heart disease, etc.

2. Considerations by Situation:

    o Location: Management guidelines for non-home environments like daycare centers or during travel.

    o Time: Response strategies when medical access is limited, such as at night or on weekends.

3. Considerations for Different Caregivers:

    o Key information to be conveyed to non-primary caregivers like grandparents or babysitters.

Requirements:

- For each fact, please provide evidence where possible (e.g., 'According to the American

148 Academy of Pediatrics guidelines...').

- Please specify areas of controversy or uncertainty where expert opinions differ or more research is needed as 'Controversial Area' or 'Uncertainty'."

2. Prompt for Perplexity

Objective: To gather the latest clinical guidelines, systematic reviews, and key clinical summaries with a strong emphasis on citing credible sources.

"I am developing a digital health algorithm for parents/caregivers about pediatric febrile seizures. Please find and summarize comprehensive information with sources on the following topics.

1. Latest Clinical Guidelines (Focusing on 2020-2025)

- American Academy of Pediatrics (AAP) guidelines
- UK National Institute for Health and Care Excellence (NICE) guidelines
- Korean Pediatric Neurological Society guidelines
- Comparison of recommendations from major national pediatric societies (focusing on differences)

2. Results from Systematic Reviews and Meta-Analyses

- Epidemiology, risk factors, recurrence rates
- Management and treatment strategies (e.g., use of antipyretics, anticonvulsants)
- Long-term prognosis (neurodevelopment, risk of transitioning to epilepsy)

3. Summary of Key Clinical Information

- Clear diagnostic criteria for simple vs. complex febrile seizures
- Red flag symptoms requiring immediate medical attention
- Evidence-based first-aid methods that can be done at home

4. Parent Education Materials

- Parent guidance materials distributed by major medical institutions (e.g., university hospitals)

- Common misconceptions and medical facts about febrile seizures

Requirements:

- Please cite the source (institution name, year of publication) for all information.
- If there are conflicting contents between guidelines, please highlight them specifically."

3. Prompt for Consensus

Objective: To obtain a high-level summary of the scientific consensus on the topic, explicitly highlighting areas of debate or gaps in evidence.

"Please provide a comprehensive literature review covering all of the following items on pediatric febrile seizures.

- Core concepts and definitions
- Key clinical features
- Typical case presentations
- Evidence-based medical responses and management methods
- Common misconceptions
- Summarize the consensus among researchers, and also highlight any areas of controversy or gaps in the evidence."

Multimedia Appendix 2. Full Text of the Cross-Validation Synthesis Report

Analysis Objective: To establish reliable medical facts about 'pediatric febrile seizures' by cross-validating information generated by four AI models (Consensus, ChatGPT, Gemini, Perplexity), and to lay the groundwork for the development of a clinical decision algorithm. This report is the output of Phase 2 of the S-ACAD workflow and served as the 'Single Source of Truth' in the subsequent algorithm development process.

1. High Consensus

This section contains key information that was consistently identified across all sources or is commonly recommended in major international guidelines.

- **Definition and Diagnostic Criteria**
    - A febrile seizure is a convulsion in a child aged 6 months to 5 years (60 months) that occurs with a fever of 38°C (100.4°F) or higher. [Confidence: Very High]
    - It is a diagnosis of exclusion, meaning there must be no history of central nervous system infection, metabolic disorders, or afebrile seizures. [Confidence: Very High]
    - It is the most common convulsive disorder in childhood, occurring in 2-5% of all children. [Confidence: Very High]
- **Classification Criteria**
    - Simple febrile seizure: Lasts less than 15 minutes, is generalized, and occurs once within a 24-hour period. [Confidence: Very High]
    - Complex febrile seizure: Meets one or more of the following: lasts 15 minutes or longer, is focal, or recurs within 24 hours. [Confidence: Very High]
    - Febrile Status Epilepticus: A seizure lasting 30 minutes or longer. [Confidence: High]
- **First Aid and Treatment Principles**
    - If a seizure lasts longer than 5 minutes, it is essential to call emergency medical services (911/119) immediately. [Confidence: Very High]
    - Lay the child on their side to secure the airway, do not put anything in their mouth, and do not forcibly restrain them. [Confidence: Very High]

151
- Antipyretics are intended to relieve discomfort from fever and are not effective in preventing the recurrence of febrile seizures. [Confidence: Very High]
- Daily or intermittent use of prophylactic anticonvulsants is not recommended due to the risk of side effects. [Confidence: Very High]
- Tests such as EEG or brain imaging (CT/MRI) are not routinely necessary after a simple febrile seizure. [Confidence: High]
- A lumbar puncture is performed selectively in cases with signs of meningitis or in unvaccinated infants aged 6-12 months. [Confidence: High]

2. Conflicting Information

This section details topics where criteria differ or conflict between sources or national guidelines.

- **Topic:** Diagnostic Criteria (Age and Fever)
  - **Claim A (AAP/NICE, etc. international standard):** 6 months to 5 years, 38.0°C or higher.
  - **Claim B (Korean Pediatric Neurological Society, etc.):** 3 months to 6 years, 37.8°C or higher.
  - **Analysis:** A clear discrepancy exists in the diagnostic criteria. For a global application, the international standard (e.g., AAP) should be the default, with local standards considered for regional localization.
- **Topic:** Prophylactic Diazepam Use
  - **Claim A (AAP/NICE):** Routine prophylactic use is not recommended due to the risk of side effects.
  - **Claim B (Japanese guidelines, etc.):** Intermittent use during fever is recommended for high-risk groups (e.g., complex type, family history).
  - **Analysis:** There is a clear difference in stance between Western and Eastern guidelines.
- **Topic:** Short-term Recurrence Prevention with Antipyretics
  - **Claim A (All guidelines):** No long-term recurrence prevention effect.
  - **Claim B (2021 Finnish RCT - single study):** Suggested the possibility of

reducing short-term recurrence 'within the same febrile episode.'

- **Analysis:** This is not a standard recommendation but is a controversial area that requires explanation during parent education.

3. Unique Information & Needs Verification

This section contains information mentioned only in limited sources or with low evidence levels, requiring further verification.

- **Complex Febrile Seizures and Long-term Prognosis:** Claims exist that the risk of ADHD is 1.4 times higher (OR 1.4) or IQ is 4.5 points lower on average in the group with a history of complex febrile seizures, but this has not been consistently confirmed in large-scale studies and is controversial. [Confidence: Medium]

- **Micronutrient (Iron, Zinc) Deficiency:** Some studies have reported a correlation suggesting that iron or zinc deficiency may be a risk factor, but causality is uncertain, and there is insufficient evidence for the preventive effect of supplementation. [Confidence: Low]

- **Use of Digital Education:** Information that some hospitals in Australia mandate text/video education before discharge is specific to certain institutions and difficult to generalize. [Confidence: Low]

4. Overall Assessment and Conclusion

- **Overall Reliability:** High. All sources show a high degree of agreement on core definitions, classifications, and first-aid principles.

- **Key Issues Proposed for Expert Validation:**

  - Standardization of Diagnostic Criteria

  - Guidelines for Prophylactic Anticonvulsant Use

  - Management Protocol for Infants Under 12 Months

  - Redefining the Role of Antipyretics

  - Inclusion of Uncertain Information on long-term prognosis.

Multimedia Appendix 3. Complete Flowchart of the Final Algorithm (v1.0)

Objective: To establish a standardized response protocol to support caregiver decision-making and ensure clinical safety when a pediatric febrile seizure occurs. **Target Audience:** Global users, with a focus on the U.S. (units like temperature are dual-notated). **Evidence Level Notation Guide:**

- **Consensus:** The level of agreement among the multiple sources provided (A: High, B: Medium, C: Conservative Recommendation).
- **Evidence:** The type of medical evidence (Guideline, SR/Meta-analysis, Cohort/Observational Study, Expert Consensus/Review).

Prerequisites (Input Data): [Child's Name], [Age (in months)], [Underlying Conditions and Risk Factors (Boolean/Array)], [History of Febrile Seizures (Boolean)], [Prescribed Emergency Anticonvulsant (Boolean)], [Recent Vaccination History (String/Date)].

START: Module Execution

[Decision 0: Initial Risk Factor Screening]

- **[Question]** Does the child belong to any of the high-risk groups below?
    - Neurodevelopmental delay (e.g., cerebral palsy)
    - Direct family history of epilepsy
    - Other severe underlying conditions (e.g., congenital heart disease)
- **[Judgment]**
    - → **[Yes]**
        - FLAG: set high_risk = true.
        - **[Notification]** "[Child's Name] requires more careful observation due to an underlying condition. It is recommended to consult a doctor or visit the emergency room immediately if a seizure occurs."

Rationale: High-risk groups may have different causes and prognoses for seizures compared to non-high-risk groups, thus a more conservative approach (medical intervention) is induced from the start. [Consensus: A | Evidence: Cohort/Observational Study]

- → **[No]**

154

- FLAG: set high_risk = false.
- **[Proceed]** Go to Decision 1

[Decision 1: Immediate Top-Priority Emergency Filtering (Life-Threatening Red Flags)]

- **[Question]** Does any of the following absolute emergency situations apply right now?
  - A. Is the child having difficulty breathing, or are their face/lips turning blue (cyanosis)?
  - B. Has the seizure already been going on for more than 5 minutes?
- **[Judgment]**
  - → **[Yes]** → **[Final Action]** 🔴 **Call 911 Immediately** (localize emergency number by country)

Rationale: A seizure lasting more than 5 minutes can progress to status epilepticus, and hypoxia is an emergency requiring immediate intervention. [Consensus: A | Evidence: Guideline]

  - → **[No]** → **[Proceed]** Go to Decision 2

[Decision 2: Age Appropriateness Check]

- **[Question]** Is the [Child's Age] between 6 months and 60 months (5 years) old?
- **[Judgment]**
  - → **[Under 6 months]** → **[Final Action]** 🔴 **Call 911 Immediately**

Rationale: Seizures in infants under 6 months have a very high risk of serious underlying conditions like meningitis. [Consensus: A | Evidence: Guideline]

  - → **[Over 60 months (5 years)]**
    - **[Additional Question]** Have they been diagnosed with a febrile seizure in the past?
    - **[Yes]** → **[Final Action]** 🟡 **Recommend ER Visit**
      - "Although over 5 years old, a history of febrile seizures warrants an ER visit to determine the cause. If the child is breathing stably and is clearly conscious after the seizure, you may consider going directly to the ER. However, if you are unsure or anxious, call 911

155 for advice."

- **[No] → [Final Action] 🔴 Call 911 Immediately**

Rationale: A first seizure after the age of 5 is likely not a febrile seizure. [Consensus: A | Evidence: Guideline]

- → **[6 to 60 months] → [Proceed]** Go to Decision 3

[Decision 3: Fever and Vaccination Association Check]

- **[Question 3-1]** Was there a 'fever' of ≥ 38.0°C (100.4°F) during the febrile illness at the time of or during the seizure?
  - → **[Definitely no (normal temperature)] → [Final Action] 🔴 Call 911 Immediately**

Rationale: An 'Afebrile Seizure' requires immediate neurological evaluation. [Consensus: A | Evidence: Guideline]

  - → **[Couldn't measure/Uncertain]**
    - **[Additional Question]** Did the child's body feel hot before or after the seizure?
    - **[Yes] → [Proceed]** Go to Decision 3-2
    - **[No] → [Final Action] 🟡 Contact Medical Facility in Steps**
      - "Since it's uncertain whether there was a fever, an accurate evaluation is needed. Please follow the steps below."
      - "Step 1: First, try to contact your primary care physician's (PCP) after-hours on-call line within 5 minutes."
      - "Step 2: If you cannot reach them, or if you are advised to go to the ER immediately, go to the ER right away."
  - → **[Yes (38.0°C or higher)] → [Proceed]** Go to Decision 3-2

- **[Question 3-2]** Has the child been vaccinated within the last 48 hours?
  - → **[Yes]**
    - **[Information]** "A seizure can be triggered by a fever that occurs after

156 vaccination. This is a reaction to the fever, not the vaccine itself, and most cases have a good prognosis. This information is very important for the medical staff, so please be sure to tell them." [Consensus: A | Evidence: Observational Study/Review]

        - **[Proceed]** Execute [Situation Protocol: Seizure in Progress]

    - → **[No]**

        - (No information provided)

        - **[Proceed]** Execute [Situation Protocol: Seizure in Progress]

[Situation Protocol: Seizure in Progress]

- **[Action 1] ✅ Top Priority Safety Measures:** Lay the child on their side and clear the surrounding area. [Consensus: A | Evidence: Guideline]
- **[Action 2] ⏱️ Time and Observe:** Start a timer and observe the seizure's characteristics (record a video if possible). [Consensus: A | Evidence: Expert Consensus]
- **[Action 3] ❌ Absolute Don'ts:** Do not put anything in the mouth, do not forcibly restrain, etc. [Consensus: A | Evidence: Guideline]

[Decision 4: Response if Seizure Lasts 5 Minutes]

- **[Question]** Has it been 5 minutes since the seizure started?
    - → **[No]** → **[Proceed]** Wait until the seizure stops
    - → **[Yes]**
        - **[Question 4-1]** Do you have an emergency anticonvulsant (e.g., rectal Diazepam) prescribed by a doctor?
        - **[No]** → **[Final Action] 🔴 Call 911 Immediately**
        - **[Yes]** → **[Final Action] 🔴 Call 911 Immediately** and say the following:
            - "My child has been seizing for more than 5 minutes, and I have an emergency medication (like Diazepam) prescribed by a doctor. Please give me instructions on its administration."

Rationale: The use of emergency medication is safest when done under the real-time guidance of a 911 dispatcher or medical professional. [Consensus: A | Evidence: Expert

Consensus/Guideline]

[Situation Protocol: Seizure Stopped]

[Decision 5: Post-ictal State Evaluation]

- **[Question 5-1]** Is the child breathing normally?
    - → **[No]** → **[Final Action]** 🔴 **Call 911 Immediately**
- **[Question 5-2]** What is the child's level of consciousness? (Apply AVPU Scale)
    - A (Alert): Awake, aware of surroundings, and makes eye contact with parents?
    - V (Verbal): Responds to speech (babbling) or sounds?
    - P (Pain): Responds only to mild stimuli (like a pinch)?
    - U (Unresponsive): No response to any stimuli?
- **[Judgment]**
    - → **[P or U state]** → **[Final Action]** 🔴 **Call 911 Immediately**
    - → **[V state continues for >10 mins]** → **[Final Action]** 🟡 **Visit the ER**
        - "A slow recovery of consciousness after a seizure requires an ER visit. If the child is gradually improving and breathing is stable, you can go directly to the ER. However, if they worsen at all or you are anxious, call 911 immediately."
    - → **[A or V state (improving)]** → **[Proceed]** Go to Decision 6

Rationale: Delayed recovery of consciousness after a seizure may indicate a serious underlying condition. The AVPU scale is a tool that caregivers can use for objective assessment.
[Consensus: A | Evidence: Guideline]

[Decision 6: Determining the Level of Hospital Visit]

- **[Question]** The seizure has stopped, and the child is stabilizing. Which of the following applies?
    - **A. Cases Requiring an Immediate ER Visit:**
        - [ ] Seizure lasted more than 15 minutes (complex febrile seizure)

158

        - [ ] Focal seizure (e.g., only one arm/leg shaking) (complex febrile seizure)
        - [ ] Seizure occurred more than once within 24 hours (complex febrile seizure)
        - [ ] Infant 6-12 months old + incomplete Hib/pneumococcal vaccination
        - [ ] Incomplete recovery of consciousness or child seems very lethargic after the seizure
        - [ ] high_risk flag is true (underlying condition)
    - **B. Cases Requiring a Same-Day Visit (Urgent Care or PCP):**
        - [ ] First-ever febrile seizure (with full recovery afterward)
        - [ ] Parent is worried about the child's condition and wants medical consultation
    - **C. Cases Where a Next-Day PCP Visit is Possible:**
        - [ ] Previously diagnosed with a simple febrile seizure, the current episode was similar, and the child has fully returned to their usual condition afterward
        - [ ] Parent fully understands the child's condition and feels reassured
- **[Judgment]**
    - **→ [Any item in A applies] → [Final Action] 🟡 Visit the ER**
        - "The features mentioned suggest a 'complex febrile seizure' or the possibility that other causes need to be ruled out. Please go to the ER now for an evaluation."
    - **→ [Applies to B or C] → [Final Action] ✅ Check Time-Based Care Guidance**
        - "The child's condition appears stable. Please choose the most appropriate medical facility based on the current time."
        - **[Weekdays (8 AM - 6 PM)]**
            - If B applies (Same-Day Visit): 🟢 Contact a nearby PCP or Urgent Care for a visit. This may be faster and more economical.

159

        - If C applies (Next-Day Visit): 🔵 Contact your PCP to inform them of the seizure and schedule a visit for the next day.

    - **[Nights / Weekends]**

        - If B or C applies: 🟢 The best first step is to contact your PCP's on-call line for advice. If that's difficult, check for pediatric Urgent Care centers that are open. If neither of these options is feasible, you should consider visiting the ER.

- 💡 **Insurance & Accessibility Note**
    - **Cost Consideration:** In general, medical costs are highest at the ER, followed by Urgent Care, and then the PCP. Insurance coverage may vary depending on your individual plan.
    - **Regional Differences:** Depending on your location (e.g., urban vs. rural), there may be no available Urgent Care, or ER wait times could be very long. Please make your decision considering these practical constraints.
- ✅ **Post-Consultation Follow-Up**
    - After receiving medical care for the seizure, please check the following to continue managing your child's health.
    - [ ] **Share Records:** If you visited an Urgent Care or ER, be sure to forward the medical records to your child's PCP.
    - [ ] **Vaccination Plan:** If any vaccinations were delayed due to the seizure, consult with your PCP to reschedule them safely.
    - [ ] **Recurrence Action Plan:** In case a seizure happens again, create an action plan in advance with your PCP, including the potential need for a prescription for an emergency anticonvulsant.
- **[Common Safety Alert]**
    - ⚠️ **Call 911 or visit the ER immediately if any of the following occur:**
        - Another seizure occurs
        - Consciousness worsens
        - Persistent vomiting

- Stiff neck
- Seems to have difficulty breathing
- Your parental intuition tells you something is wrong

- 💡 When in doubt, always choose a higher level of care. Your child's safety is the top priority.

Multimedia Appendix 4. Detailed Log of the AI Sparring and Iterative Development Process

This appendix provides a detailed record of the *AI Sparring* and iterative improvement process, which is the core of the S-ACAD workflow. It shows, step-by-step, how the initial AI-generated draft was progressively refined through feedback from an AI critic and a human expert to become the final algorithm (v1.0). The quantitative data described in section 3.4 of the main text, 'AI Sparring Effectiveness Analysis,' is based on the detailed analysis below.

Iteration 1: From Initial Draft (v0.1) to User-Centric Design (v2.0)

- **Step 1.1:** Initial Draft Generation (v0.1 by Gemini)
    - **Input:** 4 foundational data documents + 1 synthesis report
    - **Prompt:** See Multimedia Appendix 1 - Request to generate an evidence-based algorithm draft for parents.
    - **Output (v0.1 Summary):** Medically accurate, but a text-heavy, list-based structure. It was divided into [Before Seizure], [During Seizure], and [After Seizure], with Red Flag conditions scattered throughout.
- **Step 1.2:** Critical Review (by Claude) & Quantitative Analysis
    - **Input:** Algorithm v0.1
    - **Prompt:** Request for a multi-faceted critique covering logical flow, user experience, safety, and edge cases.
    - **Output (Summary of 16 Suggestions):** The Claude model identified a total of 16 significant improvement suggestions. After review by a human expert, 14 of these (87.5%) were adopted into the final algorithm.

Table: Review and Implementation Results of AI-Generated Suggestions

Category	Suggestion Content	Adoption Status	Rationale and Outcome
Logical Flaws (3/3)	1. Confusion in 119 call timing (mixing 5-minute standard and immediate call conditions)	Adopted (O)	Starting from v2.0, an 'Immediate 911' section was placed at the very top to clarify

162			priority.
	2. Ambiguity in priority between 119 call for first seizure and hospital visit guidelines	Adopted (O)	From v2.0, 'first seizure' was elevated to an 'Immediate 911' condition, then clarified to 'ER visit' after the seizure.
	3. Omission of exception rules for infants under 6 months (mentioned in body)	Adopted (O)	Explicitly included in the age filter logic from v3.1.
UX Improvement (5/7)	4. Use of specialized terms like "complex febrile seizure" (mentioned in body)	Adopted (O)	Replaced with simple explanations like "lasts more than 15 minutes..." from v2.0.
	5. Use of non-intuitive terms like "photosensitivity"	Adopted (O)	Changed to "cries excessively or avoids bright lights" from v2.0.
	6. Presentation of unmeasurable criteria like "when the neck is stiff"	Adopted (O)	Specified as "when the chin cannot touch the chest" from v2.0.
	7. Text-heavy structure reduces readability in urgent situations	Adopted (O)	Icon system (🚨, ✅, etc.) introduced from v2.0.
	8. Equal visual importance of DOs and DON'Ts makes priority identification difficult	Adopted (O)	Changed to a format that emphasizes 'Top Priority Safety Actions' from v2.0.

[163]	9. All text in black only makes it difficult to distinguish urgency	Not Adopted (X)	Judged that the icon system adequately expresses urgency.
	10. Entire algorithm length is long, leading to severe scroll pressure (accordion menu suggestion)	Not Adopted (X)	Judged that having all information expanded is safer in an emergency.
Missing Scenarios (4/4)	11. Handling cases where a child falls into a deep sleep after a seizure and is difficult to wake (mentioned in body)	Adopted (O)	Included in the 'Consciousness Check' step with the criterion "if they don't open their eyes after 10 minutes" from v2.0.
	12. Response protocol for seizures occurring during vehicle travel	Adopted (O)	Included in 'Special Situation Management' from v4.0.
	13. Response protocol for seizures occurring during bathing	Adopted (O)	Included in 'Special Situation Management' from v4.0.
	14. Response protocol for seizures occurring in public places	Adopted (O)	Included in 'Special Situation Management' from v4.0.
Safety Enhancement (2/2)	15. Insufficient emphasis on not putting anything in the mouth during a seizure (mentioned in body)	Adopted (O)	Strongly warned with a ❌ icon in the 'Absolute Don'ts' section from v2.0.

164	16. Failure to reflect the practical difficulty of time measurement (suggestion for 119 dispatcher assistance)	Adopted (O)	Added the phrase "Call 911 and ask the operator to help you keep time" from v2.0.

- **Step 1.3: Revision and Output (v2.0)**
    - **Action:** Fully adopted Claude's critique to redesign the algorithm with a focus on user experience (UX).
    - **Key Changes in v2.0:**
        - **Structure Overhaul:** Placed absolute situations requiring an 🚨 immediate 911 call at the very top to clarify priority.
        - **Visualization Introduced:** Implemented an icon system: 🚨 (Immediate 911), ✅ (Safety Action), ⏱️ (Time Check), ❌ (Absolute Don't).
        - **Simplification:** Simplified the complex [After Seizure] decision-making into three steps: Breathing → Consciousness → Hospital Visit Decision.
        - **Terminology Improvement:** Changed to measurable and easy expressions like "when the chin cannot touch the chest."

Iteration 2: Personalization and Contextualization (v2.0 -> v3.1)

- **Step 2.1: New Requirements (by Human Expert)**
    - **Feedback:** "Since this is a module within the Fevercoach app, it must utilize pre-saved user information (age, history, etc.). Specifically, the definition of a febrile seizure (6-60 months, accompanied by fever) must be clearly filtered at the start of the algorithm."
- **Step 2.2: Revision and Output (v3.1)**
    - **Action:** Added personalization logic using pre-saved information and a core definition filter.
    - **Key Changes in v3.1:**
        - **Personalization:** Dynamically displays the child's name, as in "A

personalized guide for [Child's Name]'s parents."

- **Age Filter Added (Decision 1):** Checks the 6-60 month range at the start of the algorithm, setting a path to call 911 immediately if outside this range.
- **Fever Filter Added (Decision 2):** Checks if there was a 'fever' during the seizure, branching to an immediate 911 call for 'afebrile seizures' without fever.

Iteration 3: Expert Review and Globalization (v3.1 -> v4.2)

- **Step 3.1: Expert-level Review**
  - **Input:** Algorithm v3.1
  - **Feedback:** "As a global app, it must follow international standards (AAP). The temperature criteria, consciousness recovery assessment criteria, etc., need to be refined more precisely, and screening for high-risk groups like those with underlying conditions is necessary."
- **Step 3.2: Revision and Output (v4.2)**
  - **Action:** Reflected expert feedback to enhance clinical precision and compliance with global standards.
  - **Key Changes in v4.2:**
    - **Global Standard Applied:** Temperature criterion corrected to 38.0°C (100.4°F).
    - **Risk Stratification (Decision 0):** Added an initial risk factor screening step to identify high-risk groups beforehand.
    - **Objective Assessment Tool Introduced:** Applied the AVPU Scale for consciousness recovery assessment.
    - **Observation Items Detailed:** Specified seizure characteristics such as Tonic/Clonic/Atonic, bilateral/unilateral, etc.

Iteration 4: Final Safety Enhancements and Nuanced Guidance (v4.2 -> v1.0 Final)

- **Step 4.1: Final Safety & Usability Check (by Human Expert)**

166
- **Input:** Algorithm v4.2
- **Feedback:** "The association with vaccination must be specified, and the protocol for using prescribed emergency anticonvulsants must be modified in the safest way possible. Also, rather than having all situations result in a 'Call 911,' the guidance should be segmented into 'ER Visit,' 'Same-Day Visit,' etc., based on the urgency of the situation, to promote the efficient use of medical resources and reduce the user's burden."

- **Step 4.2: Revision and Finalization (v1.0)**
  - **Action:** Completed the algorithm by reflecting final safety and usability review comments.
  - **Key Changes in v1.0:**
    - **Vaccination History Added (Decision 3.2):** Checks for recent vaccinations and guides the user to report this information to medical staff.
    - **Emergency Medication Protocol Modified (Decision 4):** Instead of the app directly instructing medication administration, the path was modified to 🚨 call 911 immediately, inform them of the medication possession, and follow the medical professional's instructions, minimizing legal/safety risks.
    - **Hospital Visit Level Segmented (Decision 6):** For patients stable after a seizure, the guidance was segmented into ER visit, same-day visit, and next-day visit, with customized guidance for different times (day/night/weekend) to increase practicality.
    - **Evidence Level Notation Completed:** Added dual [Consensus | Evidence] notation and citations to all decision points.

Through this multi-stage process of iterative validation and refinement, the initial idea was able to evolve into a final algorithm (v1.0) that meets the complexity and safety requirements of the actual clinical setting.

Multimedia Appendix 5: Comparative Analysis of F-ACAD Methodology and Results

F-ACAD represents an experimental baseline configuration constructed on a specific platform (Genspark) to illustrate the risks of ungoverned automation, rather than a definitive representation of all fully automated approaches.

A5.1 F-ACAD System Architecture

Overview

The F-ACAD (Fully Autonomous Clinical Algorithm Development) system consists of 14 specialized AI agents operating autonomously to develop clinical algorithms without human intervention beyond initial task specification and final review.

Agent Configuration

1. Search Agents (4):

   - US Guideline Specialist: Focuses on AAP, ACEP guidelines

   - International Guideline Specialist: NICE, ILAE, global standards

   - Systematic Review Specialist: Cochrane, PubMed meta-analyses

   - Emerging Research Specialist: Recent studies (2020-2025), controversial topics

2. Validation & Synthesis Agents (3):

   - Verification Agent: Source and reference validation

   - Synthesis Agent: Data integration and guideline comparison

   - Quality Control Agent: Consistency and completeness review

3. Design Agents (2):

   - Algorithm Design Agent: Decision tree generation

   - Content Generation Agent: Parent-friendly educational modules

4. Critic Agents (4):

   - US Pediatrician AI: Clinical accuracy and standard compliance

   - Pediatric Emergency AI: Emergency classification and response

   - Parent UX AI: Language clarity and user experience

   - Medical Informatics AI: Logical completeness and data structure

5. Improvement Management Agent (1):

- Prioritizes critiques and applies improvements iteratively

A5.2 Detailed Time Logs Comparison

Table A5-1. Phase-by-Phase Time Comparison Between S-ACAD and F-ACAD (Single-Run Operational Estimates)

Development Phase	S-ACAD Time (single-run estimate)	F-ACAD Time (single-run estimate)	Relative difference (illustrative)
Phase 1: Data Collection	≈30 min	≈10 min 22 sec	65.4%
Phase 2: Validation & Synthesis	≈125 min	≈9 min 33 sec	92.4%
Phase 3: Algorithm Generation and Refinement	≈60 min	≈18 min 03 sec*	69.9%
Phase 4: Final Expert Review	≈30 min	≈30 min**	0%
Total	**245 min**	**67 min 58 sec**	**72.3%**

General note: All time values (including those reported for S-ACAD) represent approximate, single-run operational estimates derived from one end-to-end execution and are provided to characterize relative workflow burden rather than statistically representative performance benchmarks.

* For F-ACAD, 'Algorithm Generation' (10 min 15 sec) and 'Iterative Refinement' (7 min 48 sec) are aggregated to align with S-ACAD's Phase 3 ('Algorithm Generation and Refinement') for an illustrative phase-level comparison.

** For F-ACAD, 'Final Expert Review' time is assumed to match S-ACAD's Phase 4, reflecting the second human touchpoint described in the methodology; this assumption is used solely to support workflow characterization.

A5.3 AI Agent Specifications and Performance

Table A5-2. F-ACAD Agent Activity Log

169 Agent Category	Number of Agents	Total API Calls	Processing Time
Search Agents	4	47	10 min 22 sec
Validation Agents	3	23	9 min 33 sec
Design Agents	2	18	10 min 15 sec
Critic Agents	4	31	7 min 48 sec
Management Agent	1	12	Throughout

A5.4 Comprehensive S-ACAD vs F-ACAD Comparison

Table A5-3. Detailed Methodology Comparison

Aspect	S-ACAD	F-ACAD
Process Control	Human-guided at each phase	Fully autonomous
Decision Points	19 human interventions	0 (post-initial prompt)
Quality Assurance	Human expert validation	AI sparring validation
Error Detection	Human clinical judgment	Multi-agent critique
Contextual Understanding	Deep, nuanced	Surface-level
Safety Considerations	Proactive, comprehensive	Reactive, rule-based
Adaptability	High (human insight)	Limited (predefined logic)

A5.5 F-ACAD AI Sparring Results

Critical Issues Identified by AI Critics (Total: 17)

High Priority (9):

1. Inconsistent emergency response timing (3-5 minute ambiguity)
2. Missing U.S. liability considerations
3. Inadequate differentiation between ER vs. PCP visits
4. Complex medical terminology without lay explanations
5. Insufficient empathetic tone for distressed parents
6. Medication dosing guidelines too generic
7. Missing age-specific nuances for <12 months
8. Incomplete post-seizure monitoring guidance
9. Structural redundancy in decision nodes

Medium Priority (8):

1. Inconsistent terminology (epilepsy vs. seizure disorder)
2. Limited visual guidance references
3. Missing vaccination history integration
4. Unclear time-of-day considerations
5. Insufficient multi-language considerations
6. Limited caregiver type differentiation
7. Missing insurance/cost guidance
8. Incomplete follow-up timeline specifications

A5.6 Key Insights from Comparison

Where F-ACAD Excelled:

- Rapid parallel data collection
- Comprehensive source coverage (25 vs. 18 sources)
- Consistent structure generation
- Automated cross-validation

Where S-ACAD Proved Superior:

- Clinical judgment integration
- Safety-first approach
- Contextual nuance capture
- Parent-centric communication
- Liability and ethical considerations

Convergence Analysis: Both approaches produced plausible algorithm drafts. However, the quality gap became evident during AI sparring: F-ACAD generated 17 critical issues (including 9 high-priority safety/accuracy concerns) requiring correction, whereas many of these issues were addressed earlier through 19 human expert interventions in S-ACAD. Overall, S-ACAD took ~3.6× longer than F-ACAD (245 min vs 67 min 58 sec), but it incorporated continuous expert validation and an independent clinical review checkpoint prior to any deployment. This was consistent with external safety verification reporting no medically inaccurate sections or critical safety errors requiring mandatory correction.

A5.7 Recommended Hybrid Approach

Based on the comparison, an optimal hybrid approach would:

1. Use F-ACAD for Phase 1 (Data Collection) - 10 minutes
2. Apply human review at Phase 2 (Validation) - 30 minutes
3. Use F-ACAD for Phase 3 (Draft Generation) - 10 minutes
4. Apply human-guided refinement at Phase 4 - 30 minutes
5. Automate Phase 5 (Documentation) - 9 minutes

Estimated Total Time: ~90 minutes (63% reduction from S-ACAD, with quality preservation)

Multimedia Appendix 6: The Expert Evaluation Form

External Expert Review Form: [Pediatric Febrile Seizure Response Algorithm]

1. Overview of the Review Request

- **Research Title:** Rapid Development of Parent-Actionable Pediatric Febrile Seizure Guidance Using a Human-AI Collaborative Workflow (S-ACAD): A Prospective Proof-of-Concept Study
- **Purpose of Review:** This document aims to systematically verify the clinical validity, safety, and completeness of the 'Pediatric Febrile Seizure Response Algorithm (v1.0),' which was rapidly developed through human-AI collaboration, from an independent, external expert's perspective.
- **Review Materials:** Pediatric Febrile Seizure Response Algorithm (v1.0) text, with the flowchart diagram attached separately.
- **Use of Review Results:** Your review comments will be **anonymized** and utilized as quantitative and qualitative data to substantiate the reliability of this algorithm and ensure the objectivity of the manuscript.
- **Contact:** *Corresponding Author (Anonymized for Peer Review)*

Thank you for taking your valuable time to provide this expert review.

2. Reviewer Information

- **Name:** _____
- **Affiliation:** _____
- **Specialty:** Board-Certified Pediatrician (Yes/No) / **Subspecialty:** _____
- **Clinical Experience:** Total _____ years
- **Date of Review:** _____ / _____ / 2025

Part A: Quantitative Validity Assessment

Please rate your agreement with each item on a 5-point scale.

(1: Strongly Disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly Agree)

173 Assessment Domain	Detailed Assessment Item	Scale (1-5)
1. Clinical Accuracy	1-1. The algorithm's decision-making criteria are consistent with current clinical guidelines and medical evidence.	1—2—3—4—5
	1-2. The Red Flag symptoms for identifying emergencies are clinically valid and appropriately included.	1—2—3—4—5
	1-3. The recommendations at each step (e.g., first aid, tests) are medically accurate.	1—2—3—4—5
2. Completeness	2-1. Most major clinical scenarios that a caregiver might experience (e.g., first seizure, recurrence) are included.	1—2—3—4—5
	2-2. Exceptional situations (e.g., under 6 months, over 5 years) are appropriately handled.	1—2—3—4—5
	2-3. There are no apparent missing steps or logical flaws in the decision-making pathways.	1—2—3—4—5
3. Safety	3-1. The criteria for identifying life-threatening situations are appropriately placed with the highest priority.	1—2—3—4—5

174 Assessment Domain	Detailed Assessment Item	Scale (1-5)
	3-2. The response guidelines for potentially dangerous situations are sufficiently conservative and safely designed.	1—2—3—4—5
	3-3. Safeguards to prevent potential harm (e.g., warnings for prohibited actions) are effectively included.	1—2—3—4—5
4. Usability for Parents	4-1. Medical terms are explained in simple language that a caregiver can understand.	1—2—3—4—5
	4-2. The action guidelines are clear and practically executable in an emergency.	1—2—3—4—5
	4-3. The overall decision-making flow is logical and easy to follow.	1—2—3—4—5

Part B: Qualitative In-depth Review

We would appreciate your in-depth expert opinion. Please describe freely.

B-1. What do you consider to be the greatest strength of this algorithm?

B-2. Are there any parts that you believe are medically inaccurate or potentially misleading and require mandatory correction? (If yes, please specify the node number and suggest a revision.)

( ) None

( ) Yes (Details below):

Node #___: _____

Node #___: _____

B-3. Are there any important clinical scenarios or Red Flag symptoms missing from the current algorithm? (If yes, please specify.)

( ) None

( ) Yes (Details below):

_____

Part C: Overall Assessment & Final Recommendation

C-1. Please rate the overall clinical validity of this algorithm on a scale of 1 to 10.

Overall Score: _____ / 10

C-2. Compared to existing written/verbal educational materials for parents, how would you rate the usefulness of this algorithm?

( ) Much Better

( ) Somewhat Better

( ) About the Same

( ) Somewhat Worse

( ) Much Worse

C-3. Assuming it successfully passes usability testing with caregivers, what is your opinion on using this algorithm for the development of an actual parent-facing service?

( ) Usable in its current state

( ) Usable after minor revisions

( ) Requires major revisions

( ) Not suitable for use

C-4. If you have any other comments or suggestions, please describe them freely.

_____

Thank you again for providing your expert review.

Multimedia Appendix 7: Anonymized Qualitative Feedback from External Experts

This appendix contains the full, anonymized text of the qualitative feedback (Parts B and C) provided by the two independent external reviewers. The original responses in Korean were translated into English by the author.

Reviewer A (Pediatric Emergency Physician, 15 years of experience)

B-1. What do you consider to be the greatest strength of this algorithm?

"It provides comprehensive and accurate guidance for febrile seizures, which can be relatively common in children, especially for situations that require attention."

B-2. Are there any parts that require mandatory correction?

None.

B-3. Are there any important clinical scenarios or Red Flag symptoms missing?

None.

C-4. Other comments or suggestions?

"Although the differences in medical access and delivery systems between the U.S. and South Korea must be considered, I have doubts about whether a non-medical caregiver can simply observe a convulsing child for more than 5 minutes without taking special measures."

Reviewer B (Primary Care Pediatrician, 15 years of experience)

B-1. What do you consider to be the greatest strength of this algorithm?

"When a child has a seizure at home, it provides parents with a way to respond, allowing them to observe the patient safely, which can likely provide reassurance to the caregiver."

B-2. Are there any parts that require mandatory correction?

None.

B-3. Are there any important clinical scenarios or Red Flag symptoms missing?

None.

C-4. Other comments or suggestions?

(No comments provided)

Multimedia Appendix 8: External Review Raw Data (Quantitative Scores)

This appendix provides the itemized, anonymized quantitative scores from the two independent external reviewers for transparency. These raw data supplement the qualitative findings and safety verification discussed in the main manuscript's Results section. The evaluation instrument can be found in Multimedia Appendix 6.

Disclaimer: Due to the small sample size (N=2), these scores lack the statistical power required for generalization. They should be interpreted primarily as qualitative indicators within this proof-of-concept study, rather than statistically validated metrics.

Part A: Quantitative Validity Assessment (5-Point Likert Scale)

Assessment Domain	Detailed Assessment Item	Reviewer A Score	Reviewer B Score
1. Clinical Accuracy	1-1. Consistency with clinical guidelines and evidence.	5	5
	1-2. Validity of Red Flag symptoms for emergencies.	5	5
	1-3. Medical accuracy of recommendations.	5	5
	Domain Average	**5.0**	**5.0**
2. Completeness	2-1. Inclusion of major clinical scenarios.	5	5
	2-2. Appropriate handling of	5	4

180 Assessment Domain	Detailed Assessment Item	Reviewer A Score	Reviewer B Score
	exceptional situations.		
	2-3. Absence of missing steps or logical flaws.	5	5
	Domain Average	**5.0**	**4.67**
3. Safety	3-1. Appropriate prioritization of life-threatening situations.	5	5
	3-2. Conservative and safe design of guidelines.	5	5
	3-3. Effective inclusion of safeguards and warnings.	5	5
	Domain Average	**5.0**	**5.0**
4. Usability for Parents	4-1. Simple language and clear explanation of terms.	5	5
	4-2. Practical executability of action guidelines.	**3**	5

181 Assessment Domain	Detailed Assessment Item	Reviewer A Score	Reviewer B Score
	4-3. Logical and easy-to-follow decision-making flow.	5	5
	Domain Average	**4.33**	**5.0**

Note: The domain averages presented in the main manuscript (e.g., Completeness: 4.8/5.0, Usability: 4.7/5.0) are the rounded averages of the two reviewers' domain-specific average scores.

Part C: Overall Assessment & Final Recommendation

Assessment Item	Reviewer A (Emergency Physician)	Reviewer B (Primary Care Pediatrician)
C-1. Overall Clinical Validity (out of 10)	9.0 / 10	9.5 / 10
C-2. Usefulness vs. Existing Materials	Much Better	Much Better
C-3. Recommendation for Actual Service	Usable after minor revisions	Usable after minor revisions

Note: The average overall clinical validity score reported in the manuscript (9.25/10) is the mean of the two reviewers' scores.

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract clearly states S-ACAD as a governance framework for AI-assisted clinical algorithm development, with explicit limitations (proof-of-concept, single case study).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4.3 discusses limitations including n-of-1 design, researcher bias, and lack of inter-rater reliability.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results requiring proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Complete prompts, AI outputs, and development logs are provided in Supplementary Materials (Appendices 1–4).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All prompts, outputs, and logs are provided in supplementary materials for full reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 2 specifies all AI models, versions, access dates, and workflow procedures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This is a single-case proof-of-concept study; statistical analysis is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All AI interactions used subscription-based web interfaces; no local compute resources required.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study used only publicly available medical information; reviewer feedback was de-identified.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 4 discusses positive impacts (faster knowledge translation) and potential risks (equity concerns, accountability).

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release models or datasets with high misuse risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All AI models were accessed via official subscription services; clinical guidelines are publicly available.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The S-ACAD workflow and prompts are documented in supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Expert review involved voluntary professional consultation, not crowdsourcing or human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No patient data or human subjects research; expert review feedback was de-identified.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.