Original Article

Cephalalgia Reports

Cephalalgia Reports
Volume 5: 1–6

DOI: 10.1177/25158163221115391
journals.sagepub.com/home/rep
SAGE

Interrater agreement in headache diagnoses

Maria Susanne Neumeier[1], Miranda Stattmann[1], Susanne Wegener[1], Andreas R Gantenbein[1,2], and Heiko Pohl[1]

Abstract

Background: Diagnosing headache disorders comprises the collection and interpretation of information. This study estimates agreement and bias in the latter.

Methods: Physicians and medical students diagnosed eight patients' headaches using the International Classification of Headache Disorders. We calculated Cohen's Kappa for all participants and subgroups (board-certified neurologists, physicians working in a neurology department). Moreover, we asked how sure they felt about their diagnoses. Finally, participants estimated the number of different headache diagnoses a patient receives when consulting many physicians for the same headache and indicated the highest acceptable number.

Results: The data of 63 participants entered the analysis, of whom 18 were neurologists (18/63, 28.6%), and 41 were currently working at a neurology clinic (41/63, 66.7%). Cohen's Kappa decreased (0.706, 0.566, and 0.408) with increasing levels of the classification hierarchy. Interrater agreement was highest among neurologists. Physicians not working in a neurology clinic tended to diagnose secondary headaches more often were less confident about their diagnoses.

Conclusions: Physicians with less experience in headache disorders struggle more to diagnose headaches than neurologists do; they suspect secondary headaches, disagree, and feel insecure more often. Thus, interpreting a headache history is prone to error and bias.

Keywords

diagnosis, headache classification, interrater reliability, noise, primary headache, secondary headache

Date received: 16 March 2022; Received revised June 30, 2022; accepted: 3 July 2022

Introduction

Assessing headaches is not easy. Many patients treated in an emergency department do not receive one specific diagnosis.[1] In addition, patients with cluster headache often receive several different diagnoses from different practitioners before the correct one.[2] Thus, the diagnoses vary and 'scatter' around the 'true diagnosis'. However, assuming there is only one correct headache diagnosis, all disagreement is unwanted and implies an error. Whether these variations are due to differing approaches to data collection, interpretation, or both is unknown.

The International Headache Classification provides an operational definition for all headache disorders but, in addition, requires other diagnoses not to explain the patient's symptoms any better.[3] Thus, due to this latter restriction, diagnoses are not made by just ticking off diagnostic criteria but also require non-standardised interpretation of the available information. Diagnosing headache disorders requires a two-step approach that comprises the collection and interpretation of information. This study focuses on the latter.

[1] Department of Neurology, University Hospital Zurich, Zurich, Switzerland
[2] Department of Neurology, Zurzach Care, Bad Zurzach, Switzerland

Corresponding author:
Heiko Pohl, Department of Neurology, University Hospital Zurich, Frauenklinikstrasse 26, 8091 Zurich, Switzerland.
Email: heiko.pohl@usz.ch

We aim to estimate agreement and bias in participants' interpretation of different headache histories. Notably, this study does not seek to assess whether the diagnoses are correct.

Methods

Study design

In preparation of the study, the authors of this article created eight case vignettes of different headache disorders (four primary headache disorders, three secondary headache disorders, one facial pain). Half of the cases were set up in the emergency room and half of them in the outpatient clinic. In the latter case, most patients had already received additional examinations.

Moreover, the case vignettes contained no pitfalls or 'red herrings' intended to set the participants on the wrong track. On the contrary, each case contained all information necessary to make one headache diagnosis according to the third edition of the International Classification of Headache Disorders (ICHD-3).[3] Please find the questionnaire provided as supplementary materials.

Having provided their informed consent, we first asked all participants to estimate the average number of diagnoses headache patients would receive if they were to consult a large number of physicians and to report the highest number of diagnoses they would deem acceptable for one single headache.

Next, the participants read and diagnosed the cases. Firstly, they classified the pain as primary or secondary headache or facial pain. Then, in two subsequent questions, they assigned the headache to a category and a subcategory of the ICHD-3. Along with each case, we provided a link to the International Headache Classification and encouraged the participants to look up any diagnosis as needed. Next, the participants indicated how sure they felt about the correctness of the assigned diagnosis.

Moreover, participants also indicated their age, sex, professional situation (medical student, assistant doctor, specialist, or other), specialisation, working experience (in years) and the average number of headache patients they usually treat (daily, more than once weekly, once weekly, at least once per month but not weekly, less than once per month, never).

Recruitment procedure and inclusion criteria

We included German-speaking physicians and medical students and recruited them through flyers, e-mails, and personal contacts. All participants enrolled explicitly for this research project from September 2021 to March 2022; we did not use their data in any other study.

We had not aimed for a specific number of participants, as no preliminary data had been available allowing power calculations. Hence, the available data determine the sample size.

Ethical clearance

As participation was anonymous, the study did not require ethical approval according to Swiss legislation and received a waiver from the local ethics committee (REQ-2021-00802).

Data analysis

First, we report the participants' beliefs regarding the average number of diagnoses a headache patient receives consulting many doctors and the highest acceptable number of different diagnoses for one headache as averages and standard deviations. Then, we used the Mann-Whitney *U* test to estimate the influence of binary variables and Spearman's Rho to assess the correlation of continuous and ordinal variables with these estimates.

Second, we analysed interrater agreement with Cohen's Kappa. As this method only calculates agreement for two individuals but not for a large collective, we calculated Cohen's Kappa for all possible pairs of participants and then took the arithmetic mean of the resulting values as Hallgren suggested.[4] We repeated this analysis for the following subgroups of the participants: physicians working in a neurology department, board-certified neurologists, and participants treating headache patients at least once per week. We interpreted the strength of agreement according to Landis and Koch. A Cohen's Kappa <0.00 implies 'poor agreement', 0.00 to 0.20 'slight agreement', 0.21 to 0.40 'fair agreement', 0.41 to 0.60 'moderate agreement', 0.61 to 0.80 'substantial agreement', and 0.81 to 1.00 'almost perfect agreement'.[5]

Furthermore, we analysed the participants' ratings regarding how sure they were about their diagnoses. Again, we used the Mann-Whitney *U* test to assess the influence of dichotomous variables on these ordinal ratings.

Next, we analysed whether subgroups of the participants were biased towards a part of the classification (i.e., primary or secondary headaches and facial pain) calculating odds ratios.

Lastly, we analysed the proportion of missing values for each case vignette.

With the significance level set at 0.05, we performed the statistical analysis using IBM SPSS version 26; missing values were not imputed.

Data availability

The data collected for this study are available from the corresponding author upon reasonable request.

Results

Sixty-nine individuals participated in the survey, of whom two did not meet the inclusion criteria because they had indicated being neither physicians nor medical students. In addition, we excluded four respondents, as the provided

Table 1. Interrater agreement measured with Cohen's kappa for all participants and different subgroups regarding the part of third edition of the International Classification of Headache Disorders (e.g., primary headache, secondary headache), the category (e.g., migraine, trigeminal autonomic headache), and the subcategory (e.g., chronic migraine, cluster headache).

Participants	Interrater agreement about the part	Interrater agreement about the category	Interrater agreement about the subcategory
Working in a neurology clinic	0.784	0.674	0.585
Board-certified neurologists	0.931	0.798	0.673
Attending to headache patients at least once weekly	0.658	0.512	0.412
All	**0.706**	**0.566**	**0.408**

The bold-faced values indicate the interrater agreement of all participants.

incomplete information precluded the validation of the inclusion criteria.

The data of the remaining 63 participants entered further analysis. They had an average age of 37 $\pm$ 10 years (range 24 to 62; 1 missing value), and 9 $\pm$ 10 years (range 0 to 37) of working experience. Thirty-three were female (33/63, 52.4%), 29 were male (29/63, 46.0%), and 1 identified as non-binary (1/63, 1.6%). Only 39 participants answered all questions (39/63, 61.9%); we used the available data for our analyses.

About half of the participants were assistant doctors, i.e. they had not completed their specialisation yet (31/63, 49.2%); 18 were board-certified neurologists (18/63, 28.6%), and 6 were medical students (6/63, 9.5%). Most of the participants were currently working at a neurology clinic (41/63, 66.7%).

More than half of the participants treated headache patients at least once weekly (36/63, 57.1%) and four of them even daily (4/63, 6.3%); only eight participants indicated never attending to headache patients (8/63, 12.7%).

On average, the participants estimated that patients received 5 $\pm$ 3 different diagnoses (range 1 to 15) if they consulted a vast number of physicians for the same headache. The duration of professional experience, the age of the respondents and the number of headache consultations did not influence this estimate ($p = 0.850$, $p = 0.581$, $p = 0.956$, respectively). However, participants working at a neurology clinic (4 $\pm$ 2 vs 6 $\pm$ 3 diagnoses, $p = 0.007$), and board-certified neurologists expected the number to be lower than other participants did (4 $\pm$ 2 vs 5 $\pm$ 3 diagnoses, $p = 0.017$).

The participants deemed 2 $\pm$ 1 diagnoses (range 1 to 5) for one headache disorder acceptable. Again, the duration of professional experience, the age of the respondents and the number of headache consultations did not influence this estimate ($p = 0.199$, $p = 0.172$, $p = 0.152$, respectively). Likewise, the estimates of participants working in a neurology clinic and board-certified neurologists did not differ statistically significantly from other participants ($p = 0.333$, and $p = 0.440$, respectively).

On average, the assigned diagnoses fell into 2 $\pm$ 1 different parts of the classification (e.g., 'part 1, the primary headaches'), 5 $\pm$ 1 categories (e.g., '1. Migraine', '3. Trigeminal Autonomic Headache'), and 8 $\pm$ 1 subcategories (e.g., '1.3 Chronic Migraine', '3.1 Cluster Headache').

Table 1 summarises Cohen's Kappa for all participants, including subgroups across all case vignettes.

When participants reported how sure they were about their diagnoses, the most commonly given answer was 'relatively sure' (150/329, 45.6%; if all participants had answered all questions, then 8*63 = 504 ratings would have been provided; thus, 175/504 = 34.7% answers were missing). Participants working in a neurology clinic ($p < 0.001$), board-certified neurologists ($p < 0.001$), and those attending to at least one headache patient per week ($p = 0.026$) were more confident about their diagnoses than those not in that group. Figure 1 depicts the frequencies of the participants' ratings.

Next, we studied whether specific subgroups were biased towards one part of the classification. Table 2 summarises the results.

Finally, we analysed the patterns of missing values. The proportion of participants who did not respond was 9.5% (6/63) in the first case, 28.6% (18/63) in the second, 33.3% (31/63) in the third and fourth, 36.5% (23/63) in the fifth, and 38.1% (24/63) in the remaining vignettes.

Discussion

This study analysed how the medical speciality and experience in treating headache patients affects interrater agreement about and bias towards specific headache diagnoses. Our main findings are that neurologists and resident physicians working in a neurology clinic generally agree well about their headache diagnoses and feel relatively sure about them. Other physicians, however, agreed less, felt less sure, and – perhaps consequently – tended to diagnose secondary headaches more often than other respondents.

The participants estimated that a patient who consults many different physicians for the same headache would receive five different diagnoses. This number seems to reflect the reality of patients with cluster headache remarkably well, as one study reported them to receive an average of 3.9 different diagnoses before the correct one.[2] Of course, the number of different diagnoses is likely

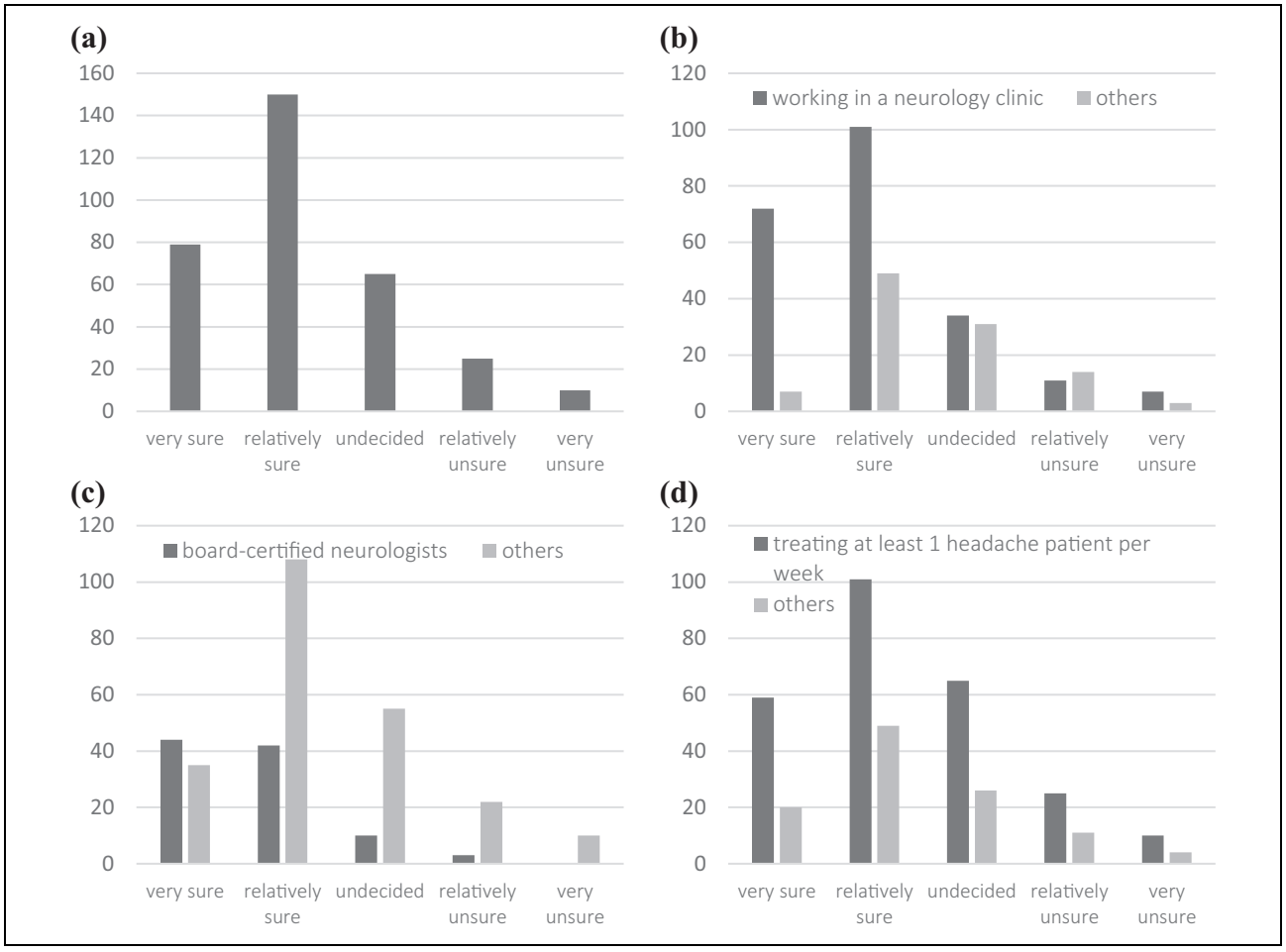

Figure 1. Bar diagram indicating how sure the participants felt about their diagnoses. The respondents provided one rating for each case; the diagrams lump together all their replies. The *y*-axis indicates the number of ratings per category. Of 504 possible ratings, 175 were missing. (a) depicts how sure all participants felt about their diagnoses. (b) compares participants working in a neurology clinic with those working in another speciality. (c) compares board certified neurologists with all other participants. (d) compares physicians attending to at least one headache patient per week with those who treat headache patients less frequently.

Table 2. The odds ratios indicate whether members of the subgroup indicated in the first column were more prone to diagnosing primary or secondary headache disorders or facial pain compared with other respondents not falling into these categories. An odds ratio >1 indicates that respondents in the subgroup indicated in the first column were more likely than others to diagnose a primary or secondary headache or facial pain.

Participant	Probability of diagnosing a primary headache	Probability of diagnosing a secondary headache	Probability of diagnosing facial pain
Working in a neurology clinic	OR = 1.30395% CI: 0.826–2.056p = 0.298	OR = 0.66595% CI: 0.420–1.051p = 0.100	OR = 1.47195% CI: 0.667–3.244p = 0.450
Board-certified neurologists	OR = 1.17895% CI: 0.741–1.872p = 0.555	OR = 0.78295% CI: 0.486–1.258p = 0.339	OR = 1.20695% CI: 0.578–2.515p = 0.700
Attending to headache patients at least once weekly	OR = 1.03895% CI: 0.666–1.619p = 0.910	OR = 0.88595% CI: 0.565–1.387p = 0.646	OR = 1.25195% CI: 0.593–2.640p = 0.712

smaller in patients with migraine and tension type headache.

The number of diagnoses that our participants estimated indicates that they consider diagnosing headaches challenging and prone to error. Interestingly, neurologists expected this number to be lower, suggesting that diagnosing headaches feels less difficult for them.

In our study, the participants assigned the case vignettes on average to five different categories of the ICHD-3. (The classification lists several categories, e.g. 'Migraine' and

'Trigeminal Autonomic Headache'.) Thus, the estimate that headache patients receive an average of five different diagnoses when consulting many physicians seems accurate. However, we had not asked the participants to take the history but only to interpret it, and the former may be equally or more prone to error. Therefore, the number of different diagnoses would likely have been even higher if they had taken the history themselves. While this may be due to the high number of physicians diagnosing the same headache, it might also imply that this study reflects reality only to a certain degree (see limitations).

According to the participants, the highest acceptable number of different headache diagnoses for one type of headache was two. Since most headache diagnoses are mutually exclusive,[3] the highest desirable number of diagnoses would be, of course, one. The participants' estimates support the hypothesis of them perceiving ambiguity as inherent in headache disorders. However, perhaps, the participants also considered that some patients sometimes have two headache disorders at the same time, e.g. a primary headache and a medication overuse headache.

The overall interrater agreement was 'substantial' regarding the 'part' of the headache classification (i.e., 'part 1, primary headache', 'part 2, secondary headache', and 'part 3, facial pain') and 'moderate' regarding the category and subcategory according to the graduation of Landis and Koch.[5] This finding is particularly meaningful because correctly classifying headaches as primary and secondary is vital given the momentous potential consequences of mistaking. One must bear in mind that the classification does not help make that distinction; primary and secondary headaches are discerned implicitly by making a headache diagnosis.

Classifying the headaches into a category or a subcategory of the classification was more difficult. The overall interrater agreement dropped considerably; solely board-certified neurologists maintained relatively favourable Kappa values similar to a previous study.[6] To our surprise, respondents reporting treating patients at least once per week had a relatively low interrater agreement. We conclude that experience in treating headache patients cannot replace training for making a headache diagnosis.

Because of the sheer number of persons experiencing headache attacks each year[7] who will certainly not all seek a neurologist's advice, these numbers must give us pause (see Table 1). If chance substantially influences diagnoses made by non-neurologists, then the classification is less helpful in their hands. Since 'Red Flags' have their weaknesses, too,[8] making a correct headache diagnosis may rely more on chance than we ought to accept. Consequently, all physicians should acquire expertise in diagnosing and treating headache disorders.

The analysis of how sure the participants were about their diagnoses (see Figure 1) indicates that neurologists felt, generally, more confident than others. Thus, participants who are not experienced in or not educated about treating headaches are well aware of their shortcomings. Consequently, the high number of unnecessary imaging studies ordered for headache patients may be due to felt insecurity.[9,10] After all, additional examinations may do more to reduce physicians' anxiety than that of patients.[11]

We also assessed whether physicians working in a neurology clinic, board-certified neurologists, and physicians treating headaches at least once weekly were biased towards primary and secondary headaches or facial pain compared with participants not falling into these categories (see Table 2). Although none of our findings was statistically significant, all calculated odds ratios pointed in the same direction and may therefore be meaningful, nonetheless.

Physicians working in a neurology clinic, board-certified neurologists, or physicians treating headache patients were, generally, less likely to suspect a secondary headache disorder. Conversely, less experienced participants suspected secondary headaches more frequently. Perhaps they attempt to prevent missing a secondary headache, as it might potentially lead to negative consequences quickly because they feel less secure about their diagnoses (see above). Finally, the bias towards a secondary headache might also be the result of learning and internalisation of the prior probabilities of different working environments. For example, secondary headaches are relatively rare in tertiary headache centres,[12,13] but occur more frequently in emergency departments or general practitioners' practices.[14,15]

Limitations

The artificial situation we set up to gain these insights may somewhat limit the study's value. Perhaps many participating physicians did not take diagnosing the invented patients as seriously as they would have in actual patients. The analysis of missing values indicates that many participants quit the survey early and thus supports this suspicion. Thus, we may have underestimated the interrater agreement.

On the other hand, many participants may rarely have a complete headache history and the results of additional examinations at their disposal when making a diagnosis; in daily practice, patients may struggle to provide all necessary information, especially under time pressure. In addition, several case vignettes were probably less most complex than most patients that physicians are confronted with in daily work. Thus, this study might also overestimate interrater agreement. Moreover, neurologists might immediately have recognised which diagnoses we had in mind when creating the case vignettes and might have quickly chosen the apparent answer. Hence, we could have overestimated the agreement of specific subgroups.

Finally, of course our convenience sample is unlikely to be representative of all physicians.

Conclusion

Neurologists generally agree well on headache diagnoses. However, physicians less experienced in treating headaches frequently struggle; they disagree and feel insecure about their diagnoses more often. Thus, taking a headache history is not the only difficulty in making a headache diagnosis – interpreting the collected information is also prone to error.

Despite all its undisputed strengths, the international headache classification does not support classifying a headache disorder to infrequent users. However, as neurologists are not the only ones confronted with headache patients, an additional and user-friendly decision tree may be warranted to guide through the diagnostic process, reduce uncertainty, and increase agreement.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: MSN and HP were funded by the Werner Dessauer Stiftung. The funding source had no influence on the content of this article.

ORCID iDs

Maria Susanne Neumeier https://orcid.org/0000-0001-6410-5547
Heiko Pohl https://orcid.org/0000-0002-2778-6790

Supplemental material

Supplemental material for this article is available online.

References

1. Friedman BW, Hochberg ML, Esses D, et al. Applying the International Classification of Headache Disorders to the emergency department: an assessment of reproducibility and the frequency with which a unique diagnosis can be assigned to every acute headache presentation. *Ann Emerg Med* 2007; 49: 409–419.
2. Klapper JA, Klapper A and Voss T. The misdiagnosis of cluster headache: a nonclinic, population-based, Internet survey. *Headache* 2000; 40: 730–735.
3. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders, 3rd edition. *Cephalalgia* 2018; 38: 1–211.
4. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol* 2012; 8: 23–34.
5. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
6. Granella F, D'Alessandro R, Manzoni GC, et al. International Headache Society classification: interobserver reliability in the diagnosis of primary headaches. *Cephalalgia* 1994; 14: 16–20.
7. Stovner LJ, Nichols E, Steiner TJ, et al. Global, regional, and national burden of migraine and tension-type headache, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* 2018; 17: 954–976.
8. Pohl H. Red flags in headache care. *Headache* 2022; 62(4): 534–535. DOI: 10.1111/head.14273.
9. Callaghan BC, Kerber KA, Pace RJ, et al. Headaches and neuroimaging: high utilization and costs despite guidelines. *JAMA Intern Med* 2014; 174: 819–821.
10. Gilbert JW, Johnson KM, Larkin GL, et al. Atraumatic headache in US emergency departments: recent trends in CT/MRI utilisation and factors associated with severe intracranial pathology. *Emerg Med J* 2012; 29: 576–581.
11. Howard L, Wessely S, Leese M, et al. Are investigations anxiolytic or anxiogenic? A randomised controlled trial of neuroimaging to provide reassurance in chronic daily headache. *J Neurol Neurosurg Psychiatry* 2005; 76: 1558–1564.
12. Song TJ, Kim YJ, Kim BK, et al. Characteristics of elderly-onset (≥65 years) headache diagnosed using the International Classification of Headache Disorders, third edition beta version. *J Clin Neurol* 2016; 12: 419–425.
13. Jensen R, Zeeberg P, Dehlendorff C, et al. Predictors of outcome of the treatment programme in a multidisciplinary headache centre. *Cephalalgia* 2010; 30: 1214–1224.
14. Frese T, Druckrey H and Sandholzer H. Headache in general practice: frequency, management, and results of encounter. *Int Sch Res Notices* 2014; 2014: 169428.
15. Munoz-Ceron J, Marin-Careaga V, Pena L, et al. Headache at the emergency room: Etiologies, diagnostic usefulness of the ICHD 3 criteria, red and green flags. *PLoS One* 2019; 14: e0208728.