Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(2):593-601

Evaluation of a guideline-anchored retrieval-augmented generation (RAG) system for headache diagnosis and management: A case-vignette study

Gulcin Babaoglu

Ankara Bilkent City Hospital, Department of Pain Medicine, Ankara, Türkiye

Received 10 March 2026; Accepted 08 April 2026
Available online 22 April 2026 with doi: 10.5455/medscience.2026.03.055

Abstract

Headache disorders are frequently misdiagnosed, because of diagnostic criteria complexity. This study aimed to evaluate the diagnostic accuracy, safety, and usability of a guideline-anchored retrieval-augmented generation (RAG) clinical decision support workflow for headache diagnosis and management. Thirty-five fictional case vignettes representing primary headache disorders, secondary headache disorders, and neuropathic or facial pain syndromes were developed. NotebookLM was indexed with the International Classification of Headache Disorders (ICHD-3) and major headache treatment guidelines. Two independent pain specialists assigned gold-standard diagnoses and evaluated model outputs. Outcomes included diagnostic accuracy, treatment concordance (5-point Likert scale), red-flag detection using SNNOOP10 criteria, usability using the System Usability Scale (SUS) and Healthcare Systems Usability Scale (HSUS), and readability using Flesch Reading Ease (FRE) and Flesch–Kincaid Grade Level (FKGL). The system correctly classified secondary versus non-secondary headache categories in all vignettes (35/35, 100%). Subtype-level diagnostic agreement with expert evaluation was also perfect (κ=1.00). Acute treatment recommendations were rated as fully concordant in 62.9% and mostly concordant in 37.1% of cases, with no unsafe recommendations identified. Preventive therapy suggestions were applicable in 20 cases and achieved high concordance (Likert score ≥4 in 100%). Red-flag detection sensitivity was 87.5% overall and 93.3% for secondary headache cases, with 100% specificity. Usability scores were favorable (mean SUS 77.5; HSUS 4.5/5). However, readability analysis indicated highly technical output with very low FRE scores. A guideline-constrained RAG workflow demonstrated high diagnostic agreement and strong guideline concordance in headache case vignettes. While the system shows promise as a supervised clinical decision support tool, further validation in real-world clinical settings is required before routine implementation.

Keywords: Headache disorders, clinical decision support systems, artificial intelligence, natural language processing, diagnostic accuracy

Introduction

Headache disorders are among the most prevalent neurological conditions worldwide; however, accurate classification and management remain challenging. Epidemiological studies indicate that although 74% of patients with migraine reported receiving a physician's diagnosis, 36% wait more than two years before diagnosis is established [1]. Moreover, a cross-sectional study conducted across six countries found that fewer than 50% of respondents with migraine reported receiving a migraine diagnosis [2]. These diagnostic gaps persist across multiple levels of care and underscore the need for consistent application of standardized criteria. One contributing factor is that the International Classification of Headache Disorders (ICHD-3) is hierarchical and highly granular; using sequential numeric codes that integrate clinical and laboratory features and require clinicians to satisfy multiple criteria [3]. In routine clinical practice, clinicians may not apply every criterion systematically and overlapping symptom profiles further blur the boundary between headache disorders.

Alongside these diagnostic challenges, digital tools are reshaping healthcare. Artificial intelligence (AI) now supports tasks ranging from radiology and ophthalmology to hematology, yet adoption remains cautious because patients and clinicians worry about safety and factual reliability [4].

Large language models (LLMs) can summarize medical literature, answer clinical questions, and draft clinical

CITATION
Babaoglu G. Evaluation of a guideline-anchored retrieval-augmented generation (RAG) system for headache diagnosis and management: A case-vignette study. Med Science. 2026;15(2):593-601.

Corresponding Author: Gulcin Babaoglu, Ankara Bilkent City Hospital, Department of Pain Medicine, Ankara, Türkiye
Email: gulcinpektasli@gmail.com

DOI: 10.5455/medscience.2026.03.055

documentation; however, they may also hallucinate due to limitations in training data and the auto-regressive generation process [5]. Such hallucinations can yield inaccurate diagnostic or therapeutic suggestions and reinforce flawed reasoning [5]. The risk increases when prompts omit key clinical details, and conventional LLMs rarely ask clarifying questions [5].

The reliability of LLMs in healthcare is under increasing scrutiny. In a large-scale assessment of LLM-generated clinical notes, investigators reported hallucinations and omissions that could be reduced—but not eliminated—through prompt refinement [6]. Separate analyses of adversarially fabricated clinical prompts demonstrate that state-of-the-art models can elaborate on false details at high rates, underscoring persistent safety risks in clinical environments [7].

Retrieval-augmented generation (RAG) combines a retrieval component that searches authoritative sources with a generator that conditions its output on retrieved evidence. By grounding responses in guidelines and evidence sources, RAG may reduce hallucinations and improve transparency [8]. Recent medical reviews highlight that RAG architectures integrate external knowledge to produce more reliable and contextually relevant responses [9]. Dual-RAG systems that combine dense and sparse retrieval strategies may incorporate additional safeguards while maintaining high accuracy [8]. For example, a RAG system evaluated using synthetic questions about clinical laboratory regulations performed well when responses were grounded in retrieved regulatory sources [10].

NotebookLM is a retrieval-augmented language model that allows users to upload documents and interact with them conversationally. By conditioning responses on uploaded source materials, the system can cite specific guideline sections and, in principle, minimize hallucinations. These features make it a plausible clinical decision support system (CDSS) for headache medicine, where diagnostic criteria are complex and misclassification is common. Before such systems can be integrated into routine clinical workflows, their accuracy, safety, and usability must be assessed rigorously.

The objective of this study was to determine whether a knowledge-augmented clinical decision support workflow could accurately classify primary and secondary headache disorders and generate guideline-concordant treatment plans. We hypothesized that NotebookLM, seeded with the ICHD-3 diagnostic criteria and treatment guidelines for headache, would achieve high diagnostic accuracy and provide safe, comprehensive recommendations when evaluated by pain specialists.

Material and Methods

Study Design

This vignette-based observational study evaluated guideline-constrained decision support outputs for headache diagnosis and management. The unit of analysis was the individual vignette, and all outcomes were assessed at the level of model outputs relative to expert-derived standards. The focus was not real-time clinical workload but the usability, diagnostic accuracy, and clinician-perceived quality of AI-generated outputs in a vignette setting.

Ethics

Because the study used fictional case vignettes and did not involve identifiable patient data, formal ethics approval and informed consent were not required.

Case Vignettes

Thirty-five fictional headache case vignettes were constructed by one expert to represent a broad diagnostic spectrum, including primary headache disorders, secondary headache disorders, and selected neuropathic/facial pain syndromes (Supplementary Table S1). Each vignette included patient demographics, symptom chronology, associated features, relevant medical history, and examination findings. The gold-standard diagnosis and ideal treatment approach based on ICHD-3 and the European Academy of Neurology and International Headache Society guidelines were assigned by two different headache specialists who were not involved in vignette construction and were blinded to each other's initial ratings; disagreements were resolved by consensus. Vignettes were balanced to include common migraine and tension-type phenotypes, clinically urgent secondary causes, and unilateral facial pain syndromes that can mimic headache emergencies.

Knowledge Base and NotebookLM Prompt

The CDSS was implemented using NotebookLM (Google LLC; powered by the Gemini 3.0 Pro architecture at the time of access; Google does not publicly disclose specific version identifiers or update logs for NotebookLM, and the system may receive incremental updates without user notification; accessed February 2026). The knowledge base consisted of (i) the full ICHD-3 diagnostic criteria and classification hierarchy and (ii) the European Academy of Neurology and International Headache Society treatment guidelines [11-16]. These documents were uploaded into NotebookLM in PDF format. For CDSS, a structured prompt was submitted: For each case, evaluate headache burden, activity limitation, and impact on the patient using the available metrics; calculate six-item Headache Impact Test (HIT-6) and Migraine Disability Assessment Scale (MIDAS) scores. Based on symptoms, list a differential diagnosis and the necessary investigations. To reach a final diagnosis, ask additional guideline-based questions or request tests if needed, if not give final diagnosis. Evaluate each case for red flags; if present, summarize them, and if not, state "no red flags." Summarize evidence-based treatment options for the final diagnosis. Limit recommendations to guideline content

and state which section is cited. If evidence is insufficient, state "insufficient information in the guideline."

Each vignette was submitted in a separate, fresh NotebookLM session to prevent carryover between cases and minimize context contamination. NotebookLM's responses were captured by direct copy–paste into a structured spreadsheet (no manual transcription). Readability metrics (Flesch Reading Ease (FRE) and Flesch–Kincaid Grade Level (FKGL)) were computed automatically using textstat library in Python (version 3.11).

Expert Evaluation

Two pain specialists with expertise in headache management, who were not involved in vignette construction, independently and blindly reviewed all cases. During initial scoring, each rater was blinded to the other rater's evaluations. They assigned the ground-truth diagnosis for each vignette and rated the corresponding NotebookLM output according to the outcome measures described below. In addition, they noted any unsafe recommendations or omissions and whether model reasoning aligned with guideline logic. Disagreements were adjudicated by discussion, and consensus ratings were used for analysis.

Outcome Measures

This study assessed the performance of a structured, guideline-anchored decision support workflow using validated outcome instruments rather than model-centric metrics. Outcome domains spanned diagnostic accuracy, treatment quality, usability, scoring patient burden metrics, red-flag detection, and readability. Two independent blinded pain specialists assigned gold-standard diagnoses and burden scores; inter-rater agreement and consensus resolution were recorded.

Diagnostic performance was assessed at the level of major triage category (secondary vs non-secondary) and at the level of specific subtypes. Sensitivity, specificity, positive and negative predictive values, and overall accuracy were calculated. Agreement between NotebookLM's diagnosis and the gold-standard diagnosis was quantified using Cohen's κ [17].

Treatment plan quality Treatment plan quality was judged against pre-defined guideline concordant templates. These templates were derived from the cited guidelines by the two headache specialists prior to model evaluation, and the full 5-point Likert anchors for acute and preventive therapy are provided in the Supplementary Material (Table S2). Metrics included the frequency distribution of 5-point acute and preventive scores and the frequency of contraindicated or unsafe suggestions. Preventive recommendations were pre-specified as not applicable for secondary-headache vignettes and were coded as missing for inferential analyses.

Usability After completing all 35 vignettes, each clinician completed the System Usability Scale (SUS) and the Healthcare Systems Usability Scale (HSUS). These instruments provide standardized assessments of perceived usability and served as primary usability outcomes [18,19]. Headache-specific disability (HIT-6 and MIDAS) To contextualize clinical relevance, patient burden was assessed using two validated instruments. The HIT-6 yields reliable scores across a wide range of headache burdens and demonstrates acceptable test–retest reliability [20]. The MIDAS provides internally consistent and stable assessments of days of missed or reduced productivity due to headache [21]. HIT-6 and MIDAS were scored only in clinically eligible primary-headache vignettes (n=13); these instruments were not applied to secondary headache vignettes, trigeminal autonomic cephalalgias or facial pain phenotypes where disease-specific burden scoring is not recommended.

Detection of red-flag features suggesting possible secondary pathology was evaluated against the SNNOOP10 checklist, which summarizes key warning signs [22]. Experts recorded whether red-flag features were present in each vignette and whether NotebookLM appropriately highlighted them and recommended relevant investigations.

Readability of decision support outputs Readability of explanatory text and recommendations was assessed using FRE and FKGL Level formulas, which are widely used tools for health communication [23].

Statistical Analysis

Diagnostic accuracy metrics were calculated with 95% confidence intervals. Cohen's κ was interpreted using conventional thresholds (e.g., 0.61–0.80 substantial agreement). Inter-rater agreement for gold-standard diagnosis and burden scoring was reported with Cohen's κ and 95% CIs where estimable; for perfect agreement in small samples, asymptotic κ CIs can be degenerate, so CIs were not reported in those instances. Differences in treatment quality scores between acute and preventive recommendations were compared using paired non-parametric tests (Wilcoxon signed-rank) applied to paired 5-point expert Likert ratings in applicable cases only (n=20). Preventive scores marked not applicable were excluded from inferential analyses by design. Correlations between diagnostic accuracy and treatment concordance or burden metrics were explored using Spearman's ρ. No a priori power calculation was performed due to the vignette-based design; estimates should be interpreted as exploratory. SUS and HSUS scores were summarized descriptively given the small number of raters (n=2). Analyses were performed using Python v3.11 with significance set at $P < .05$.

Results

Thirty-five fictional headache case vignettes were independently evaluated by two pain medicine specialists. Of the 35 vignettes, 17 represented primary headache disorders, 15 represented secondary headache disorders, and 3 represented neuropathic or facial pain syndromes (e.g., trigeminal neuralgia, painful

DOI: 10.5455/medscience.2026.03.055

trigeminal neuropathy). For the purposes of triage-level classification and preventive treatment analysis, the 17 primary headache and 3 neuropathic/facial pain cases were grouped as non-secondary (n=20), while the 15 secondary headache cases formed the secondary group (n=15). Vignette characteristics are listed in the Supplementary Material (Table S1).

NotebookLM correctly identified the major diagnostic category (secondary vs non-secondary) in 35/35 vignettes (100.0%, 95% CI 90.0–100.0%). Subtype-level top-1 diagnostic accuracy was also 35/35 (100.0%, 95% CI 90.0–100.0%). Agreement between NotebookLM and gold-standard subtype diagnosis was perfect (Cohen's κ=1.00) (Table 1).

Table 1. Diagnostic performance of NotebookLM vs. gold-standard diagnosis

Group	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95%CI)	NPV % (95%CI)
Secondary vs non-secondary (overall)	100.0 (78.2–100.0)	100.0 (83.2–100.0)	100.0 (78.2–100.0)	100.0 (83.2–100.0)
Non-secondary class	100.0 (83.2–100.0)	100.0 (78.2–100.0)	100.0 (83.2–100.0)	100.0 (78.2–100.0)
Migraine (all)	100.0 (69.2–100.0)	100.0 (86.3–100.0)	100.0 (69.2–100.0)	100.0 (86.3–100.0)
Tension-type headache	100.0 (29.2–100.0)	100.0 (89.1–100.0)	100.0 (29.2–100.0)	100.0 (89.1–100.0)

CI: confidence interval, PPV: positive predictive value, NPV: negative predictive value

Acute treatment quality scores were high: score-5 in 22/35 vignettes (62.9%) and score-4 in 13/35 (37.1%). No acute recommendation scored below 4. Preventive treatment recommendations were evaluated exclusively in the 20 non-secondary vignettes. Among these, preventive recommendations were score-5 in 15/20 (75.0%) and score-4 in 5/20 (25.0%).

Unsafe or contraindicated recommendations were not observed (0/35, 0.0%; 95% CI 0.0–10.0%). The correlation between diagnostic correctness and acute treatment quality was not estimable because diagnostic correctness was constant (100%) (Table 2).

Table 2. Treatment plan quality: 1–5 Likert score distribution

Likert category	Acute (n/N)	Acute (%)	Preventive (n/N)[a]	Preventive (%)
5 (fully concordant)	22/35	62.9	15/20	75.0
4 (mostly concordant)	13/35	37.1	5/20	25.0
3 (partially concordant)	0/35	0.0	0/20	0.0
2 (low concordance)	0/35	0.0	0/20	0.0
1 (discordant/unsafe)	0/35	0.0	0/20	0.0
Not applicable	—	—	15/35	42.9
Median likert (IQR)	5.00 (4.00–5.00)		5.00 (4.75–5.00)	

[a]: Preventive percentages are calculated among clinically applicable cases (N =20); secondary headache cases are not applicable for preventive planning in this framework

Red-flag detection varied according to the analytical denominator. Across all vignettes with a recorded red flag, sensitivity was 87.5% (21/24; 95% CI 67.6–97.3%). Specificity was 100.0% (11/11; 95% CI 71.5–100.0%), with no false-positive alerts in red-flag-negative cases. In the prespecified secondary-headache subset, red-flag sensitivity was 93.3% (14/15; 95% CI 68.1–99.8%) (Table 3).

Table 3. Red-flag detection performance (SNNOOP10)

Outcome	n/N	% (95% CI)
Sensitivity (all red-flag-positive vignettes)	21/24	87.5 (67.6–97.3)
Specificity (all red-flag-negative vignettes)	11/11	100.0 (71.5–100.0)
Sensitivity (secondary-headache subset)	14/15	93.3 (68.1–99.8)

DOI: 10.5455/medscience.2026.03.055

HIT-6 and MIDAS analyses were performed only in clinically eligible vignettes (n=13). HIT-6 Agreement Expert HIT-6 scores ranged from 44 to 68 (mean ± SD:57.13 ± 6.27), while NotebookLM HIT-6 scores ranged from 44 to 68 (57.27 ± 6.22). Categorical HIT-6 agreement was 12/13 (93.3%; 95% CI 68.1–99.8%), corresponding to substantial-to-almost-perfect agreement (κ=0.90) (Table 4).

Expert MIDAS totals ranged from 0 to 52 (18.33 ± 16.46), and NotebookLM MIDAS totals ranged from 0 to 52 (18.27 ± 16.46). MIDAS grade agreement was 13/13 (100.0%; 95% CI 78.2–100.0%), with perfect κ (κ=1.00). Because agreement was perfect in a small sample, a CI for κ was not estimated (Table 4).

Table 4. Patient burden assessment: expert vs. NotebookLM (eligible cases, n=13)

Measure	Expert Mean ± SD	NotebookLM Mean ± SD	Mean absolute difference	Cohen's κ (95% CI)
HIT-6 total score	57.13 ± 6.27	57.27 ± 6.22	0.13	—
HIT-6 category agreement	—	—	—	0.90 (CI not estimated[a])
MIDAS total score	18.33 ± 16.46	18.27 ± 16.46	0.07	—
MIDAS grade agreement	—	—	—	1.00 (CI not estimated[a])

[a]CI for κ not estimated because of near-perfect/perfect agreement with small sample size. HIT-6, headache impact test; MIDAS, migraine disability assessment scale

Usability was assessed by two raters after completion of all 35 vignettes (two observations per instrument). SUS was 85 and 70 (mean 77.5), above the commonly used 68-point benchmark. HSUS overall was 5 and 4 (mean 4.5/5), with similar values across subdomains.

Readability metrics were calculated for all 35 NotebookLM outputs. FRE was very low (mean ± SD: -19.55 ± 22.60; range -48.2 to 23.4), and FKGL remained very high (21.50 ± 3.26; range 15.9–26.34).

Discussion

In this study, NotebookLM achieved perfect agreement at both the major triage level (secondary vs non-secondary) and the exact subtype level (35/35; Cohen's κ=1.00). Because of the limited sample size, the findings should be interpreted as robust proof-of-concept evidence rather than definitive estimates of real-world error rates.

The diagnostic pattern is directionally consistent with prior headache CDSS literature. Dong et al. reported sensitivity above 80% and specificity above 95% for guideline-based primary headache diagnosis, and the recent Head.AI validation also showed high top-rank performance in an ICHD-3-structured setting [24,25]. In migraine-focused work, a multicenter computer-based diagnostic engine reported κ=0.83, sensitivity 90.1%, and specificity 95.8% [26]. In a prospective use scenario, AI assistance improved non-specialist headache diagnostic accuracy from 46.0% to 83.2%, with parallel agreement gains [27]. Against this background, the 100% subtype agreement observed here is encouraging, but these findings must be interpreted within the context of spectrum bias. The use of expert-constructed vignettes often results in "textbook" clinical presentations that lack the semiological "noise," diagnostic ambiguity, or fragmented histories common in real-world practice. Consequently, the observed 100% accuracy represents an upper performance bound that likely reflects the superior performance of RAG systems in medical AI and controlled vignette conditions. However, real-world consultations may involve incomplete histories and atypical presentations that would likely reduce diagnostic concordance. Therefore, the 100% accuracy reported here should be regarded as an upper performance bound under idealized conditions.

Recent RAG evaluations in other clinical domains provide convergent context for these findings. A systematic review and meta-analysis reported a pooled odds ratio of 1.35 (95% CI 1.19–1.53) favoring RAG-enhanced systems over baseline LLM approaches in biomedicine [28]. Additional domain studies have shown meaningful gains in factual alignment and decision performance, including ophthalmology and emergency message triage [29,30].

The mechanistic difference between standard LLM workflows and RAG likely explains part of this effect. Standard LLMs generate outputs primarily from parametric memory and next-token probabilities, which can yield fluent but unsupported statements when details are missing, ambiguous, or outdated. RAG introduces an explicit retrieval layer that injects query-relevant external evidence into generation, improving grounding and source traceability [8,9,28]. NotebookLM follows this retrieval-grounded logic by conditioning outputs on uploaded source documents. However, RAG does not eliminate the risk of hallucinations; failure can still arise through retrieval miss, retrieval noise, or incorrect synthesis of correctly retrieved content [5-7].

The treatment-related findings also require careful clinical interpretation. Acute and preventive recommendations clustered at Likert 4–5, and no contraindicated or unsafe recommendation

was identified in this dataset. This pattern is compatible with guideline-constrained prompting and retrieval grounding but may also reflect the structured nature of vignette inputs. Importantly, strong diagnostic agreement does not guarantee treatment safety in routine care, where comorbidities, drug interactions, pregnancy status, and longitudinal treatment context are often incompletely captured at first contact.

CDSS implementation literature in pain medicine and diagnostic workflows shows that desk-based performance may degrade at the bedside unless accompanied by iterative safety review, workflow integration, and explicit escalation logic [31,32]. Accordingly, supervised use with independent safety checks remains essential.

Red-flag performance also benefits from denominator-aware interpretation. Sensitivity was 93.3% within secondary headache cases and 87.5% when all red-flag-positive vignettes were pooled, while specificity remained 100% among red-flag-negative cases. This profile indicates reliable capture of high-risk secondary patterns, with residual false-negative risk when warning features are subtle or distributed across complex symptom narratives. Because SNNOOP10 is intended as a structured safety screen, criterion-by-criterion reporting should be enforced in both prompts and output templates, ideally with explicit urgency stratification (emergent/urgent/nonurgent) [22].

Agreement for burden scales was high in clinically eligible cases: MIDAS grade agreement was perfect and HIT-6 category agreement was near-perfect. This pattern is clinically plausible because MIDAS maps directly to countable disability days, whereas HIT-6 requires broader multi-domain severity judgments and is therefore slightly more sensitive to wording and contextual inference [20,21].

Usability was favorable (mean SUS 77.5; HSUS overall 4.5/5), with scores above the commonly used SUS acceptability threshold and consistently positive HSUS domain ratings [18,19]. This suggests that raters perceived the system as both usable and clinically aligned in safety/workflow terms. Nonetheless, interpretation remains cautious because only two raters participated and assessments were conducted in a controlled vignette setting; response-style effects cannot be excluded. Human-factors evidence indicates that desk-based acceptability does not automatically translate into reliable workflow performance in live emergency or outpatient environments without iterative implementation testing [31,32].

Readability metrics indicated specialist-level complexity and poor suitability for direct patient-facing communication. Generated text was highly technical, corresponding to specialist-level prose that may support clinician-to-clinician communication but is inappropriate for direct patient counseling [23]. The RAG system was grounded in highly sophisticated materials, so these might be expected. However, for practical deployment, a dual-output strategy may be preferable: one guideline-dense clinical output and one constrained plain-language output with explicit readability targets, followed by comprehension-focused validation in end users.

This was a vignette-based study and therefore cannot fully replicate real-world diagnostic uncertainty, missing data, multitasking pressure, or comorbidity complexity. The dataset was relatively small for subgroup inference, particularly for rare subtypes with n=1 or n=2. Case construction by a single expert may introduce framing effects in clinical presentation, even though outcome assessment was performed independently by two blinded specialists with consensus adjudication. This design reduces incorporation bias compared with same-author/same-rater workflows, but residual spectrum and framing bias remain possible. The reliance on only two raters for usability assessment (SUS and HSUS) represents another limitation that severely restricts the generalizability of these findings; future studies should incorporate a larger and more diverse panel of end-users across different clinical settings to establish robust usability estimates.

Several outcome domains were intentionally restricted by clinical applicability: HIT6/MIDAS were not calculated for all headache cases, and preventive treatment quality was not scored for secondary-pathology workflows. These restrictions are methodologically appropriate but reduce denominator comparability across endpoints. In addition, hallucination alone was not quantified as a prespecified standalone endpoint, however all output was carefully reviewed for suitability with guidelines. These implementation dimensions should be addressed alongside prospective external validation. At present, the most defensible role for this approach is supervised decision support that augments–rather than replaces–specialist judgment [24,25].

Conclusion

This guideline-constrained RAG workflow showed perfect diagnostic agreement and high treatment-quality scores in clinically applicable domains. Red-flag detection remained high in secondary headache cases, and burden estimation accuracy was strong where HIT-6/MIDAS were appropriate. Usability was favorable. Readability remained poor for patient-facing use. Overall, NotebookLM appears suitable as a supervised specialist-support tool and should be validated prospectively in real clinical workflows before broader implementation.

Ethical Approval

Because the study used fictional case vignettes and did not involve identifiable patient data, formal ethics approval and informed consent were not required. All procedures adhered to the principles of the Declaration of Helsinki.

Patient Consent

Not necessary for this manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Funding

The authors received no financial support for this study.

References

1. Overeem LH, Ulrich M, Fitzek MP, et al. Consistency between headache diagnoses and ICHD-3 criteria across different levels of care. J Headache Pain. 2025;26:6.
2. Adams AM, Buse DC, Leroux E, et al. Chronic migraine epidemiology and outcomes - international (CaMEO-I) study: methods and multi-country baseline findings for diagnosis rates and care. Cephalalgia. 2023;43:3331024231180611.
3. Headache Classification Committee of the International Headache Society (IHS). The international classification of headache disorders, 3rd edition. Cephalalgia. 2018;38:1-211.
4. Wang X, Zhang NX, He H, et al. Safety challenges of AI in medicine in the era of large language models. arXiv. 2024;arXiv:2409.18968.
5. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. 2025;14:e59823.
6. Asgari E, Montana-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of large language models for medical text summarisation. npj Digit Med. 2025;8:274.
7. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med. 2025;5:330.
8. Lee J, Cha H, Hwangbo Y, Cheon W. Enhancing large language model reliability: minimizing hallucinations with dual retrieval-augmented generation based on the latest diabetes guidelines. J Pers Med. 2024;14:1131.
9. Gargari OK, Habibi G. Enhancing medical AI with retrieval-augmented generation: a mini narrative review. Digit Health. 2025;11:20552076251337177.
10. Nanua S, Steward R, Neely B, et al. Retrieval-augmented generation for interpreting clinical laboratory regulations using large language models. J Pathol Inform. 2025;19:100520.
11. Bendtsen L, Zakrzewska JM, Abbott J, et al. European academy of neurology guideline on trigeminal neuralgia. Eur J Neurol. 2019;26:831-49.
12. Bendtsen L, Evers S, Linde M, et al. EFNS guideline on the treatment of tension-type headache - report of an EFNS task force. Eur J Neurol. 2010;17:1318-25.
13. May A, Evers S, Goadsby PJ, et al. European academy of neurology guidelines on the treatment of cluster headache. Eur J Neurol. 2023;30:2955-79.
14. Ornello R, Caponnetto V, Ahmed F, et al. Evidence-based guidelines for the pharmacological treatment of migraine. Cephalalgia. 2025;45:3331024241305381.
15. Puledda F, Sacco S, Diener HC, et al. International headache society global practice recommendations for preventive pharmacological treatment of migraine. Cephalalgia. 2024;44:3331024241269735.
16. Puledda F, Sacco S, Diener HC, et al. International headache society global practice recommendations for the acute pharmacological treatment of migraine. Cephalalgia. 2024;44:3331024241252666.
17. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22:276-82.
18. Brooke J. SUS: a "quick and dirty" usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, eds. Usability evaluation in industry. London: Taylor & Francis; 1996:189-94.
19. Ghorayeb A, Darbyshire JL, Wronikowska MW, Watkinson PJ. Development and validation of a new healthcare systems usability scale (HSUS) for clinical decision support systems: a mixed-methods approach. BMJ Open. 2023;13:e065323.
20. Houts CR, McGinley JS, Wirth RJ, et al. Reliability and validity of the 6-item headache impact test in chronic migraine from the PROMISE-2 study. Qual Life Res. 2021;30:931-43.
21. Zandifar A, Asgari F, Haghdoost F, et al. Reliability and validity of the migraine disability assessment scale among migraine and tension type headache in Iranian patients. Biomed Res Int. 2014;2014:978064.
22. Do TP, Remmers A, Schytz HW, et al. Red and orange flags for secondary headaches in clinical practice: SNNOOP10 list. Neurology. 2019;92:134-44.
23. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res. 2010;468:2572-80.
24. Dong Z, Yin Z, He M, et al. Validation of a guideline-based decision support system for the diagnosis of primary headache disorders based on ICHD-3 beta. J Headache Pain. 2014;15:40.
25. Clares de Andrade JBC, Costa TBS, Vasconcelos JL, et al. Validation of an AI-based platform for structured diagnosis of headache disorders using ICHD-3 criteria (Head.AI). BMC Neurol. 2025;25:489.
26. Cowan RP, Rapoport AM, Blythe J, et al. Diagnostic accuracy of an artificial intelligence online engine in migraine: a multi-center study. Headache. 2022;62:870-82.
27. Katsuki M, Narita N, Matsumori Y, et al. Developing an artificial intelligence-based headache diagnostic model and its utility for non-specialists' diagnostic accuracy. Cephalalgia. 2023;43:331024231156925.
28. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32:605-15.
29. Luo MJ, Chiu M, Li DQ, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol. 2024;142:798-805.
30. Liu S, Wright AP, McCoy AB, et al. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. J Am Med Inform Assoc. 2025;32:1032-39.
31. Carayon P, Hoonakker P, Hundt AS, et al. Application of human factors to improve usability of clinical decision support for diagnostic decision-making: a scenario-based simulation study. BMJ Qual Saf. 2020;29:329-40.
32. Trafton J, Martins S, Michel M, et al. Evaluation of the acceptability and usability of a decision support system to encourage safe and effective use of opioid therapy for chronic, noncancer pain by primary care providers. Pain Med. 2010;11:575-85.

DOI: 10.5455/medscience.2026.03.055

Supplementary Table 1. Case vignette characteristics

ID	Headache type / Subtype	Category	Key characteristics
1	**Episodic migraine**	**Primary**	Recurrent unilateral throbbing attacks (12–24 h) with nausea, photophobia/phonophobia, and anxiety comorbidity.
2	**Episodic migraine**	**Primary**	Exam stress and sleep deprivation triggered unilateral migraine attacks with nausea and photophobia.
3	**Episodic migraine**	**Primary**	Long severe migraine attacks (24–48 h) with vomiting and menstrual exacerbation.
4	**Episodic migraine**	**Primary**	Pregnancy-associated episodic migraine with marked functional limitation.
5	**Episodic migraine**	**Primary**	Stress- or sleep-change–associated unilateral throbbing migraine attacks.
6	**Migraine with aura**	**Primary**	Typical visual aura (scintillating scotoma) preceding unilateral migraine headache.
7	**Migraine with aura**	**Primary**	Hemianopic aura followed by severe unilateral migraine; menstrual association.
8	**Migraine with aura**	**Primary**	Transient positive visual aura followed by unilateral throbbing migraine pain.
9	**Migraine with aura**	**Primary**	Bilateral visual aura with severe migraine symptoms and nausea/vomiting.
10	**Migraine with aura**	**Primary**	Brief visual aura (~40 min) followed by unilateral throbbing migraine attacks.
11	**Chronic tension-type headache**	**Primary**	Near-daily bilateral pressing headache without migrainous features.
12	**Chronic tension-type headache**	**Primary**	Chronic bilateral pressure-type headache worsened by prolonged computer work.
13	**Chronic tension-type headache**	**Primary**	Daily bilateral pressing headache with neck/shoulder tension and stress sensitivity.
14	**Cluster headache**	**Primary**	Strictly unilateral orbital attacks (30–90 min) with ipsilateral cranial autonomic symptoms.
15	**Cluster headache**	**Primary**	Nocturnal unilateral orbital attacks (45–120 min) with cranial autonomic features.
16	**Meningitis**	**Secondary**	Acute severe diffuse headache with fever and meningeal signs.
17	**Subarachnoid hemorrhage**	**Secondary**	Very severe progressive diffuse headache with meningeal irritation.
18	**Idiopathic intracranial hypertension**	**Secondary**	Progressive daily headache with transient visual obscurations, diplopia, and papilledema.
19	**Cerebral venous thrombosis**	**Secondary**	Postpartum progressive headache with vomiting and focal neurological deficit suggestive of CVT.
20	**Giant cell arteritis (temporal arteritis)**	**Secondary**	New temporal headache with jaw claudication and visual symptoms.
21	**Meningitis**	**Secondary**	Subacute diffuse headache with low-grade fever and mild neck stiffness.
22	**Subarachnoid hemorrhage**	**Secondary**	Explosive “worst headache of life” with vomiting, neck stiffness, and reduced consciousness.
23	**Idiopathic intracranial hypertension**	**Secondary**	Obesity-associated progressive daily headache with papilledema and diplopia.
24	**Intracranial infection**	**Secondary**	Immunocompromised patient with progressive headache and focal neurological deficits.
25	**Giant cell arteritis (temporal arteritis)**	**Secondary**	Temporal headache with elevated inflammatory markers, jaw pain, and visual blurring.
26	**Meningitis**	**Secondary**	Acute febrile meningeal syndrome consistent with bacterial meningitis.
27	**Subarachnoid hemorrhage**	**Secondary**	Classic thunderclap headache with meningeal irritation and altered mental status.
28	**Idiopathic intracranial hypertension**	**Secondary**	Progressive daily pressure-type headache with papilledema and visual symptoms.
29	**Cerebral venous thrombosis**	**Secondary**	Progressive bilateral headache with vomiting and focal weakness compatible with CVT.
30	**Giant cell arteritis (temporal arteritis)**	**Secondary**	Subacute temporal headache with jaw claudication and visual disturbance.
31	**Trigeminal neuralgia**	**Neuropathic/ Facial pain**	Brief shock-like V2/V3 facial pain triggered by touch or speaking.
32	**SUNCT**	**Primary**	Very short unilateral periorbital attacks with prominent cranial autonomic features.
33	**Post-herpetic trigeminal neuralgia**	**Neuropathic/ Facial pain**	Post-zoster unilateral V1 neuropathic pain with allodynia and sensory change.
34	**Secondary trigeminal neuralgia**	**Neuropathic/ Facial pain**	Trigeminal neuralgia pattern with interictal pain and mild sensory deficit.
35	**Paroxysmal hemicrania**	**Primary**	Frequent short unilateral orbital autonomic attacks with episodic clustering.

DOI: 10.5455/medscience.2026.03.055

Supplementary Table 2. Five-point Likert rubric for treatment concordance (acute and preventive recommendations)

Domain	Likert	Concordance definition	Example rating anchor
Acute therapy	5	Fully concordant with guideline recommendations; subtype-appropriate option(s), contraindication screening, and correct urgency level are all present.	Migraine attack: recommends a guideline-supported triptan or NSAID, checks vascular contraindications, and gives appropriate use limits.
Acute therapy	4	Mostly concordant; correct core treatment is present, with only minor omission that does not create immediate safety risk.	Gives appropriate acute medication but does not explicitly mention antiemetic or timing details.
Acute therapy	3	Partially concordant; contains a mix of correct and incomplete elements, with clinically relevant gaps.	Suggests a reasonable drug class but omits key safety screening or gives vague dosing/monitoring language.
Acute therapy	2	Low concordance; major mismatch with guideline logic, likely to reduce effectiveness or delay appropriate care.	In suspected secondary headache, focuses on symptomatic treatment and fails to prioritize urgent diagnostic evaluation.
Acute therapy	1	Discordant/unsafe; includes contraindicated advice or explicitly guideline-discouraged management.	Recommends vasoconstrictive therapy despite clear contraindication, or advises discharge despite red-flag findings.
Preventive therapy	5	Fully concordant; prevention is correctly indicated and the recommended option is guideline-supported and patient-appropriate, with safety considerations.	High-frequency migraine with disability: recommends a guideline-supported preventive option and documents contraindication checks.
Preventive therapy	4	Mostly concordant; appropriate preventive strategy with minor incompleteness that does not invalidate treatment direction.	Selects a suitable preventive class but does not fully specify follow-up interval or titration detail.
Preventive therapy	3	Partially concordant; prevention is mentioned but indication, selection, or safety framing is incomplete or mixed.	Proposes prevention without clearly linking to attack frequency/disability threshold, or mixes evidence-based and non-guideline options.
Preventive therapy	2	Low concordance; preventive plan is poorly aligned with subtype or clinical context.	Suggests prevention with weak subtype fit and without meaningful contraindication review.
Preventive therapy	1	Discordant/unsafe; contraindicated preventive advice or replacement of necessary causal/urgent secondary-headache management.	Recommends preventive therapy as primary plan in a secondary-headache scenario requiring urgent etiologic work-up.

Scores are assigned on a 5-point Likert scale (5=best concordance, 1=unsafe/discordant) based on guideline alignment, completeness, and safety. Rubric anchors were defined a priori by the expert panel and applied18 uniformly across vignettes