**Abstract**

**Introduction**

When doctors are presented with a patient, they must integrate the information they have so far, align their mental models of the patient with other doctors and decide the following:

1. Do they have enough information to make a determination of the patient’s condition?
2. If not, what extra information do they need? Are there further tests that need to be performed?
3. As per their most likely diagnosis, what actions should they start taking to treat the patient?

Generally, doctors gather information through tests, patient documentation and physical examinations to generate a model of the patient’s condition, through which they can formulate a hypothesis for what could be the underlying condition of a patient. Based on this hypothesis, the doctor can then prescribe the most suitable treatment.

A report from the US Institute of Medicine (McGlynn, McDonald & Cassel, 2015) concluded that most patients will experience a diagnostic error within their lifetime. Bhise et al (2017) found that as much as 55.5% of patients experienced diagnostic error. Diagnostic errors have also been found to lead to longer hospital stays and even increased patient mortality (Hautz et al, 2019). The Quality in Australian Health Care Study found that 20% of adverse events were due to delayed diagnosis (Wilson et al, 1999). Crucially for this research, 32% of clinical errors have been found to be caused by the clinician’s failure to weigh up competing diagnoses (Schiff et al, 2009).

Individual diagnostic errors are by no means the sole causes of medical incidents. There are a number of factors tied to the wider work environment, culture and technology that can contribute to incidents and errors. A lot of these factors are challenging to isolate and emulate in an experimental setting. One could intuit however that an error in diagnosis can have knock-on effects later on in the medical timeline. A misdiagnosis increases the likelihood of inappropriate treatment, which in turn increases the likelihood of an adverse patient event. Gaining a greater understanding of diagnosis as a cognitive process can have important implications for future interventions within healthcare settings.

One account of diagnostic errors is that they can stem from cognitive biases during decision making. For example, a clinician may have recently experienced a patient with a particular condition and, upon seeing another patient with what are perceived to be similar symptoms, is then more likely to choose the same diagnosis again. While it seems intuitive that classical decision making biases affect those in healthcare too (Restrepo et al, 2020), the empirical evidence is scant, particularly when showing that these biases contribute to medical errors (van den Berge & Mamede, 2013). One example of past literature looking at classical decision biases attempted to automatically detect uses of heuristics and biases by dermatologists, with examples of satisficing bias (premature closure) and anchoring were found, but with very few examples of others such as availability and representative biases found (Crowley et al, 2012).

This makes confidence an interesting area of study. Confidence is viewed within the cognitive psychology literature as one’s subjective probability of their own decisions being correct (Fleming & Daw, 2017). Confident individuals also tend to be more influential on others in a group (Zarnoth & Sniezek, 1997) and can even causally increase the confidence of other observers (Cheng et al, 2021). This behaviour has been observed in mock jury trials, during which participants hear eyewitness testimonies presented with high confidence and then perceive as those testimonies as more credible than testimonies provided with low confidence (Cutler, Penrod & Dexter, 1989, Roediger, Wixted & DeSoto, 2012). Confidence is a commonly used predictor of another person’s accuracy, especially when feedback is not readily available of an individual’s true accuracy. Confidence also varies across individuals with what may be considered a ‘subjective fingerprint’ (Ais et al, 2016), such as if individuals are systematically underconfident or overconfident. Confidence has been explained computationally as the difference in the strength of evidence for a decision alternative compared to other alternatives (Vickers & Packer, 1982). After a decision is made, we continue to process evidence, meaning that we continue to think about a decision after the decision is made. This means that having ‘second thoughts’ or changes of mind are more likely with a lower level of confidence.

One is said to be well-calibrated with regards to their confidence if their internal likelihood of being correct is predictive of their true accuracy. However, confidence can become decoupled from the true accuracy of one’s decisions. This decoupling is known as ‘miscalibration’. One would show miscalibration of confidence if they were confident when incorrect (overconfidence) or uncertain when they are correct (underconfidence).

In a task that involved diagnosing ultrasound scans, it was found that overconfidence was negatively associated with the amount of clinical experience that the clinicians/participants had (Schoenherr, Waechter & Millington, 2018). However, it has also been found that underconfidence can be more prevalent than overconfidence, especially when comparing medical students to residents (Friedman et al, 2005). Similarly, Yang and Thompson (2010) found that experienced nurses exhibited similar performance to nursing students, but were more confident in their judgements, showing differences in confidence calibration across experience levels. More broadly, highly confident members within a group could unknowingly reduce the chance of less confident members speaking up about potential errors, which is a common problem within healthcare (Hémon et al, 2020). Overconfidence has also been linked to a lower likelihood of sufficient patient management and clinical effort as per a field study in Senegal (Kovacs, Lagarde & Cairns, 2019).

We would argue that mapping and building on the current research landscape of diagnostic confidence is important. There may be a tacit assumption that others will be metacognitively aware and calibrate their confidence with their true accuracy, meaning that heeding high confidence advice or judgements would be an optimal strategy for maximising accuracy. However, this can be a serious issue when high confidence errors lead others astray. Put simply, without objective feedback (a doctor may never know for certain if they diagnosed a patient correctly), a clinician’s confidence is one of the only markers available for other clinicians and for patients when making key medical decisions. Seniority or speciality experience may also be markers, but they may not suffice as they are not related to the specific decision being made.

**Linking Confidence and Information Seeking**

We can infer that the relationship between confidence and information seeking could have wide-reaching consequences within healthcare. In other words, seeking too much information can lead to unnecessary wastage of time and resources within the healthcare system, whilst too little information can lead to overcommitting to certain diagnoses too early, increasing the likelihood of diagnostic error.

Medical decisions have been thought of as ‘ideal’ when using the hypothetico-deductive process (Kuipers & Kassirer, 1984), whereby hypotheses are formulated based on specific features of a patient and are then linked to established criteria for a diagnosis, with further information gathering to test these hypotheses (Higgs et al, 2008).

There are interesting questions here around how clinicians generate hypotheses and then gather information to reduce the space of hypotheses. One should ideally eliminate hypotheses from consideration only when it makes sense given the incoming evidence. By the same token, they should also not continue attaching themselves to a hypothesis when there is overwhelming evidence to the contrary. One conclusion of Wason (1960) was that individuals struggle to remove a hypothesis from consideration even if they receive evidence against it. Understanding how individuals generally reason about a possible space of hypotheses is interesting for understanding how the reasoning process works differentially for novices and experts, especially in a specialised domain such as medicine. One question that is worth investigating is how the ‘process of elimination’ affects confidence.

A question here is whether hypothesis elimination is related to information seeking patterns and subjective confidence. Looking at information seeking and its relationship with hypothesis generation and elimination could help our understanding of confidence within medical decision making.

The link between confidence and information seeking has been previously investigated in cognitive psychology research. Desender, Boldt & Yeung (2018) manipulated the variance of a visual stimulus and found that higher variability was associated with lower confidence and higher information seeking. Information can be gathered that is either in support of or against an individual’s beliefs or decisions, with information being used to accumulate strength of evidence in favour of different decision alternatives (Vickers & Packer, 1982). However, the mere quantity of information, even if that information favours the non-preferred option, may increase confidence in of itself (Ko, Feuerriegel, et al, 2022).

However, this relationship may depend on one’s metacognitive awareness, or how closely one’s confidence relates to their objective accuracy. For example, pathologists with better metacognitive awareness were found to request more information, such as second opinions or ancillary tests, when unconfident in their judgements (Clayton et al, 2022). In a sample of 118 physicians presented with patient vignettes, it was found that higher confidence (as well as a higher difficulty) was associated with a decreased amount of diagnostic tests being ordered, even if confidence and accuracy were larger decoupled/miscalibrated (Meyer et al, 2013). When considering the decision process past the mere request of information, it has also been observed previously that physicians may ‘distort’ neutral or inconclusive evidence to be interpreted as supporting prior beliefs (Kostopolou et al, 2012). Similarly, it has been found that a patient’s case history that suggests a particular diagnosis prompts selective processing of clinical features that favour the initial diagnosis (Leblanc, Brooks & Norman, 2002). Together, these findings indicate interesting implications for how clinicians may seek and integrate evidence when making decisions and how patterns of receiving information could affect decision confidence and in turn confidence calibration.

There is a need for the teaching and assessment of non-technical skills and human factors in healthcare (Higham et al, 2019), which is currently not addressed in a widespread standardised manner in speciality curricula (Grieg, Higham & Vaux, 2015). Curricula within medicine also place little emphasis on how uncertainty is communicated and approached in medical decision making (Hall, 2002). Hence, this research informs medical education of non-technical skills such as management and communication of uncertainty.

We hypothesised that differences in diagnostic ability would be reflected in information seeking patterns and confidence, and diagnostic ability is related to how medical students weight up competing differentials/hypotheses.

**Methods**

**Participants**

The study was conducted online, with participants able to run the experiment in their browser. We recruited fifth or sixth (foundation) year medical students within the UK. The UK Medical Schools Council distributed the study to UK medical students using a mailing list. Participants were emailed with a study information sheet and a link to access the experiment, where they first provided consent via an anonymous online form. After doing so, the participant provided demographic information (age, gender and years of medical experience). The age ranged between 22-34 (M = 24.2). 85 medical students completed the study, including 32 males, 52 females and 1 participant who self-reported as non-binary. Participants were recruited between July 11th 2022 and April 6th 2023. We also recruited 7 experienced clinicians to complete the study.

**Materials**

This study involved the usage of patient vignettes, which have been adapted from actual past cases. We adapted scenarios from a bank of patient cases from Friedman (2004). These vignettes were developed by a team of researchers based in the US, meaning that certain medical terms (eg medication names, tests etc) had to be ‘translated’ into the vernacular used by doctors based in the UK.

Our study involved 6 patient cases, each with a true underlying condition. These conditions were: Aortic Dissection (AD), Guillain-Barre Syndrome (GBS), Miliary TB (MTB), Temporal Arteritis (TA), Thrombotic Thrombocytopenic Purpura (TTP) and Ulcerative Colitis (UC). The order in which the cases were presented was randomised for each participant. We also included a practice case (Colon Cancer) to familiarise the participants with the experimental procedure and the interface.

The procedure of a single case is as follows. The participant is asked to imagine that they are working in a busy district hospital and they encounter patients in a similar way to how they would in their real medical practice. At the start of each case, the participant is shown a description of a patient, which includes the patient’s gender, age and their presenting complaint (e.g. “patient is a 68 year old male presenting with fever and arthralgia”). Each case is split into three information stages: Patient History, Physical Examination and Testing.

At each stage, the participant sees pieces of information or tests that they can request, with results provided after a short delay. The set of information requests for each stage is the same for all cases. 29 possible tests can be requested across the three information stages. It was also emphasised during the task instructions that participants should only request information that they believe will help them with diagnosing the patient for that specific case.

At any point, the participant can choose to stop gathering information for that stage. They then report a list of all differential diagnoses that they are considering for that patient at that stage. For each differential, participants report a “level of concern” on a 4 point scale, which we describe as how concerned the participants would be for that patient if this differential really was the patient’s underlying condition. Participants also reported a likelihood rating for each differential on a 10 point scale, ranging from 1 (very unlikely) to 10 (certain). Participants update concern/likelihood ratings and add/remove differentials from the list at each stage.

After recording their differentials, participants are then asked to report their confidence that they are “ready to start treating the patient” on a 100 point scale, ranging from fully unconfident to fully confident. Participants are also able to indicate using a checkbox that they are ready to start treating the patient, at which a text box appears for them to report what further tests they would perform, any escalations they would make to other medical staff and treatments they would start administering for the patient. At the end of all six patient cases, participants are told the true underlying conditions for all the patients.

**Outcome Measures**

• **Confidence:** Initial Confidence refers to the reported confidence after the first stage of information seeking (Patient History), whilst Final Confidence refers to the reported confidence after the third/last stage of information seeking (Testing). We can then use these two variables to calculate Confidence Change, by subtracting the participants' Initial Confidence from their Final Confidence. A positive value for Confidence Change means that the participant has gained confidence over the course of the patient case.

• **Number of Differentials:** we record the number of items in the list of differentials at each stage. Initial Differentials refer to the number of differentials after the first stage of information seeking (Patient History), whilst Final Differentials refer to the number of differentials after the third and last stage of information seeking (Testing).

• **Accuracy:** Responses were coded for correctness manually with help from a medical consultant. Our accuracy measure is then computed by taking the likelihood value assigned to a correct differential if it is included in the reported list. Accuracy hence ranges from 1-10 when a correct differential is included and has a value of 0 when a correct differential is not included.

• **Proportion of Information Seeking:** we take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage or including tests that had been requested before during that stage) and divide this by the number of possible tests available.

• **Information Seeking Variance:** We compute binary vectors of length 29, where for each piece of information available for a case, a value of 1 is assigned if that information is requested and 0 is assigned if that information is not requested during the case. We calculate the Dice dissimilarity coefficient between each case. We then compute the mean distance value across cases. A lower mean value indicates that participants sought similar information across the cases whilst a higher value indicates that information seeking is varied more by case.

• **Information Seeking Value:** We use the information seeking of the experienced clinicians as a reference to compare the student participants against. We take each of the 29 pieces of information in turn by case and split trials into two groups: trials of that case where that information was sought and trials of that case where that information was not sought by experienced clinicians. For each group, we compute the proportion of trials where the experienced clinicians included a correct differential, and then take the difference between these two values. A positive value would indicate that participants were more likely to identify the correct condition with that information than without it. This difference is considered the information’s ‘value’. We then calculate the mean ‘value’ across all information that each student sought. This gives an overall measure of how useful the information was that students sought.

**Results**

**Overall Performance**

We first look at the number of differentials that participants report at each stage. Participants overall increased the number of the differentials they reported as they received more information (F(1, 107) = 94.02, η2G = .08, p < .001). Participants reported fewer differentials during the Patient History stage (M = 3.20, SD = 1.11) than during the Physical Examination (M = 3.88, SD = 1.33) and Testing stages (M = 4.12, SD = 1.43). We especially note that out of 85 participants, 74 did not decrease their number of differentials between Patient History and Testing on any case. Confidence also increased as participants received more information (F(1, 123) = 75.45, η2G = .15, p < .001). Participants reported lower confidence during the Patient History stage (M = 0.30, SD = 0.15) than during the Physical Examination (M = 0.41, SD = 0.17) and Testing stages (M = 0.47, SD = 0.19). The Proportion of Information Seeking decreased with each information stage (F(2, 151) = 122.0, η2G = .30, p < .001). Participants sought more of the available information during the Patient History stage (M = 0.85, SD = 0.20) than during both during the Physical Examination (M = 0.59, SD = 0.24) and Testing stages (M = 0.50, SD = 0.22).

We now report the proportion of trials where participants include a correct differential within their set of differentials (henceforth referred to as Differential Accuracy). Participants increased their Differential Accuracy overall with more information (F(2, 128) = 59.52, η2G = .08, p < .001). Participants had lower Differential Accuracy during the Patient History stage (M = 0.54, SD = 0.23) than during the Physical Examination (M = 0.66, SD = 0.22) and Testing stages (M = 0.69, SD = 0.21).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Case** | **Differential Accuracy** | **Accuracy** | **Perceived Difficulty** | **Final Confidence** |
| UC | 0.92 | 7.32 | 5.30 | 61.08 |
| GBS | 0.69 | 5.44 | 6.88 | 37.03 |
| TA | 0.66 | 6.68 | 6.16 | 48.96 |
| TTP | 0.55 | 5.50 | 6.80 | 40.86 |
| AD | 0.53 | 4.67 | 5.92 | 48.87 |
| MTB | 0.42 | 5.67 | 6.69 | 44.91 |

A chart of different levels of performance

Description automatically generated

We find that confidence is fairly well calibrated to objective performance except for a slight deviation between the two during the Testing stage (t(84) = 2.40, MDiff = 0.056, p = .02). This could imply that past findings on miscalibration of confidence are dependent on how objective performance is operationalised.

**Information Seeking**

Using information seeking data, we sought to identify if differences in information seeking patterns were predictive of differences in diagnostic accuracy on this task. To investigate the differences in information seeking, we first trained a binary classification algorithm using a generalised logistic regression model to predict accuracy. We aimed to identify if participants of different diagnostic ability on this task exhibited differences in information seeking. To do this, we first split all trials into high and low ability participant trials with a median split of participants by their average accuracy across the six cases. We train the classifier by treating the 29 binary variables for each information as predictors (with a 1 signifying that the information was sought for that case and 0 when the information was not sought) to predict the binary outcome of whether the participant is a low or high accuracy participant. We used Leave One Out Cross Validation, such that each trial is predicted by training the algorithm on all other trials. By plotting an ROC curve of our classifier, we find an area under the curve (AUC) value of 0.71 (with p < .001 when comparing the ROC curve to AUC = 0.5). We also use a Classification Tree Algorithm to classify trials by participant accuracy in the same way and find an ROC AUC of 0.64. This indicates overall that differences in information seeking are indeed predictive of a difference in participant ability. Next, we identify and better characterise the specific differences in information seeking that contribute to diagnostic ability.

A graph of a curve

Description automatically generated

When conducting a Pearson’s Correlation test, we found evidence for a positive correlation between the change in confidence (the difference in confidence in the first and last stages) and the proportion of information sought (r(83) = 0.24, p = .03), such that seeking more information was associated with higher gains in confidence. We also find an association between the number of differentials generated from the Patient History and the amount of information sought during a given case (r(83) = 0.30, p = .005). One can surmise here that higher information seeking is associated with the consideration of more diagnostic differentials.

We then look at how ability on the task relates to information seeking behaviour. We find marginal evidence for a negative correlation between overall accuracy (across cases) and information seeking variance (r(83) = -0.21, p = .06). We find that students with higher overall accuracy exhibitied a lower variance in information seeking. In other words, students with a higher diagnostic ability are found to have varied the information they sought across cases less. While not strong evidence, this indicates a broad pattern that being more standardised in information seeking is associated with higher diagnostic accuracy.

We apply a similar analysis to look at how information value varies as a function of participant ability. We find evidence for a positive relationship between accuracy and information value (r(83) = 0.25, p = .02), but not between confidence and information value (r(83) = 0.15, p = .18). We also find that the proportion of available information sought is not shown to correlate with accuracy on at the final stage (r(83) = 0.17, p = .11) but does correlate with the participants’ change in confidence, which is the difference in confidence between the first and final stages (r(83) = 0.24, p = .03). While seeking more information may imbue students with a greater level of confidence, it does not necessarily translate into more accurate diagnoses. This is important to note, it indicates that being selective in information seeking is a good marker of performance and having all available information does not necessarily translate into accurate diagnoses. In addition, students with a higher diagnostic ability sought better information but also approach each case in a more similar manner.

This has interesting implications for medical practice, as the ordering of unneeded tests or patient examinations may not contribute to better decisions. Given the constraints within most hospitals and healthcare to obtain certain tests, being selective with information seeking is already a frequent necessity and results from this study seem to show evidence that it is also a good marker of diagnostic performance.

**Discussion**

There are a few limitations with our study. We did not use more naturalistic stimuli, such as images of scans/test results or audio cues (such as the sound of lung auscultation) and instead used solely textual results for all tests. While this may make the experiment more ecologically valid, it takes away the interpretation of complex stimuli which could affect information seeking. For example, if two students saw a chest X-ray, they may interpret the scan in different ways, adding a potential confound to this study. That is why if a student requested a chest X-ray, they instead see a result like “no abnormalities found”, such that the interpretation of the image is constant for all students.

Our experiment also assumed that all tests were equal in terms of how long they take for results to be shown. If the tests were analogous to real medical practice, certain tests would take longer to produce results after being requested. We should also note that our experiment was run via an internet browser, meaning that study participants were taken out of the setting within which they would usually make these decisions. This means that participants may act differently than they might do in their regular medical practice.

Within this discussion, it is worth mentioning a few general observations in the data. Firstly, participants did not tend to use the ability to remove differentials from their list. This may indicate an attachment to hypotheses and an unwillingness to remove them from consideration. It is likely that students are taught not to disregard diagnostic differentials easily.

Another aspect of note is the manner in which participants reported their differentials. Given that differentials were provided via free text, there is a lot of freedom in the diagnostic differentials that participants can report. What this can mean however is there are differences in the specificity of differentials provided. For example, one participant may report “lymphoma” as a differential whilst another may report “Hodgkin’s Lymphoma”, “Non-Hodgkin’s Lymphoma” and “Chronic Lymphocytic Leukaemia” within the list (all of which are different types of lymphoma). Both participants essentially capture the same ‘differential’ but do so in different manners. When looking at the number of differentials however, the former produces one differential whilst the latter produces three. This example illustrates that participants differ in how specific they are when reporting their differentials and how this affects our ability to analyse the number of differentials that participants report.

Previous research that allows participants to report multiple diagnostic differentials would consider a trial correct if the correct condition is within the list of differentials. However, accuracy would be conflated with the number of differentials provided. If a participant simply continues to add more differentials, they are more likely to be correct. Therefore, we calculated Accuracy using a more fine-grained measure that takes into account likelihood ratings and how students weigh up differentials against each other.

**Participants**

Due to the richness of qualitative data, we recruited a smaller sample of participants. In total, 16 participants were recruited for this study. Participants had to be 5th or 6th year medical students at Oxford university in order to take part. Participants were recruited using posters in John Radcliffe Hospital in Oxford and via a mailing list for students managed by the Medical Sciences department at the University of Oxford. The study was conducted onsite at John Radcliffe hospital. Participants were recruited between July 5th 2023 and December 1st 2023.

**Materials**

The same set of cases and interface from Study 1 was also used for this study. The study was conducted onsite using a laptop, with actions on screen recorded on video and the audio of participants’ thinking aloud recorded via a microphone.

**Procedure**

The general procedure was very similar to that of study 1, whereby participants were shown six patient scenarios and were tasked with diagnosing the patient. However, we removed the screen where participants record their list of differentials. Instead, the experiment was run in-person so that participants could think aloud as they were doing the task. Participants were given the following instructions at the start of the study:

“Whilst you are doing the task, you will be asked to think aloud. This means that you verbalise what you are thinking about, especially how you interpret the information you receive and what conditions or diagnoses you are considering or are concerned about for each patient case. If you have nothing to say or nothing on your mind, there’s no need to say anything but do say whatever is on your mind once it pops up. If you are unsure about anything you see or do not know about what something means, you will not receive any help but verbalise when you are unsure about anything during the task. Please make sure that you speak clearly ‘to the room’.”

The researcher in the room was to remain mostly silent, aside from asking the participant “can you tell me what you are thinking?” if there is a period of long silence and asking the participant “can you tell me more?” if the participant says something vague that may warrant further detail. This was so that any utterances by participants were not swayed or prompted to say certain things by the researcher. The audio of the participants’ verbalisations was recorded and then transcribed.

At the end of the experiment, the researcher administered a semi-structured interview to better understand what the participants feel their diagnostic reasoning approach tends to be. The main questions are indicated below, each with a corresponding follow-up question in case they are not answered by responses to the main questions:

Data Analysis

Once the audio recordings had been transcribed, the transcriptions for specific categories of utterances. Hence, this study adopted a theory-driven semantic thematic analysis (as per definitions detailed by Braun and Clarke, 2006), given that we use pre-existing theory to produce codes a priori to be applied to our data. This kind of thematic analysis is also suitable given that our qualitative data is from a structured experiment, rather than a dataset with a looser structure (e.g. interview recordings).

Firstly, we code all utterances related to the main research areas of interest in this project, namely information seeking, confidence and differential/hypothesis generation. Respectively, we code utterances for whenever students:

* Mention a new condition that they are considering
* Rule out or eliminate a condition from consideration
* Mention an increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition
* Mention a decreased likelihood of a previously mentioned condition, or that information seems to contradict with a condition

We also define a group of codes that indicate reasoning strategies, which pertains to how the participants generate and considered differentials. These codes are based around the work of Coderre et al (2003), which had also used a think-aloud protocol for a diagnostic reasoning task. Their paper defines three different diagnostic reasoning strategies: hypothetico-deductive reasoning, scheme-inductive reasoning and pattern recognition. Adapted from their definitions, we define each as follows:

* **Hypothetico-Deductive Reasoning** - prior to selecting the most likely diagnosis, the participant analysed any alternative differentials one by one through something akin to a process of elimination.
* **Scheme Inductive Reasoning** - participant structures their diagnosis by anatomical systems or categories of conditions (e.g infective vs cardiovascular causes) to determine root causes of patient symptoms rather than focusing on specific conditions.
* **Pattern Recognition** - participant considers only a single diagnosis with only perfunctory attention to the alternatives, or makes reference to pattern matching when using a prototypical condition to match its symptoms against the current observed symptoms for the patient (e.g “these symptoms sound like X” or “this fits with a picture of Y”).

In addition to coding each case under one of these strategies, we also code participants on an overall level based on their subjective perception of how they make diagnostic decisions. This is based on responses provided during the debrief interview (as described in the Procedure section). Hence, reasoning strategy codes are at the case level and also at the participant level. Finally, we apply thematic analysis to the responses from the debrief interview to find other themes that seem to recur in the data.

**Results**

**Discussion**

**Overall Discussion**

**References**

1. Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. Cognition, 146, 377-386.
2. Allwood, C. M. (1984). Error detection processes in statistical problem solving. Cognitive science, 8(4), 413-437.
3. Audley, R. J. (1960). A stochastic model for individual choice behavior. Psychological review, 67(1), 1.
4. Baker, G. R., Norton, P. G., Flintoft, V., Blais, R., Brown, A., Cox, J., ... & Tamblyn, R. (2004). The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. Cmaj, 170(11), 1678-1686.
5. Bang, D., & Frith, C. D. (2017). Making better decisions in groups. Royal Society open science, 4(8), 170193.
6. Barnett, M. L., Boddupalli, D., Nundy, S., & Bates, D. W. (2019). Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. JAMA network open, 2(3), e190096-e190096.
7. Berwick, D. M., & Hackbarth, A. D. (2012). Eliminating waste in US health care. Jama, 307(14), 1513-1516.
8. Bhise, V., Meyer, A. N., Singh, H., Wei, L., Russo, E., Al-Mutairi, A., & Murphy, D. R. (2017). Errors in diagnosis of spinal epidural abscesses in the era of electronic health records. The American Journal of Medicine, 130(8), 975-981.
9. Brannon, L. A., & Carson, K. L. (2003). Nursing expertise and information structure influence medical decision making. Applied Nursing Research, 16(4), 287-290.
10. Carlebach, N., & Yeung, N. (2020). Subjective confidence acts as an internal cost-benefit factor when choosing between tasks. Journal of Experimental Psychology: Human Perception and Performance, 46(7), 729.
11. Carpenter, C. R., Raja, A. S., & Brown, M. D. (2015). Overtesting and the downstream consequences of overtreatment: implications of “preventing overdiagnosis” for emergency medicine. Academic Emergency Medicine, 22(12), 1484-1492.
12. Carrigan, A. J., Curby, K. M., Moerel, D., & Rich, A. N. (2019). Exploring the effect of context and expertise on attention: is attention shifted by information in medical images?. Attention, Perception, & Psychophysics, 81, 1283-1296.
13. Chapman, G. B., Bergus, G. R., & Elstein, A. S. (1996). Order of information affects clinical judgment. Journal of Behavioral Decision Making, 9(3), 201-211.
14. Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive psychology, 4(1), 55-81.
15. Cheng, J. T., Anderson, C., Tenney, E. R., Brion, S., Moore, D. A., & Logg, J. M. (2021). The social transmission of overconfidence. Journal of Experimental Psychology: General, 150(1), 157.
16. Chew, K. S., Durning, S. J., & Van Merriënboer, J. J. (2016). Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study. Singapore medical journal, 57(12), 694.
17. Chi, M. T., Glaser, R., & Farr, M. J. (Eds.). (2014). The nature of expertise. Psychology Press.
18. Clark, L., Robbins, T. W., Ersche, K. D., & Sahakian, B. J. (2006). Reflection impulsivity in current and former substance users. Biological psychiatry, 60(5), 515-522.
19. Clayton, D. A., Eguchi, M. M., Kerr, K. F., Miyoshi, K., Brunyé, T. T., Drew, T., ... & Elmore, J. G. (2023). Are Pathologists Self-Aware of Their Diagnostic Accuracy? Metacognition and the Diagnostic Process in Pathology. Medical Decision Making, 43(2), 164-174.
20. Cohen, A. J., Lui, H., Zheng, M., Cheema, B., Patino, G., Kohn, M. A., ... & Breyer, B. N. (2021). Rates of serious surgical errors in California and plans to prevent recurrence. JAMA network open, 4(5), e217058-e217058.
21. Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., ... & Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. Advances in Health Sciences Education, 18, 343-363.
22. Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1989). The eyewitness, the expert psychologist, and the jury. Law and Human Behavior, 13(3), 311-332.
23. Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. Psychological science, 29(5), 761-778.
24. Ellis, M. V., Robbins, E. S., Schult, D., Ladany, N., & Banker, J. (1990). Anchoring errors in clinical judgments: Type I error, adjustment, or mitigation?. Journal of Counseling Psychology, 37(3), 343.
25. Ely, J. W., Graber, M. L., & Croskerry, P. (2011). Checklists to reduce diagnostic errors. Academic Medicine, 86(3), 307-313.
26. Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. Psychological science, 17(4), 311-318.
27. Featherston, R., Downie, L. E., Vogel, A. P., & Galvin, K. L. (2020). Decision making biases in the allied health professions: a systematic scoping review. PLoS One, 15(10), e0240716.
28. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological review, 124(1), 91.
29. Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., ... & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. Journal of General Internal Medicine, 20, 334-339.
30. Frotvedt, T. F., Bondevik, Ø., Seeligmann, V. T., & Sætrevik, B. (2020). Primacy, Congruence and Confidence in Diagnostic Decision-Making.
31. Greig, P. R., Higham, H., & Vaux, E. (2015). Lack of standardisation between specialties for human factors content in postgraduate training: an analysis of specialty curricula in the UK. BMJ quality & safety, 24(9), 558-560.
32. Hall, K. H. (2002). Reviewing intuitive decision‐making and uncertainty: the implications for medical education. Medical education, 36(3), 216-224.
33. Hautz, W. E., Kämmer, J. E., Hautz, S. C., Sauter, T. C., Zwaan, L., Exadaktylos, A. K., ... & Schauber, S. K. (2019). Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. Scandinavian journal of trauma, resuscitation and emergency medicine, 27(1), 1-12.
34. Hémon, B., Michinov, E., Guy, D., Mancheron, P., & Scipion, A. (2020). Speaking up about errors in routine clinical practice: a simulation-based intervention with nursing students. Clinical Simulation in Nursing, 45, 32-41.
35. Higgs, J., Jones, M. A., Loftus, S., & Christensen, N. (2008). Clinical reasoning in the health professions E-book. Elsevier Health Sciences.
36. Higham, H., Greig, P. R., Rutherford, J., Vincent, L., Young, D., & Vincent, C. (2019). Observer-based tools for non-technical skills assessment in simulated and real clinical environments in healthcare: a systematic review. BMJ Quality & Safety, 28(8), 672-686.
37. Hoffman, R. R. (1998). How can expertise be defined? Implications of research from cognitive psychology. In Exploring expertise: Issues and perspectives (pp. 81-100). London: Palgrave Macmillan UK.
38. Jaspan, O., Wysocka, A., Sanchez, C., & Schweitzer, A. D. (2022). Improving the relationship between confidence and competence: implications for diagnostic radiology training from the psychology and medical literature. Academic Radiology, 29(3), 428-438.
39. Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. Educational technology research and development, 45(1), 65-94.
40. Kämmer, J. E., Schauber, S. K., Hautz, S. C., Stroben, F., & Hautz, W. E. (2021). Differential diagnosis checklists reduce diagnostic error differentially: a randomised experiment. Medical education, 55(10), 1172-1182.
41. Katz, J. (1984). Why doctors don't disclose uncertainty. Hastings Center Report, 35-44.
42. Ko, Y. H., Feuerriegel, D., Turner, W., Overhoff, H., Niessen, E., Stahl, J., ... & Bode, S. (2022). Divergent effects of absolute evidence magnitude on decision accuracy and confidence in perceptual judgements. Cognition, 225, 105125.
43. Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. Journal of experimental psychology: general, 139(4), 665.
44. Kostopoulou, O., Rosen, A., Round, T., Wright, E., Douiri, A., & Delaney, B. (2015). Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. British Journal of General Practice, 65(630), e49-e54.
45. Kostopoulou, O., Russo, J. E., Keenan, G., Delaney, B. C., & Douiri, A. (2012). Information distortion in physicians’ diagnostic judgments. Medical Decision Making, 32(6), 831-839.
46. Kovacs, R. J., Lagarde, M., & Cairns, J. (2020). Overconfident health workers provide lower quality healthcare. Journal of Economic Psychology, 76, 102213.
47. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of personality and social psychology, 77(6), 1121.
48. Kuipers, B., & Kassirer, J. P. (1984). Causal reasoning in medicine: analysis of a protocol. Cognitive Science, 8(4), 363-385.
49. Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. Science, 208(4450), 1335-1342.
50. Lawton, R., Robinson, O., Harrison, R., Mason, S., Conner, M., & Wilson, B. (2019). Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England. BMJ Quality & Safety, 28(5), 382-388.
51. Leape, L. L., Brennan, T. A., Laird, N., Lawthers, A. G., Localio, A. R., Barnes, B. A., ... & Hiatt, H. (1991). The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. New England journal of medicine, 324(6), 377-384.
52. LeBlanc, V. R., Brooks, L. R., & Norman, G. R. (2002). Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. Academic Medicine, 77(10), S67-S69.
53. Mann, D. (1993). The Relationship between Diagnostic Accuracy and Confidence in Medical Students.
54. McGlynn, E. A., McDonald, K. M., & Cassel, C. K. (2015). Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine. Jama, 314(23), 2501-2502.
55. Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians’ diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine, 173(21), 1952-1958.
56. Monteiro, S. D., Sherbino, J., Patel, A., Mazzetti, I., Norman, G. R., & Howey, E. (2015). Reflecting on diagnostic errors: taking a second look is not enough. Journal of general internal medicine, 30, 1270-1274.
57. Morgan, P. J., & Cleave‐Hogg, D. (2002). Comparison between medical students' experience, confidence and competence. Medical education, 36(6), 534-539.
58. Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. Decision making in action: Models and methods, 1, 3-20.
59. Oskamp, S. (1965). Overconfidence in case-study judgments. Journal of consulting psychology, 29(3), 261.
60. Patel, V. L., & Medley-Mark, V. (1986). Relationship between representation of textual information and underlying problem representation in medicine. Cognitive Research Series, Montreal: McGill University, Centre for Medical Education.(CME86-CS1).
61. Persky, A. M., & Robinson, J. D. (2017). Moving from novice to expertise and its implications for instruction. American journal of pharmaceutical education, 81(9).
62. Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. Attention, Perception, & Psychophysics, 80, 134-154.
63. Restrepo, D., Armstrong, K. A., & Metlay, J. P. (2020). Annals clinical decision making: avoiding cognitive errors in clinical decision making. Annals of internal medicine, 172(11), 747-751.
64. Roediger, H. L., Wixted, J. T., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. Memory and law, 84.
65. Royce, C. S., Hayes, M. M., & Schwartzstein, R. M. (2019). Teaching critical thinking: a case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. Academic Medicine, 94(2), 187-194.
66. Salem-Schatz, S. R., Avorn, J., & Soumerai, S. B. (1990). Influence of clinical knowledge, organizational context, and practice style on transfusion decision making: Implications for practice change strategies. Jama, 264(4), 476-483.
67. Schiff, G. D., Hasan, O., Kim, S., Abrams, R., Cosby, K., Lambert, B. L., ... & McNutt, R. A. (2009). Diagnostic error in medicine: analysis of 583 physician-reported errors. Archives of internal medicine, 169(20), 1881-1887.
68. Schoenherr, J. R., Waechter, J., & Millington, S. J. (2018). Subjective awareness of ultrasound expertise development: individual experience as a determinant of overconfidence. Advances in Health Sciences Education, 23, 749-765.
69. Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. Trends in cognitive sciences, 18(4), 186-193.
70. Sherbino, J., Kulasegaram, K., Howey, E., & Norman, G. (2014). Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. Canadian Journal of Emergency Medicine, 16(1), 34-40.
71. Sibbald, M., Monteiro, S., Sherbino, J., LoGiudice, A., Friedman, C., & Norman, G. (2022). Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. BMJ Quality & Safety, 31(6), 426-433.
72. St Pierre, M., & Nyce, J. M. (2020). How novice and expert anaesthetists understand expertise in anaesthesia: a qualitative study. BMC medical education, 20(1), 1-11.
73. Tabak, N., Bar-Tal, Y., & Cohen-Mansfield, J. (1996). Clinical decision making of experienced and novice nurses. Western Journal of Nursing Research, 18(5), 534-547.
74. van den Berge, K., & Mamede, S. (2013). Cognitive diagnostic error in internal medicine. European journal of internal medicine, 24(6), 525-529.
75. Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. Acta psychologica, 50(2), 179-197.
76. Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12(3), 129-140.
77. Wei, L. (2021). Confidence and information seeking in decision making under uncertainty (Doctoral dissertation, University of Oxford).
78. Wilson, R. M., Harrison, B. T., Gibberd, R. W., & Hamilton, J. D. (1999). An analysis of the causes of adverse events from the Quality in Australian Health Care Study. Medical Journal of Australia, 170(9), 411-415.
79. Wood, G., Batt, J., Appelboam, A., Harris, A., & Wilson, M. R. (2014). Exploring the impact of expertise, clinical history, and visual search on electrocardiogram interpretation. Medical Decision Making, 34(1), 75-83.
80. Yang, H., & Thompson, C. (2010). Nurses’ risk assessment judgements: A confidence calibration study. Journal of Advanced Nursing, 66(12), 2751-2760.
81. Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. Journal of Experimental Social Psychology, 33(4), 345-366.

**Appendices**

TA Debrief Questions

1. What's your general approach to making diagnoses?

Follow-Up: Do you have those cognitive aids or frameworks you use?

2. Do you tend to keep a broad set of differentials in mind?

Follow-Up: Are there particular situations where having a narrower set would be more useful?

3. How do you decide what information or tests to get on a patient?

Follow-Up: Would you say you tend to seek information to confirm or to rule out differentials that you have in mind?

4. How similar was your diagnostic reasoning on this task versus how you would approach diagnosis in real life?

Follow-Up: Was there anything that prevented you from approaching the task as you would in real life?