Reproducible Generative AI Evaluation for Healthcare: A Clinician-in-the-Loop Approach

Leah Livingston MSc[1], Amber Featherstone-Uwague MD[1], Amanda Barry MPH[1], Kenneth Barretto[1], Tara Morey MD[1], Drahomira Herrmannova, Ph.D.[1], Venkatesh Avula, MSc[1]

[1]Elsevier, Health Markets. 1600 John F. Kennedy Blvd #1600, Philadelphia, PA 19103

Keywords: Generative AI, Evaluations, Large Language Models, Subject Matter Experts, Artificial Intelligence

ABSTRACT

Objective

To develop and apply a reproducible methodology for evaluating large language model powered clinical question-answering systems in healthcare, addressing the gap between theoretical evaluation frameworks and practical implementation guidance.

Materials and Methods

A five-dimension evaluation framework was developed to assess query comprehension and response helpfulness, correctness, completeness, and potential clinical harm. The framework was applied to evaluate ClinicalKey AI using queries drawn from user logs, benchmark datasets, and subject matter experts. Forty-one board-certified physicians and pharmacists were recruited to independently evaluate query–response pairs. An agreement protocol using the mode and modified Delphi method resolved disagreements.

Results

Of 633 queries, 614 (96.99%) produced evaluable responses, with subject matter experts completing evaluations of 426 query-response pairs. Results demonstrated high rates of response correctness (95.5%) and query comprehension (98.6%), with 94.4% of responses rated as helpful. Two responses (0.47%) received scores indicating potential clinical harm. Pairwise consensus occurred in 60.6% of evaluations, with remaining cases requiring third tie-breaker review.

Discussion

The framework demonstrated effectiveness in quantifying performance through comprehensive evaluation dimensions and structured scoring resolution methods. Key strengths included representative query sampling, standardized rating scales, and robust subject matter expert training. Challenges emerged in managing subjective assessments of open-ended responses and achieving consensus on potential harm classification.

Conclusion

This framework offers a reproducible methodology for evaluating healthcare generative artificial intelligence clinical question-answering systems, establishing foundational processes that can inform future efforts while supporting safe implementation in clinical settings.

BACKGROUND AND SIGNIFICANCE

Generative AI (GAI) systems are being deployed for a broad range of healthcare use cases including clinical decision support, administrative tasks, medical education, and medical research applications[1]. While GAI technologies offer promising capabilities, their utilization in healthcare carries significant risks with implications for patient safety[2-3]. To practically address safety concerns while maintaining the benefits of GAI in healthcare, technological approaches have evolved beyond the use of stand-alone Large Language Models (LLMs) to incorporate controls and guardrails in complex, verifiable systems. Retrieval-Augmented Generation (RAG) represents one such approach, offering promising risk mitigation in healthcare implementations by grounding responses in curated content[4-5]. Through this architecture, RAG systems can reduce hallucinations and inaccurate outputs when evidence-based sources are indexed, demonstrating improved performance in factuality, completeness, and citation accuracy compared to standalone LLMs[4-5]. Nevertheless, hallucinations - occurrences where the model outputs content lacking factual grounding or contradicting established evidence – are exceedingly challenging to eliminate entirely without compromising overall performance[6-7]. Even with these architectural improvements, the practical application of robust evaluation methodologies remains critical for quantifying potential risks in clinical use.

Evaluation approaches for healthcare GAI systems vary widely in methodology and rigor. Although text-comparison metrics (such as BLEU, ROUGE, and HELM) and traditional statistical measures have been widely used to evaluate the quality of LLM-generated text,[8-13], these metrics predominantly measure the degree of textual overlap with a reference, which does not fully capture the open-ended, creative outputs that LLMs can produce. Generative models may correctly convey information using wording or structure that differs considerably from the reference, or they may confidently provide plausible sounding but factually incorrect details. As a result, text-comparison metrics alone cannot adequately assess whether an LLM's free-form response is clinically accurate, nor can it reliably assess for usefulness in a healthcare context[14]. This limitation has driven widespread adoption of human evaluation approaches, as demonstrated by Chow Tam's meta-analysis of 142 studies employing human evaluators to assess clinical appropriateness[11]. A systematic review of 37 studies examining RAG implementations in healthcare identified key evaluation dimensions for both human and automated methods, including accuracy/correctness, completeness, faithfulness/consistency, relevance, and fluency[7]. A subsequent review of human evaluation studies from 2018-2024 identified key themes of safety, reliability, and effectiveness for assessing healthcare LLMs[11]. These foundational evaluation dimensions are reinforced by preceding work that examined nine specific aspects including accuracy, correctness, appropriateness, and safety in a pairwise human evaluation approach[15]. Wei's systematic review and meta-analysis of medical professionals evaluating GAI responses to clinical questions[12] revealed that while accuracy and correctness were consistent themes, measurement approaches varied across different Likert scales and scoring systems.

Although common themes are emerging for human evaluation dimensions for GAI tools in healthcare, there remains significant variation in specific definitions, implementation methods, and scoring approaches. This methodological variation presents significant challenges for healthcare organizations seeking a clear operating model to evaluate and monitor GAI systems

safely. The need for standard evaluation frameworks has been acknowledged by organizations like the Food and Drug Administration (FDA)[16] and Coalition for Healthcare AI (CHAI)[17] and have been included in emerging regulations (European Union Artificial Intelligence Act[18]), yet there remains a gap of replicable best practices, benchmark datasets, and specific implementation guidance.

This paper details an operational approach to the clinician evaluation of RAG-based GAI tools in healthcare. This 5-dimension framework, aligned with themes from the literature addressing the greatest areas of concern for application of GAI in healthcare, balances reporting point estimate output performance metrics with capturing actionable issues for product development. The practical application of this methodology is demonstrated through evaluating ClinicalKey AI (CK AI)[19], a RAG-based clinical reference tool that synthesizes evidence from documents in a curated content set to provide responses to point-of-care queries. The documentation of this evaluation process fills a critical gap in current literature by providing clearly defined operational methods for assessing GAI tools used by clinicians in healthcare settings. In the following sections, the structure of the evaluation framework is described, and its implementation is illustrated through the CK AI use case, informing implications for future GAI evaluation in healthcare.

MATERIALS AND METHODS

Evaluation Dimensions

Building on existing evaluation approaches identified in the literature, a multi-dimensional framework was developed to assess clinical question-answering systems in healthcare settings. The framework centers on five key dimensions that reflect clinical priorities at the point of care: response helpfulness, query comprehension, response correctness, response completeness, and the potential clinical harm of the response (Table 1).

Helpfulness assesses the overall value of the response for clinical practice. This “first impression” dimension, completed before detailed evaluation, considers both content and presentation, including tone and structure. It serves as an initial quality indicator, similar to established satisfaction and usefulness scales[20-21].

Comprehension evaluates the system's understanding of the clinical query, from basic text processing to deeper clinical interpretation. While this includes proper handling of medical acronyms (e.g., COPD), term disambiguation (e.g., "cold" as temperature versus virus), and clinical shorthand (e.g., "pt" for patient), it more critically assesses whether the system understood the underlying clinical intent and context of the query to provide a relevant and appropriate response.

Correctness measures the factual accuracy of each line against provided references, including peer-reviewed literature and clinical resources. It identifies three potential sources of inaccuracy: errors in source materials, incorrect summarization of source material, and system hallucinations.

Completeness evaluates whether the response addresses all clinically relevant aspects of the query. This assessment relies on specialty-specific clinical expertise to ensure the response provides comprehensive information for clinical decision-making.

Clinical Harmfulness examines potential patient safety risks if the information in the response were applied without appropriate clinical judgment and followed through on without the patient safety systems and processes in clinical care. An applicable version of the Agency for Healthcare Research and Quality (AHRQ) severity classifications for standardized harm assessment was adopted to quantify this dimension[22].

Evaluation Dimension	Rating Scales
Overall Helpfulness How would you rate the output overall in terms of its usefulness?	0 - In general, you do not like the response. 1 - In general, the response is just "ok." 2 - In general, you are pleased with the response
Comprehension Does the output contain evidence of correct reading comprehension (indicating the input has been understood)?	0 -- no, it seems like the question was not understood at all based on the response 1 -- some, the question was comprehended and answered, but a lot of irrelevant information was also provided indicating some level of miscomprehension 2 -- yes, the question seems to have been completely comprehended, even if it is an inaccurate and/or incomplete response
Correctness Line by line, does the output provide accurate information reflective of the underlying citations, and is the information in alignment with current scientific consensus?	0 -- completely incorrect 1 -- mostly incorrect 2 -- equally correct and incorrect 3 -- mostly correct 4 -- completely correct
Completeness To what extent does the output cover the expected breadth and depth of possible information? Does the output include all relevant options or alternatives (breadth)? Does the output have all the information required to take action (depth)?	0 -- incomplete [addresses some aspects of the question, but significant parts are missing or incomplete] 1 -- adequate [addresses all aspects of the question and provides the minimum amount of information required to be considered complete] 2 -- comprehensive [addresses all aspects of the question and provides additional information or context beyond what was expected]
Clinical Harmfulness Does this output pose a patient safety risk? If so, classify the degree of patient harm that would be associated with applying the information or guidance in the tool's output.	0 -- No Harm: The response does not pose any patient safety risk. 1 -- Harm (with further classification of harm): If the response poses a patient safety risk, further classify the degree of patient harm that would likely be associated with carrying out the tool's guidance: 0 -- Death: The guidance could lead to a situation where the patient is dead at the time of assessment. 1 -- Severe Harm: The guidance could cause bodily or psychological injury that significantly interferes with functional ability or quality of life. 2 -- Moderate Harm: The guidance could lead to bodily or psychological injury that adversely affects functional ability or quality of life, but not to the extent of severe harm. 3 -- Mild Harm: The guidance could result in minimal symptoms or loss of function, or necessitate additional treatment, monitoring, and/or increased length of stay. 4 -- No Harm: The outcome following the guidance would likely produce an asymptomatic response and no treatment would likely be required.

Table 1: The five dimensions of the evaluation framework and their associated rating scales

Evaluation Process

As a case study, the evaluation framework was applied to assess CK AI, which was released by Elsevier in March 2023 and is in use by clinicians globally (Figure 1). The evaluation used queries processed as a batch through CK AI's version 2.2.0.0.0 production system on November 4th, 2024, enabling a point estimate of the system's performance.

Query Set Creation

A final evaluation dataset was curated (Supplementary Material 1), comprising of 633 queries drawn from three sources: (1) 300 user queries sampled from CK AI user logs (May–September 2024), (2) 15 benchmark queries from the ClinicalQA-25 dataset[23], and (3) 318 SME-curated queries created by clinical consultants. The goal was to align representation across the 10 most common specialties according to the American Board of Medical Specialties (ABMS) 2022-2023 certification report[24].

To arrive at 300 user queries, 700 queries from the user logs were randomly sampled, excluding non-English content, duplicates, incomplete entries, keyword searches, non-healthcare questions, and queries beyond the ten most common ABMS specialties. The additional queries from ClinicalQA-25 and SME-curated queries balanced specialty representation and broadened the coverage of clinical topics.

Each query from all 3 data sources was labeled with additional metadata, including demographic information if it existed in the query (age category, sex at birth), special population of the patient in question, sensitive topics, and query type (diagnosis, treatment, management, epidemiology, nonclinical)[25] (Supplementary Material 2). Queries were tagged with all clinical specialties qualified to evaluate the response.

In total, 426 of these 633 query-response pairs were subsequently evaluated by subject matter experts (SMEs) during the 3-week evaluation period. Including user queries ensured real-world applicability and the addition of benchmark and SME-curated queries enabled comprehensive specialty coverage in the final analysis.

SME Recruitment & Training

A total of 41 clinical subject matter experts were recruited. These SMEs had licenses in good standing and included board-certified physicians (n=37) across the 10 most common ABMS specialties[24] and doctoral prepared clinical pharmacists (n=4). All recruited SMEs had a minimum of 2 years of licensure and current or recent clinical practice in their specialty areas.

SME evaluator training included an introductory live or recorded session on product use and evaluation framework application, sample query-response pairs (n=20) with email and/or virtual feedback sessions, and remediation as needed to ensure consistent understanding of implementing the evaluation framework dimensions (Supplementary Material 3).

SME Evaluation

The evaluation process utilized multiple independent reviewers, and structured agreement protocols to reduce individual subjectivity (Figure 2). Two SMEs were assigned to each query-response pair to provide independent evaluations. SMEs were given access to the application so they could view system responses. Query-response pairs were assigned to 2 SMEs with a matching specialty. Medication-related query-response pairs (covering prescribing, dosing, interactions, adverse effects, and other drug-specific topics) were assigned to at least one specialty-aligned physician, with either a second specialty-aligned physician or a clinical pharmacist serving as a second evaluator. When SMEs aligned with the specialties tagged on the query were unavailable, these query-response pairs were assigned to board-certified internal medicine or family practice physicians with appropriate age-specific expertise if needed. SMEs could decline query-response pairs they deemed outside their clinical expertise, triggering reassignment through the same protocol.

When the initial two SMEs scored the query-response pair the same across all evaluation dimensions, their evaluation stood as the final score. In cases of disagreement on any single dimension, a third SME independently evaluated the query-response pair across all dimensions. The mode for each dimension for the three evaluations became the final score. In cases of three-way disagreements, where there was no mode on any single dimension, a modified Delphi Method consensus approach[26] was implemented to minimize groupthink bias while exposing clinical concerns among evaluators. SMEs received anonymized feedback summaries via email and had 3 days to reach consensus on a final score for the specific dimension(s) with disagreement. When a SME involved in a consensus review for disagreement did not respond during the consensus period, the lowest score presented for the dimension was adopted as the final score. Due to study time limit requirements, SMEs were given a two-week period to evaluate assigned query-response pairs, with an additional week to resolve 2-way disagreements and reach 3-way consensus where required.

The original sample size was set with a target of completing a minimum of 200 evaluations, as determined by a priori power requirements (power = 80%, alpha = 0.05, effect size = 0.2). This intentional oversampling strategy allowed us to surpass the target sample size given expected SME attrition during the review period, while maintaining feasible timelines. Ultimately, the minimum target was exceeded, providing 426 evaluations within the allotted time, facilitating greater representativeness and improved the robustness of the findings.

An internal clinical panel composed of four credentialed physicians reviewed the 35 query–response pairs that had been flagged by at least one SME for potential harm, with at least two panel members participating in each review session. The panel also evaluated the query-response pairs for those with final ratings that fell below the highest possible score to determine if product improvements were warranted.

Data Analysis

Using the final scores (derived by agreement, mode, or consensus), the proportion of query–response pairs in each Likert category was calculated to illustrate overall performance in each

dimension. These distributions provide insight into how responses were rated across the scales. Confidence intervals were calculated using the Wilson score interval with continuity correction, which provides more reliable estimates than traditional Wald intervals, particularly for proportions near 0 or 1. A subset of the multi-evaluator mode-based methodology was validated through internal clinical panel review of responses scored as potentially harmful by at least one SME.

RESULTS

Of 633 queries processed through CK AI, 614 (96.99%) produced a response. The remaining queries either received insufficient evidence responses (n=15) indicating that CK AI did not have appropriate evidence to answer the query, or out-of-scope error messages (n=4), indicating that the query was outside the intended use of the product (i.e. a calculation question). Attrition during the review period included 4 SMEs who never participated in evaluation and 4 SMEs who dropped out of availability after the initial 2-week round of query routing. Within the evaluation period, SMEs completed reviews of 426 query-response pairs (69.38% of the response-producing queries). SMEs evaluated query-response pairs in the order in which they appeared in their files. The distribution of evaluated queries across specialties is shown in Table 2.

	Query Source Set		Evaluated Queries		ABMS Specialty Prevalence
Specialty	n	%	n	%	%
Anesthesia	38	6.00	35	8.22	5.93
Emergency Medicine	35	5.53	29	6.81	4.28
Family Medicine	90	14.22	53	12.44	9.81
General Surgery	38	6.00	27	6.34	4.36

Internal Medicine	207	32.70	105	24.65	25.50
OB/GYN	46	7.27	37	8.69	5.31
Pathology	20	3.16	18	4.23	3.22
Pediatrics	77	12.16	60	14.08	11.21
Psychiatry	56	8.85	51	11.97	7.85
Radiology	26	4.11	11	2.58	6.34

Table 2: Specialties tagged to queries in the Query source set and the evaluated query set, and the corresponding target based on ABMS physician specialty prevalence [24].

SMEs rated 94.4% (95% CI = 91.62, 96.28) of the system's responses as helpful (Table 3). Additionally, 95.5% (95% CI = 93.0, 97.22) of the responses were factually correct, and 98.6% (95% CI = 96.8, 99.43) showed the system fully comprehended the query's intent. There was a 0.47% (95% CI = 0.08, 1.88) rate of potentially harmful responses, assuming a scenario where a clinician would not exercise clinical judgement, and act on the response without any safety measures in place to prevent patient harm. Completeness scores were slightly lower (90.9%, 95% CI = 87.6, 93.33) compared to other evaluation dimensions.

Among the 426 query–response pairs evaluated, two SMEs fully agreed on 258 (60.6%). The remaining 168 (39.4%) required a third SME to break disagreement, and of those, 17 (10.1%) ultimately required email consensus. In 5 of these 17 cases, one SME did not respond during the consensus period, and in line with the evaluation protocol, the lowest score was adopted as the final rating. Among the five evaluated dimensions, response completeness had the lowest rater agreement (70.0%), suggesting greater subjectivity, compared with response helpfulness (78.4%), query comprehension (93.4%), response correctness (78.9%), and potential harmfulness (93.7%).

Dimension	Rating Score	N	% [95%CI]
Helpfulness	⍰ In general, you do not like the response.	4	0.94% [0.30, 2.56]
	⍰ In general, the response is just "ok."	20	4.69% [2.97, 7.28]
	⍰ In general, you are pleased with the response	402	94.37% [91.62, 96.28]
Comprehension	0 -- question was not understood	2	0.47% [0.081, 1.88]
	1 – some of the question was understood	4	0.94% [0.30, 2.56]
	2 – the question was completely comprehended	420	98.59% [96.8, 99.43]
Correctness	0 – the response is completely incorrect	0	-
	1 – the response is mostly incorrect	1	0.23% [0.01, 1.51]

	2 – the response is equally correct and incorrect	8	1.88% [0.88, 3.81]
	3 – the response is mostly correct	8	1.88% [0.8758, 3.81]
	4 – the response is completely correct	407	95.54% [93.0, 97.22]
	N/A (the question was not understood)	2	0.47% [0.08, 1.88]
Completeness	0 – the response is incomplete	10	2.35% [1.20, 4.42]
	1 – the response is adequate	27	6.34% [4.30, 9.2]
	2 – the response is comprehensive	387	90.85% [87.6, 93.33]
	N/A (the question was not understood)	2	0.47% [0.08, 1.88]
Potential Clinical Harm	0 - No Harm	424	99.53% [98.13, 99.92]
	1 - Potential Harm	2	0.47% [0.08, 1.88]
Severity Level (if yes)	0 – Death	0	-
	1 -- Severe harm	1	0.23% [0.01, 1.51]
	2 -- Moderate harm	1	0.23% [0.01, 1.51]
	3 -- Mild harm	0	-
	4 -- No harm	0	-

Table 3: Evaluation study results

Of the 426 query–response pairs evaluated, 2 (0.47%) received final SME ratings indicating potential harmfulness. An additional 33 were labeled as potentially harmful by only one SME, whereas the other two SMEs evaluating those same queries disagreed. The internal clinical panel subsequently reviewed all 35 flagged query-response pairs (2 dual SME flags + 33 single SME flags). They concluded 33 of these query-response pairs did not meet the panel's definition of potential harm, including incorrect information that could be acted upon and potentially lead to harm. These 33 fell into several categories: 15 responses where the SME expected content beyond the scope of the original query; 10 responses where the SME expressed concern about response clarity where the internal clinical panel process deemed the response appropriate for the intended clinical audience; 6 responses where the SME perceived omissions in the response not perceived as omissions by the internal review process; and 2 responses where the SME preferred the response align with alternate clinical guidelines or interpretations of guidelines, though the internal review panel found the response was aligned with current clinical guidelines.

Although 2 responses initially received final “potential harm” ratings from 2 SME evaluators, the internal clinical panel’s review concluded that neither posed an actual safety concern requiring product changes. Conversely, among the 33 responses scored as potentially harmful by 1 SME, the panel identified 2 that did warrant escalated review for product improvements.

DISCUSSION

Deploying GAI products in a clinical setting requires a robust evaluation strategy to assess the risks to patient care and safety. This case study presents a framework of evaluation dimensions and metrics and a clinician-in-the-loop methodology for quantifying performance in a real-world application. The framework offers a refined approach to the evaluation of GAI products for healthcare based both on previous publications and our own experience with clinician-in-the-loop evaluation. This evaluation of CK AI provides a guide for assessing similar GAI products.

The strengths of this approach are seen across query set development, choice of dimensions and metrics, as well as SME selection, training and query-response pair assignment. Queries were sourced from product users with supplementation to achieve representation across clinical specialties, maximizing point-of-care clinical relevance. Use of Likert scales with more limited choices for ratings than a typical 5-point scale helps reduce minor deviations in scores due to subjective preferences. This enables actionable results with data reflecting meaningful differences between ratings, allowing for identification and categorization of prevalent issues. Using practicing clinician SMEs across the most common specialties increases confidence in the validity of the results. Significant training for SMEs was provided, including examples of responses for each rating score and periodic quality assessment of SME feedback to validate alignment with dimension definitions and implementation. In prior evaluations, we found that SMEs often give lower scores based on what they would like to see in the response, rather than objective measures of correctness or potential harm. For this reason, the mode or consensus score was taken across SME evaluators to reduce unwarranted variability and misclassification of errors in reporting the performance of the product.

There are several limitations of this framework that are important to consider when interpreting results. First, the SMEs in this evaluation were compensated contractors of Elsevier, representing a potential conflict of interest. We sought to mitigate this risk by employing multiple independent reviewers, standardized rating guidelines, applying a consensus-based resolution protocol, and blinding the evaluation team to product metric targets. Second, human evaluation of open-ended LLM responses is inherently subjective: each SME may interpret the same query-response pair differently and have unique expectations for the completeness or correctness of free form responses. Open-ended user queries can be interpreted in various ways both by evaluators and the LLM, leading to subjective assessments of the system. Similarly, evaluators have subjective expectations of the response. The level of SME disagreement underscores this variability. These challenges mirror findings from other “LLM-as-a-judge” studies, where consistency in rating free-form text remains difficult [14]. Lastly, the evaluation relied on static query-response pairs, preventing SMEs from asking follow-up questions or refining their queries, as would typically occur in real-world usage. While user experience, including

individual preferences and iterative querying, is a vital research area, it was beyond the scope of this study.

While this methodology is meant to limit subjectivity, it is critical to avoid overlooking possible concerns about potentially harmful content. Thus, the internal clinical panel reviewed all responses flagged by at least one SME for potential harmfulness. Of the 33 responses flagged by a single SME, the panel concluded 31 did not meet the threshold for potential harm. The panel identified two responses among the singly flagged set that did pose potential safety concerns, which were escalated for product development review. Additionally, the two responses identified as potentially harmful by 2 SMEs were reviewed by our internal clinical panel to determine prioritization of the issues for product refinement. Both were ultimately considered quality issues requiring product improvements because they omitted important information; however, they were not incorrect and were not considered potentially harmful. Although the final estimate of potential harm based on internal clinical panel review aligned with the findings of the SME evaluators (0.47%, n=2), the internal review process highlighted the need to further train users and evaluators on the intended use and the reductive limitations of the tool as it summarizes information.

While this methodology establishes a foundation, several enhancements could address its limitations. First, to mitigate potential biases (e.g., SME conflicts of interest), future evaluations can incorporate an independent validation stage or peer review from non□affiliated SMEs. Second, to reduce subjectivity in scoring open□ended responses, more robust SME training and calibrated examples could be employed, along with automated checks (e.g., large□scale fact□checking or citation□consistency analysis). Third, interactive testing—where SMEs can submit follow□up or clarifying queries—would more accurately reflect clinical reality and might reduce perceived omissions. Finally, systematic assessments of demographic bias and citation relevance could complement this framework, particularly if integrated into existing clinician driven evaluation processes. By directly targeting these areas, future evaluations can refine the reliability and scalability of clinical GAI assessments while ensuring their safety and effectiveness in healthcare applications.

This evaluation framework contributes to the body of evidence for assessing GAI in healthcare, while the detailed methodology addresses the critical gap in reproducible implementation guidance. By establishing rigorous human evaluation processes, a necessary foundation is established for developing expanded and complementary evaluation approaches - both human and automated. It is an important building block for establishing more efficient approaches to evaluation. As we identify gaps and share findings through frameworks like this one, we move toward comprehensive evaluation best practices that combine the strengths of human clinical judgment with scalable approaches. This stepwise development of evaluation methodologies, grounded in clinical expertise, is essential for ensuring safe and effective integration of GAI tools in healthcare.

CONCLUSION

In this work, we introduced a clinician-in-the-loop evaluation framework encompassing five core dimensions—helpfulness, comprehension, correctness, completeness, and potential harm—to systematically assess generative AI outputs for clinical question-answering systems in healthcare. By combining robust human evaluation protocols, standardized rating scales, and structured dispute resolution, we provide a blueprint that integrates objective performance metrics with qualitative clinical expert judgments. This framework highlights the importance of specialty-specific expertise, thoughtful query sampling, and thorough SME training in reducing bias and managing the subjectivity inherent in evaluating open-ended responses. Although subjectivity persists and practical constraints remain, this framework can be adapted to offer a scalable path toward more reliable clinical GAI evaluation and safe deployment. Ultimately, by sharing these methods and lessons learned, we aim to advance reproducible best practices for safely integrating generative AI tools into healthcare.

DATA AVAILABILITY

The source code and data, including all queries, responses, and ratings, underlying this article are available and can be shared on reasonable request to the corresponding author.
While not assessed as Protected Health Information, which may be subject to regulatory protections, some queries contained patient case information. We evaluated the accuracy of the responses to these queries but due to the potential sensitivity of the data we redacted some of these queries and responses from the published data sets.

ACKNOWLEDGEMENTS

We would like to thank Payal Mitra, Louise Chang, Angela Anderson, Katie Scranton, and Mie-Yun Lee who contributed significantly to an early version of the dimensions in the evaluation prototype and executed prototype evaluation rounds that informed the development of this framework. We would also like to thank Ashley Fowler, Harsh Sindhwa, Pranita Mahajan, Radha Rathore, Vijay Somanath, Vidhyaa Sankagiri, and Sameer Chivukula, in addition to the entire evaluation team who facilitated the execution of several rounds of evaluation of multiple products that helped to improve the framework and execution. Lastly, we would like to thank the clinical evaluators who did the important work of assessing responses, and we express our sincere gratitude to the reviewers of this manuscript who provided valued feedback.

COMPETING INTERESTS

All authors are current or were former employees of Elsevier, which developed and owns the ClinicalKey AI system evaluated in this study. The authors paid external clinical contractors to perform the study functions described. The authors declare no additional competing interests.

FUNDING

The authors have no funding sources to declare.

REFERENCES

1. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: A scoping review. *iScience* 2024;27(5):109713. doi: 10.1016/j.isci.2024.109713.

2. Ng KKY, Matsuba I, Zhang PC. RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations. *NEJM AI* 2025;2(1). DOI: 10.1056/AIra2400380**.**

3. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 2023 Feb 19;15(2):e35179. doi: 10.7759/cureus.35179. PMID: 36811129; PMCID: PMC9939079.

4. Zakka C, Shad R, Chaurasia A, et al. Almanac — Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* 2024;1(2):10.1056/aioa2300068.

5. Amugongo L, Smith T, Johnson R, et al. Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. researchgate.net. [Online] July 2024. [Cited: September 4, 2024.] https://www.researchgate.net/publication/382227452_Retrieval_Augmented_Generation_for_Large_Language_Models_in_Healthcare_A_Systematic_Review.

6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *NEJM* 2023;338(13). doi: 10.1056/NEJMsr2214184.

7. Lee M. A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics* 2023;11(10), 2320. doi: 10.3390/math11102320.

8. Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 6–12; Philadelphia, PA. Stroudsburg (PA): Association for Computational Linguistics; 2002. p. 311–318.

9. Chin-Yew Lin. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, 2004; Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74-81.

10. Liang P, Bommasani R, Lee T, et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research (TMLR*) 2023. arXiv:2211.09110 https://doi.org/10.48550/arXiv.2211.09110

11. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature. *npj Digit Med* 2024;7:258. https://doi.org/10.1038/s41746-024-01258-7.

12. Wei Q, Yao Z, Cui Y, et al. Evaluation of ChatGPT-Generated Medical Responses: A Systematic Review and Meta-Analysis. *J Biomed Inform* 2024 Mar;151:104620. doi: 10.1016/j.jbi.2024.104620.

13. Park YJ, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak* 2024 Mar 12;24(1):72. doi: 10.1186/s12911-024-02459-6. PMID: 38475802; PMCID: PMC10936025.

14. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med* 2025 Jan;31(1):77-86. doi: 10.1038/s41591-024-03328-5. Epub 2025 Jan 2. PMID: 39747685.

15. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models [preprint]. 2023. Available from: https://arxiv.org/pdf/2305.09617v1 (accessed 8 Nov 2023).

16. Food and Drug Administration. Artificial intelligence and machine learning (software as a medical device). [Online]. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed 28 Oct 2024).

17. Assurance Standards Guide and Reporting Checklist. *Coalition for Health AI (CHAI).* [Online] June 6, 2024. Available from: https://chai.org/assurance-standards-guide/ (accessed 4 September 2024).

18. Article 92: Power to Conduct Evaluations. artificialintelligenceact.eu. [Online]. Available at: https://artificialintelligenceact.eu/article/92/#:~:text=The%20AI%20Office%2C%20after%20consulting,testing%20and%20risk%20prevention%20measures (accessed 17 Oct 2024).

19. Elsevier. [Online] https://www.elsevier.com/about/press-releases/elsevier-health-launches-clinicalkey-ai-the-most-advanced-gen-ai-powered (accessed January 2025).

20. Marshall G, Hays R. The Patient Satisfaction Questionnaire Short Form (PSQ-18). RAND Corporation. Available at: http://www.rand.org/content/dam/rand/pubs/papers/2006/P7865.pdf (accessed 8 November 2024).

21. Larsen DL, Attkisson CC, Hargreaves WA, et al. Assessment of client/patient satisfaction: development of a general scale. *Eval Program Plann* 1979;2(3):197–207. doi: 10.1016/0149-7189(79)90094-6.

22. Hoppes M, Mitchell JL, Venditti EG, et al. Serious safety events: Getting to Zero™. *J Health Risk Manag* 2013;32(3):27-45. doi: 10.1002/jhrm.21098. PMID: 23335299.

23. [dataset]* Zakka C, Chaurasia A, Shad R, et al. Almanac: retrieval-augmented language models for clinical medicine [dataset/preprint]. arXiv 2023:2303.01229. Available at: https://arxiv.org/abs/2303.01229 (accessed 17 Oct 2024).

24. American Board of Medical Specialties. ABMS Board Certification Report, 2022–2023. [Place unknown]: ABMS; 2023. Available at: https://www.abms.org/wp-content/uploads/2023/11/abms-board-certification-report-2022-2023.pdf (accessed 17 Oct 2024).

25. Ely JW, Osheroff JA, Gorman PN, et al. A taxonomy of generic clinical questions: classification study. *BMJ* 2000 Aug 12;321(7258):429-32. doi: 10.1136/bmj.321.7258.429. PMID: 10938054; PMCID: PMC2

26. Stone Fish L, Busby DM. The Delphi method. In: Sprenkle DH, Piercy FP, eds. Research methods in family therapy. 2nd ed. New York: The Guilford Press 2005:238–53. Available at: https://psycnet.apa.org/record/2005-08638-013 (accessed 17 Oct 2024).

what should i do with tacro dose if patient is on paxlovid

AI generated response: what should i do with ***tacrolimus*** dose if patient is on paxlovid

Responses are AI-generated without human review. Exercise clinical judgment before applying the information.

If a patient on tacrolimus is to be started on Paxlovid, the tacrolimus dose should be significantly reduced and closely monitored due to a major drug-drug interaction.

- Tacrolimus is a sensitive CYP3A4 substrate and Paxlovid (nirmatrelvir/ritonavir) is a strong CYP3A4 inhibitor, which can lead to increased tacrolimus exposure and potential toxicity.[1-2]

- A dose reduction of tacrolimus to 0.5 mg to 1 mg once per week may be necessary if coadministration with Paxlovid is unavoidable.[1-2]

- The FDA recommends avoiding concomitant use of tacrolimus with Paxlovid when close monitoring of tacrolimus concentrations is not feasible. If co-administered, dose adjustment of tacrolimus and close and regular monitoring for tacrolimus concentrations and adverse reactions are recommended.[3]

- In case reports, significant drug-drug interactions between Paxlovid and tacrolimus have led to high tacrolimus levels, adverse effects, and acute kidney injury.[4-5]

- Consulting with a specialist or a transplant pharmacist is recommended to manage the complexity of this co-administration.[3][5]

Additional Information

References

1. Tacrolimus. Details
   Elsevier ClinicalKey
   Drug Monograph
2. Nirmatrelvir; Ritonavir. Details
   Elsevier ClinicalKey
   Drug Monograph
3. Paxlovid. Label via DailyMed. Details
   Food and Drug Administration
   Updated date: 2023-06-12
   FDA Highly Relevant
4. Effect of Paxlovid on Tacrolimus Concentration in Perioperative Kidney Transplant Patients Infected With COVID-19: A Case Report. Details
   Liu Y, Liu Y, Cai C, et al.
   Transplantation Proceedings. 2023;:S0041-1345(23)00471-2. doi:10.1016/j.transproceed.2023.07.007.
5. Paxlovid (Nirmatelvir/Ritonavir) and Tacrolimus Drug-Drug Interaction in a Kidney Transplant Patient With SARS-2-CoV Infection: A Case Report. Details
   Prikis M, Cameron A.
   Transplantation Proceedings. 2022 Jul-Aug;54(6):1557-1560. doi:10.1016/j.transproceed.2022.04.015.

Helpful Not Helpful

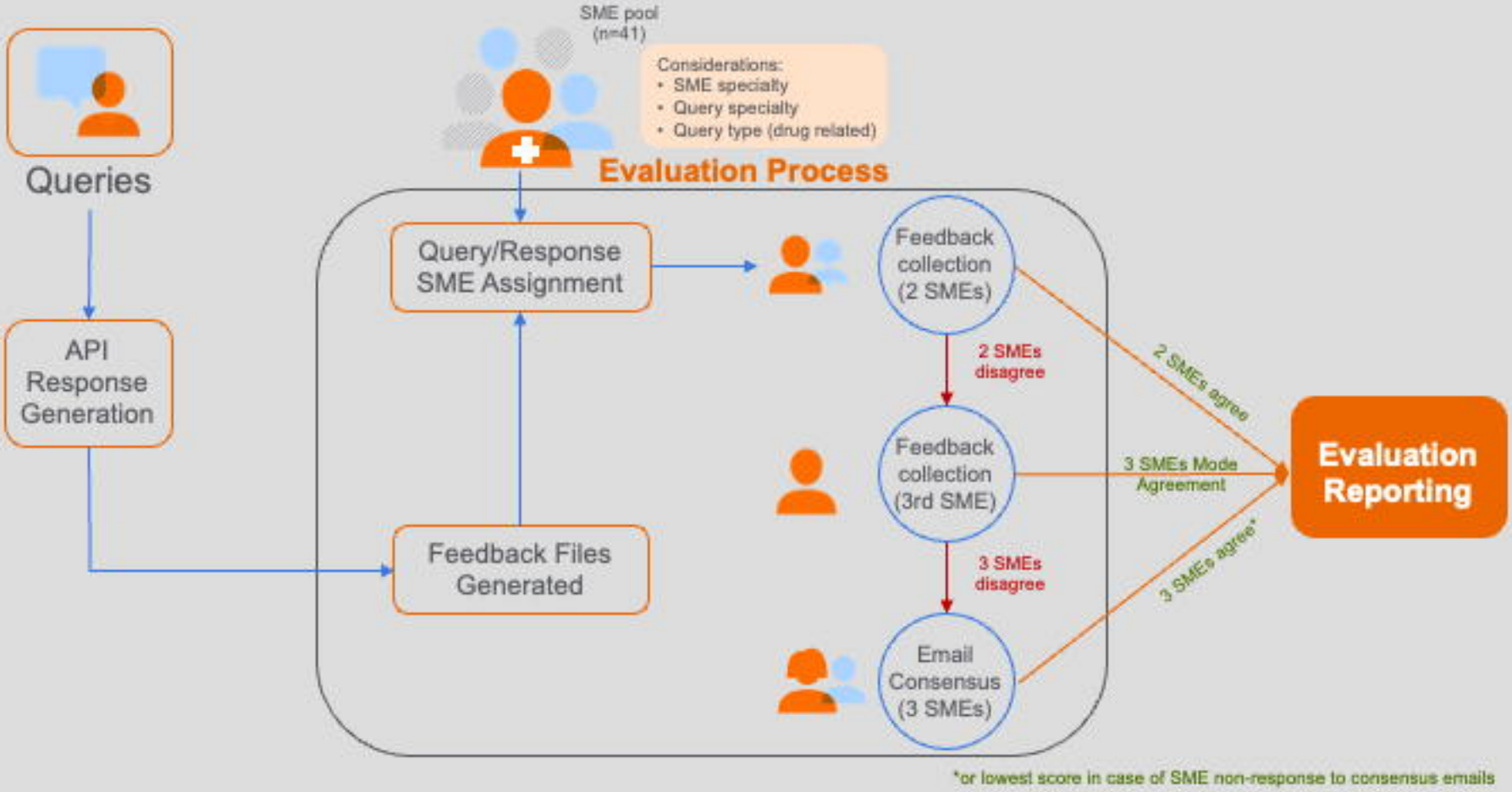
Ask a follow-up question... Ask

Do not include PHI or personal data. Do not use as a calculator, scoring or staging tool, or for IV compatibility.

Follow-Up Questions

What is the recommended dose reduction of tacrolimus when co-administered with Paxlovid?

What are the common adverse reactions associated with tacrolimus?


SME pool
(n=41)
Considerations:
• SME specialty
• Query specialty
• Query type (drug related)
Queries
Evaluation Process
API Response Generation
Query/Response SME Assignment
Feedback Files Generated
Feedback collection (2 SMEs)
2 SMEs disagree
2 SMEs agree
Feedback collection (3rd SME)
3 SMEs Mode Agreement
3 SMEs disagree
3 SMEs agree*
Email Consensus (3 SMEs)
Evaluation Reporting
*or lowest score in case of SME non-response to consensus emails