Original article

Cephalalgia
An International Journal of Headache

International Headache Society

Cephalalgia
2026, Vol. 46(4) 1–13

DOI: 10.1177/03331024261441574
journals.sagepub.com/home/cep

Sage

Comparative diagnostic agreement of a supervised machine learning model and a general-purpose, zero-shot, non-domain-adapted large language model for classifying headache disorders using structured questionnaires

Masahito Katsuki[1,2,3,4], Kieran Moran[2,5,6], Siobhán O'Connor[2], Tomás Ward[3,7], Yutaro Fuse[8], Omid Kohandel Gargari[9], Marina Romozzi[10,11], Alicia Gonzalez-Martinez[12,13,14], Miguel Á Huerta[15,16,17], Woo-Seok Ha[18], Jackson TS Cheung[19], and Yasuhiko Matsumori[20]

Abstract

Background: Accurate diagnosis of headache disorders is essential in clinical practice. Supervised machine learning models trained on structured clinical data have shown good performance, whereas the diagnostic ability of large language models (LLMs) for headache disorders has not been evaluated. This study compared a validated machine learning classifier with a general-purpose, zero-shot, non-domain-adapted LLM using the same structured patient questionnaire data, focusing on their agreement with specialist-confirmed diagnoses as the ground truth. This study was designed to reflect current real-world use scenarios, in which clinicians may apply off-the-shelf LLMs for diagnostic purposes without few-shot prompting, domain-specific fine-tuning, or adaptation, rather than to assess the theoretical upper limits of LLM capabilities.

Methods: We analyzed 1818 patients from an independent hold-out test cohort who completed a 22-item structured headache questionnaire and received specialist-confirmed diagnoses. A previously developed machine learning model and a general-purpose, non-domain-adapted LLM (GPT-4.1 with zero-shot prompting) each generated five-class International Classification of Headache Disorders, 3rd edition (ICHD-3)-based predictions: migraine and/or medication-overuse headache (MOH), tension-type headache (TTH), trigeminal autonomic cephalalgias (TACs), other primary headache

[1]Physical Education and Health Center, Nagaoka University of Technology, Nagaoka, Japan
[2]School of Health and Human Performance, Dublin City University, Dublin, Ireland
[3]Insight Research Ireland Centre for Data Analytics, Dublin City University, Dublin, Ireland
[4]Department of Biostatistics, Graduate School of Medicine, Saitama Medical University, Saitama, Japan
[5]Insight Research Ireland Centre for Data Analytics, Maynooth University, Kildare, Ireland
[6]Department of Sport Science and Nutrition, Maynooth University, Kildare, Ireland
[7]School of Computing, Dublin City University, Dublin, Ireland
[8]Department of Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan
[9]Headache Department, Iranian Center of Neurological Research, Neuroscience Institute, Tehran University of Medical Sciences, Tehran, Iran
[10]Dipartimento Universitario di Neuroscienze, Università Cattolica del Sacro Cuore, Rome, Italy
[11]Neurologia, Dipartimento di Scienze dell'invecchiamento, Neurologiche, Ortopediche e della Testa-Collo, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy
[12]Neurology Department, La Princesa University Hospital, Madrid, Spain
[13]Translational Research Group in Multimodal Biomarkers in Neurological Diseases, IIS-Princesa, Madrid, Spain
[14]Autonomous University, Madrid, Spain
[15]Department of Pharmacology, University of Granada, Granada, Spain
[16]Biosanitary Research Institute ibs.GRANADA, Granada, Spain.
[17]Department of 17harmacology, University of Cambridge, Cambridge, UK
[18]Department of Neurology, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[19]UCL Faculty of Medical Sciences, London, UK
[20]Sendai Headache and Neurology Clinic, Sendai, Japan

Corresponding author:
Masahito Katsuki, MD PhD, Physical Education and Health Center, Nagaoka University of Technology 1603-1, Kamitomiokamachi, Nagaoka, Niigata 9402188, Japan.
Email: ktk1122nigt@gmail.com

disorders, and secondary headaches. Agreement with the specialist's diagnosis and diagnostic performance metrics were calculated. Class-wise sensitivity and specificity were compared using McNemar's test.

Results: The machine learning classifier showed significantly higher diagnostic agreement with the specialist than the LLM (Cohen's κ: 0.46 vs. 0.26; 95% confidence interval of the difference: 0.15–0.25). Although the LLM showed slightly higher macro-averaged sensitivity (balanced accuracy) than the machine learning model, the machine learning classifier showed higher macro-averaged precision, specificity, and F-value. Class-wise analysis showed that the machine learning model demonstrated greater sensitivity for migraine and/or MOH and secondary headaches, while the LLM showed higher sensitivity for TTH. Regarding specificity, the machine learning model outperformed the LLM in TTH, TACs, and other primary headache disorders, whereas the LLM showed higher specificity only for migraine and/or MOH.

Conclusions: A supervised machine learning model trained on real-world clinical data showed better agreement with a specialist-confirmed diagnosis than a general-purpose, zero-shot, non-domain-adapted LLM. These findings indicate that, in its current off-the-shelf configuration under this experimental setting, the diagnostic agreement between a general-purpose LLM and specialists can be limited for headache disorders.

Keywords

artificial intelligence (AI), ChatGPT, headache diagnosis, large language models (LLMs), machine learning (ML)

Received: 21 December 2025; accepted: 14 March 2026

Introduction

Primary headache disorders represent a prevalent public health issue, with a severe burden.[1,2] Accurate diagnosis of specific headache disorders is essential for appropriate management and prevention of chronicity.[3] Misdiagnosis can lead to ineffective treatments, unnecessary investigations, and prolonged patient suffering.[4,5] However, access to headache specialists remains limited in many regions, and primary care physicians often struggle to distinguish among the numerous headache subtypes defined in the International Classification of Headache Disorders, 3rd edition (ICHD-3).[6] These challenges have motivated efforts to develop artificial intelligence (AI)-based tools to support headache diagnosis.[7–9]

AI has increasingly been explored as a tool to support diagnosis, drug selection, and other complicated analyses.[10–14] AI refers to computational systems designed to perform tasks that typically require human cognitive abilities. Supervised machine learning is a subset of AI that learns patterns from structured data to make predictions or classifications.[15] Regarding the diagnosis of headache disorders, prior studies have shown that machine learning models trained on structured clinical data can achieve high diagnostic performance for ICHD-3-based classifications, in some cases approaching specialist-level performance.[7,8] These models were developed from directly learning patterns embedded in well-curated clinical datasets composed of consistent symptom structures and clearly defined diagnostic labels.

In contrast, general-purpose, non-domain-adapted large language models (LLMs) such as ChatGPT, Gemini, Claude, and related systems, have rapidly expanded their capabilities through training on vast amounts of documents on the Internet.[16] These LLMs can be useful for education[12] and for providing clinical information in headache practice.[17] However, LLMs show characteristic limitations: these models are typically general-purpose systems not optimized for specific diagnostic tasks unless explicitly fine-tuned or adapted to a given clinical domain. In addition, they may produce inconsistent outputs, rely more heavily on texts they learned than structured clinical features, and exhibit diagnostic biases, particularly when used as independent diagnostic tools.[16]

Despite growing interest in their clinical potential, general-purpose, non-domain-adapted LLMs have not been evaluated for diagnostic performance in headache disorders, and no prior study has directly compared their diagnostic performance with that of machine learning models in headache classification against a specialist-confirmed diagnosis. This study compared a general-purpose, zero-shot, non-domain-adapted LLM with a previously validated machine learning model[8] using the same structured patient questionnaire data, focusing on their agreement with specialist-confirmed diagnoses as the ground truth.[8] Importantly, this study was designed to reflect current real-world use scenarios, in which clinicians may apply off-the-shelf, general-purpose LLMs for diagnostic purposes without few-shot prompting, domain-specific fine-tuning, or adaptation, rather than to assess the theoretical upper limits of LLM capabilities. We hypothesized that a domain-specific machine learning model would outperform the general-purpose, zero-shot, non-domain-adapted LLM in classifying headache disorders into five categories according to ICHD-3 criteria.

Materials and methods

Ethical considerations

The Itoigawa General Hospital Ethics Committee approved this study (approval numbers: 2022-2 and 2022-10).

Because the investigation was retrospective, the committee waived the written informed consent requirement. Instead, an opt-out notice was posted on the hospital's website for individuals who preferred not to participate. All procedures adhered to the Declaration of Helsinki.

Overall procedure

This study aimed to compare the diagnostic performance metrics of a conventional supervised machine learning classifier and an LLM using the headache specialist's diagnoses as the ground truth. First, we used a previously developed five-class headache machine learning classification model[8] that predicts one of the following five categories based on a structured headache questionnaire: Migraine and/or medication-overuse headache (MOH) (Class 1), Tension-type headache (TTH) (Class 2), Trigeminal autonomic cephalalgias (TACs) (Class 3), Other primary headache disorders (Class 4), and Secondary headaches (Class 5). The general-purpose, zero-shot, non-domain-adapted LLM then evaluated the same test dataset used for this machine learning model to generate corresponding diagnostic predictions. Both the machine learning and the LLM predictions were generated for the entire test dataset as a hold-out method, and diagnostic performance metrics were computed, using the specialist's headache diagnosis as the ground truth.

Dataset and ground truth

We used a retrospective dataset consisting of consecutive patients who visited the Sendai Headache and Neurology Clinic between January 2020 and December 2022. During this period, 6058 patients completed a structured headache questionnaire on their first visit, and all patients were diagnosed by a board-certified headache specialist. The questionnaire comprised 22 structured items covering demographic information, headache characteristics, temporal patterns, associated symptoms, triggers, medication use, and family history (Table 1). Because the questionnaires were completed under the supervision of clinic staff familiar with headache care, no missing values were present.

For the original development of the machine learning model, the entire dataset was randomly divided into a training cohort (n = 4240) and an independent hold-out test cohort (n = 1818).[8] In the present study, the same hold-out test cohort was used for model comparison between the machine learning classifier and the LLM.

The ground-truth diagnosis for each patient was defined as the final clinical headache diagnosis determined by the headache specialist after comprehensive clinical evaluation, with reference to ICHD-3. This evaluation included, but was not limited to, a structured headache questionnaire, clinical interviews, neurological examinations, and additional radiological and laboratory investigations to exclude secondary headaches. The diagnosis was not determined solely by whether patients fulfilled ICHD-3 criteria based on questionnaire responses. For example, patients who did not strictly meet ICHD-3 criteria according to the questionnaire but were clinically diagnosed with migraine by the specialist were labeled as migraine. Conversely, patients who fulfilled migraine criteria based on the questionnaire but were ultimately diagnosed with a secondary headache were labeled as secondary headache in the ground truth.

Diagnoses were categorized into five classes according to ICHD-3[6]: (Class 1) Migraine and/or MOH; (Class 2) TTH; (Class 3) TACs; (Class 4) Other primary headache disorders; and (Class 5) Secondary headaches. When multiple diagnoses were present, the primary diagnosis was used for analysis.

Machine learning model

A previously developed supervised machine learning-based headache diagnosis model was used for comparison in this study.[8] The machine learning-based model was constructed using the training cohort of 4240 patients, with preprocessing, hyperparameter optimization, and internal 10-fold cross-validation performed using PyCaret (version 3.0.0). The model incorporated all 22 structured questionnaire variables and classified patients into the same five diagnostic categories defined in the present study. Among multiple candidate algorithms, the final model (gradient boosting classifier) was selected based on cross-validated performance metrics, and hyperparameters were further tuned to maximize diagnostic sensitivity. The remaining 1818 patients, unseen during model development, served as an independent hold-out test dataset, and the performance metrics for the test dataset were evaluated as a hold-out method.

In the current study, we used the predictions generated by this previously reported machine learning model as the benchmark against which the LLM's diagnostic performance metrics were compared, referring to the headache specialist's diagnosis as ground truth. No retraining or modification of the machine learning classifier was performed for this analysis.

To enhance model interpretability and to assess the relative contributions of demographic and symptom-based questionnaire variables to classification performance, we additionally computed Shapley Additive Explanations (SHAP) values for the machine learning classifier.[18] SHAP was used to quantify each input variable's contribution to the model's predictions. Global feature importance was summarized using mean absolute SHAP values across the test dataset, and class-specific feature effects were visualized using SHAP beeswarm plots.

Table 1. Headache questionnaire sheet.

Questions	Answers
01. Age	( ) years old (y.o.)
02. Biological sex	Male/Female
03. Height	( ) cm
04. Weight	( ) kg
05. Dominant hand	Right, Left, Other ( )
06. Regular alcoholic consumption	No, Sometimes, Everyday
07. Regular smoking habit	No, Previous, Current
08. Bedtime	a.m./p.m. (:)
09. Wake-up time	a.m./p.m. (:)
10. Headache onset age	( ) y.o., ( ) days/months/years ago
11. Headache frequency	( ) times per min/h/month/year, every day
12. Headache duration	Always, ( ) days, 1 day, Half a day, ( ) h, ( ) min, Moment
13. Site of headache	Unilateral (right/left), bilateral, center, different site, around the eye, front, back, side, top, craniocervical transitional, ear, chin, nose, teeth
14. Headache characteristics	Pulsating, constricting, stabbing, tingling, grasped, gouged out, racking, dull, others ( )
15. Headache severity	Numerical Rating Scale (/10) Needs rest, disturbing daily life without rest, not disturbing
16. Presence of aggravation or improvement by exercise	Aggravation, improvement, no change
17. Concomitant symptoms	Nausea, vomiting, photophobia, phonophobia, osmophobia, red eye, lacrimation, runny nose, dizziness, fatigue, stiff shoulders, numbness in the extremities, others ( )
18. Presence of aura or prodrome	Absent, scintillating scotoma, numbness in the extremities, increased appetite, edema, sleepy, frequent urination, nausea, vomiting, photophobia, phonophobia, osmophobia
19. Times when headaches are most likely to occur	Wake up, morning, afternoon, evening, sleeping
20. Triggers of headache	None, lack of sleep, too much sleep, tired, drinking, smoking, bathing, weather, light, loudness, smell, holiday, crowd, weather
21. Use of acute medication	Drug's name: ( ) Frequency: ( ) times per day/month/year Effectiveness: very effective, mildly effective, not effective, different at times
22. Family history of headache disorders	Yes/No (Mother, Father, Son, Daughter, Grandmother, Grandfather, Brother, Sister)

Note: Ask patients to check or fill out each item on the questionnaire.

LLM model and prompting

We evaluated diagnostic performance using an LLM accessed through the ChatGPT application programming interface (API) (OpenAI, San Francisco, CA, USA; GPT-4.1; accessed on 10 November 2025). To avoid any influence from previous interactions or user-specific context, the memory function was disabled, and the model was run in a stateless configuration throughout the study. The model received the same structured questionnaire data used for the machine learning classifier. Each patient's 22 questionnaire variables were converted into a standardized text format and incorporated into a fixed prompt. All comma-separated values-based variables, including binary and one-hot encoded items, were directly placed into predefined textual fields without further transformation.

The LLM was intentionally evaluated in a zero-shot, off-the-shelf configuration without domain-specific fine-tuning. This design choice was made to reflect a real-world use scenario in which clinicians may apply general-purpose LLMs without task-specific optimization, and to maintain a clear methodological distinction from supervised machine learning models trained on labeled clinical data.

The LLM was provided with no additional training, fine-tuning, or examples, and all predictions were generated in a zero-shot setting (full prompt is shown in Supplementary File 1). The prompt format was in Japanese and identical across all patients, and the output was parsed to extract the five-class diagnostic label of headache disorders. The temperature, which controls the randomness of the model's output during text generation, was set to 0 to minimize sampling randomness. Only the final predicted class label was used for performance evaluation.

LLM reproducibility assessment

To evaluate the reproducibility of LLM predictions, a random sample of 200 cases from the test dataset was selected, stratified to preserve the original proportion of the five diagnostic categories. These 200 cases were processed five times independently using the same prompts and settings. Agreement across the five outputs was quantified using simple percentage agreement and Fleiss' kappa (κ). This

assessment aimed to determine the degree of stochastic variability in LLM-based diagnoses, even under deterministic prompting conditions, and to confirm stability before comparing the ground truth with the LLM and the machine learning model. Reproducibility was assessed to evaluate output consistency across repeated runs and was not intended as a measure of diagnostic performance or validity.

Evaluation metrics

Diagnostic performance of both the LLM and the machine learning model were evaluated using balanced accuracy, sensitivity (recall), specificity, precision, and F-value. Balanced accuracy was defined as the mean of the sensitivity and specificity. For the multiclass setting, balanced accuracy corresponds to the macro-averaged sensitivity, calculated as the unweighted mean of class-wise true positive rates across all diagnostic categories. Sensitivity (recall) was defined as the proportion of true cases of a given class that were correctly identified as that class: true positives divided by (true positives + false negatives). Specificity was defined as the proportion of non-cases correctly recognized as not belonging to that class: true negatives divided by (true negatives + false positives). Precision represented the proportion of predicted cases that were true cases: true positives divided by (true positives + false positives). The F-value was calculated as the harmonic mean of sensitivity and precision, providing a single metric that balances false-negative and false-positive errors.

Class-wise sensitivity and specificity were calculated using a one-vs-rest approach. Because five diagnostic classes were evaluated, we computed macro-averaged metrics. Macro-averaging first calculates each metric for every class and then takes their unweighted mean, giving equal weight to all classes regardless of prevalence. Confusion matrices were generated for both models to further characterize misclassification patterns.

Statistical analysis

All statistical analyses were performed using Python (version 3.10) on Google Colaboratory, PyCaret 3.0.0, and IBM SPSS Statistics 31.0.0 (IBM Corp., Armonk, NY, USA). Variables with normal distribution are expressed as mean (standard deviation), and categorical variables as counts and percentages. No imputation was required because the dataset contained no missing values. All statistical tests were two-sided, and p-values <0.05 were considered statistically significant.

Diagnostic performance indicators included sensitivity and specificity. Concordance between each model and the ground truth was assessed using unweighted Cohen's κ, because the diagnostic categories were nominal. Interpretation of κ values followed Cohen's conventional thresholds: ≤ 0 indicated no agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement.[19]

To compare the agreement with the ground truth between the LLM and the machine learning model, we calculated Cohen's κ coefficients for each model (ground truth vs. machine learning, and ground truth vs. LLM). Then we computed their difference (κ_machine_learning − κ_LLM). The statistical significance of this difference was assessed using nonparametric bootstrapping with 10,000 resamples, yielding a 95% confidence interval (CI) for the difference in κ.

To compare the sensitivity and specificity of the machine learning model and the LLM, class-wise sensitivity and specificity were compared between the two models using McNemar's test under a one-vs-rest formulation. For each diagnostic class, we constructed a 2×2 contingency table based solely on discordant paired outcomes: cases in which the machine learning model correctly classified the patient while the LLM misclassified it, and cases in which the LLM was correct while the machine learning model was incorrect. McNemar's test evaluates whether these two discordant counts differ significantly, thereby providing a direct paired comparison of model performance for that class.

The McNemar effect size (φ) was defined as the difference between the two types of discordant pairs, divided by the square root of their sum.[20] A larger value of this effect size indicates a greater imbalance between these two types of discordant outcomes. A positive signed value indicates that the machine learning model produced more correct classifications than the LLM among the discordant cases, whereas a negative value indicates that the LLM outperformed the machine learning model.

Results

Dataset characteristics

Of the total, 6058 patients completed the structured headache questionnaire and were included in the dataset, 4240 formed the training cohort, and 1818 formed the independent hold-out test cohort. The mean (standard deviation) age was 34.7 (14.5) years, and 3906 patients (65.8%) were female. No missing values were present in any questionnaire items.

The distribution of ground truth diagnoses was not statistically different between the training and test sets. Overall, 4829 patients (79.7%) had migraine and/or MOH, 834 (13.8%) had TTH, 78 (1.3%) had TACs, 38 (0.6%) had other primary headache disorders, and 279 (4.6%) had secondary headaches. Baseline characteristics did not differ significantly between the training and test datasets. Detailed diagnosis information and additional baseline variables are provided in the original article on the model.[8]

Table 2. Diagnostic performance of machine learning model compared to the headache specialist.

Ground truth by the headache specialist	Prediction by machine learning model						Performance metrics				
	Class 1: Migraine and/or MOH	Class 2: TTH	Class 3: TACs	Class 4: Other primary headache disorders	Class 5: Secondary headaches	Total	Sensitivity (recall), %	Precision, %	Specificity, %	F-value, %
Class 1: Migraine and/or MOH	1396	49	6	1	11	1463	88.8	95.4	50.4	92.0
Class 2: TTH	119	112	1	4	9	245	60.5	45.7	91.9	52.1
Class 3: TACs	8	1	9	0	2	20	56.3	45	99.4	50.0
Class 4: Other primary headache disorders	2	3	0	0	2	7	0	0	99.6	–
Class 5: Secondary headaches	47	20	0	2	14	83	36.8	16.9	96.1	23.1
Total	1572	185	16	7	38	1818	48.5*	87.5*	40.6*	54.3*

Abbreviations: MOH, medication-overuse headache; TACs, trigeminal autonomic cephalalgias; TTH, tension-type headache. *, Macro-averaged metrics were calculated as the unweighted mean of class-wise values. Balanced accuracy corresponds to the macro-averaged sensitivity.

Performance of the machine learning model

As reported previously,[8] the gradient boosting classifier model achieved high performance on the independent test dataset. Macro-average sensitivity (balanced accuracy), specificity, precision, and F-values were 48.5%, 87.5%, 40.6%, and 54.3% (Table 2). Cohen's κ comparing the machine learning model predictions with the ground truth diagnosis was 0.46, indicating moderate agreement.

SHAP-based feature importance in the machine learning model

To interpret the decision-making process of the machine learning model, SHAP analysis was performed. Figure 1 shows the global feature importance based on mean absolute SHAP values. Age at consultation and pain intensity (visual analog scale) were identified as the most influential features overall. Timing of headache: not specific, site: around the eye, and attack frequency (4–14 days/month) also contributed substantially to model predictions. Detailed SHAP values are provided in Supplementary Table 1, and class-specific SHAP beeswarm plots with corresponding values are shown in Supplementary Figures 1–5 and Supplementary Table 2. These analyses enabled transparent assessment of the relative contributions of demographic and symptom-based variables to the model's predictions across headache subtypes.

Reproducibility of LLM predictions

To evaluate the intra-model consistency of the LLM, we selected a random sample of 200 cases from the test dataset using stratified sampling to preserve the original proportion of the five diagnostic categories. These 200 cases were then processed five times using the same zero-shot prompt. The LLM produced identical predictions across all five runs in 80.0% of cases. Fleiss' κ was 0.83, indicating more than substantial inter-run agreement.

Performance of the LLM

Regarding the LLM's performance, the macro-average sensitivity (balanced accuracy), specificity, precision, and F-values were 50.4%, 87.1%, 32.1%, and 34.6% (Table 3). Cohen's κ comparing the LLM predictions with the ground truth diagnosis was 0.26, indicating fair agreement.

Comparison of diagnostic performance between the LLM and the supervised machine learning model

The LLM (Table 3) showed slightly higher macro-averaged sensitivity (balanced accuracy) than the supervised machine learning model (Table 4). On the other hand, the machine

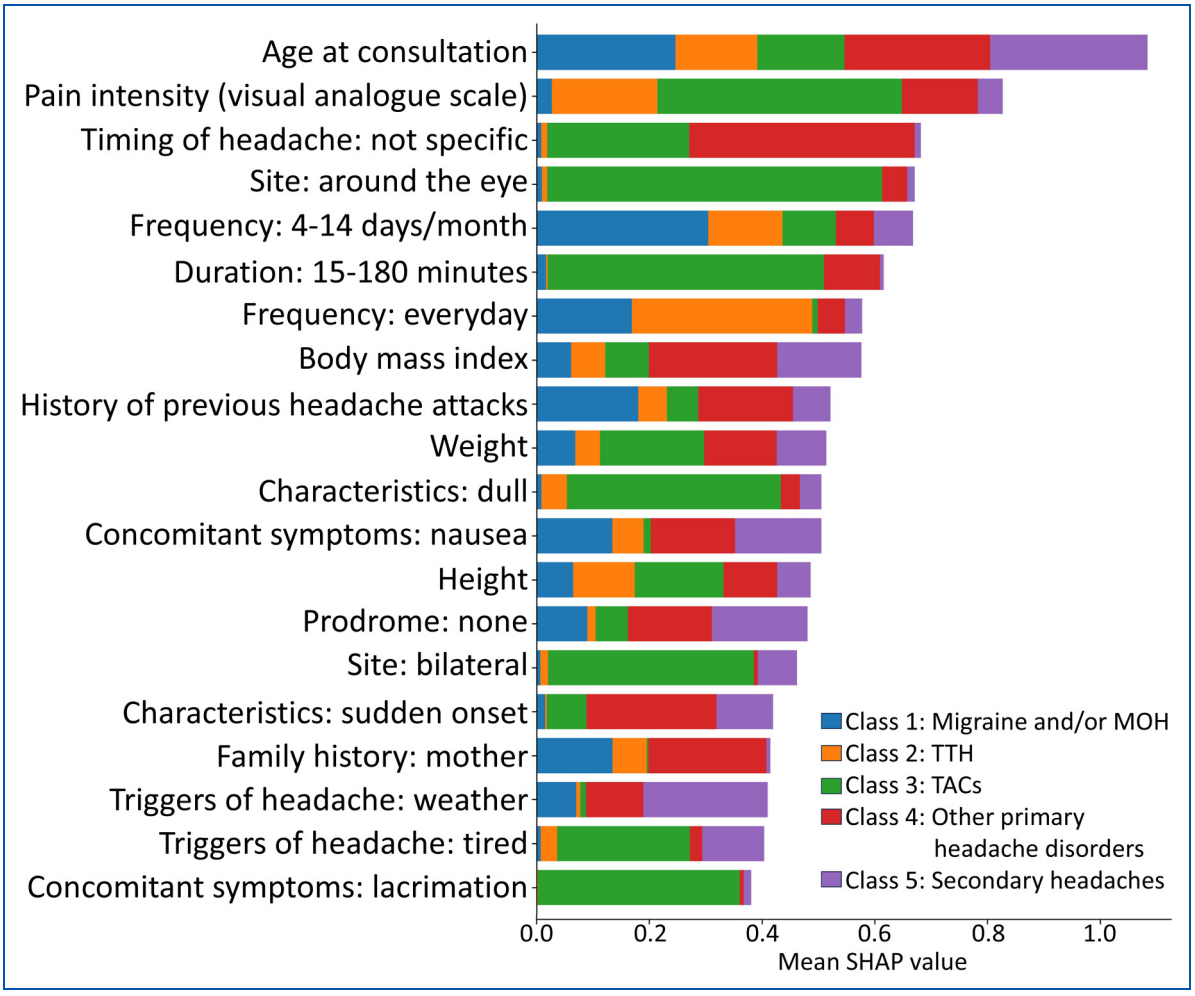

Figure 1. Global feature importance across headache classes. Global feature importance for the multiclass headache classification model based on mean absolute SHapley Additive exPlanations (SHAP) values. Horizontal stacked bars represent the contribution of each feature to the prediction of five headache classes: Migraine and/or medication-overuse headache (MOH) (Class 1), Tension-type headache (TTH) (Class 2), Trigeminal autonomic cephalalgias (TACs) (Class 3), Other primary headache disorders (Class 4), and Secondary headaches (Class 5). Features are ordered by overall importance, and colors indicate each feature's relative contribution to its respective class. Global feature importance is the sum of feature importance values across the model and does not account for direction (positive or negative contribution). Therefore, this metric reflects how important each feature was to the model's overall decision-making, but it does not indicate whether the feature increased or decreased the likelihood of a specific headache class. Details of the value are shown in Supplementary Table 1.

learning classifier showed higher macro-averaged precision, specificity, and F-value, although these metrics were not statistically compared.

To statistically compare agreement with the ground truth between the two models, we performed a paired bootstrap analysis (10,000 resamples) of Cohen's κ. The difference in κ (machine learning minus LLM) was significantly positive at 0.20 (95% CI: 0.15–0.25), and the bootstrap-based p-value was <0.001. These results indicate that the machine learning classifier demonstrated significantly higher diagnostic agreement with the ground truth than the LLM.

Class-wise sensitivity and specificity were further compared between the two models using McNemar's test under a one-vs-rest framework (Table 4). For sensitivity, the machine learning model significantly outperformed the LLM in Migraine and/or MOH (Class 1) (φ = 18.08, 95% CI [17.44–18.60]) and Secondary headaches (Class 5) (φ = 2.83 [1.41–4.24]). Conversely, the LLM showed significantly higher sensitivity than the machine learning classifier in TTH (Class 2) (φ = -4.04 [-5.70 to −2.18]).

For specificity, the machine learning model demonstrated significantly better performance in TTH (Class 2) (φ = 15.85 [14.95–16.64]), TACs (Class 3) (φ = 5.67 [5.00–6.00]), and Other primary headache disorders (Class 4) (φ = 6.86 [6.30–7.14]). In contrast, the LLM exhibited significantly higher specificity in Migraine and/or MOH (Class 1) (φ = -5.53 [-7.26 to −3.80]).

Table 3. Diagnostic performance of LLM compared to the headache specialist.

	Prediction by LLM						Performance metrics			
Ground truth by the headache specialist	Class 1: Migraine and/or MOH	Class 2: TTH	Class 3: TACs	Class 4: Other primary headache disorders	Class 5: Secondary headaches	Total	Sensitivity (recall), %	Precision, %	Specificity, %	F-value, %
Class 1: Migraine and/or MOH	1075	308	26	32	22	1463	73.5	89.3	63.7	80.6
Class 2: TTH	86	142	5	5	7	245	58.0	29.6	78.5	39.2
Class 3: TACs	7	0	12	1	0	20	60.0	24.5	97.9	34.8
Class 4: Other primary headache disorders	3	0	0	4	0	7	57.1	7.6	97.3	13.3
Class 5: Secondary headaches	33	30	6	11	3	83	3.6	9.4	98.3	5.2
Total	1204	480	49	53	32	1818	50.4*	87.1*	32.1*	34.6*

Abbreviations: LLM, large language model; MOH, medication-overuse headache; TACs, trigeminal autonomic cephalalgias; TTH, tension-type headache. *, Macro-averaged metrics were calculated as the unweighted mean of class-wise values. Balanced accuracy corresponds to the macro-averaged sensitivity.

Table 4. Effect sizes (φ) from McNemar analysis comparing machine learning model and LLM performance for each headache category.

	Sensitivity		Specificity	
Ground truth by the headache specialist	φ (95% CI)	p value	φ (95% CI)	p value
Class 1: Migraine and/or MOH	18.08 (17.44–18.60)	<0.001*	−5.53 (−7.26 to −3.80)	<0.001*
Class 2: TTH	−4.04 (−5.70 to −2.18)	<0.001*	15.85 (14.95–16.64)	<0.001*
Class 3: TACs	−1.00 (−2.33–0.33)	0.508	5.67 (5.00–6.00)	<0.001*
Class 4: Other primary headache disorders	−2.00 (−2.00 to −2.00)†	0.125	6.86 (6.30–7.14)	<0.001*
Class 5: Secondary headaches	2.83 (1.41–4.24)	0.008*	1.26 (0.14–3.22)	0.262

The signed McNemar effect size (φ) represents the directional imbalance between discordant classifications made by the two models. A positive value indicates that the machine learning model made more correct classifications than the LLM among discordant cases, whereas a negative value indicates the opposite. Values near zero reflect minimal difference between the models. CIs were obtained through bootstrap resampling of discordant pairs (10,000 iterations), and p-values were derived from the two-sided binomial form of McNemar's test. Abbreviations: CI, confidence interval; LLM, large language model; MOH, medication overuse headache; TACs, trigeminal autonomic cephalalgias; TTH, tension-type headache; *, $p < 0.05$; †, For classes with very small sample sizes, sensitivity analyses had limited statistical power. Because a signed effect size was used, confidence intervals entirely above or below zero indicate consistent directional differences rather than estimation bias.

Discussion

Overall summary of findings

In this study, we directly compared the diagnostic performance metrics of the previously validated supervised machine learning classifier with those of the general-purpose, non-domain-adapted, zero-shot LLM using identical structured headache questionnaire data and specialist-confirmed diagnoses as the ground truth. The machine learning model demonstrated higher overall diagnostic agreement with the ground truth than the LLM, as reflected by Cohen's κ and macro-averaged performance metrics. In contrast, the LLM exhibited distinct classification tendencies, including higher sensitivity for TTH, indicating that it was less likely to miss TTH cases, but lower sensitivity for migraine and secondary headaches, suggesting a greater tendency to overlook these conditions. It also showed higher specificity for migraine, meaning it was less likely to overdiagnose migraine, but lower specificity for TTH, TACs, and other primary headache disorders, reflecting a greater likelihood of assigning these diagnoses erroneously.

Although the LLM showed high intra-model reproducibility under zero-shot prompting, its diagnostic outputs were less consistent with the ground truth by the headache specialist than those of the supervised machine learning model. These findings highlight that using a general-purpose, zero-shot, non-domain-adapted LLM in this experimental configuration for headache diagnosis should be approached with caution and that it was inferior to a well-trained, domain-specific machine learning model. This comparison was not intended to represent the upper bound of LLM capability, but rather to reflect a realistic scenario in which clinicians may apply general-purpose LLMs without few-shot prompting or domain-specific optimization.

Interpretation of macro-averaged metrics and class imbalance

Macro-averaged metrics were appropriate in this study because they gave equal weight to all diagnostic classes, including clinically important minority categories such as secondary headaches, TACs, and other primary headache disorders. Unlike micro-averaged metrics, which are dominated by the highly prevalent migraine and/or MOH category, macro-averaging better addressed our research question of comparing the diagnostic behavior of the supervised machine learning model and the LLM across all five headache categories in this imbalanced dataset.

From this perspective, the machine learning model showed an important advantage, particularly in identifying secondary headaches. This may reflect the supervised machine learning model's ability to learn clinically relevant combinations of structured questionnaire features from labeled data, even within a heterogeneous category such as secondary headaches. It may also reflect the clinical setting of the dataset, which was derived from a headache specialty outpatient clinic largely consisting of walk-in patients; in this context, the machine learning model may have been better adapted to the types of secondary headaches encountered in specialist ambulatory practice, whereas a general-purpose LLM may have relied on broader medical representations that implicitly span a wider range of clinical scenarios, including emergency department and other acute care presentations.

Nevertheless, the machine learning model performed best in categories with relatively distinctive symptom patterns, such as migraine and TACs. In contrast, performance was lower in more heterogeneous categories, particularly other primary headache disorders and secondary headaches. In addition, the marked imbalance in class size in the present dataset, with very small sample numbers in some categories, may have limited the stability of performance estimates for these groups. Future studies with larger datasets enriched for rare headache subtypes will be required to more accurately assess model performance in these categories.

Previous research on LLMs in clinical practice for headache disorders

LLMs have only recently begun to be explored in clinical practice for headache disorders.[17,21–25] Although our study is the first to directly test the diagnostic performance metrics of an LLM using the structured headache questionnaire, several previous investigations have examined the utility of LLMs for other headache-related tasks, such as extracting headache frequency[21] and answering questions on headache disorders.[22,23]

Chiang et al.[21] developed a few-shot GPT-2-based generative natural language processing model fine-tuned on clinical notes and calculations, which accurately extracted headache frequency with 92% accuracy from electronic health records and outperformed traditional natural language processing approaches. It demonstrates the growing usefulness of LLMs for handling real-world clinical text in the medical records on headache disorders. Also, their zero-shot GPT-2 model demonstrated markedly lower performance compared with fine-tuned few-shot models, highlighting that LLMs generally require task- or domain-specific tuning to achieve optimal accuracy. This suggests that the absence of fine-tuning in our LLM evaluation may have contributed to the relatively lower diagnostic performance compared with the machine learning classifier.

Garcia et al.[22] showed that ChatGPT-4o can provide generally accurate and clinically coherent answers to migraine-related questions, with most responses rated as satisfactory but some limited by reference errors and insufficient technical depth. Raposio and Baldelli[23] found that generative AI tools with LLMs could provide rapid, generally accurate, and scientifically reliable descriptions of the outcomes and complications of migraine surgery. Similar studies have described the educational and informational uses of LLMs in headache medicine, and all have concluded that these tools can be helpful in providing accessible, clinically relevant knowledge.[17,24,25] Taken together, these prior studies indicate that LLMs show promise as supportive tools for headache information dissemination and education, although their role in headache diagnosis has mainly remained unexplored until now.

Comparison between machine learning models and LLMs in clinical practice

Comparative evaluations of LLMs and conventional machine learning approaches have now been reported across a wide range of clinical domains.[26] Some studies have shown that LLMs can perform as well as, or even better than, machine learning models, especially when LLMs are fine-tuned on domain-specific data, or used with few-shot prompting or multimodal information.[27,28] However, most reports still indicate that supervised machine learning models outperform general-purpose, zero-shot, non-domain-adapted

LLMs when the task involves structured prediction from tabular clinical data.[26,29]

Previous studies comparing LLMs with conventional machine learning in clinical diagnosis have shown that LLMs often have characteristic diagnostic biases. For example, general-purpose, zero-shot, non-domain-adapted LLMs can broaden the differential diagnosis, but may still fail to reliably identify the most likely diagnosis, even when the correct diagnosis appears somewhere in the list.[30] Also, a randomized trial using complex clinical vignettes showed that access to a general-purpose, zero-shot, non-domain-adapted LLM as a diagnostic support did not significantly improve physicians' diagnostic reasoning or final diagnosis accuracy compared with conventional resources alone, even though the LLM itself outperformed physicians.[16] Furthermore, LLMs may perform well for common cases but less reliably for more complex ones.[31]

These patterns are similar to our results. In our study, the LLM showed higher sensitivity for TTH, which has more general symptom patterns, but lower sensitivity for migraine and secondary headaches, which require more specific clinical features. Regarding specificity, the LLM demonstrated higher specificity for migraine but lower specificity for TTH, TACs, and other primary headache disorders, suggesting a tendency to misclassify these categories. In contrast, the machine learning model trained on structured real-world data showed higher sensitivity for migraine and secondary headaches and better specificity across several classes, indicating more stable, calibrated diagnostic behavior in headache disorder-specific conditions.

Task-specific machine learning models, such as a diagnostic model, can outperform general-purpose, zero-shot, non-domain-adapted LLMs when clear diagnostic criteria and structured variables are required.[29,32] Although LLMs can provide useful medical information and may support clinical education, their diagnostic performance is still influenced by the nature of their training text. Therefore, even with the rapid evolution of LLMs, continued collection of high-quality clinical data and refinement of domain-specific machine learning models will remain essential for developing reliable headache diagnostic tools.[32–34] To that end, it will also be necessary to develop systems that efficiently collect large volumes of high-quality clinical data through app-based consultations and treatment progress tracking.[35–37] At the same time, our analysis was conducted using a general-purpose, zero-shot, non-domain-adapted LLM. Few-shot prompting and domain-specific fine-tuning can substantially improve diagnostic performance.[27,28] Therefore, future work should also evaluate these enhanced LLM configurations to assess their potential utility for better headache diagnosis.

Limitations

This study has several limitations. First, the findings are based on data from a single headache clinic, and the models may have learned diagnostic tendencies specific to one specialist, potentially reflecting local diagnostic habits. External validation in primary care settings, non-headache-specialized hospitals, and international cohorts with differing epidemiology and clinical practices is required to assess generalizability. Also, the machine learning model used both the training and test data from the same clinic, which will have artificially improved its performance metrics, making the difference with the LLM larger. Ideally, model comparison should be performed using external test data from other institutions, both domestic and abroad, to provide a fairer assessment of relative performance.

Second, marked class imbalance was present in the dataset, with migraine and/or MOH accounting for approximately 80% of cases, while TACs and other primary headache disorders comprised only about 1–2% each. This imbalance limits the stability and interpretability of class-specific and macro-averaged performance estimates, particularly for rare headache categories, and likely contributes to the relatively low macro-averaged sensitivity and precision observed in these classes. In addition, several clinically distinct headache subtypes were grouped into broader diagnostic categories, potentially obscuring subtype-specific features and introducing classification bias for both the supervised machine learning model and the LLM. Diagnostic heterogeneity and subtype aggregation can substantially influence model performance in headache classification tasks.[38,39] Therefore, continued data collection and careful consideration of model design, including the handling of rare classes and subtype heterogeneity, will be necessary to improve the robustness and generalizability of future models.

Third, the analysis relied exclusively on structured questionnaire data without neurological findings, vital signs, laboratory results, or imaging information. Moreover, although the questionnaire was originally completed in Japanese, the LLM processed the information through a translated structured prompt. As most current LLMs are predominantly trained on English-language corpora, subtle linguistic or cultural biases in symptom interpretation cannot be excluded.

Fourth, our evaluation of the LLM was intentionally limited to a general-purpose, off-the-shelf configuration without few-shot prompting and domain-specific adaptation. This design was chosen to reflect current real-world use scenarios in which clinicians may apply LLMs as diagnostic aids without appropriate optimization. Consequently, the present findings should not be interpreted as representing the theoretical upper limits of LLM performance in headache diagnosis, but rather as an assessment of performance under this constrained and commonly encountered configuration.

Fifth, primary headache disorders are defined by symptom-based diagnostic criteria rather than objective or

definitive biological markers. Therefore, both the supervised machine learning model and the LLM rely on patient-reported symptoms, which imposes an inherent ceiling on achievable diagnostic performance. Moreover, as migraine diagnostic criteria remain an active area of discussion,[40] continuous data collection and model updating will be necessary.

Finally, the diagnostic performance of both approaches may vary depending on methodological choices. For the machine learning model, alternative feature preprocessing, class-imbalance handling, or algorithm selection could yield different results. Similarly, LLM performance depends on the underlying model architecture, training data, language, and prompting strategy. In particular, few-shot prompting or domain-specific fine-tuning may substantially improve LLM performance; however, these approaches were not evaluated in the present study, as they would shift the focus toward prompt engineering or model optimization and introduce additional concerns related to reproducibility and data leakage. Nevertheless, evaluating LLM performance with few-shot prompting and domain-specific fine-tuning will be necessary in future studies to more fully assess LLMs' intrinsic diagnostic capabilities. Besides, our evaluation used a single LLM version in the structured-text format in Japanese and may not reflect the performance of other models or languages. Furthermore, while high intra-model reproducibility indicates output stability under identical prompting conditions, it does not imply diagnostic correctness or clinical validity. We also disclose that this study is for research purposes and is not intended for clinical implementation.

Conclusions

In this study, a supervised machine learning model trained on real-world headache data showed higher diagnostic agreement with specialist-confirmed diagnoses than a general-purpose, zero-shot, non-domain-adapted LLM when classifying headache disorders based on structured questionnaire data in this experimental configuration. Future work should evaluate whether few-shot prompting, domain-specific fine-tuning, improved prompt engineering for LLMs, or multimodal data integration can further improve LLM performance for headache diagnosis.

Article highlights

- Using headache questionnaire data, we compared a supervised machine learning model with a general-purpose, zero-shot, non-domain-adapted large language model (LLM) for five-class ICHD-3-based diagnosis.
- Diagnostic agreement with the specialist-confirmed diagnosis was higher for the machine learning model than for the LLM (Cohen's κ 0.46 vs. 0.26).
- These findings highlight that the use of a general-purpose, zero-shot, non-domain-adapted LLM in this experimental configuration for headache diagnosis should be approached with caution and was inferior to a well-trained, domain-specific machine learning model.

Abbreviations

AI	artificial intelligence
API	application programming interface
CI	confidence interval
ICHD-3	International Classification of Headache Disorders, 3rd edition
LLM	large language model
MOH	medication-overuse headache
TACs	trigeminal autonomic cephalalgias
TTH	tension-type headache

Acknowledgements

We are thankful for the medical staff.

ORCID iDs

Masahito Katsuki https://orcid.org/0000-0002-0192-5430
Kieran Moran https://orcid.org/0000-0003-2015-8967
Siobhán O'Connor https://orcid.org/0000-0002-2001-0746
Omid Kohandel Gargari https://orcid.org/0000-0002-8182-0582
Marina Romozzi https://orcid.org/0000-0001-6016-3141
Alicia Gonzalez-Martinez https://orcid.org/0000-0002-1228-1503
Miguel Á Huerta https://orcid.org/0000-0003-2842-0085
Woo-Seok Ha https://orcid.org/0000-0003-1188-449X
Yasuhiko Matsumori https://orcid.org/0000-0001-5852-0146

Ethical considerations

The Itoigawa General Hospital Ethics Committee approved this study (approval numbers: 2022-2 and 2022-10). Instead, an opt-out notice was posted on the hospital's website for individuals who preferred not to participate. All procedures adhered to the Declaration of Helsinki.

Consent for publication

The authors agree to publish with Cephalalgia if the manuscript is accepted.

Consent to participate

Because the investigation was retrospective, the committee waived the written informed consent requirement.

Author contributions

Conceptualization, MK and MR; methodology, MK, YF, OKG, software, MK; validation, MK, YF, OKG; formal analysis, MK; investigation, MK; resources, YM, SOC, TW, KM; data curation, MK, YM; writing – original draft preparation, MK; writing – review and editing, MK, SOC, TW, MK, AGM; visualization, MK, MAH, JTSC; supervision, MK; project administration, TW, MK; funding acquisition, MK, SOC, TW, MK. All authors have read and agreed to the published version of the manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This publication has emanated from research jointly funded by Taighde Éireann – Research Ireland under Grant number 12/RC/2289_P2, the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101034252. This work was also supported by the Instituto de Salud Carlos III (JR23/00005 and PI24/01085) and co-funded by the European Union (FEDER/European Regional Development Fund-"A way to make Europe") through funding granted to A.G.M.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: MK has a patent related to headache diagnosis. YF has another patent application pending related to headache diagnosis. AGM has received speaker honoraria from TEVA, Lilly, and Altermedica. The other authors declare no conflicts of interest.

Data availability statement

The datasets and codes from this study are available from the corresponding author upon reasonable request.

Open practices

Not applicable.

Supplemental material

Supplemental material for this article is available online.

References

1. GBD 2021 Nervous System Disorders Collaborators. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet Neurol* 2024 Apr; 23: 344–381.
2. Matsumori Y, Ueda K, Komori M, et al. Burden of migraine in Japan: results of the observational survey of the epidemiology, treatment, and care of migraine (OVERCOME [Japan]) study. *Neurol Ther* 2022 Mar; 11: 205–222.
3. Pozo-Rosich P, Caronna E, Sacco S, et al. Early treatment in migraine – a call to shift prevention from attacks to disease progression: a position statement from the International Headache Society. *Cephalalgia* 2025 Oct; 45: 3331024251387721.
4. Tatsuno Y, Katsuki M, Kawata Y, et al. Understanding delays and diagnostic shifts in primary headaches: evidence from Japanese Health Insurance Claims. *Cureus* 2025 May 28; 17: e85005.
5. Katsuki M, Matsumori Y, Ichihara T, et al. Treatment patterns and characteristics of headache in patients in Japan: a retrospective cross-sectional and longitudinal analysis of health insurance claims data. *Cephalalgia* 2024 Jan; 44: 3331024231226177.
6. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders. *Cephalalgia* 2018; 38: 1–211.
7. Katsuki M, Shimazu T, Kikui S, et al. Developing an artificial intelligence-based headache diagnostic model and its utility for non-specialists' diagnostic accuracy. *Cephalalgia* 2023 May; 43: 3331024231156925.
8. Katsuki M, Matsumori Y, Kawamura S, et al. Developing an artificial intelligence-based diagnostic model of headaches from a dataset of clinic patients' records. *Headache* 2023 Sep; 63: 1097–1108.
9. Katsuki M, Narita N, Matsumori Y, et al. Preliminary development of a deep learning-based automated primary headache diagnosis model using Japanese natural language processing of medical questionnaire. *Surg Neurol Int* 2020 Dec 29; 11: 75.
10. Danelakis A, Stubberud A, Tronvik E, et al. The emerging clinical relevance of artificial intelligence, data science, and wearable devices in headache: a Narrative Review. *Life (Basel)* 2025 Jun 4; 15: 909.
11. Ihara K, Dumkrieger G, Zhang P, et al. Application of artificial intelligence in the headache field. *Curr Pain Headache Rep* 2024 Oct; 28: 1049–1057.
12. Romozzi M, García-Azorín D, Rubio-Beltran E, et al. Generative chatbots in headache education and research: A narrative review. *Cephalalgia* 2025 Sep; 45: 3331024251372117. DOI: 10.1177/03331024251372117
13. Petrušić I, Ha W-S, Labastida-Ramirez A, et al. Influence of next-generation artificial intelligence on headache research, diagnosis and treatment: the junior editorial board members' vision - part 1. *J Headache Pain* 2024 Sep 13; 25: 51.
14. Petrušić I, Chiang C-C, Garcia-Azorin D, et al. Influence of next-generation artificial intelligence on headache research, diagnosis and treatment: the junior editorial board members' vision - part 2. *J Headache Pain* 2025 Jan 2; 26: 2.
15. Solomonides AE, Koski E, Atabaki SM, et al. Defining AMIA's artificial intelligence principles. *J Am Med Inform Assoc* 2022 Mar 15; 29: 585–591.

16. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open* 2024 Oct 1; 7: e2440969.
17. Schütz P, Lob S, Chahed H, et al. ChatGPT as an information source for patients with migraines: a qualitative case study. *Healthcare (Basel)* 2024 Aug 10; 12, 1594.
18. Lundberg S and Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; 30: 4768–4777.
19. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; 20: 37–46.
20. Fagerland MW, Lydersen S and Laake P. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Med Res Methodol* 2013 Jul 13; 13: 91.
21. Chiang C-C, Luo M, Dumkrieger G, et al. A large language model-based generative natural language processing framework fine-tuned on clinical notes accurately extracts headache frequency from electronic health records. *Headache* 2024 Apr; 64: 400–409.
22. Garcia LB, Ferreira AJ, Hussein MA, et al. What does ChatGPT know about Migraine? A comparative-descriptive analysis. *Cephalalgia* 2025 Oct; 45: 3331024251387684.
23. Raposio E and Baldelli I. Reliability and accuracy of generative artificial intelligence tools in providing general information on migraine surgery. *Plast Reconstr Surg Glob Open* 2025 Oct 2; 13: e7176.
24. Chaulagain A, Aujla S, Priyadarsini A, et al. A cross-sectional comparison of patient information guides generated by chatgpt versus google gemini for Alzheimer's disease, parkinsonism, and migraine. *Cureus* 2025 May 20; 17: e84507.
25. Li L, Li P, Wang K, et al. Benchmarking state-of-the-art large language models for migraine patient education: performance comparison of responses to common queries. *J Med Internet Res* 2024 Jul 23; 26: e55927.
26. Yildiz Y, Nenadic G, Jani M, et al. Will large language models transform clinical prediction? *Diagn Progn Res* 2025 Nov 6; 9: 28.
27. Ben Shoham O and Rappoport N. CPLLM: clinical prediction with large language models. *PLOS Digit Health* 2024 Dec 6; 3: e0000680.
28. Mansoor I, Abdullah M, Rizwan MD, et al. Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. *Health Inf Sci Syst* 2025 Nov 26; 14: 6.
29. Brown KE, Yan C, Li Z, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *J Am Med Inform Assoc* 2025 May 1; 32: 811–822.
30. Ríos-Hoyo A, Shan NL, Li A, et al. Evaluation of large language models as a diagnostic aid for complex medical cases. *Front Med (Lausanne)* 2024 Jun 20; 11: 1380148.
31. Dinc MT, Bardak AE, Bahar F, et al. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. *JAMIA Open* 2025 Jun 12; 8: ooaf055.
32. Su H, Sun Y, Li R, et al. Large language models in medical diagnostics: scoping review with bibliometric analysis. *J Med Internet Res* 2025 Jun 9; 27: e72062.
33. Shool S, Adimi S, Saboori Amleshi R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak* 2025 Mar 7; 25: 17.
34. Zahn A, Strauss S and Zwanzig D. Towards community-based evaluation of AI in neurology: development of a headache diagnosis dataset for large language models. *Stud Health Technol Inform* 2025 Oct 2; 332: 237–241.
35. Katsuki M, Tatsumoto M, Kimoto K, et al. Investigating the effects of weather on headache occurrence using a smartphone application and artificial intelligence: a retrospective observational cross-sectional study. *Headache* 2023 May; 63: 585–600.
36. Stubberud A, Ingvaldsen SH, Brenner E, et al. Forecasting migraine with machine learning based on mobile phone diary and wearable data. *Cephalalgia* 2023 May; 43: 3331024231169244.
37. Goadsby PJ, Constantin L, Ebel-Bitoun C, et al. Multinational descriptive analysis of the real-world burden of headache using the Migraine Buddy application. *Eur J Neurol* 2021 Dec; 28: 4184–4193.
38. Petrušić I, Savić A, Mitrović K, et al. Machine learning classification meets migraine: recommendations for study evaluation. *J Headache Pain* 2024 Dec 5; 25: 15.
39. Petrušić I, Messina R, Pellesi L, et al. Application of machine learning in migraine classification: a call for study design standardization and global collaboration. *J Headache Pain* 2025 Oct 2; 26: 200.
40. 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 Jan 9; 26: 6.