Physicians' metacognition in medical decision-making

Camille Lakhlifi, Rayan Kouzy, François-Xavier Lejeune, Nicolas Beuzon, Jérôme Mawet, Caroline Roos, Mehdi Khamassi, Benjamin Rohaut, Marion Rouault

Physicians' metacognition in medical decision-making

AUTHORS

Lakhlifi Camille[1,2], Kouzy Rayan[1] , Lejeune François-Xavier[1,3], Beuzon Nicolas[1], Mawet Jérôme[5], Roos Caroline[2,5], Khamassi Mehdi[4], Rohaut Benjamin[1,6,#], Rouault Marion[1,#]

[#] equal contribution

[1]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France.
[2]Université Paris Cité, Paris, France.
[3]Paris Brain Institute - ICM, Inserm, CNRS, Data Analysis Core, Paris, France.
[4]Institute of Intelligent Systems and Robotics, Sorbonne Université, CNRS, Paris, France.
[5]AP-HP, Hôpital Lariboisière, Emergency Headache Center (Centre d'Urgences Céphalées), Department of Neurology, Paris, France.
[6]AP-HP, Hôpital de la Pitié Salpêtrière, DMU Neurosciences, Paris, France.

Correspondence: benjamin.rohaut@sorbonne-universite.fr, marion.rouault@gmail.com

KEYWORDS

Clinical decision-making, metacognition, headache

ABSTRACT

A large literature reports that physicians are overconfident when making medical decisions. However, the validity of these findings is questioned due to methodology pitfalls leading to poor ecological validity. Moreover, most previous studies do not distinguish between i) metacognitive bias, the general tendency to report high or low confidence irrespectively of actual accuracy, and ii) metacognitive sensitivity, the capacity to discriminate between one's own correct and incorrect decisions. Here, we aim to overcome these limitations by drawing from state-of-the-art experimental paradigms and robust tools from the cognitive science of metacognition. To examine overconfidence and disentangle metacognitive bias from metacognitive sensitivity, we developed a carefully-controlled set of case-vignettes of patients with headaches that was completed by 52 physicians. We found that, although physicians are overconfident, they maintain relatively good insight into the accuracy of their decisions. In addition, and unlike previous reports, we found limited inter-individual variability depending on specialty, gender, or seniority level. These results shed new light on the (meta)cognitive properties of medical decision-making and carry practical implications for medical education.

INTRODUCTION

Decision-making is a cornerstone of medical practice, profoundly influenced by a multitude of factors such as time constraints, stress, cognitive and emotional load, personality traits, fatigue, and uncertainty. These elements shape the professional environment of physicians, impacting their clinical reasoning and decision-making processes [1]. A significant challenge in a context of uncertainty is a lack of direct, immediate feedback, preventing physicians from refining their future medical strategies. In this context, metacognition, our ability to evaluate and reflect on our own cognitive capacities and decisions, is crucial. Metacognition enables the self-assessment of physicians' medical performance, for instance by generating explicit or implicit confidence judgments that in turn influence a number of processes, such as exploring alternative options or seeking a second opinion [2]. Biased confidence judgments can have important negative consequences, e.g., when an overconfident doctor becomes blind to signs or symptoms that go against their initial diagnosis. While advances in medical technology and artificial intelligence offer promising tools to mitigate such biases, they are not to entirely replace the nuanced decision-making processes that experienced physicians bring to the table; in particular, this capacity for accurate confidence seems still bound to humans [3].

Overconfidence in physicians has been repeatedly reported over the past decades [4]. A systematic review identified overconfidence as one of the most common cognitive biases in medical contexts, with a prevalence between 46,3 to 70% [5]. To examine the link between physicians' objective performance and their subjective perception of it, classic experiments consist of surveys with knowledge questions, clinical case-vignettes or clinical procedures, each associated with a retrospective or concurrent confidence rating. However, analysis strategies and computed indicators vary from one study to another (correlations, calibration/bias, and to a lesser extent discrimination/sensitivity), and results about the link between accuracy and confidence remain unclear [6–11].

However, the variability and complexity of these empirical results is related to a number of theoretical and practical experimental limitations. A first issue lies in the low ecological validity of experiments, where overconfidence may be an artifact resulting from artificial stimuli whose content, difficulty, and frequency poorly represent real-world situations [12]. A second hurdle is the partially intertwined nature of objective performance and subjective confidence [13]. Confidence partly stems from the same inputs that determine performance, e.g., the objective difficulty of the case. To put it simply, easier cases will lead both to better performance and (rightfully so) higher confidence. In contrast, behavioral studies of metacognition have typically relied on comparing subjective confidence estimates to objective accuracy across a sequence of decisions to derive various indicators of

metacognitive abilities [13–17]. Two main families of indices have been proposed: metacognitive bias reflecting one's general tendency to report high or low confidence (notions of calibration, overconfidence), and metacognitive sensitivity, reflecting one's capacity to discriminate between their own correct and incorrect decisions (notions of resolution, discrimination, metacognitive accuracy, metacognitive efficiency) [13–19]. However, the analysis of confidence in clinical decision-making has almost exclusively focused on the former and overlooked the latter. This second notion may indeed reflect a crucial facet of metacognitive ability for accurate medical decisions, for instance being able to seek additional evidence or a second opinion when confidence is rightfully low. Moreover, different combinations of metacognitive bias and metacognitive sensitivity can arise. While showing an overall high (overconfidence) or low (underconfidence) metacognitive bias, individuals can at the same time either adequately assess their accuracy by expressing stronger confidence in correct than incorrect decisions (high metacognitive sensitivity), or confound the quality of their decisions by reporting overlapping confidence judgments for correct and incorrect decisions (low metacognitive sensitivity). A third pitfall arises from analysis strategies [20–22]; moreover, the strength of the overconfidence effect depends on what is defined as the optimal level of what confidence should be.

Although these considerations are extensively discussed in metacognition research, they have not yet been addressed in clinical decision-making research (though see [8,10] for exceptions). This indicates a need for a better integration of theoretical insights and methodological improvements into protocols studying physicians' decision-making and their associated confidence.

Hence, we developed a new experimental design to avoid a bias toward demonstrating overconfidence and to enable the study of metacognitive sensitivity in clinical decisions. We developed 60 two-alternative forced-choice case-vignettes of patients with acute headaches, which were administered online to 52 physicians. We designed the task, instructions, confidence scales, and vignette content to be as ecological and representative of real-world scenarios as possible [23]. Critically, our protocol allows for the application of type-I and -II signal detection theory analyses to disentangle metacognitive bias from metacognitive sensitivity [19]. We found that physicians were not as overconfident as previously estimated, and that their confidence could discriminate between their own correct and incorrect judgements reasonably well. Contrary to previous reports and popular belief, we found limited variations according to medical specialty, gender, and seniority level. These findings not only provide solid evidence for overconfidence in physicians, but also highlight the practical importance of cultivating metacognitive sensitivity in medical training to enhance diagnostic accuracy and patient outcomes.

METHODS

Study design

This preregistered experiment (https://osf.io/q7zrm) consisted of an online questionnaire chunked into six sessions (Fig. 1). This research received approval from our local ethics committee (CER-Sorbonne Université; protocol CER-2022-Lakhlifi-METADECIDOC) and complies with European General Data Protection Regulation. Participants gave their informed consent.

Participants registered their contact data and demographic profile, before being presented with instructions and an example. They reported their gender, age, level of experience (number of years of practice since the end of residency), additional qualifications, type of practice (public/private, hospital/medical office), medical specialty (neurologists, emergency physicians or general practitioners), research activity (and if any, the percentage of research in their professional time). Participants were also asked to indicate the average number of patients presenting with headaches they see per month. This step allowed us to check whether potential participants met inclusion criteria before continuing (see **Exclusion criteria** section below). Following the validation of their profile, participants received a daily email with a link to the survey, with a reminder of task instructions followed by one of the five sessions of cases. Each session was composed of 12 case-vignettes to examine. The 60 case-vignettes were pseudo-randomly distributed across 5 sessions (with a balance of 25% to 75% of signal-present(/absent) trials within session). The order of sessions was randomized across participants (Latin square design). The order of the 12 vignettes was randomized within each session for each participant. The collection of identifying data and the management of email invitations were handled via REDCap [24,25] while other demographic data and responses to case-vignettes were collected via LimeSurvey [26]. Participant recruitment was supported by the Paris Brain Institute PRISME Core Facility (RRID:SCR_026394), Paris, France.

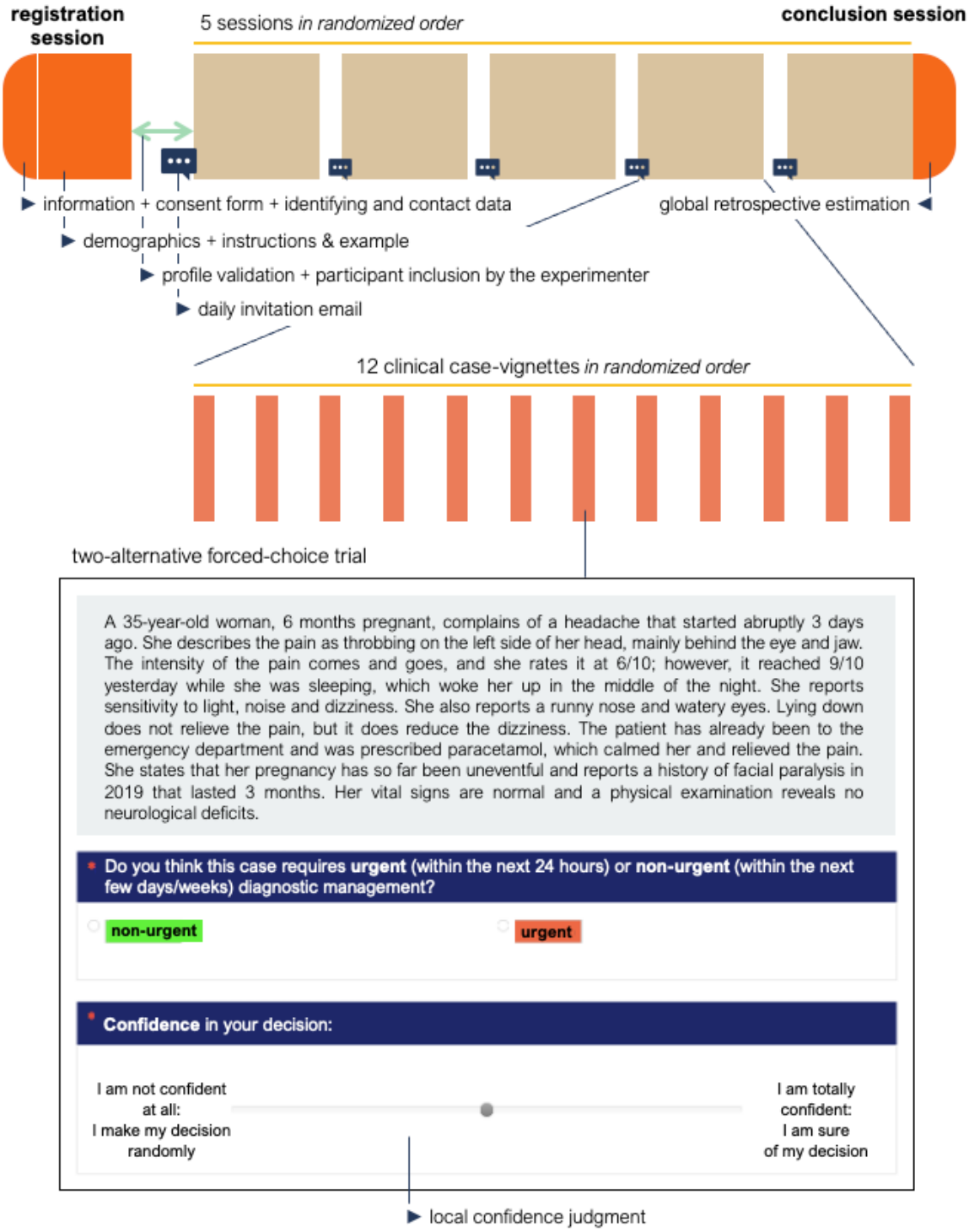

Figure 1. Experimental paradigm. 52 physicians were presented with case-vignettes describing an acute non-traumatic headache. The text could contain 0, 1, or 2 red-flags according to official guidelines, for which the expected response was non-urgent (0) and urgent (1 or 2) respectively. All case-vignettes were validated by neurologists specializing in headaches. Participants indicated their decision (binary) together with their decision confidence on a continuous visual analog scale ranging from “I am not confident at all: I make my decision randomly” to “I am totally confident: I am sure of my decision”. At the end of the survey, participants were further asked for a global retrospective estimate of confidence, that is their expected proportion of correct responses overall (see Methods).

Case-vignettes

Our case-vignettes and protocol were inspired by previous work evaluating general practitioners' referral rates based on signal detection theory [23]. We conducted a pilot study evaluating an initial set of 30 vignettes based on the International Classification of Headache Disorders' official diagnostic guidelines and academic publications on a convenience sample of 9 residents and attendants in neurology and emergency medicine. We focused on headache since it is a frequent symptom, often disabling, with numerous possible causes, ranging from benign to life-threatening thus offering variability in the required response. We then controlled for the signal (clinically relevant information) and for the noise (irrelevant information) to build a set of 60 literature-based ecologically relevant and experimentally tailored vignettes. The final set of vignettes was based on a mix of SNNOOP10-list [27], the International Classification of Headache Disorders, 3rd edition (ICHD-3) [28], and guideline recommendations (e.g. S08-P01-C13-III by Mawet et al. [29], and Moisset et al. [30,31]). Each of these 60 vignettes contains multiple pieces of information detailing the symptom sets of a patient suffering from an acute nontraumatic headache. In light of available information about the patient's profile and symptoms, participants were instructed to make a binary triage decision about whether the case required or not urgent investigation management (urgent: within the next 24h, non-urgent: in the next day(s)/weeks). According to official guidelines on which the case-vignettes were created, the presence of typical red-flags in the patient's presentation should lead physicians to consider the case as urgent. Half of our vignettes were considered "urgent" as they entailed at least one red-flag (signal-present trials) and could in majority be suspected of being secondary serious headaches (high likelihood of a serious intracranial underlying cause) [27,29–37]. These vignettes could have one, two, or three red-flags, which constituted different difficulty levels. The other half was mainly composed of cases of primary headaches (signal-absent trials) that did not represent emergency situations, for which patients could be safely sent home with instructions to have an examination in the following days/weeks if the pain persisted. Participants were not informed of the half-half distribution.

Cues introduced as red-flags in signal-present vignettes included: (i) age above 50, (ii) onset of headache is sudden or abrupt, (iii) positional headache, headache with exertion, (iv) pattern change of headache, new headache, (v) systemic symptoms (e.g., fever); (vi) neurological deficit or dysfunction (e.g., decreased consciousness), (vii) neck stiffness or neck pain, (viii) pathology of the immune system, immunosuppression, (ix) eye ptosis or pain, (x) pregnancy, (xi) neoplasm, (xii) potential infection (e.g., travel abroad). Examples of "urgent" and "non-urgent" vignettes are provided in the supplementary materials. We intended that the set was balanced between urgent and non-urgent cases in terms of the case patient's gender (Chi-square test, $p = 0.796$), age (two-sided t-test, $p = 0.765$) and

vignette text length (two-sided t-test, $p = 0.179$). We also verified the impact of the vignette patient's demographics. Unsurprisingly, the patient's age was linked to participants' accuracy (Spearman correlation, rho = 0.26, $p = 0.045$). The patient' gender did not affect participants' accuracy (Wilcoxon-Mann-Whitney test, W = 542.5, $p = 0.17$), nor did vignette's length (Spearman correlation, rho = 0.0095, $p = 0.94$). Mean confidence in each vignette was not related to the patient's age (Spearman correlation, rho = 0.11, $p = 0.41$), vignette's length (Spearman correlation, rho = -0.17, $p = 0.20$) or gender (Wilcoxon-Mann-Whitney test, W = 539.5, $p = 0.19$).

Data collection

Participants were asked to examine and solve the case-vignettes as they would have done in their daily practice in their usual work environment. Case-vignettes were two-alternative forced choices presented one at a time. Each vignette included a short reminder of instructions, the main question to be answered ("Do you think that this case requires urgent (within the next 24 hours) or non-urgent (within the next few days/weeks) diagnostic management?") followed by two response buttons, "urgent" and "non-urgent" (Fig. 1). After each decision and on the same screen, participants were asked to explicitly report their confidence in their decision on a visual analog-scale ranging from "I am not confident at all: I make my decision randomly" (collected as 50%) to "I am totally confident: I am sure of my decision" (collected as 100%) (Fig. 1). The use of two-options questions combined with a confidence judgment between 50% and 100% is a frequent practice in judgment and decision-making studies [38] and has already been used in studies evaluating physicians' confidence [10]. This confidence scale match accuracy units, ranging from chance level to maximum confidence. No numerical values were indicated on the scale to prevent any anchoring biases. Both the triage decision and the confidence judgment were compulsory to proceed to the next case. Participants had unlimited time to answer each case. At the end of the last session, participants were asked to give an overall global retrospective estimation of the proportion of trials they believed to have correctly solved [12] henceforth referred to as "global confidence judgment".

Participants

Emergency physicians, general practitioners and neurologists with an official national registration number and practicing in France were recruited between March 7th, 2023 and June 11th, 2023 using professional mailing-lists (direct contacts, Haute Autorité de Santé, Fédération Française de Neurologie, Collège de la Médecine Générale, Société Française de Médecine d'Urgence, Collège Français de Médecine d'Urgence) and social media. Our communication strategy emphasized the need for more research to better understand triage

decisions while facing patients with acute headaches, without mentioning our interest in confidence. Participants who completed the whole study were compensated 50€.

Sample size

Our design relies on 60 vignettes, a higher number than all previous studies exploring physicians' confidence in their decisions (most had fewer than 10 cases), and that typically tested 25 to a few hundred participants. Considering this and practical constraints (e.g., time pressure, limited pool of potential participants), a sample size goal was set at 50 to 60 participants to ensure reasonable statistical power with a balanced distribution of medical specialties (at least 15 GPs, 15 neurologists, and 15 emergency physicians) [39].

Exclusion criteria

No session met our pre-registered exclusion criterion about response time (sessions with less than 10 seconds to answer to at least 8 case-vignettes out of 12, as this represents a strong clue of rapid and random completion without reading the case). According to a binomial law with $p = 0.5$ and 60 trials, the probability of obtaining the lowest accuracy rate 38/60 observed among our participants by responding randomly to all vignettes is 0.026. Therefore, we did not exclude any participant based on accuracy. Three participants reported confidence judgments above 95% for 50 or more decisions out of 60; another participant gave confidence judgments above 99% for all decisions. Considering that the accuracy of these four participants was high (at least 45/60 correct responses), their high confidence could be justified, and we thus decided not to exclude them from the analyses.

Statistical analysis of behavior

We computed several pre-registered (https://osf.io/q7zrm) indicators from our measured variables to depict all possible relevant facets of physicians' metacognitive abilities.

First, for accuracy, analyses of metacognitive capacities require cases to have an objectively correct answer. However, clinical decision-making and especially triage decisions are inherently complex tasks with a variable level of uncertainty. Therefore, we preregistered two possible definitions for correctness, according to: (i) official guidelines (with verification from experts in the field); (ii) the participants' majority response (>50% of participants). We expected minimal discrepancies in responses' correctness according to these two definitions, which was indeed the case (supplementary materials). Based on signal detection theory, we computed for each participant their type-1 correct responses according to official guidelines by pooling type-1 hit rates ("urgent" response for signal-present stimulus, i.e., with at least one red-flag) and type-1 correct rejections ("non-urgent" response for signal-absent stimulus, with zero red-flag). Similarly, type-1 incorrect responses were computed from type-

1 misses (“non-urgent” response for signal-present stimulus, with at least one red-flag) and type-1 false alarms (“urgent” response for signal-absent stimulus, with zero red-flag). Type-1 discrimination (d’) and type-1 bias (c criterion) were calculated using type-1 hit (HR1) and false alarm rates (FA1): type-1 d′ = z(HR1) − z(FA1); type-1 c = -0.5 x (z(HR1) + z(FA1), with z the inverse of a cumulative normal distribution. For four participants, the real number of false alarms was 0 but was forced to 1 for calculation purposes. All other participants had at least one false alarm.

Second, for each participant we extracted metacognitive bias (also referred to as calibration index in the literature) as the difference between their average confidence and their average accuracy across decisions [19,40]. This is to compare their percentage of accuracy (with 50% representing chance level) to their average confidence judgment (ranging from 50 to 100%).

Third, for each participant we computed a “global metacognitive bias” (analogous to the one above based on local confidence), as the difference between their unique global retrospective estimate provided at the end of all sessions and their average accuracy.

Fourth, to examine whether participants’ confidence reports were reliably associated with their objective performance, we focused on metacognitive sensitivity using two indices. A first index, discrimination, was calculated as the difference between participants’ mean confidence in their correct and incorrect responses. We also compared confidence between type-1 hits, correct rejections, misses and false alarms. Meta-d’, a second index of metacognitive sensitivity based on signal detection theory was computed: meta-d’ indicates how well each participant’s confidence discriminates between their own correct and incorrect decisions, independently from their tendency to use high or low confidence reports on the scale (which corresponds to the metacognitive bias [39]). We employed a hierarchical Bayesian framework for fitting meta-*d’*, with all convergence values $\hat{R} < 1.006$ indicating satisfactory convergence [39]. We directly report the ratio meta-d’/d’, called metacognitive efficiency, which corresponds to metacognitive sensitivity divided by objective d’ (Fig. 5B).

All statistical analyses were conducted using R version 4.2.2 [41]. Continuous variables were reported as mean and standard deviation (SD), and range (min and max) for normally distributed data, or as median and interquartile range (IQR) for non-normally distributed data. Intergroup comparisons were performed using Welch’s two-sample t-test, paired t-test, Wilcoxon-Mann-Whitney test, Wilcoxon signed-rank test for paired data, Chi-square test, one-way analysis of variance (ANOVA) and Kruskal-Wallis test as appropriate. All statistical tests were two-sided and the level of statistical significance was set at $p < 0.05$.

Data and code availability

The complete stimuli set (in French) and the dataset are available upon reasonable request.

RESULTS

Experimental design

52 participants (31 women and 21 men) between 26 and 64 years old (median = 35.5 [IQR = 31.75 – 42]) completed the experiments. There were 19 general practitioners (GPs), 18 emergency physicians and 15 neurologists with no gender or age imbalance across medical specialties (all $p > 0.05$, Table 1). Most of them work in a public hospital (29, mainly emergency physicians and neurologists) or a medical office (19, mainly GPs). 19 participants had less than 5 years of professional experience, 13 between 5 and 10 years, 9 between 11 and 15 years and other participants above 16 years of experience. The majority of participants (38) reported devoting all their professional time to clinical practice; among the few (14 participants) who are involved in research, 10 declared spending more than 10% of their time doing research. Participants reported varying levels of exposure to patients presenting with acute nontraumatic headaches, with a mean of 23.8 (SD = 26.3) cases per month. Participants took a median duration of 57 min [IQR = 44 – 80] to complete the 5-seesions survey, with a median of 11 min [IQR = 8 – 15] for each session of 12 case-vignettes. Most participants respected the suggested planning of doing one session per day for five consecutive days (median time between two sessions: 1.5 days [IQR = 1.25 – 2]).

	medical specialty			
	GPs N = 19	**emergency Ps** N = 18	**neurologists** N = 15	*p-value*
gender				*0.894*
Women	12 (63.2%)	10 (55.6%)	9 (60%)	
Men	7 (36.8%)	8 (44.4%)	6 (40%)	
age (years)				*0.984*
	38.6 (9.2)	39.2 (10.9)	38.8 (10.3)	
experience				*0.946*
less than 5 years	8 (42.1%)	6 (33.3%)	5 (33.3%)	
5 to 10 years	3 (15.8%)	5 (27.8%)	5 (33.3%)	
11 to 15 years	4 (21.1%)	3 (16.7%)	2 (13.3%)	
16 years or more	4 (21.1%)	4 (22.2%)	3 (20%)	
practice				*<0.001*
Other	1 (5.3%)	2 (11.1%)	1 (6.7%)	
medical office	18 (94.7%)	0 (0%)	1 (6.7%)	
public hospital	0 (0%)	16 (88.9%)	13 (86.7%)	
research				*0.121*
Yes	2 (10.5%)	7 (38.9%)	5 (33.3%)	
No	17 (89.5%)	11 (61.1%)	10 (66.7%)	

Table 1. Participants' demographics. 52 physicians completed the survey. P-values were calculated using one-way analysis of variance (ANOVA) for age and chi-square test for other categorical variables.

Decision accuracy

52 physicians were presented with case-vignettes describing a patient consulting for an acute non-traumatic headache. Accuracy varied across the vignettes: for each vignette, the median number of physicians answering correctly according to the guidelines was 44 out of 52 (84.6%; [IQR = 36.5 – 49], range = 14 – 52). We obtained virtually identical results when defining accuracy as the group's majority response, instead of the official guidelines, as pre-registered (again 84.6%; [IQR = 37.75 – 49], range = 27 – 52). Consequently, we focus the remainder of the paper on accuracy defined according to official guidelines and report all results with our alternative definition of accuracy in the supplementary materials. Accuracy was relatively high, with a mean of 78.6% correct and a median of 78.3% correct [IQR = 75.0% – 83.3%] (range = 63.3% – 88.3%). Individual decision sensitivity (hit rate) and 1 - specificity (false alarm rate) coordinates were plotted, together with theoretical receiver operating characteristic curves for different d' values for visualisation purposes (Fig. 2). Mean d' across participants was 1.769, confirming a good discriminability.

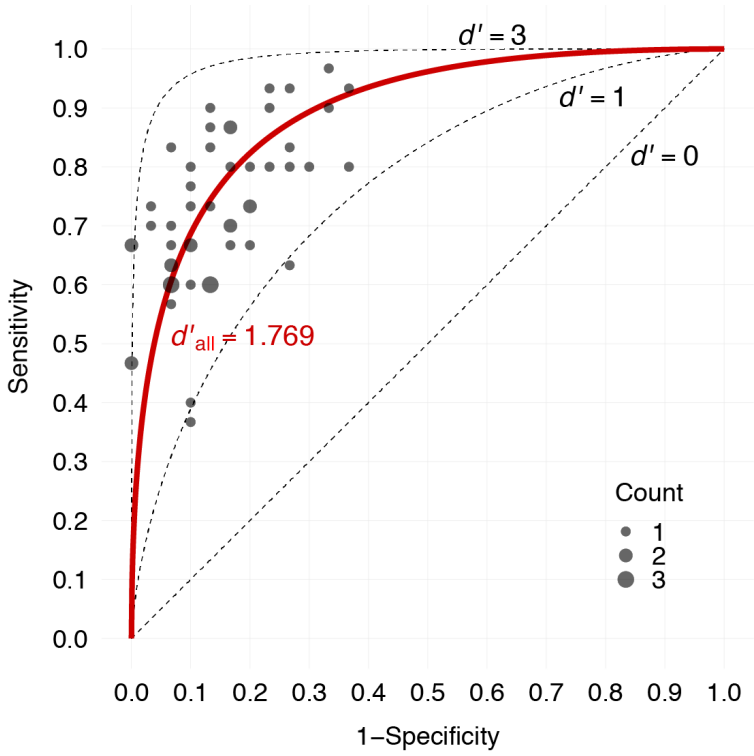

Figure 2. Scatterplot of decision sensitivity (hit rate) and 1 – specificity (false alarm rate) for the 52 participants with the group mean d' (red line) and theoretical receiver operating characteristic curves produced by different values of d' (black dotted lines). In our design, d' could theoretically vary from 0 to 3.67. Each dot represents 1, 2, or 3 participants (N=52).

As expected, the number of red-flags reflecting the strength of signal for each stimulus (i.e., vignette) had a significant impact on participants' accuracy (Kruskal-Wallis test, $p = 0.017$). Participants were less accurate when facing vignettes containing only 1 red-flag (median = 57.7% [IQR = 42.3% – 84.6%]) compared to vignettes with 0 (median = 86.5% [IQR = 82.7%

– 94.2%]; Dunn's post-hoc test with Benjamini-Hochberg [BH] multiple test correction, p = 0.020) and 2 or more red-flags (median = 87.5% [IQR = 70.7% – 95.7%]; BH-Dunn's post-hoc test, p = 0.027), plausibly facing a higher level of uncertainty with vignettes entailing a unique red-flag compared to vignettes with no or several red-flags.

Local confidence

Participants overall reported high confidence judgments in their triage decisions with a mean of 87.5% (sd = 5.8) and a median of 88.2% [IQR = 83.3% – 90.6%] (range = 73.3% – 99.5%). Local median confidence across vignettes (reported on a scale ranging from 50 to 100%) was 87.2 ([IQR = 85.6 – 89.0]. This reported mean confidence in each case-vignette was not affected by the number of red-flags (Kruskal-Wallis test, $\chi 2$ = 1.75, df = 2, p = 0.42) or the vignette text length (Spearman correlation, rho = -0.17, p = 0.20).

Metacognitive bias

Participants' level of subjective confidence was significantly higher than their objective accuracy (one-sided paired Wilcoxon signed-rank test, V = 1290, p = 2.3e-8) (Fig. 3A). We summarised this information in a metacognitive bias index (difference between average confidence and average accuracy; see Methods). For most participants, metacognitive bias was positive, with a mean of 8.9 (sd = 8.1) and a median of 8.5 [IQR = 2.6 – 14.5] (range = -8.4 – 24.8), reflecting participants' miscalibration in the form of overconfidence. At the group level, however, we did not observe any correlation between participants' accuracy and their average confidence at the group level (Spearman correlation, rho = -0.07, p = 0.40), indicating variable degrees of overconfidence.

Metacognitive bias negatively correlated with participants' accuracy (Fig. 3B). The higher doctors' performance, the lower their overconfidence (Pearson correlation, R = -0.69, p = 1.5e-8).

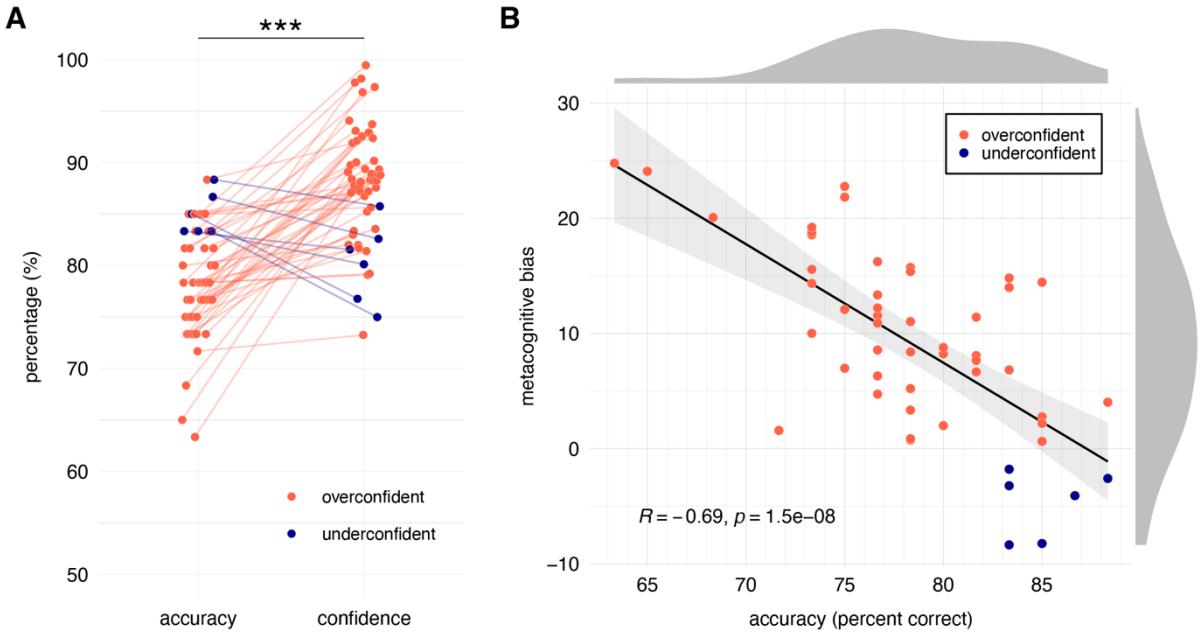

Figure 3. (A) Participants' accuracy (percentage of correctly solved trials) (left) and mean confidence over the 60 vignettes (right) (N=52). Both variables are expressed in percentages between 50% and 100%. Overconfident participants (N=46) are in orange; underconfident participants (N = 6) are in blue. ***p < 0.001. **(B).** Scatter plot of participants' individual metacognitive bias as a function of their accuracy. Each dot represents a participant (N=52); dots with a metacognitive bias above zero represent participants with overconfidence. The line represents a linear regression fit, with the shaded gray area indicating the 95% confidence interval. Marginal density plots on the top and right margins illustrate the distributions of accuracy and metacognitive bias respectively. Metacognitive bias correlated with accuracy on the 60 vignettes (R and p indicate Pearson's correlation coefficient and the statistical significance).

Global retrospective confidence: an alternative to local confidence judgments

Complementing local confidence judgments for each case, we also asked participants retrospectively, at the end of the survey, to estimate the proportion of trials they believed they had correctly answered overall. This measure is an alternative form of confidence that participants had in their knowledge or decisions, at a global scale rather than a local, decision level. The comparison of this estimate with the actual average accuracy provided an alternative 'global' metacognitive bias/calibration index.

Global median confidence across participants was 83.3 [IQR = 83.3 - 91.7]. As highlighted in previous work [12], in our sample also overconfidence disappeared when measured at a global level. This analysis revealed that participants significantly underestimated their performance at the global level (paired Wilcoxon signed-rank test, V = 1331, p = 5.2e-9) (Fig. 4). Indeed, the indices calculated as the difference between their estimated and real accuracy were mostly negative with a mean index of -9.6 (sd = 19.1) (one sample t-test for negativity, p = 0.00035) and a median of -10.0 [IQR = -18.3 – 2.1]), in contrast to metacognitive bias at the local level (Fig. 3).

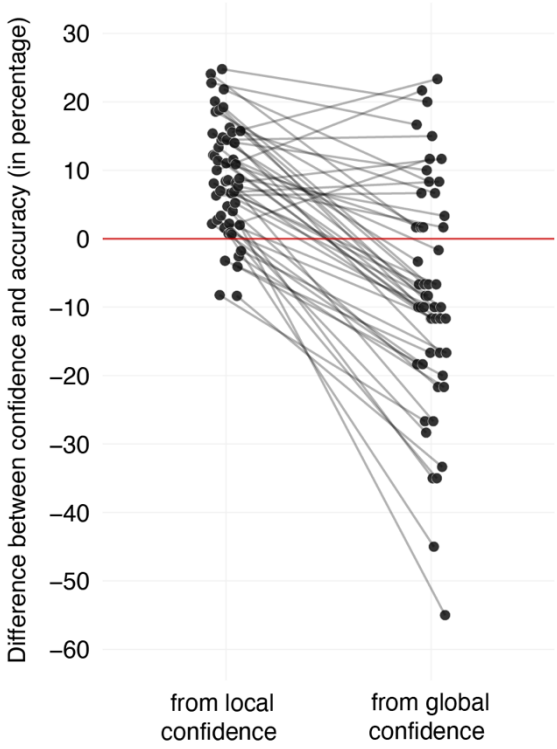

Figure 4. Two forms of confidence. The vertical axis represents the difference between confidence (local decision by decision, left; or retrospective global, at the end of the survey, right) and accuracy. Each dot represents one participant (N=52).

Metacognitive sensitivity

Besides physicians' metacognitive biases at the local and global confidence levels, a key goal of our study was to evaluate their metacognitive sensitivity, i.e., their capacity to discriminate between their correct and their incorrect decisions by reporting higher confidence in the former than in the latter. We used a two-way repeated measures ANOVA with a linear mixed-effects model to compare the individual mean confidence judgments across the types of response defined by correctness (trial correctly or incorrectly solved) and urgency (urgent or non-urgent objective categorization). We further included the urgency factor in case it modulates the effect of correctness on confidence, since prior studies have shown that yes/no answers can indeed influence metacognitive sensitivity [42]. Type-II Wald Chi-square tests revealed a significant difference in mean confidence judgments between correct and incorrect answers ($\chi 2 = 85.63$, df = 1, $p < 2.2e\text{-}16$), with higher reported confidence in correct as compared to incorrect answers (Fig. 5A). In addition, we found a significant difference in mean confidence judgments between objectively urgent (hits and misses) and objectively non-urgent (correct rejections and false alarms) cases ($\chi 2 = 11.13$, df = 1, $p = 0.00085$) (Fig. 5A). Mean confidence (estimated marginal means) for hits, correct rejections, misses and false alarms were respectively 89.0, 87.7, 84.2 and 81.7. There was no statistically significant interaction between these two factors ($\chi 2 = 1.12$, df = 1, $p = 0.29$),

meaning that higher confidence judgments were reported towards objectively urgent cases relatively to non-urgent cases, independently from correctness.

We further computed two indices to formally quantify this metacognitive sensitivity: discrimination ability (difference between confidence in correct vs. incorrect answers at the individual level) and meta-d'/d' (see Methods). First, median participants' discrimination ability was 4.0 [IQR = 1.6 – 6.3] at the group level. Metacognitive sensitivity measured as discrimination ability correlated with participants' accuracy (Pearson correlation, R = 0.29, *p* = 0.034) (Fig. 6A). This means that a higher objective accuracy made it a bit easier for participants to detect their errors.

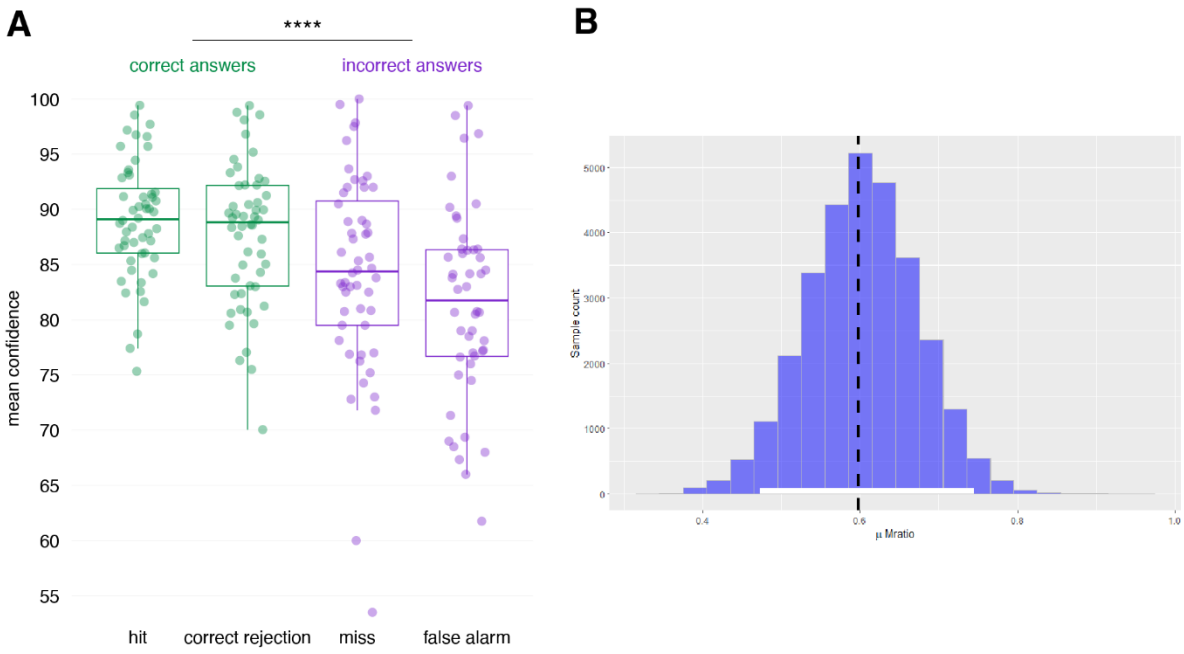

Figure 5. Physicians' metacognitive sensitivity. (A) Boxplots of participants' mean confidence according to correct (hits and correct rejections, green) and incorrect (misses and false alarms, purple) answers. Each dot represents one participant (N=52). Hits and false alarms represent cases categorized as urgent by participants while correct rejections and misses were identified as non-urgent. Hits and misses are cases that were objectively urgent according to official guidelines (at least one red-flag); correct rejections and false alarms are cases that were objectively non-urgent (no red-flag). ****$p<0.0001$ based on type II Wald chi-square test for linear mixed-effects model fit of mean confidence values. **(B)** Group-level metacognitive efficiency (H-Mratio = meta-d'/d') estimated hierarchically (see Methods).

Second, we estimated metacognitive efficiency by computing the standard index of meta-d'/d' using a hierarchical approach [19,39]. We found an average M-ratio of 0.598 over the group (Fig. 5B) with satisfactory convergence of mean $\hat{R}$=1.006. This is in line with typical metaperception studies, and smaller than typical metamemory studies - possibly related to generally high accuracy and confidence levels observed at the group level in the present study.

Finally, we examined the association between metacognitive bias and metacognitive sensitivity since previous reports (outside the clinical context) revealed that they were not

independent [19]. As expected, we found a significant correlation between participants' metacognitive bias and their metacognitive sensitivity (Pearson correlation, R = -0.45, p = 0.00078) (Fig. 6B). This correlation remains significant when removing three outlier participants (rightmost dots on Fig. 6B) (Pearson correlation, R = -0.33, p = 0.020). Participants with higher overconfidence had a weaker capacity to discriminate between their own correct and incorrect responses (and had a lower accuracy, Fig. 3B).

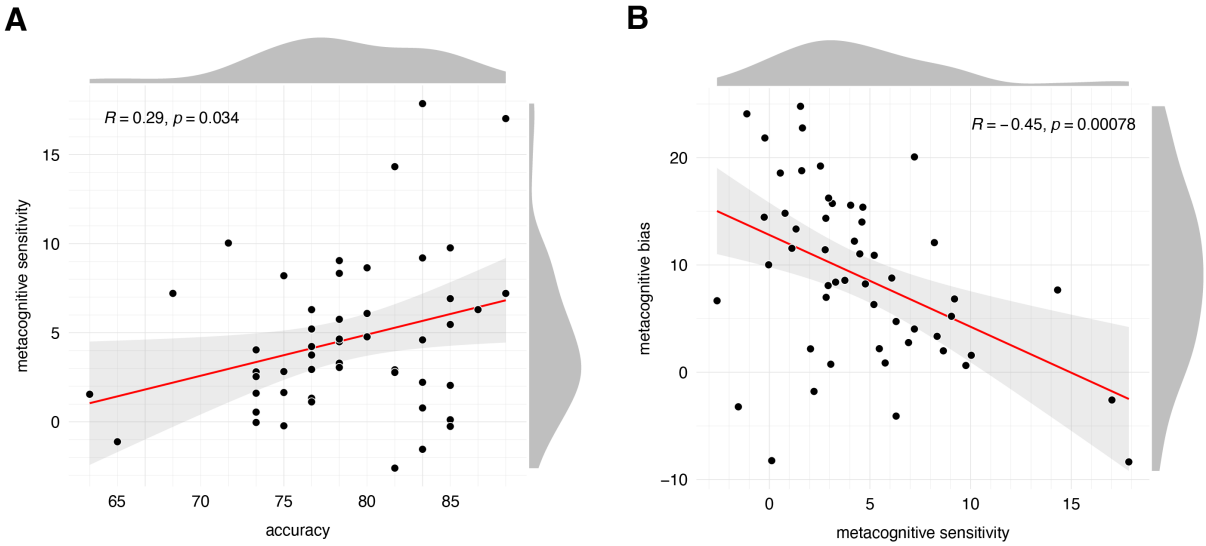

Figure 6. (A) Participants' metacognitive sensitivity as a function of their accuracy. **(B)** Participants' metacognitive bias as a function of their metacognitive sensitivity. Metacognitive sensitivity corresponds to the measure of discrimination ability in both these plots. Each dot represents one participant (N=52). Red lines represent a linear regression fit with a shaded gray area illustrating 95% confidence intervals. Marginal density plots on the top and right margins illustrate the distributions of accuracy, metacognitive bias and metacognitive sensitivity. R and p respectively indicate Pearson's correlation coefficient and the statistical significance.

Inter-individual variability

Even though it was not the primary focus of our study, and therefore our statistical power is limited, we examined the influence of participants' demographic characteristics (gender, seniority level, medical specialty and research activity) on their accuracy, metacognitive bias and metacognitive sensitivity. To anticipate our findings, overall, we found none to little associations.

At the level of decision, mean d' did not significantly vary across medical specialties (p = 0.072) (Table 2). c criterion values were also extracted for each participant. Kruskal-Wallis test and subsequent post-hoc comparisons revealed that c criterion significantly differed between specialties (p < 0.001), with a slightly more conservative bias among neurologists compared to emergency physicians and GPs (Table 2).

At the level of metacognition, none of our demographic parameters had any influence on local confidence. As for the global calibration index, it was mostly negative, with a trend of women underestimating more their performance than men (women: median = -11.7 [IQR = -

24.2 – -2.5], men: median = -6.7 [IQR = -11.7 – 8.3], Wilcoxon-Mann-Whitney test, W = 220, p = 0.0498). None of the participants' demographics significantly impacted their metacognitive sensitivity.

measure	**all (52)**	**medical specialty**			
		neurologists (N=15)	**EPs (N=18)**	**GPs (N=19)**	*p-value*[†]
urgent (/60)	26.1	32.5	24.8	22.3	
hits	21.6	25.7	20.5	19.4	
misses	8.4	4.3	9.5	10.6	
FAs	4.5	6.8	4.3	2.8	
CRs	25.5	23.2	25.7	27.2	
d'					*0.072*
mean (sd)	1.769 (0.384)	1.921 (0.378)	1.624 (0.385)	1.786 (0.355)	
c criterion					*< 0.001*
mean (sd)	0.240 (0.397)	-0.158 (0.314)	0.303 (0.294)[a]	0.495 (0.291)[a]	

Table 2. Mean number of cases identified as urgent, hits, misses, false alarms (FAs), correct rejections (CRs). Mean and standard deviation for d' and c criterion of all participants and by medical specialty. [†]Kruskal-Wallis test, [a]p<0.001 significant difference with the neurologists group based on Dunn's post hoc test with a Benjamini-Hochberg correction.

DISCUSSION

Our study aimed at bridging the gap between state-of-the-art methods in metacognition research and the real-life context of medical triage decision-making. We took particular care to establish hypotheses and pre-register our analysis plan, elaborate ecologically relevant stimuli, and conceive a robust experimental paradigm with carefully chosen parameters. This allowed us to test several hypotheses regarding physicians' confidence in their triage decisions for patients with acute non-traumatic headaches. This is to our knowledge the first study deciphering metacognitive sensitivity in medical practitioners using signal detection theory and characterizing decision-by-decision fluctuations in local confidence. We found that physicians had overall high accuracy but were slightly overconfident. We also established that physicians had insight into the accuracy of their triage decisions reflected in a degree of metacognitive sensitivity. Unlike popular belief, we found little variations in self-evaluation of medical decision-making according to gender, medical specialty or seniority level. These findings carry important practical implications for clinical decision-making.

Many previous studies investigating clinicians' perception of their own performance rely on computing correlation coefficients between accuracy and confidence, or calculating calibration indices of the difference between average accuracy and average confidence. At least three studies using experimental paradigms similar to the present work already concluded that physicians exhibit overconfidence, as reflected by significantly higher confidence compared to their actual accuracy [9,43], even when a significant positive relationship between accuracy and confidence was observed [44]. Consistently, our data replicate this miscalibration of confidence among physicians. Our individual (mis)calibration indices were characterized by an important effect size (mean and median difference between confidence and accuracy around 8 points, on a scale of 50 to 100%).

As previously reported [12], an alternative way to probe confidence based on a single retrospective global estimation of the perceived number of correctly solved vignettes - rather than local confidence judgments elicited after each decision – interestingly led to the disappearance of the observed overconfidence at the local level and instead showed revealed underconfidence at the global level. Nevertheless, maintaining appropriate subjective confidence in each individual decision made under uncertainty remains crucial for physicians, as it helps minimize the risk of suboptimal decisions and adverse events.

The key goal of this study was to disentangle metacognitive bias from metacognitive sensitivity - a key probe into physicians' self-evaluation abilities. We found evidence that physicians were able to discriminate between correct and incorrect answers using their confidence judgments. To our knowledge, only two studies investigated metacognitive

sensitivity so far in the medical context, highlighting physicians' ability to discriminate between their correct and incorrect answers through their confidence judgments [8,10], which are in line with our results. However, and albeit significant, the magnitude of the difference in reported confidence between correct and incorrect answers is relatively small (3 to 7 points of percentage). This is potentially due to the overall high levels of accuracy in our study, meaning that participants had little space to express confidence differences on the scale. Nevertheless, these small effect sizes raise questions about the real-life behavioral consequences of variations in subjective confidence. For instance, a previous study suggested that high confidence was related to decreased requests for additional diagnostic tests [9]. Further research is needed to evaluate the required level of metacognitive sensitivity for physicians to effectively ask for a second opinion from a colleague, seek additional information, or request complementary tests. Moreover, among correct answers, hits elicited higher confidence than correct rejections. This difference might be due to finding a red-flag potentially strengthening participants' confidence, whereas when finding no red-flag, it might be difficult to assess the space of all possible absences. Among incorrect answers, misses elicited a higher confidence than false alarms. Attention should be devoted to this potentially problematic observation, considering the cost asymmetry between these two types of error for patients.

Our findings indicate that physicians generally adhere to official guidelines, with minor variations in accuracy and confidence across medical specialties. Overconfidence in clinicians can hinder the pursuit of verification or assistance, potentially fostering biases such as confirmation bias [45–49]. However, confidence can be beneficial in high-stakes or emergency scenarios, where a lack of confidence might impede decision-making. Interventions aimed at enhancing metacognition, particularly those focused on reducing overconfidence, should be carefully considered to avoid unintended consequences on performance and physician well-being [4,50–53]. Additionally, expressing doubts may also impact credibility and patient-physician relations [4,54].

The observed difference of c-criterion among medical specialties suggest that neurologists tend to adopt a more conservative approach compared to GPs and EPs. This conservatism may stem from neurologists' exposure to a higher proportion of cases requiring urgent investigation, leading them to minimize false negatives [55]. It is possible that they also assess the probability to find any other relevant information when deciding on further exploration of the patient. In contrast, GPs and EPs, who often serve as the first point of contact for patients, must balance efficient triage with the risk of over-referral, resulting in a more liberal c-criterion. Understanding these calibration differences is crucial for tailoring training and support systems to enhance the accuracy and reliability of clinical judgments across the healthcare spectrum.

Enhancing metacognition is often proposed to reduce cognitive biases in medical decisions and improve care quality and patient safety [55–59]. Solutions such as a longitudinal teaching curriculum aimed at medical professionals to identify their biases and promote reflective reasoning, decisional aids and mnemonic checklists can foster metacognition and critical thinking in medical reasoning [60–65].

Despite our carefully controlled experimental design, this work has some limitations. A first limitation concerns how representative of real-life cases our vignettes are. Although clinical case-vignettes were carefully designed, for our research purpose using signal-detection theory, hard or tricky cases might be overrepresented. Accuracy and confidence can be tied to difficulty [65], exemplified by the “hard-easy” effect in which individuals can be underconfident in easy tasks and overconfident in harder ones [38].

A second limitation concerns the ecological relevance of case-vignettes to assess medical decision-making. As most studies focusing on confidence in clinical decision-making [66], our survey relied on clinical case-vignettes, previously reported as valid instruments to investigate physicians’ decision-making [67]. Their formatting and framing are widely used for medical students, residents and physicians’ initial and continuous training and certification. They resemble emails or phone calls in daily practice when a colleague is asking for help in the management of a patient’s case. Despite these elements and drastically increasing the number of vignettes compared to previous studies, it remains possible that our settings created artificial and simplistic contexts prone to suboptimal reasoning [68]. We cannot rule out that the same content, if provided instead through actors or real-life patients, could lead to better calibrated and more sensitive confidence, if not increased accuracy. Moreover, we did not provide any feedback on decision accuracy or confidence. While many contexts of medical practice also lack immediate feedback, our experimental does not allow feedback-based learning and might not be fully representative of real-life situations [68].

A third limitation concerns the impact of explicitly asking for subjective confidence reports on downstream decisions. Physicians do not receive any training during the medical curriculum in consciously producing and expressing metacognitive judgments of their own performance. We devoted special attention to the framing of our instructions, questions, texts, and scale selection to probe confidence using literature-informed arbitrations and iteratively piloting and refining these aspects [38,69]. Nevertheless, variability arises when eliciting subjective, intuitive and sometimes unconscious feelings at the individual level, and there may be additional inter-individual variability in interpreting and reporting on the scale.

A last limitation is that there are reciprocal interactions by which accuracy and confidence may influence each other [18]. Eliciting confidence judgments during the task sometimes impact accuracy itself, a phenomenon known as confidence reactivity; either

enhancing performance by activating participants' metacognition or diminishing performance by diverting cognitive resources to the metacognitive judgements [69–71]. The pathway of production of metacognitive evaluations is not yet fully understood, whether one common and unified metacognitive ability applies to various cognitive processes [18,72], or metacognitive judgments arising from the same evidence as decision accuracy [73–75]. Since accuracy and confidence are related [18,76–81], we must be cautious in extending our conclusions to different types of medical decisions, targeted physicians, medical specialties, or to draw practical recommendations. The quality of care is also dependent on many other factors besides the quality of the medical decision itself as studied here at the individual level, including working conditions and available equipment.

Future studies relying on similar paradigms will be useful to evaluate metacognitive sensitivity of clinicians facing different types of medical decisions such as diagnostic and treatment choice, and different decision parameters (e.g., signal strength, emotional valence, stakes, uncertainty). This will allow us to better understand the potential factors affecting the production of confidence judgements, such as experience level, sleep deprivation, working conditions, social context, and presence of feedback. Prior to any corrective implementations, further research is required to characterize and quantify the behavioral consequences of given levels of confidence on subsequent decision-making, and to increase the number of studies investigating physicians’ confidence in their real-life decisions [66].

REFERENCES

1. Croskerry P. The Rational Diagnostician and Achieving Diagnostic Excellence. JAMA. 2022;327:317–8.

2. Boldt A, Schiffer AM, Waszak F, et al. Confidence predictions affect performance confidence and neural preparation in perceptual decision making. Sci Rep. 2019;9:1–17.

3. Griot M, Hemptinne C, Vanderdonckt J, et al. Large Language Models lack essential metacognition for reliable medical reasoning. Nat Commun. 2025;16:642.

4. Berner ES, Graber ML. Overconfidence as a Cause of Diagnostic Error in Medicine. Am J Med. 2008;121:S2–23.

5. Saposnik G, Redelmeier D, Ruff CC, et al. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak. 2016;16:138.

6. Friedman CP, Gatti GG, Franz TM, et al. Do physicians know when their diagnoses are correct?: Implications for decision support and error reduction. J Gen Intern Med. 2005;20:334–9.

7. Davis DA, Mazmanian PE, Fordis M, et al. Accuracy of Physician Self-assessment Compared With Observed Measures of Competence: A Systematic Review. JAMA. 2006;296:1094.

8. Lakhlifi C, Lejeune FX, Rouault M, et al. Illusion of knowledge in statistics among clinicians: evaluating the alignment between objective accuracy and subjective confidence, an online survey. Cogn Res Princ Implic. 2023;8:23.

9. Meyer AND, Payne VL, Meeks DW, et al. Physicians' Diagnostic Accuracy, Confidence, and Resource Requests: A Vignette Study. JAMA Intern Med. 2013;173:1952.

10. Naguib M, Brull SJ, Hunter JM, et al. Anesthesiologists' Overconfidence in Their Perceived Knowledge of Neuromuscular Monitoring and Its Relevance to All Aspects of Medical Practice: An International Survey. Anesth Analg. 2019;128:1118–26.

11. Richardson ML, Amini B, Beckmann NM, et al. Measuring and teaching confidence calibration among radiologists: a multi-institution study. J Am Coll Radiol. 2020;17:1314–21.

12. Gigerenzer G. How to make cognitive illusions disappear: Beyond "heuristics and biases." Eur Rev Soc Psychol. 1991;2:83–115.

13. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn. 2012;21:422–30.

14. Norman E, Pfuhl G, Sæle RG, et al. Metacognition in Psychology. Rev Gen Psychol. 2019;23:403–24.

15. Rahnev D, Desender K, Lee ALF, et al. The Confidence Database. Nat Hum Behav. 2020;4:317–25.

16. Rouault M, Dayan P, Fleming SM. Forming global estimates of self-performance from local confidence. Nat Commun. 2019;10:1141.

17. Schraw G. A conceptual analysis of five measures of metacognitive monitoring. Metacognition Learn. 2009;4:33–45.

18. Fleming SM, Dolan RJ. The neural basis of metacognitive ability. Philos Trans R Soc B Biol Sci. 2012;367(1594):1338–49.

19. Fleming SM, Lau HC. How to measure metacognition. Front Hum Neurosci. 2014;8:443.

20. Dunkel CS, Nedelec J, van der Linden D. Reevaluating the Dunning-Kruger effect: A response to and replication of Gignac and Zajenkowski (2020). Intelligence. 2023;96:101717.

21. Gignac GE, Zajenkowski M. The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. Intelligence. 2020;80:101449.

22. Olsson H. Measuring overconfidence: Methodological problems and statistical artifacts. J Bus Res. 2014;67:1766–70.

23. Kostopoulou O, Nurek M, Cantarella S, et al. Referral Decision Making of General Practitioners: A Signal Detection Study. Med Decis Making. 2019;39:21–31.

24. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42:377–81.

25. Harris PA, Taylor R, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform. 2019;95:103208.

26. LimeSurvey: An open source survey tool. Version 5.3.31. Hamburg, Germany: LimeSurvey GmbH; 2022. Available from: https://www.limesurvey.org/

27. Do TP, Remmers A, Schytz HW, et al. Red and orange flags for secondary headaches in clinical practice: SNNOOP10 list. Neurology. 2019;92:134–44.

28. Headache Classification Committee of the International Headache Society (IHS). The International Classification of Headache Disorders, 3rd edition. Cephalalgia. 2018;38:1–211. doi: 10.1177/0333102417738202

29. Mawet J, Roos C, Ducros A. Démarche diagnostique devant une céphalée aigue. In: Traité de médecine. Guillevin L., Mouthon L., Lévesque H. (Eds). 2018.

30. Moisset X, Mawet J, Guegan-Massardier E, et al. French Guidelines For the Emergency Management of Headaches. Rev Neurol (Paris). 2016;172:350–60.

31. Moisset X, Mawet J, Guegan-Massardier E, et al. Recommandations pour la prise en charge d'une céphalée en urgence. Douleurs Éval-Diagn-Trait. 2018;19:4–16.

32. Cutrer FM, Wippold II FJ, Edlow JA. Evaluation of the adult with nontraumatic headache in the emergency department. UpToDate, Post TW (Ed). Available at: https://www.uptodate.com/contents/evaluation-of-the-adult-with-nontraumatic-headache-in-the-emergency-department . Accessed April 22, 2022

33. Detsky ME, McDonald DR, Baerlocher MO, et al. Does This Patient With Headache Have a Migraine or Need Neuroimaging? JAMA. 2006;296:1274.

34. Hainer BL, Matheson EM. Approach to Acute Headache in Adults. Am Fam Physician. 2013;87:682–7.

35. Locker TE, Thompson C, Rylance J, et al. The Utility of Clinical Features in Patients Presenting With Nontraumatic Headache: An Investigation of Adult Patients Attending an Emergency Department. Headache J Head Face Pain. 2006;46:954–61.

36. Olesen J. International classification of headache disorders. Lancet Neurol. 2018;17:396–7.

37. Wootton RJ, Wippold F, Whealy M. Evaluation of headache in adults. UpToDate Com. 2021;

38. Juslin P, Winman A, Olsson H. Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. Psychol Rev. 2000;107:384.

39. Fleming SM. HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. Neurosci Conscious, 2017;1-4.

40. Rouault M, Will GJ, Fleming SM, et al. Low self-esteem and the formation of global self-performance estimates in emerging adulthood. Transl Psychiatry. 2022;12:1–10.

41. R Core Team. R: A language and environment for statistical computing. Version 4.2.2. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/

42. Lee ALF, Ruby E, Giles N, et al. Cross-Domain Association in Metacognitive Efficiency Depends on First-Order Task Types. Front Psychol. 2018;9.

43. Silveira SQ, da Silva LM, Gomes RF, et al. An evaluation of the accuracy and self-reported confidence of clinicians in using the ASA-PS Classification System. J Clin Anesth. 2022;79:110794.

44. Cheung T, Harianto H, Spanger M, et al. Low accuracy and confidence in chest radiograph interpretation amongst junior doctors and medical students. Intern Med J. 2018;48:864-8

45. Koriat A, Lichtenstein S, Fischhoff B. Reasons for confidence. J Exp Psychol. 1980;6:107-118

46. Martin JM. Confirmation bias in the therapy session: The effects of expertise, external validity, instruction set, confidence and diagnostic accuracy. The University of Memphis; 2000.

47. Miller DJ, Spengler ES, Spengler PM. A meta-analysis of confidence and judgment accuracy in clinical decision making. J Couns Psychol. 2015;62:553–67.

48. Owen J. The nature of confirmatory strategies in the initial assessment process. J Ment Health Couns. 2008;30:362–74.

49. Strohmer DC, Shivy VA, Chiodo AL. Information processing strategies in counselor hypothesis testing: The role of selective memory and expectancy. J Couns Psychol. 1990;37:465.

50. Johnson DDP, Fowler JH. The evolution of overconfidence. Nature. 2011 Sep;477:317–20.

51. Eva KW, Norman GR. Heuristics and biases − a biased perspective on clinical reasoning. Med Educ. 2005;39:870–2.

52. Graber M. Metacognitive training to reduce diagnostic errors: ready for prime time? Acad Med. 2003;78:781.

53. Norman E. Why Metacognition Is Not Always Helpful. Front Psychol. 2020;11.

54. Tenney ER, Small JE, Kondrad RL, et al. Accuracy, confidence, and calibration: how young children and adults assess credibility. Dev Psychol. 2011;47:1065.

55. Zwaan L, Hautz WE. Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. BMJ Qual Saf. 2019;28:352–5.

56. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 2: impediments to and strategies for change. BMJ Qual Saf. 2013;22:ii65–72.

57. O'Sullivan ED, Schofield SJ. A cognitive forcing tool to mitigate cognitive bias–a randomised control trial. BMC Med Educ. 2019;19:1-8.

58. Royce CS, Hayes MM, Schwartzstein RM. Teaching critical thinking: a case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. Acad Med. 2019;94:187–94.

59. Stark M, Fins JJ. The ethical imperative to think about thinking: Diagnostics, metacognition, and medical professionalism. Camb Q Healthc Ethics. 2014;23:386–96.

60. Reilly JB, Ogdie AR, Von Feldt JM, et al. Teaching about how doctors think: a longitudinal curriculum in cognitive bias and diagnostic error for residents. BMJ Qual Saf. 2013;22:1044–50.

61. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Med Educ. 2008;42:468–75.

62. Mamede S, Schmidt HG, Rikers RM, et al. Conscious thought beats deliberation without attention in diagnostic decision-making: at least when you are an expert. Psychol Res. 2010;74:586–92.

63. Chew KS, Durning SJ, Van Merriënboer JJ. Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study. Singapore Med J. 2016;57:694.

64. Chew KS, van Merrienboer JJ, Durning SJ. Perception of the usability and implementation of a metacognitive mnemonic to check cognitive errors in clinical setting. BMC Med Educ. 2019;19:18.

65. Schraw G, Roedel TD. Test difficulty and judgment bias. Mem Cognit. 1994;22:63–9.

66. Nagendran M, Chen Y. Real-time confidence of clinical decision making: a systematic review. Future Heal J. 2019;6:82–82.

67. Mohan D, Fischhoff B, Farris C, et al. Validating a Vignette-Based Instrument to Study Physician Decision Making in Trauma Triage. Med Decis Making. 2014;34:242–52.

68. Lejarraga T, Hertwig R. How experimental methods shaped views on human competence and rationality. Psychol Bull. 2021;147:535–64.

69. Double KS, Birney DP. Reactivity to measures of metacognition. Front Psychol. 2019;10:2755.

70. Double KS, Birney DP. Do confidence ratings prime confidence? Psychon Bull Rev. 2019;26:1035–42.

71. Mitchum AL, Kelley CM, Fox MC. When asking the question changes the ultimate answer: Metamemory judgments change memory. J Exp Psychol Gen. 2016;145:200.

72. Rouault M, Lebreton M, Pessiglione M. A shared brain system forming confidence judgment across cognitive domains. Cereb Cortex. 2023;33:1426–39.

73. van Den Berg R, Anandalingam K, Zylberberg A, et al. A common mechanism underlies changes of mind about decisions and confidence. Elife. 2016;5:e12192.

74. Webb TW, Miyoshi K, So TY, et al. Natural statistics support a rational account of confidence biases. Nat Commun. 2023;14:3992

75. Zylberberg A, Fetsch CR, Shadlen MN. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. Elife. 2016;5:e17688.

76. Jansen RA, Rafferty AN, Griffiths TL. A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. Nat Hum Behav. 2021;5:756–63.

77. Kao YC, Davis ES, Gabrieli JD. Neural correlates of actual and predicted memory formation. Nat Neurosci. 2005;8:1776–83.

78. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol. 1999;77:1121.

79. Kunimoto C, Miller J, Pashler H. Confidence and accuracy of near-threshold discrimination responses. Conscious Cogn. 2001;10:294–340.

80. Modirrousta M, Fellows LK. Medial prefrontal cortex plays a critical and selective role in 'feeling of knowing'meta-memory judgments. Neuropsychologia. 2008;46:2958–65.

81. Morgan MJ, Mason AJS, Solomon JA. Blindsight in normal subjects. Nature. 1997;385:401–2.

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

The funding for this work was provided by the Haute Autorité de Santé (CL's Ph.D. grant), INSERM and the "Investissements d'avenir" program (ANR-10-I AIHU-06) to Paris Brain Institute - ICM (BR's research annual grant). CL is a student from the FIRE Ph.D. program funded by the Bettencourt Schueller Foundation and the EURIP Graduate Program (ANR-17-EURE-0012). MR work has been supported by La Fondation des Treilles and by a postdoctoral fellowship from the AXA Research Fund.

AUTHOR CONTRIBUTIONS

MK, RK, CL, MR and BR collaborated to design the experimental paradigm. RK, NB and CL elaborated the clinical case-vignettes with the help of BR, JM and CR, and conceived the survey on LimeSurvey and RedCap with the support of the PRISME platform (Paris Brain Institute - ICM). CL, BR, JM and CR worked at the data acquisition by ensuring the survey distribution. CL, F-XL and BR analyzed the collected data and all authors collectively interpreted them. CL and MR drafted the article manuscript. BR and MK revised it. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

We are grateful to colleagues from the Paris Brain Institute (alphabetic order: Amina Ben Salah, Laurent Cohen, Fabien Hauw, Céline Louapre, Esteban Muñoz-Musat, Lionel Naccache, Aude Sangare), the Haute Autorité de Santé (Philippe Cabarrot, Marie Coniel, Emmanuel Nouyrigat, the Service des Bonnes Pratiques department from DAQSS), and other colleagues (Nicolas Dritsch, Yonathan Freund, Hélène Goullet, Yoann Launey, Claire Morgand, Anne-Caroline Papeix, Raphaël Veil and Youri Yordanov) for their help in survey distribution among physicians from our targeted medical specialties. We acknowledge Karim N'Diaye at the Paris Brain Institute PRISME Core Facility (RRID:SCR_026394) and Mathias Antunes from the Data Analysis Core (RRID:SCR_026138) for their support in establishing the survey, as well as Baptiste Crinière-Boizet from the Data Analysis Core for his contribution in performing control tests on the vignette set. Finally, we thank the 52 participants that contributed to our study. At last, we warmly thank the local, national and international teams of French Three Minute Thesis (Ma Thèse en 180 secondes) for providing opportunities to help disseminating this work worldwide and awarding CL presentation with the jury's first prizes (youtube.com/watch?v=n9UNLc5bYmw).