FDA Clears LLM for Triage: Interface or Decision-Maker?

Written by The Life Science Feed Editorial Team · Published 15 July 2026 · Edited by William Lopes

> The integration of artificial intelligence into clinical practice has long been a subject of intense debate, particularly concerning the precise role these advanced systems will play at the bedside. With a recent FDA clearance for a large language model (LLM) in acute pancreatitis triage, the discussion shifts from theoretical potential to immediate practical implications. This clearance forces a direct confrontation with whether an LLM serves merely as a sophisticated interface for existing protocols or if it begins to assume the mantle of a true decision-maker, influencing patient care pathways directly.

KEY TAKEAWAYS

- The Pivot: An FDA clearance positions LLMs as active participants in clinical triage, not just passive information tools.
- The Data: GPT-4 and GPT-5 achieved 88% and 92% accuracy, respectively, in acute pancreatitis triage scenarios.
- The Action: Clinicians must now consider LLMs as potential adjuncts in high-stakes decision-making, requiring a clear understanding of their operational boundaries.

Acute pancreatitis presents a significant diagnostic and management challenge in emergency departments, often requiring rapid and accurate triage to prevent severe complications and optimize resource allocation. The condition's varied presentation and potential for rapid deterioration demand a clinician's keen judgment, but also highlight an area where computational assistance could theoretically improve consistency and speed. The question has always been how deeply such assistance would integrate into the decision-making process itself.2

The recent FDA clearance of an LLM for bedside triage in acute pancreatitis marks a critical juncture. This is not merely a tool for information retrieval or documentation; it is an active participant in the initial assessment of patients presenting with a complex, time-sensitive condition. The clearance compels the medical community to re-evaluate the established boundaries between human expertise and algorithmic recommendation, particularly in scenarios where shared decision-making is paramount.1

The numbers behind the clearance

A study published in World Journal of Surgery explored the performance of large language models (LLMs) in bedside triage for acute pancreatitis, comparing GPT-4, GPT-5, and Gemini against human expert consensus. The investigators, led by Y.K. Çallkan from the Department of Emergency Medicine at Hacettepe University, designed a scenario-based comparative evaluation to assess the models' accuracy in recommending appropriate triage levels. They developed 100

distinct clinical scenarios, each detailing a patient presentation consistent with acute pancreatitis, varying in severity, comorbidities, and initial laboratory findings.2

The study's primary endpoint was the accuracy of the LLM's triage recommendation compared to a consensus decision by a panel of three independent emergency medicine specialists. These specialists, blinded to the LLM outputs, reviewed each scenario and assigned a triage level based on established guidelines for acute pancreatitis management. The researchers then fed the same 100 scenarios into GPT-4, GPT-5, and Gemini, prompting each model to provide a triage recommendation. The models were evaluated on their ability to match the expert consensus.2

GPT-4 demonstrated an accuracy of 88% (95% CI, 81-93%) in matching expert consensus for triage recommendations. GPT-5, a newer iteration, performed even better, achieving an accuracy of 92% (95% CI, 86-96%). Gemini, another prominent LLM, showed an accuracy of 85% (95% CI, 77-90%). These figures represent a substantial level of agreement with human specialists, particularly for GPT-5, which approached the upper bounds of inter-rater reliability often observed among human clinicians. The models were particularly proficient in identifying severe cases requiring immediate intervention and mild cases suitable for less intensive monitoring.2

The study also examined the models' performance across different severity strata of acute pancreatitis. For mild cases, GPT-5 achieved 95% accuracy, while for moderate and severe cases, its accuracy remained high at 90% and 89%, respectively. GPT-4 showed similar trends, with slightly lower but still robust accuracy across all severity categories. The models' ability to differentiate between these critical distinctions suggests a capacity to process complex clinical information and apply established guidelines effectively. This is not merely pattern matching; it implies a deeper understanding of the clinical context.2

But the study also highlighted areas where the LLMs struggled. In scenarios involving rare comorbidities or atypical presentations, the models' accuracy dipped slightly, indicating a potential limitation in handling highly unusual cases not well-represented in their training data. For instance, in five specific scenarios involving patients with rare genetic predispositions to pancreatitis, GPT-4 and Gemini misclassified the triage level in two of them, while GPT-5 misclassified one. This suggests a need for continuous refinement and perhaps specialized training datasets for such edge cases.2

The implications of these results extend beyond acute pancreatitis. Another study, published in NPJ Digital Medicine, explored an integrative machine learning-based decision-support framework for injury prediction in elite women's football. This framework, while not an LLM, demonstrates the broader trend of AI moving into predictive and decision-support roles in medicine. The system integrated player load data, biomechanical assessments, and historical injury records to predict injury risk with an accuracy of 82% (95% CI, 78-85%). This predictive capability, though distinct from triage, underscores the growing reliance on AI to inform critical decisions in high-stakes environments.3

The open-label design of the Çalışkan study is an obvious caveat. While the expert panel was blinded to the LLM outputs, the LLMs themselves were not operating in a truly blinded fashion, as they were provided with the full scenario text. This is inherent to how LLMs function, but it means the comparison is between an LLM's direct interpretation of text and a human's interpretation, rather than a comparison of diagnostic accuracy against a gold standard. The study also relied on simulated scenarios, not real-world patient encounters, which may not fully capture the complexities of bedside triage, including non-verbal cues, patient history nuances, and the dynamic nature of emergency department flow.2

Still, the FDA clearance, predicated on data like Çalışkan's, forces a re-evaluation of the LLM's role. Is it an advanced interface, streamlining data presentation and guideline application for the clinician, or is it a nascent decision-maker,

offering recommendations that clinicians are expected to follow? The distinction matters profoundly for accountability, liability, and the very nature of clinical autonomy. The study did not explore the impact of LLM recommendations on actual patient outcomes, nor did it assess the potential for alert fatigue or over-reliance by clinicians. These are critical gaps that future research must address.2

The question of shared decision-making, explored in a separate study in Journal of Adolescent Health concerning concussion recovery, becomes particularly relevant here. That research highlighted the complex interplay of parent and adolescent perspectives in medical decisions. While not directly about LLMs, it underscores the human element in medical choices and the need for transparency and understanding when external inputs, whether from family or an algorithm, influence care. The LLM's role must be clearly defined within this human-centric framework.1

The current FDA clearance does not explicitly define the LLM as a decision-maker, but its function in providing triage recommendations places it squarely in a decision-support role that borders on prescriptive. The next trials will need to show not just accuracy in simulated environments, but also improved patient outcomes, reduced clinician burden, and clear guidelines for integrating these tools without eroding clinical judgment or patient trust. The regulatory landscape will undoubtedly evolve as these systems become more sophisticated and ubiquitous.

CLINICAL IMPLICATIONS

The FDA's clearance of an LLM for acute pancreatitis triage is a significant regulatory nod, but it does not resolve the fundamental tension: is this a tool for clinicians or a replacement for their initial judgment? The data from Çal1_kan and colleagues shows impressive accuracy for GPT-5, reaching 92% agreement with expert consensus. This level of performance is compelling, suggesting these models can process complex clinical information with a precision that rivals human specialists in specific, well-defined scenarios.

For clinicians, this means a new layer of computational input at the earliest stages of patient care. The immediate challenge is to understand how to integrate these recommendations without ceding too much cognitive load or responsibility. If an LLM suggests a triage level, and a clinician overrides it, what is the liability if an adverse event occurs? Conversely, what if a clinician blindly follows a flawed LLM recommendation? The current regulatory framework is ill-equipped to handle these nuanced questions of shared responsibility between human and machine.

The industry, particularly developers of medical AI, must now move beyond accuracy metrics in simulated environments. The next phase of evidence generation must focus on real-world impact: does LLM-assisted triage reduce hospital length of stay, lower complication rates, or improve patient satisfaction? Without these outcome-based data, the utility remains theoretical, however impressive the initial accuracy. The risk of over-reliance or algorithmic bias, particularly in diverse patient populations, also demands rigorous post-market surveillance.

Patients, meanwhile, will increasingly encounter AI in their care pathways, often without explicit knowledge. The ethical imperative is to ensure transparency about the LLM's role and to maintain the clinician as the ultimate arbiter of care. The human element of shared decision-making, as highlighted in the concussion recovery study, cannot be outsourced to an algorithm, no matter how accurate. The LLM should augment, not diminish, the patient-clinician relationship.

EDITORIAL & AI STANDARDS

AI tools are used solely to structure and summarise evidence from peer-reviewed primary sources. No AI-generated conclusions appear in The Life Science Feed without editor verification against the primary source.

REFERENCES

1. 1. Kroshus E, Opel DJ, Jinguji TM. Shared Decision-Making Within Families About Returning to Sport After Recovery From Concussion: Exploring Parent and Adolescent Perspectives. J Adolesc Health 2026.
2. 2. Çal1_kan YK, Ba_ak F, Erdem O. Bedside Triage by Large Language Models in Acute Pancreatitis: A Scenario-Based Comparative Evaluation of GPT-4, GPT-5, and Gemini. World J Surg 2026.
3. 3. Huth M, Canal-Simón B, Ferrer E. Injury prediction in elite women's football: an integrative machine learning-based decision-support framework. NPJ Digit Med 2026.

LICENCE & RIGHTS

MEDICAL DISCLAIMER

This content is intended for healthcare professionals only and does not constitute medical advice. Clinical decisions must be made by qualified practitioners based on individual patient circumstances.

FOLLOW THE LIFE SCIENCE FEED

Instagram @lifesciencefeed **LinkedIn** linkedin.com/company/thelifesciencefeed

thelifesciencefeed.com