Received: 7 January 2024 | Revised: 15 March 2024 | Accepted: 3 April 2024

DOI: 10.1002/ejp.2274

ORIGINAL ARTICLE

Clinicians diagnosing virtual patients with the classification algorithm for chronic pain in the ICD-11 (CAL-CP) achieve better diagnoses and prefer the algorithm to standard tools: An experimental validation study

Ginea Hay[1] | **Beatrice Korwisi**[2] | **Norman Lahme-Hütig**[3] | **Winfried Rief**[1] | **Antonia Barke**[2]

[1]Department of Clinical Psychology and Psychotherapy, Marburg University, Marburg, Germany

[2]Clinical Psychology and Psychological Intervention, Institute of Psychology, University of Duisburg-Essen, Essen, Germany

[3]FH Münster University of Applied Sciences, Münster School of Business, Münster, Germany

Correspondence
Antonia Barke, Clinical Psychology and Psychological Intervention, Institute of Psychology, University of Duisburg-Essen, Essen, Germany.
Email: antonia.barke@uni-due.de

Abstract

Background: The ICD-11 classification of chronic pain comprises seven categories, each further subdivided. In total, it contains over 100 diagnoses each based on 5–7 criteria. To increase diagnostic reliability, the Classification Algorithm for Chronic Pain in the ICD-11 (CAL-CP) was developed. The current study aimed to evaluate the CAL-CP regarding the correctness of assigned diagnoses, utility and ease of use.

Methods: In an international online study, $n = 195$ clinicians each diagnosed 4 out of 8 fictitious patients. The clinicians interacted via chat with the virtual patients to collect information and view medical histories and examination findings. The patient cases differed in complexity: simple patients had one chronic pain diagnosis; complex cases had two. In a 2×2 repeated-measures design with the factors tool (algorithm/standard browser) and diagnostic complexity (simple/complex), clinicians used either the algorithm or the ICD-11 browser for their diagnoses. After each case, clinicians indicated the pain diagnoses and rated the diagnostic process. The correctness of the assigned diagnoses and the ratings of the algorithm's utility and ease of use were analysed.

Results: The use of the algorithm resulted in more correct diagnoses. This was true for chronic primary and secondary pain diagnoses. The clinicians preferred the algorithm over the ICD-11 browser, rating it easier to work with and more useful. Especially novice users benefited from the algorithm.

Conclusions: The use of the algorithm increases the correctness of the diagnoses for chronic pain and is well accepted by clinicians. The CAL-CP's use should be considered in routine care and research contexts.

Significance Statement: The ICD-11 has come into effect in January 2022. Clinicians and researchers will soon begin using the new classification of chronic pain. To facilitate clinicians training and diagnostic accuracy, a classification

algorithm was developed. The paper investigates whether clinicians using the algorithm—as opposed to the generic tools provided by the WHO—reach more correct diagnoses when they diagnose standardized patients and how they rate the comparative utility of the diagnostic instruments available.

1 | INTRODUCTION

In January 2022, the 11th revision of the International Classification of Diseases and Related Health Problems (ICD-11) came into effect (World Health Assembly, 2019). It includes a new classification of chronic pain, which was developed by a task force of the International Association for the Study of Pain (IASP) (Treede et al., 2015, 2019). The classification divides chronic pain into seven main categories: chronic primary pain, and six categories of chronic secondary pain. In chronic primary pain, chronic pain is considered a health condition in its own right (Nicholas et al., 2019), whereas chronic secondary pain is associated with other underlying conditions: chronic cancer-related pain (Bennett et al., 2019), chronic postsurgical or post traumatic pain (Schug et al., 2019), chronic secondary musculoskeletal pain (Perrot et al., 2019), chronic secondary visceral pain (Aziz et al., 2019), chronic neuropathic pain (Scholz et al., 2019), and chronic secondary headache or orofacial pain (Benoliel et al., 2019). Each of these main categories contains several subdiagnoses, penetrating to three or four diagnostic levels (World Health Organization, n.d.) to allow for increasing levels of diagnostic specificity. For secondary and tertiary care, diagnoses on levels 2–4 will usually be most appropriate. It has been suggested that the 7 main categories (level 1 diagnoses) may be most relevant to primary care contexts (Smith et al., 2019).

The classification of chronic pain contains about 100 different diagnoses with 5–7 diagnostic criteria each. In medicine, the use of algorithms has been shown to lead to more reliable diagnoses (Rinaldi et al., 2000), increase adherence to diagnostic criteria (Bollestad et al., 2015) and enhance diagnostic accuracy (Morgan et al., 2000). Therefore, the Classification Algorithm for Chronic Pain in the ICD-11 (CAL-CP) was developed (Korwisi et al., 2021). The CAL-CP is a linear decision tree guiding users through the diagnostic criteria of the ICD-11 chronic pain classification. In a pilot evaluation, clinicians rated the CAL-CP as useful (Korwisi et al., 2022). The present study aims to evaluate the CAL-CP with regard to diagnostic correctness and subjective utility. It will be examined whether using the CAL-CP leads to more correct and complete diagnoses than using the standard ICD-11 Browser supplied online by the World Health Organization (WHO). In addition, users will rate the CAL-CP's usefulness and ease-of-use.

The use of standardized clinical vignettes is a valid and comprehensive method of measuring the quality of health care (Peabody et al., 2000) and investigating clinicians' decision-making (Evans et al., 2015) as it allows controlling the patient variables (Keeley et al., 2016). However, text based-vignettes present all information at once and thereby do not permit to evaluate the process of information gathering. Since providing a systematic way of collecting information is seen as a profound advantage of a classification algorithm, we used computer-assisted vignettes where the clinician takes the patient's medical history by entering questions into a chat program. This allows real-time responses that better simulate patient–physician interaction and are more realistic than traditional pen-and-paper vignettes (Peabody et al., 2004).

2 | METHOD

2.1 | Ethics

The study was approved by the Ethics Committee of the Catholic University of Eichstätt-Ingolstadt (Approval No. 020-2020) and complied with the Declaration of Helsinki (World Medical Association, 2013). Before participating in the study, all participants gave their informed consent. The data were collected anonymously.

2.2 | Participants

Recruitment took place internationally through invitations to pain associations, universities, pain clinics and practising pain physicians, as well as via Prolific (www.prolific.com). The eligibility criteria for clinicians to participate were as follows: professional experience with chronic pain patients or advanced medical studies (i.e., at least in the 3rd year) and sufficient self-rated English language skills (i.e., at least 4 on a numerical rating scale (NRS) from 0 to 10). All participants were eligible for a certificate of participation. In addition, medical students received a reimbursement of 20€ for their participation. A total of 558 participants gave informed

consent, of which 74 did not provide enough information to assess eligibility and 10 did not meet the eligibility criteria. Further 279 participants terminated the study before being directed to the virtual patient platform, resulting in 195 participants interacting with the virtual patients. The final sample consisted of 195 participants from 30 countries (Figure S1 in the supplementary material shows a world map of the participants' countries). The average age was 32.0 years (±13.3 years). 63.1% of the participants were female (see Table 1 for demographic information and Figure 1 for a chart of the participant flow). A dropout analysis showed that participants who dropped out early were older [38.6 ± 15.1 vs. 32.0 ± 13.3; $t(443.7) = 4.97$, $p < 0.001$; $d = 0.46$] and had a lower level of English proficiency [8.1 ± 1.4 vs. 7.7 ± 1.7; $t(455.8) = -3.11$, $p = 0.002$; $d = -0.28$].

TABLE 1 Description of participants.

	Final sample
Number of participants	195
Sex *n* (%)	
Female	123 (63.1)
Male	68 (34.9)
Diverse	2 (1.0)
Prefer not to answer	2 (1.0)
Age	
Mean ± SD	32.0 ± 13.3
English proficiency	
Mean ± SD	8.1 ± 1.4
Professional experience *n* (%)	
Novice	125 (64.1)
Expert	70 (35.9)
Specialty	
Anesthesiology	9 (12.9)
General practice	10 (14.3)
Internal medicine	2 (2.9)
Neurology/Neurosurgery	9 (12.9)
Orthopaedics	2 (2.9)
Pain medicine	17 (24.3)
Psychiatry/Psychology	16 (22.9)
Rehabilitation	4 (5.7)
Other	10 (14.3)
Years of clinical experience (Experts)	
Median (IQR)	19.5 (16.0)
Years of experience with chronic pain (Experts)	
Median (IQR)	13.0 (15.0)

Note: English proficiency rated from 0 to 10. The novices are medical students who have not yet graduated; therefore, their professional experience is zero.

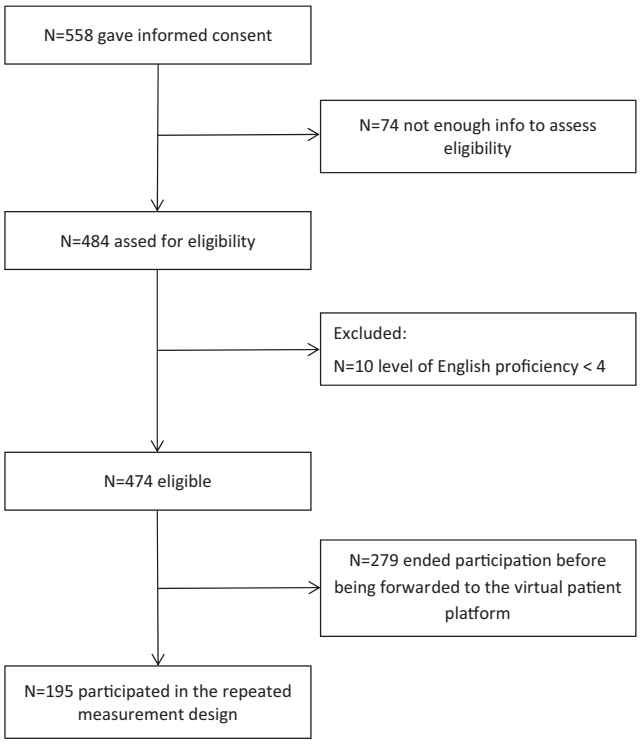

FIGURE 1 Documentation of absolute numbers of study participants.

2.3 | Procedure

The study consisted of two parts: a brief section where participant-related information was collected followed by the consultations with the virtual patients. See Figure 2 for an overview of the study procedure.

In the first part, participants completed a brief survey in which the inclusion criteria were assessed and additional information about demographic data, educational background and professional experience (e.g., area of specialty, years of clinical experience, years of experience with chronic pain) was collected. Participants rated their level of English proficiency on a NRS from 0 to 10.

The participants then received video instructions regarding the platform where they would hold the consultations with the virtual patients. In the video, the interaction with the virtual patients was explained and the functions of the consultation platform demonstrated. The instruction material was also available in written form as a download file for further reference. After answering a few test questions covering central facts about the interaction with the virtual patients, the participants were directed to the consultation platform.

In the second part, each participant carried out consultations with four of the eight virtual patients (two complex and two simple cases). In half of the cases, the diagnosis was made with the help of the algorithm; in

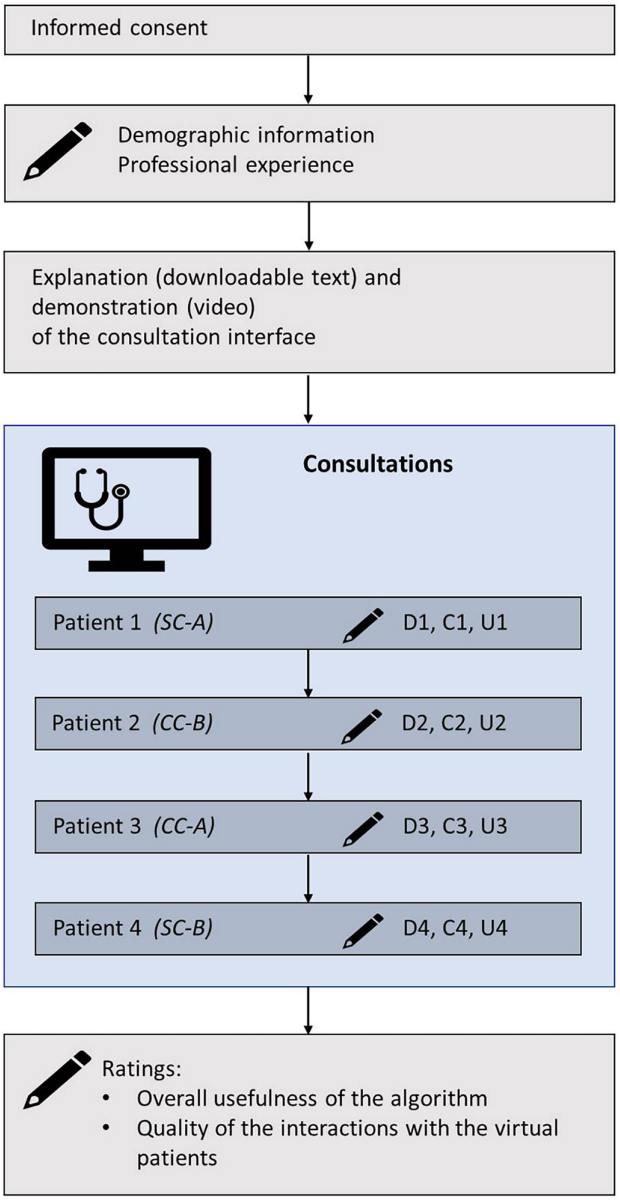

FIGURE 2 Study procedure. Four patients were randomly assigned so that each participant diagnosed two complex cases (CC) and two simple cases (SC). One complex and one simple case were diagnosed with the algorithm (A). One simple and one complex case were diagnosed with the standard WHO browser (B). (Italics show one possible sequence for demonstration purposes). For each consultation, the participant recorded the patient's diagnoses (D) and rated the subjective diagnostic certainty (C) and the utility of the diagnostic tool that was used in this instance (U). The (s)he proceeded to the next consultation.

the other half of the cases, the participants had the official ICD-11 browser at their disposal (see Figure 3). The assignment of the patients to the clinician, the order of the consultations and of the conditions (browser or algorithm) were randomized. The participants' task was to allocate the correct diagnoses to the patients. In order to do this, they had to collect the necessary information.

		Route	
		Official ICD-11 browser	Algorithm
Patient cases	Simple cases (1 diagnosis)	1	1
	Complex cases (2 comorbid diagnoses)	1	1

FIGURE 3 Experimental design of the consultations.

During each consultation, the participants could do any, or all, of the following:

1. Communicate with the patient by entering questions via the keyboard into the chat dialogue and receive the patient's answers in real time.
2. Access the stored medical file of the patient to complete their knowledge of the patient's health conditions.
3. Take notes on a virtual notepad.

The consultations lasted as long as the clinician felt necessary to reach a conclusion regarding the diagnoses. Each consultation ended with the clinician assigning one or more diagnoses to the patient. Then the clinician rated the subjective diagnostic certainty on an NRS ranging from 0 (not confident at all) to 10 (very confident). In addition, participants were asked to rate the ease of use and the utility of the diagnostic tool (algorithm or browser) they had just used in the consultation on two NRS ranging from 0 (very difficult/not useful at all) to 10 (very easy/very useful). After the last consultation, the participants rated the overall usefulness of the algorithm (e.g., perceived ease of use, personal utility, utility for novices, utility for experts and likelihood of future use) on an NRS ranging from 0 to 10. Finally, participants answered questions about the interaction with the virtual patients (e.g., how well did the virtual patients understand the clinician's questions and how well did the answers fit). Before the clinicians closed the window, they had the opportunity to give comments and suggestions for improvement.

2.4 | Material

2.4.1 | Fictitious patients

Eight fictitious patient cases, four simple and four complex ones, were created as a basis for the virtual patients and checked for plausibility by members of the IASP task force who were experts with regard to the respective pain type. In a simple case, the patient only had one pain diagnosis; in a complex case, he or she had two comorbid chronic pain diagnoses. Patient diagnoses covered all

categories included in the CAL-CP. The patient cases were balanced according to gender, pain type and complexity. For an overview of the patient cases, see Table 2.

To render the consultation with the virtual patients as realistic as possible, clinicians had the patients' medical files at their disposal during the consultations. The files included the following:

1. A brief medical history, current medication and surgical history if applicable.
2. A completed pain chart as provided by the CAL-CP (Korwisi et al., 2021) that contained information about the time of onset of the chronic pain, its temporal pattern, pain severity ratings (Treede et al., 2019) and a pain manikin showing all pain locations.
3. Additional documents.

Documents (1) and (2) were available for each participant from the outset. If necessary, further results, such as findings from MRI examinations or laboratory tests, became available once the clinician inquired about such results. Whether or not additional documents were provided depended upon the necessity for the individual case. In lieu of clinical examinations, the clinicians were instructed to ask the patient what the result would be if they conducted the clinical exam.

2.4.2 | The diagnostic tools

Algorithm

The CAL-CP is a classification algorithm for the ICD-11 chronic pain classification in which each diagnostic criterion is displayed in a decision box. The CAL-CP has been fully described elsewhere (Korwisi et al., 2021); the supplementary material of the original article also contains the full algorithm as supplementary digital material, available at http://links.lww.com/PAIN/B277. Following a linear decision tree, the user has to ascertain for each criterion whether it is fulfilled and follow the corresponding "yes" or "no" arrows to arrive at the appropriate diagnosis. In the present study, in the consultations in which the algorithm should be used, a digital version of the CAL-CP was available to participants as part of the consultation platform.

The ICD-11 browser

As a control condition, in the other half of the cases the clinicians used the ICD-11 browser and its coding tool as it is provided as the standard diagnostic tool by the WHO (https://icd.who.int/browse11/l-m/en). The ICD-11 browser was also implemented as a tab in the consultation platform.

2.5 | The consultation platform

The consultation platform (see Figure 4 for a screenshot) was custom programmed and consisted of a user interface with a stationary side panel (left) that displayed a picture of the current patient, showed the progress through the four consultations, and (in the condition in which the algorithm should be used) a section in which the current criterion of the algorithm was shown.

In the screen area to the right of the panel, the participant could choose one of four tabs: Chat, Documents, Notes, ICD-11. When the Chat Tab (shown in Figure 4) was chosen, the clinician could interact with the patient; in the Documents Tab, the clinician could look at the medical file and additional documents such as results of blood tests. In the Notes Tab, the clinician could write down notes to be at hand at the end of the consultation. In the ICD-11 Tab, the standard WHO-browser was available without leaving the consultation platform. The browser remained available for reference even in the algorithm condition. The reason for this was that the algorithm is an additional tool and everyone using it may also refer to the browser.

With regard to the chat-based interaction, each virtual patient was modelled as a set of answers to predefined questions. To enable a natural interaction with the consulting platform, natural language understanding techniques were employed. Specifically, an ensemble of Universal Sentence Encoder (Cer et al., 2018) and Sentence-Bidirectional Encoder Representations from Transformers (BERT) (Reimers & Gurevych, 2019) was used to calculate the semantic similarity of the entered question to each of the predefined questions based on the Manhattan distance. The predefined question with the highest similarity was used to look up the respective answer to present to the user.

The following information was collected for each consultation: condition (algorithm or browser), patient seen, number of questions asked, opened documents, textual diagnoses entered, ICD-11 numerical code entered, rating of the subjective diagnostic certainty, ratings of the utility and ease of use of the diagnostic tool used. In the algorithm condition, the number of clicks on the algorithm was also recorded to check whether the clinicians had followed the instruction and used it.

2.6 | Data analysis

For the evaluation of the algorithm, it was required to include only those consultations in the analysis in which the algorithm was used at least to a certain extent. Therefore, consultations in which the participants used fewer clicks in the algorithm (each click equals one step decision step in the algorithm) than one standard deviation below the

TABLE 2 Age, gender and diagnoses of the fictitious patients and number of valid consultations.

		Age, years	Gender	Diagnosis A	Diagnosis B	Number of valid consultations	Number of invalid consultations
Complex cases (2 comorbid diagnoses)	Patient 1	48	Male	Chronic visceral pain from persistent inflammation in the abdominal region	Chronic painful chemotherapy-induced polyneuropathy	63	26
	Patient 2	35	Female	Chronic pain after hysterectomy	Chronic central neuropathic pain associated with multiple sclerosis	53	43
	Patient 3	62	Female	Fibromyalgia syndrome	Chronic musculoskeletal pain associated with osteoarthritis	60	26
	Patient 4	56	Male	Chronic primary chest pain syndrome	Chronic primary low back pain	70	26
Simple cases (1 diagnosis)	Patient 5	26	Female	Chronic primary bladder pain syndrome	-	59	26
	Patient 6	47	Female	Chronic pain after burns injury	-	74	24
	Patient 7	45	Male	Chronic painful radiculopathy	-	70	19
	Patient 8	20	Male	Chronic bone cancer pain	-	71	24
					Total	520	214

Note: A consultation was classified as invalid if the participant did not open the patient documents, the number of questions asked was below the mean – 1 SD or, in the case of the algorithm condition, the number of clicks was less than the mean – 1 SD.

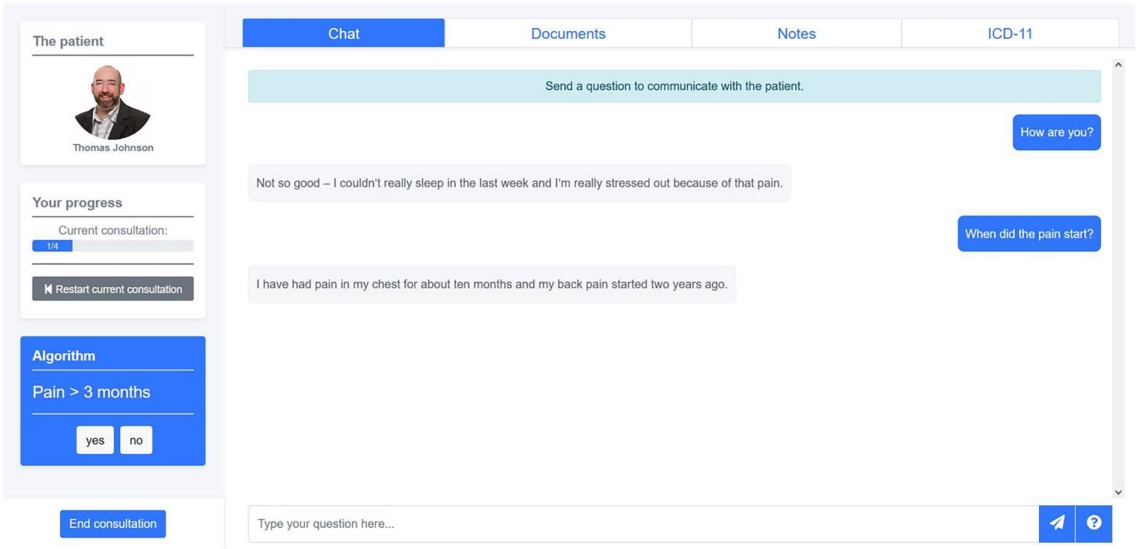

FIGURE 4 Screenshot of the consultation platform.

mean value of all participants in the condition were excluded. Similarly, we excluded consultations in both conditions (algorithm and no algorithm), in which the number of questions put to the patient was less than one standard deviation below the mean number of questions or in which the participants did not access the medical history and pain chart documents at all.

Two independent raters assessed the correctness and completeness of the diagnoses entered. They coded the diagnoses as "correct," "incorrect" or "not assigned." Cases were considered "correct" if the correct diagnosis was assigned on level 1. To count as correct, either the correct numerical code or the correct textual name of the diagnosis, or both had to be quoted. All diagnoses that had been assigned but were not correct were deemed "incorrect." If on any given level no diagnosis was attempted, it was counted as "not assigned." In complex cases there were three possible outcomes: "correct" (two correct diagnoses), "partially correct" (one correct diagnosis) and "incorrect" (no correct diagnosis). Complex cases were considered "complete" if two different pain diagnoses (not necessarily the correct ones) were assigned. The raters showed very good interrater reliability ($\kappa = 0.98$) in assigning the labels of "correct," "incorrect," "not assigned" and "complete" to the entries. While assigning these labels, they were blind to the condition (algorithm or browser). Consensus for the few cases of disagreement was formed through discussion.

To examine whether the use of the algorithm leads to better diagnostic results than the use of the browser, chi-square tests for the frequencies of correctness and completeness were calculated overall as well as separately for the complex and the simple cases and for primary and secondary pain diagnoses. Cramér's φ is reported as a measure of effect size. According to Cohen, φ between 0.10 and 0.30 is regarded as small, 0.30–0.50 as medium and ≥0.50 as a large effect (Cohen, 1992). *T*-tests for paired samples were calculated to examine whether the algorithm was estimated to be more useful than the browser and Cohen's *d* reported as measure of effect size. For Cohen's *d*, an effect of 0.20 is considered small, 0.50 medium and 0.80 large (Cohen, 1992).

Independent *t*-tests were calculated to compare the novices' and the experts' overall ratings of the algorithm.

Three repeated measures ANOVAs with the within-subject factors Tool (algorithm, browser) and Complexity (complex cases, simple cases) were calculated to compare the utility ratings (tool utility, ease of use, subjective diagnostic certainty) regarding tool and complexity.

To compare novices' (third year medical students prior to graduation) and experts' (graduated clinicians) ratings of the algorithm's and browser's utility, a mixed ANOVA was calculated with the within-subject factor Tool (algorithm, browser) and the between-subject factor Professional Experience (novice, expert).

The Eta-square is given as a measure of the effect size for ANOVAs. According to Cohen (1988), $\eta^2 = 0.01$ is a small, $\eta^2 = 0.06$ a medium and $\eta^2 = 0.14$ a large effect.

The quality of the interaction with the virtual patients was examined by analysing the corresponding ratings.

3 | RESULTS

3.1 | Participants' engagement with the consultation interface

In the algorithm condition, participants clicked on the algorithm an average of 22.7 ± 13.8 times per consultation.

All users referred to the standard supporting documents, on average 2.1 ± 1.1 times for simple cases, 2.4 ± 1.6 times for the complex cases. In cases in which additional documents were available on request, these were requested and consulted in 28.2% of the cases (132 cases of 457). An average of 14.1 ± 9.1 questions per consultation was asked in the browser condition and 13.9 ± 9.1 in the algorithm condition. For details see Table S1 in the supplemental material.

3.2 | Correctness and completeness of the diagnoses

Overall, 67.9% of diagnoses were assigned correctly, 32.1% incorrectly. For the percentages according to pain type (chronic primary or chronic secondary) and clinician status (novice or expert) see Table 3.

In the algorithm condition more correct diagnoses at level 1 were assigned than in the browser condition ($\chi^2(1) = 9.12$, $p = 0.001$, $\varphi = 0.13$). Differentiating into simple and complex cases, this was also true for the simple cases ($\chi^2(1) = 8.94$, $p = 0.001$, $\varphi = 0.18$); no significant difference was found for the complex cases ($\chi^2(1) = 2.54$, $p = 0.055$, $\varphi = 0.10$). Figure 5 shows the percentages of correct diagnoses for the browser and the algorithm. The sub-analysis for complex cases showed a significant difference between incorrect and partially correct diagnoses in favour of the algorithm ($\chi^2(1) = 8.23$, $p = 0.004$, $\varphi = 0.27$).

The number of complete diagnoses in the complex cases was $n = 103$ (78.03%) for the browser and $n = 88$ (77.19%) for the algorithm. No significant difference was found ($\chi^2(1) = 0.03$, $p = 0.44$, $\varphi = 0.01$).

Differentiating according to the nature of pain (chronic primary or chronic secondary pain), showed more correct diagnoses in the algorithm condition than in the browser condition for both, chronic primary pain ($\chi^2(1) = 4.26$, $p = 0.02$, $\varphi = 0.13$) and chronic secondary pain ($\chi^2(1) = 11.21$, $p < 0.001$, $\varphi = 0.15$). See Figure 6 for the percentage of correct diagnoses separately for primary and secondary pain.

In addition, chi-square tests were calculated separately for experts and novices. For the experts, there was no significant difference in correctness of diagnoses between the browser and the algorithm ($\chi^2(1) = 1.39$, $p = 0.12$, $\varphi = 0.09$). The novices reached more correct diagnoses when they used the algorithm ($\chi^2(1) = 8.08$, $p = 0.002$, $\varphi = 0.15$). The results are shown in Figure 7.

We also performed an exploratory analysis for the diagnoses at level 3. The total number of correct diagnoses was higher in the algorithm condition ($n = 115$; 47.33%) than in the browser condition ($n = 110$; 39.86%) ($\chi^2(1) = 2.94$, $p = 0.043$, $\varphi = 0.08$).

TABLE 3 Correctness of the assigned diagnoses according to pain type (chronic primary or chronic secondary) and clinician status (expert or novice).

	Correct diagnoses *n* (%)	Incorrect diagnoses *n* (%)
Overall	353 (67.9)	167 (32.1)
Pain type[a]		
Primary pain	195 (75.3)	64 (24.7)
Secondary pain	379 (74.8)	128 (25.2)
Experience level[b]		
Experts	110 (65.9)	57 (34.1)
Novices	243 (68.8)	110 (31.2)

[a]The values refer to the diagnosis level. As some patient cases have two diagnoses, there are higher values compared to overall.

[b]The values refer to patient case levels.

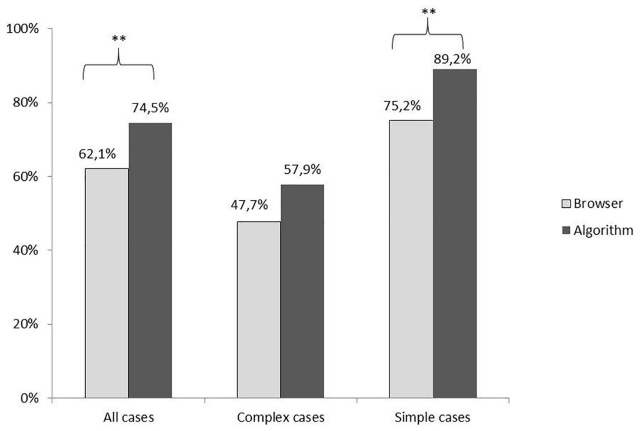

FIGURE 5 Percentage of correct diagnoses given for the browser and algorithm condition. **$p < 0.01$.

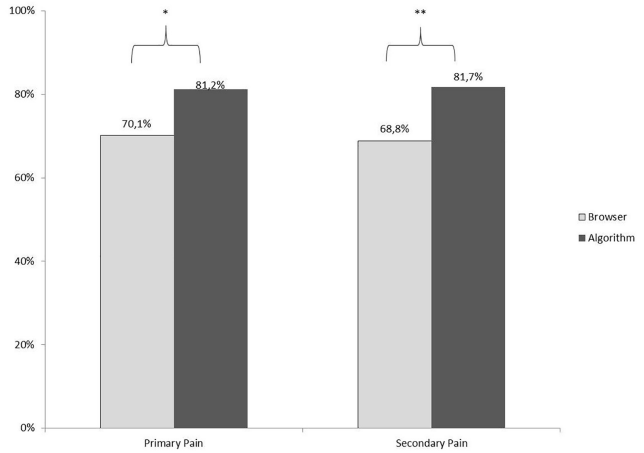

FIGURE 6 Percentage of correct diagnoses separately for primary and secondary pain diagnoses. *$p < 0.05$; **$p < 0.01$.

3.3 | Utility ratings

3.3.1 | Overall ratings of usefulness

After completing all consultations, the participants rated the overall ease of use of the algorithm on the 11-point rating scale at 7.3 ± 2.1, the overall personal utility as 6.7 ± 2.3; the use for novices as 7.3 ± 2.3, the use for experts as 5.6 ± 2.3 and the overall likelihood of future use as 7.0 ± 2.5.

Novices estimated the utility of the CAL-CP for experts lower than the experts themselves ($t(173) = 3.35$, $p = 0.001$, $d = 0.55$). In the other usefulness measures, novices and experts did not differ (Table 4).

3.3.2 | Ratings of usefulness per consultation

The participants also rated the tool's utility, the ease of use and the subjective diagnostic certainty for each patient they had diagnosed directly after each consultation.

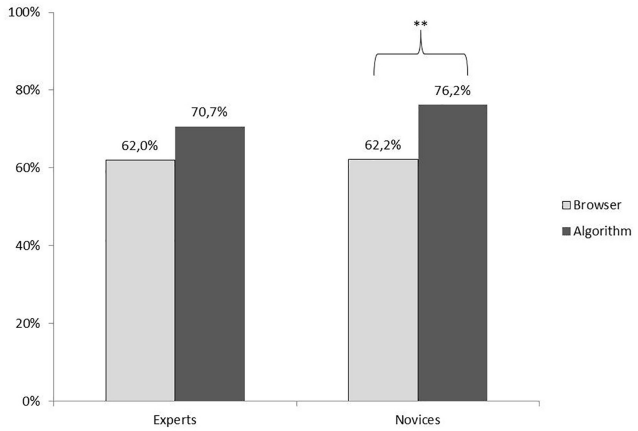

FIGURE 7 Percentage of correct diagnoses separately for experts and novices. **$p < 0.01$.

For tool utility, the ANOVA with repeated measures showed a significant interaction of the factors Tool and Complexity (F (1, 65) = 6.51, $p = 0.013$, $\eta^2 = 0.091$). There was no significant main effect for Complexity (F (1, 65) = 3.97, $p = 0.05$, $\eta^2 = 0.058$).

For ease of use, the ANOVA with repeated measures showed no significant effects.

The ANOVA with repeated measures for subjective diagnostic certainty also showed no significant effects. For details see Table 5. However, as Figure 8 shows, the algorithm tended to be preferred compared to the browser.

The mixed ANOVAs with the factors tool and professional experience revealed a main effect for Tool, favouring the algorithm [utility ($F(1, 142) = 4.70$, $p = 0.03$, $\eta^2 = 0.032$); ease of use (F (1, 142) = 9.30, $p = 0.003$, $\eta^2 = 0.061$)] over the browser. The other effects did not reach significance. For details, see Table 6 and Figure 9.

3.4 | Quality of the interaction with the virtual patients

On average, the participants rated the ease of interaction with 6.5 ± 2.2 on a scale from 0 to 10. The fit of the answers to the questions asked was 6.1 ± 2.0. The rating for ease of getting the relevant information in the interaction was 5.9 ± 2.4. The participants rated the naturalness of the interaction with 5.5 ± 2.4. The rating of the suitability of the virtual patients as training partners for learning about the new classification was 7.1 ± 2.2. Overall, 79 participants added free text comments, mostly about problems regarding question recognition or the request for voice or facial reactions.

4 | DISCUSSION

This study provides evidence that using the CAL-CP leads to more correct diagnoses than using the native ICD-11 web browser when assigning chronic pain diagnoses. In

TABLE 4 Overall ratings of usefulness of the CAL-CP separately for novices and experts, and results of significance tests.

	Experts		Novices					
	Mean	SD	Mean	SD	*t*	df	*p*	Cohen's *d*
Ease of use	7.6	2.3	7.2	1.9	1.24	81.94	0.18	0.22
Personal utility	7.1	2.2	6.6	2.2	1.32	173	0.19	0.22
Use for novices	7.3	2.5	7.2	2.1	0.07	173	0.95	0.01
Use for experts	6.5	2.3	5.2	2.2	3.35	173	0.001	0.55
Likelihood of future use	7.3	2.6	7.0	2.3	0.94	173	0.35	0.16

Note: Ease of use, Personal utility, Use for novices, Use for experts and Likelihood of future use rated from 0 to 10. CAL-CP, Classification Algorithm for Chronic Pain in the ICD-11; in case of unequal variances the Welch test was used and degrees of freedom adjusted accordingly.

TABLE 5 Results of three repeated-measures ANOVAs with the factors Tool (algorithm, browser) and Complexity (complex cases, simple cases) for the ratings of Tool utility, ease of use and subjective diagnostic certainty (assessed after each consultation).

	Algorithm		Browser		Complexity		Tool		Tool × complexity	
	M	SD	*M*	SD	$F(1, 65)$	η^2	$F(1, 65)$	η^2	$F(1, 65)$	η^2
Tool utility					0.04	0.00	3.97	0.06	6.51*	0.09
Complex cases	6.98	2.26	6.77	2.47						
Simple cases	7.44	2.46	6.21	2.66						
Ease of use					0.06	0.00	3.56	0.05	0.00	0.00
Complex cases	7.02	2.12	6.36	2.61						
Simple cases	7.08	2.24	6.43	2.54						
Subjective diagnostic certainty					0.63	0.01	1.00	0.02	0.4	0.01
Complex cases	6.60	2.02	6.55	2.25						
Simple cases	6.95	1.87	6.60	2.09						

Note: Tool utility, ease of use and subjective diagnostic certainty were rated from 0 to 10.

*$p < 0.05$.

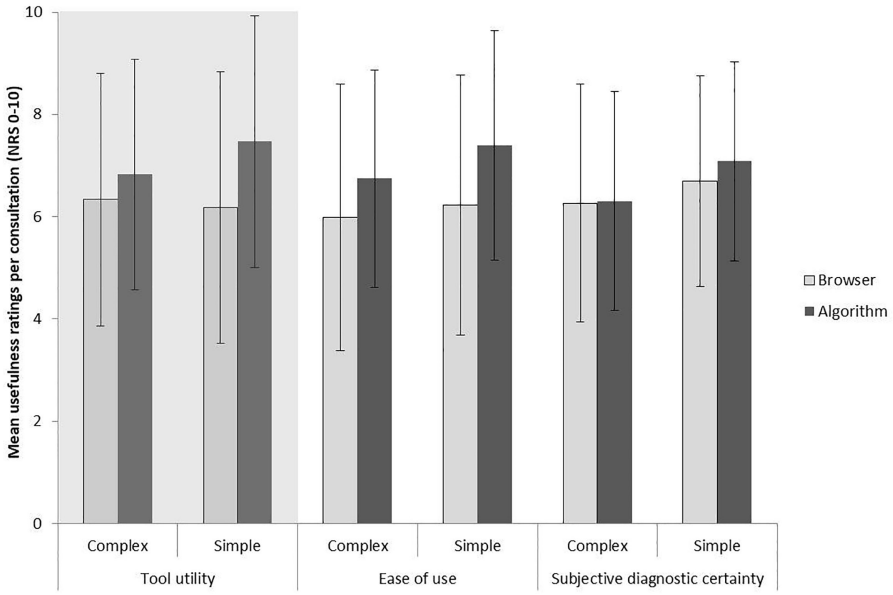

FIGURE 8 Usefulness ratings per consultation separately for complex and simple cases. This figure shows the mean values and standard errors of the usefulness ratings for the browser and the algorithm. Participants rated tool utility, ease of use and subjective diagnostic certainty from 0 to 10 after each consultation. The grey box shows a significant interaction between the factors tool and complexity.

particular, this applied to the simple cases (one diagnosis). For the complex cases, the use of the algorithm led to more partially correct diagnoses compared to the browser. The algorithm's superior performance was also sustained when cases were analysed separately for chronic primary and chronic secondary pain. Novices benefitted more from the algorithm than experts. Both experts and novices rated the algorithm as more useful and easier to use than the browser.

The absolute level of correct diagnoses was 67.7%. Given that the participants had received no training regarding the new diagnoses, except for a brief document explaining the classification's basics, and also considering the artificial consultation situation, the absolute level of correct diagnoses is noteworthy and a point in favour of the clarity of the diagnoses. The positive results are in line with earlier findings (Barke et al., 2022; Korwisi et al., 2021, 2022). Using the CAL-CP improved the rate of correct diagnoses when compared to the native browser and integrates well with a wealth of research pointing to the benefits of decision trees in diagnostic and classification contexts (Bollestad et al., 2015; Morgan et al., 2000; Rinaldi et al., 2000). Especially novices profited from the use of the CAL-CP. This accords well with other research

TABLE 6 Results of two mixed ANOVA with the within-subject factor Tool (algorithm, browser) and the between-subject factor Professional Experience (novice, expert) for the ratings of Tool utility and Ease of use (assessed after each consultation).

	Algorithm		Browser		Professional experience		Tool		Tool × rofessional experience	
	M	SD	*M*	SD	*F* (1, 142)	η^2	*F* (1, 142)	η^2	*F* (1, 142)	η^2
Tool utility					3.37	0.02	4.70*	0.03	0.53	0.00
Novice	6.93	2.13	6.09	2.40						
Expert	7.31	2.68	6.89	2.43						
Ease of use					3.44	0.02	9.29**	0.06	0.10	0.00
Novice	6.91	2.07	5.96	2.39						
Expert	7.39	2.56	6.63	2.22						

Note: Tool utility and Ease of use rated from 0 to 10.

*$p < 0.05$; **$p < 0.01$.

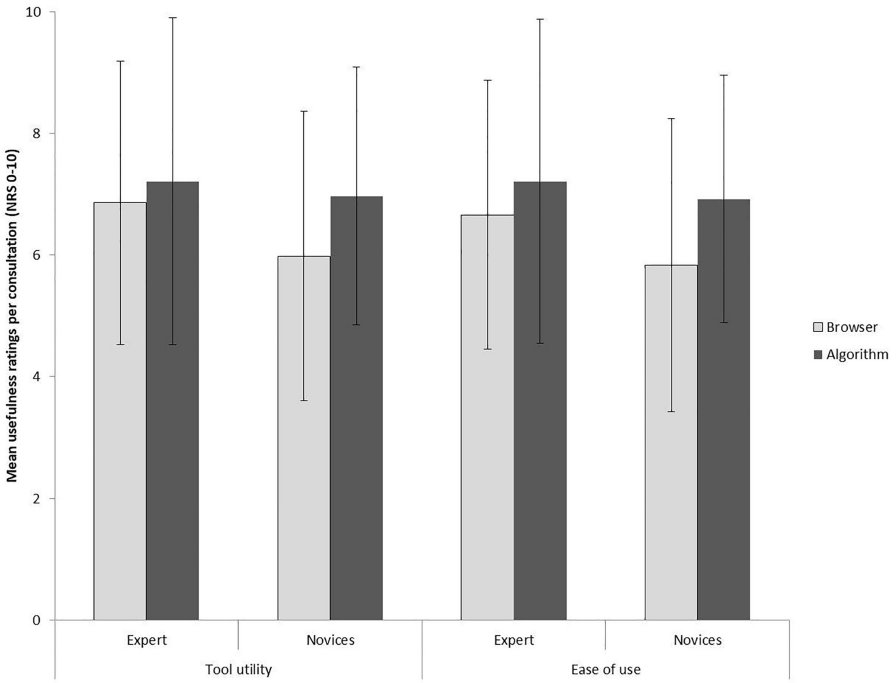

FIGURE 9 Usefulness ratings per consultation separately for novices and experts. This figure shows the mean values and standard errors of the usefulness ratings for the browser and the algorithm. Participants rated tool utility and ease of use from 0 to 10 after each consultation. A significant main effect for tool was found for both, tool utility and ease of use.

showing that novices generally benefit from the use of linear decision trees and achieve a higher number of correct diagnoses through their use than otherwise (Morgan et al., 2000). When decision trees are used to navigate through new diagnostic criteria, even minimal training is sufficient for reliable diagnoses (Malt, 1986). Thus, one of CAL-CP's uses may be in medical education and to train existing health personnel in the use of the new diagnoses.

The algorithm improves diagnostic correctness for both chronic primary and chronic secondary pain. This is a particularly important result because chronic primary pain was introduced as a new category of pain diagnoses in ICD-11 (Nicholas et al., 2019; Treede et al., 2019). For the first time, this category overcomes dualistic aetiologies defined by purely somatic or psychological causes and represents conditions in which the chronic pain, characterized by emotional distress or functional interference, is a disease in itself (Treede et al., 2019).

When testing the CAL-CP, we included complex cases. We reasoned that diagnostic decision trees, which guide and structure the collection of information, may be particularly helpful in cases in which many aspects may be relevant by ensuring that no relevant facts are missed. In the complex cases, CAL-CP tended to be superior to the

browser in terms of correctness. However, we found no advantage of the algorithm over the browser for the completeness of the diagnoses, that is, whether the clinician picked up on the fact that two diagnoses were present. One possible reason for this could be that for each patient the clinicians had at their disposal—standardly and regardless of condition—a pain chart with a pain manikin. This may have guided the clinicians' attention to all pain locations so that they did not miss the second pain diagnosis even in the browser condition.

The number of dropouts was substantial, which is a well-known problem with online studies (O'Neil et al., 2003). Almost 75% of the drop-outs took place at the point of the platform change to the virtual patients, which may have been a source of technical error and presented an additional barrier. Older participants whose English proficiency was lower dropped out more frequently. Possibly the process of having to register on a new platform with a previously self-generated code was particularly challenging for them. However, none of these participants had interacted with the material of the main study, the virtual patients. The dropout rate is therefore unlikely to exert a systematic influence on the study results.

The use of virtual patients offered many advantages for the validation study, but also incurred some limitations. Studies have shown that virtual patients can be a valid and reliable representation of real patients and that their application is more time flexible as well as more standardized compared to standardized simulation patients with equally good diagnostic results (Hubal et al., 2000; Parsons et al., 2008; Triola et al., 2006). However, the situation was artificial and unaccustomed for the clinicians: they could not conduct physical examinations but instead had to rely on documentation and written chat messages matched to the questions by a computer program. Due to the limitations of the natural language processing technologies available at the time the chat program was developed, summarizing questions such as "Have we covered everything you feel is relevant?" could not be answered by the virtual patients. To learn how the clinicians experienced the interaction with the virtual patients, we included questions about the perceived quality of the interaction. The clinicians agreed that it is a useful technology especially for training purposes. Still, the question recognition and the interaction with the virtual patients was overall in the middle range, indicating that it was possible to conduct the diagnostic interview, but not without some difficulty. This could have led to less information being collected than in a natural diagnostic situation, especially for the specific information needed for the detailed diagnoses on level 3. In addition, the virtual patients in our study did not have a voice output or non-verbal behaviour such as emotional face expressions, which are important sources of information in a diagnostic situation (Parsons et al., 2008). Considering these aspects, we would venture to suggest that in a real setting, the diagnostic success would be even higher.

We conducted the analysis at the level 1 of the classification since here we had the most extensive data source given that not all participants attempted the fine-grained diagnoses at the sub-levels. This situation reflects the situation in reality, given that most patients with chronic pain are seen by primary care physicians: About 40% of consultations with General Practitioners (GPs) are related to pain (Friessem et al., 2009; Mäntyselkä et al., 2001) and for the majority of patients with chronic pain treatment takes place in a primary care setting (Breivik et al., 2006). The GP may also refer patients further to other specialties, such as neurologists (c. 10%), orthopaedics (27%) and about 2% to a pain management specialist (Breivik et al., 2006). To determine treatment pathways in this context, the level 1 diagnoses are crucial (Treede et al., 2019). Due to limited resources and time, the assessment of chronic pain in general practice is challenging (Smith & Torrance, 2011) and the use of screening tools to identify chronic pain can be helpful (Mills et al., 2016). To date, some screening tools have proven effective in primary care. However, they were confined to specific subtypes of chronic pain, such as neuropathic pain (Haanpaa et al., 2009) or low back pain (Beneciuk et al., 2013). The CAL-CP, on the other hand, maps the entire classification of chronic pain in the ICD-11 and may therefore be a particularly useful tool for primary care.

However, exploratory analyses of the level 3 diagnoses for those participants who had attempted them, showed the same superiority of the CAL-CP for these diagnoses. At this level, the pain syndromes are described in more detail and the criteria usually require detailed examination and, in many instances, further diagnostic tests. We strove to implement these aspects with the medical records provided, but naturally there is a limit to what could be represented with the help of the virtual patients.

4.1 | Limitations

The results of the present study should be interpreted in the context of its limitations. Although great care was taken to render the diagnostic situation as natural as possible, some artificiality remained, due to the chat-medium and the particularities of the virtual patients. This meant that at the same time as using the new diagnoses the clinicians had to cope with an unaccustomed setting. The limited number of participating physicians made further subgroup analysis according to medical speciality impossible.

4.2 | Conclusion

Overall, the clinicians arrived at a high percentage of correct diagnoses. CAL-CP emerged as a useful tool for the diagnosis of chronic pain, leading to more correct diagnoses than the native browser. This applied to all diagnoses on level 1 and on level 3, and to diagnoses of chronic primary pain and chronic secondary pain analysed separately. Especially novice users benefitted from the use of CAL-CP. All users judged the CAL-CP as more useful and easier to use than the browser and preferred it over the latter. Because it leads to equally good or better results for all users, we recommend its use for clinical diagnosis, education and identification of study populations.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the IASP Taskforce who reviewed the patient cases, Kai Berkemeyer for programming the virtual patient interface, and Christiane Blöcher, Eva Driesch, Maike Klett, Anne Moser and Theresa Neumann for their support in study development and recruitment. We also acknowledge the pilot testers of the interface, the participating physicians and students. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

The research was supported by a proFOR+ grant by the Catholic University of Eichstätt-Ingolstadt. We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

CONFLICT OF INTEREST STATEMENT

BK reports that her previous position at the University of Marburg was funded through a research grant by the IASP as well as consulting fees from IASP, outside the submitted work. The other authors do not report any conflict of interest.

ORCID

Beatrice Korwisi https://orcid.org/0000-0003-1477-6742
Norman Lahme-Hütig https://orcid.org/0000-0002-1794-7095
Winfried Rief https://orcid.org/0000-0002-7019-2250
Antonia Barke https://orcid.org/0000-0002-6863-3213

REFERENCES

Aziz, Q., Giamberardino, M. A., Barke, A., Korwisi, B., Baranowski, A. P., Wesselmann, U., Rief, W., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Chronic secondary visceral pain. *Pain*, *160*, 69–76. https://doi.org/10.1097/j.pain.0000000000001362

Barke, A., Korwisi, B., Jakob, R., Konstanjsek, N., Rief, W., & Treede, R. D. (2022). Classification of chronic pain for the International Classification of Diseases (ICD-11): Results of the 2017 international World Health Organization field testing. *Pain*, *163*, e310–e318. https://doi.org/10.1097/j.pain.0000000000002287

Beneciuk, J. M., Bishop, M. D., Fritz, J. M., Robinson, M. E., Asal, N. R., Nisenzon, A. N., & George, S. Z. (2013). The STarT back screening tool and individual psychological measures: Evaluation of prognostic capabilities for low Back pain clinical outcomes in outpatient physical therapy settings. *Physical Therapy*, *93*, 321–333. https://doi.org/10.2522/ptj.20120207

Bennett, M. I., Kaasa, S., Barke, A., Korwisi, B., Rief, W., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Chronic cancer-related pain. *Pain*, *160*, 38–44. https://doi.org/10.1097/j.pain.0000000000001363

Benoliel, R., Svensson, P., Evers, S., Wang, S.-J., Barke, A., Korwisi, B., Rief, W., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Chronic secondary headache or orofacial pain. *Pain*, *160*, 60–68. https://doi.org/10.1097/j.pain.0000000000001435

Bollestad, M., Grude, N., & Lindbaek, M. (2015). A randomized controlled trial of a diagnostic algorithm for symptoms of uncomplicated cystitis at an out-of-hours service. *Scandinavian Journal of Primary Health Care*, *33*, 57–64. https://doi.org/10.3109/02813432.2015.1041827

Breivik, H., Collett, B., Ventafridda, V., Cohen, R., & Gallacher, D. (2006). Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *European Journal of Pain*, *10*, 287–333. https://doi.org/10.1016/j.ejpain.2005.06.009

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Evans, S. C., Roberts, M. C., Keeley, J. W., Blossom, J. B., Amaro, C. M., Garcia, A. M., Stough, C. O., Canter, K. S., Robles, R., & Reed, G. M. (2015). Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*, *15*, 160–170. https://doi.org/10.1016/j.ijchp.2014.12.001

Friessem, C. H., Willweber-Strumpf, A., & Zenz, M. W. (2009). Chronic pain in primary care. German figures from 1991 and 2006. *BMC Public Health*, *9*, 299. https://doi.org/10.1186/1471-2458-9-299

Haanpaa, M. L., Backonja, M. M., Bennett, M. I., Bouhassira, D., Cruccu, G., Hansson, P. T., Jensen, T. S., Kauppila, T., Rice, A. S. C., Smith, B. H., Treede, R.-D., & Baron, R. (2009). Assessment of neuropathic pain in primary care. *American Journal of Medicine*, *122*(10 Suppl), S13–S21. https://doi.org/10.1016/j.amjmed.2009.04.006

Hubal, R. C., Kizakevich, P. N., Guinn, C. I., Merino, K. D., & West, S. L. (2000). The virtual standardized patient. Simulated patient-practitioner dialog for patient interview training. *Studies in Health Technology and Informatics*, *70*, 133–138.

Keeley, J. W., Reed, G. M., Roberts, M. C., Evans, S. C., Medina-Mora, M. E., Robles, R., Rebello, T., Sharan, P., Gureje, O., First, M. B.,

Andrews, H. F., Ayuso-Mateos, J. L., Gaebel, W., Zielasek, J., & Saxena, S. (2016). Developing a science of clinical utility in diagnostic classification systems: Field study strategies for ICD-11mental and behavioural disorders. *American Psychologist*, *71*, 3–16.

Korwisi, B., Garrido Suarez, B. B., Goswami, S., Gunapati, N. R., Hay, G., Hernandez Arteaga, M. A., Hill, C., Jones, D., Joshi, M., Kleinstauber, M., Lopez Mantecon, A. M., Nandi, G., Papagari, C. S. R., Rabi Martinez, M. D. C., Sarkar, B., Swain, N., Templer, P., Tulp, M., White, N., … Barke, A. (2022). Reliability and clinical utility of the chronic pain classification in the 11th Revision of the International Classification of Diseases from a global perspective: Results from India, Cuba, and New Zealand. *Pain*, *163*, e453–e462. https://doi.org/10.1097/j.pain.0000000000002379

Korwisi, B., Hay, G., Attal, N., Aziz, Q., Bennett, M. I., Benoliel, R., Cohen, M., Evers, S., Giamberardino, M. A., Kaasa, S., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Schug, S., Smith, B. H., Svensson, P., Vlaeyen, J. W. S., Wang, S. J., … Barke, A. (2021). Classification algorithm for the International Classification of Diseases-11 chronic pain classification: Development and results from a preliminary pilot evaluation. *Pain*, *162*, 2087–2096. https://doi.org/10.1097/j.pain.0000000000002208

Malt, U. F. (1986). Teaching DSM-III to clinicians. Some problems of the DSM-III system reducing reliability, using the diagnosis and classification of depressive disorders as an example. *Acta Psychiatrica Scandinavica*, *73*, 68–75.

Mäntyselkä, P., Kumpusalo, E., Ahonen, R., Kumpusalo, A., Kauhanen, J., Viinamäki, H., Halonen, P., & Takala, J. (2001). Pain as a reason to visit the doctor: A study in Finnish primary health care. *Pain*, *89*, 175–180.

Mills, S., Torrance, N., & Smith, B. H. (2016). Identification and management of chronic pain in primary care: A review. *Current Psychiatry Reports*, *18*, 22. https://doi.org/10.1007/s11920-015-0659-9

Morgan, R. D., Olson, K. R., Krueger, R. M., Schellenberg, R. P., & Jackson, T. T. (2000). Do the DSM decision trees improve diagnostic ability? *Journal of Clinical Psychology*, *56*, 73–88. https://doi.org/10.1002/(sici)1097-4679(200001)56:1

Nicholas, M., Vlaeyen, J. W. S., Rief, W., Barke, A., Aziz, Q., Benoliel, R., Cohen, M., Evers, S., Giamberardino, M. A., Goebel, A., Korwisi, B., Perrot, S., Svensson, P., Wang, S.-J., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Chronic primary pain. *Pain*, *160*, 28–37. https://doi.org/10.1097/j.pain.0000000000001390

O'Neil, K. M., Penrod, S. D., & Bornstein, B. H. (2003). Web-based research: Methodological variables' effects on dropout and sample characteristics. *Behavior Research Methods, Instruments, & Computers*, *35*, 217–226. https://doi.org/10.3758/BF03202544

Parsons, T. D., Kenny, P., Ntuen, C. A., Pataki, C. S., Pato, M. T., Rizzo, A. A., St-George, C., & Sugar, J. (2008). Objective structured clinical interview training using a virtual human patient. *Studies in Health Technology and Informatics*, *132*, 357–362.

Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R., & Lee, M. (2000). Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of 3 methods for measuring quality. *JAMA*, *283*, 1715–1722. https://doi.org/10.1001/jama.283.13.1715

Peabody, J. W., Luck, J., Glassman, P., & Jain, S. (2004). Measuring the quality of physician practice by using clinical vignettes: A prospective validation study. *Annals of Internal Medicine*, *141*, 771–780. https://doi.org/10.7326/0003-4819-141-10-200411160-00008

Perrot, S., Cohen, M., Barke, A., Korwisi, B., Rief, W., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Chronic secondary musculoskeletal pain. *Pain*, *160*, 77–82. https://doi.org/10.1097/j.pain.0000000000001389

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing*.

Rinaldi, G., Zarrelli, M. M., Beghi, E., Apollo, F., Germano, M., Di Viesti, P., & Simone, P. (2000). The international classification of the epilepsies and epileptic syndromes: An algorithm for its use in clinical practice. *Epilepsy Research*, *41*, 223–234.

Scholz, J., Finnerup, N. B., Attal, N., Aziz, Q., Baron, R., Bennett, M. I., Benoliel, R., Cohen, M., Cruccu, G., Davis, K. D., Evers, S., First, M. B., Giamberardino, M. A., Hansson, P., Kaasa, S., Korwisi, B., Kosek, E., Lavand'homme, P., Nicholas, M., … Classification Committee of the Neuropathic Pain Special Interest, G. (2019). The IASP classification of chronic pain for ICD-11: Chronic neuropathic pain. *Pain*, *160*, 53–59. https://doi.org/10.1097/j.pain.0000000000001365

Schug, S. A., Lavand'homme, P., Barke, A., Korwisi, B., Rief, W., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Chronic postsurgical or posttraumatic pain. *Pain*, *160*, 45–52. https://doi.org/10.1097/j.pain.0000000000001413

Smith, B. H., Fors, E. A., Korwisi, B., Barke, A., Cameron, P., Colvin, L., Richardson, C., Rief, W., Treede, R.-D., & IASP Taskforce for the Classification of Chronic Pain. (2019). The IASP classification of chronic pain for ICD-11: Applicability in primary care. *Pain*, *160*, 83–87. https://doi.org/10.1097/j.pain.0000000000001360

Smith, B. H., & Torrance, N. (2011). Management of chronic pain in primary care. *Current Opinion in Supportive and Palliative Care*, *5*, 137–142. https://doi.org/10.1097/SPC.0b013e328345a3ec

Treede, R.-D., Rief, W., Barke, A., Aziz, Q., Bennett, M. I., Benoliel, R., Cohen, M., Evers, S., Finnerup, N. B., First, M. B., Giamberardino, M. A., Kaasa, S., Korwisi, B., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Scholz, J., Schug, S., … Wang, S.-J. (2019). Chronic pain as a symptom or a disease: The IASP classification of chronic pain for the international classification of diseases (ICD-11). *Pain*, *160*, 19–27. https://doi.org/10.1097/j.pain.0000000000001384

Treede, R.-D., Rief, W., Barke, A., Aziz, Q., Bennett, M. I., Benoliel, R., Cohen, M., Evers, S., Finnerup, N. B., First, M. B., Giamberardino, M. A., Kaasa, S., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Scholz, J., Schug, S., Smith, B. H., … Wang, S.-J. (2015). A classification of chronic pain for ICD-11. *Pain*, *156*, 1003–1007. https://doi.org/10.1097/j.pain.0000000000000160

Triola, M., Feldman, H., Kalet, A. L., Zabar, S., Kachur, E. K., Gillespie, C., Anderson, M., Griesser, C., & Lipkin, M. (2006). A randomized trial of teaching clinical skills using virtual and live standardized patients. *Journal of General Internal Medicine*, *21*, 424–429. https://doi.org/10.1111/j.1525-1497.2006.00421.x

World Health Assembly. (2019). The 72nd World Health Assembly Resolution for ICD-11 Adoption. Retrieved 20 May 2022 from

https://www.who.int/standards/classifications/classification-of-diseases

World Health Organization. (n.d.). WHO-FIC maintenance platform. Retrieved 20 May 2022 from https://icd.who.int/dev11

World Medical Association. (2013). World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, *310*, 2191–2194. https://doi.org/10.1001/jama.2013.281053

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hay, G., Korwisi, B., Lahme-Hütig, N., Rief, W., & Barke, A. (2024). Clinicians diagnosing virtual patients with the classification algorithm for chronic pain in the ICD-11 (CAL-CP) achieve better diagnoses and prefer the algorithm to standard tools: An experimental validation study. *European Journal of Pain*, *28*, 1509–1523. https://doi.org/10.1002/ejp.2274