**Abstract**

Medical diagnoses are complex decisions that require seeking and integration different sources of information on the part of clinicians. Past work has found that confidence and accuracy in diagnoses can significantly diverge even with more information, leading to diagnostic overconfidence that is especially impacted by increased medical experience. There has been scant work looking at information seeking patterns with diagnoses however, and how they might contribute to both confidence and accuracy. We recruited UK medical students (N = 85) in an online vignette-based diagnostic study, as well as Oxford medical students (N = 16) for an in-person think-aloud version of the same study where students performed 6 diagnostic scenarios based on past patient cases. On each scenario, students sought information about from the patient’s medical history, physical examinations and testing. With the former study, we find that diagnostic accuracy is associated with seeking more relevant information and with varying information seeking less on a case-by-case basis. However, confidence, not accuracy, was associated with seeking more information. With the latter study, we find that different reasoning strategies affect how diagnostic differentials are evaluated, in turn affecting information seeking during diagnostic decisions. Future work should hence focus on interventions for prompting suitable reasoning strategies to improve diagnostic accuracy.

**Introduction**

**Diagnosis**

*“Problems in diagnosis have…been heavily dominated by physicians with little input from the cognitive sciences. What is missing…is foundational work aimed at understanding how clinicians in actual situations take a complex, tangled stream of phenomena…to create an understanding of them as a problem.” (Wears, 2014)*

Imagine a group of doctors within a hospital’s intensive/critical care unit. They are engaged in a collective discussion about a particular patient. The patient has presented with a series of symptoms, including dizziness, breathing difficulties and eventual chest pain. She has been placed under continuous monitoring of her ‘vital signs’, including heart rate, body temperature, blood pressure, blood oxygen saturation and respiration rate. She has been recording a slow decrease in blood pressure and blood oxygen saturation. The doctors are deciding what is the most likely cause of this patient’s symptoms and how this may inform future care/treatment of that patient. It is possible that the patient is suffering from a pulmonary edema, whereby fluid is collected in the air sacs of the lungs, causing severe and sometimes fatal congestion. The symptoms could also be suggestive of a tension pneumothorax, which is when a lung collapses. Alternatively, the patient could be suffering from an anaphylactic shock, which is a severe allergic reaction that can in itself cause fluid to enter the lungs and constrict an individual’s airways. The doctors must integrate the information they have so far, align their individual mental models of the patient 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?

One of the difficulties within this scenario here is that symptoms may be indicative of multiple underlying conditions. This example is illustrative of why many medical decisions are ‘ill-structured’ problems: they present several possible courses of action, and produce disagreements over both the current hypothesis for the patient’s condition and desired end goal for that patient’s care (Jonassen, 1997). Medical staff involved in a patient case can independently formulate very different understandings of a patient’s condition and how it would be best to proceed. They have to then align their thoughts in order to align their actions as a cohesive team.

Individuals involved in clinical decision making have to frequently contend with an uncertain decision making environment (Lawton et al, 2019) as well as time pressure and personal stresses (Orasanu & Connolly, 1993), which add external factors to their decision process to be taken into account.

Clinicians have to make challenging decisions as part of their occupation, such as administration of medication, assessing risk and allocation of hospital space and delegation of responsibilities to colleagues. However, one important group of decisions is diagnosis, which is notable to study for a number of reasons. Firstly, it allows for an extension of previous research on information gathering and confidence within psychology to an ecologically valid, real-world setting. Secondly, diagnosis is an important task that has a large impact on a patient’s road to recovery as clinicians have to accurately understand the severity and nature of a patient’s condition to prescribe appropriate care. Finally, past work looking at diagnosis has received much attention due to the incidence of diagnostic errors.

A report from the US Institute of Medicine (McGlynn, McDonald & Cassel, 2015) concluded that most patients will experience a diagnostic error within their lifetime. When looking at records of new diagnoses for spinal epidural abscess in the US Department of Veteran Affairs, Bhise et al (2017) found that as high as 55.5% of patients experienced diagnostic error. The Quality in Australian Health Care Study found that 20% of adverse events were due to delayed diagnosis (Wilson et al, 1999). Around 32% of clinical errors have been found to be caused by clinician assessment, particularly the clinician’s failure to weigh up competing diagnoses (Schiff et al, 2009). Even when using the most conservative of the above estimates, this illustrates the large scale of the diagnostic error when extrapolated to the population of patients. Diagnostic errors have also been found to lead to longer hospital stays and even increased patient mortality (Hautz et al, 2019).

Diagnostic error is by no means the sole cause 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. By understanding the individual factors of the diagnostic process however, we may better understand how sociotechnical and environmental factors interacts with and amplifies individual contributors to diagnostic error. Gaining a greater understanding of the causes of diagnostic error can have important implications for future interventions within healthcare settings.

One account of diagnostic error is that they can stem from cognitive biases during the diagnostic decision making process, such as primacy (Frotvedt et al, 2020) or recency (Chapman, Bergus & Elstein, 1996) biases. 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). One type of bias that has manifested in more experimental findings is overconfidence.

**Confidence**

At this point, we shall revisit the scenario presented at the start of this section. In summary, a patient is presenting with a set of symptoms that requires doctors to assign a diagnosis in order to guide future treatment. As part of the deliberation around the diagnosis, one of the doctors presents their opinion that the patient has suffered a pneumothorax. When presenting this opinion, they do so with a high level of confidence, meaning that they describe themselves as being nearly certain that their assessment of the patient is the correct one. Due to their high confidence, this doctor’s opinion is difficult for others to disagree with.

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 building on the current research landscape of diagnostic confidence is important. If there is an assumption that others will calibrate their confidence to their true accuracy, this would mean 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. This is important, as in addition to seniority and speciality experience, a clinician’s confidence is one of the only markers available for other clinicians and for patients when making key medical decisions.

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) or eliminate others. We hence sought to extend past work on confidence within diagnosis by looking at its relationship with these two other aspects of the diagnostic process: information seeking and hypothesis (or differential) generation.

**Information Seeking and Differential Generation**

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.

The link between confidence and information seeking has been previously investigated in cognitive psychology research. 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). Desender, Boldt & Yeung (2018) found that higher variability was associated with lower confidence and higher information seeking. 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).

Pathologists with more calibrated confidence 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 was associated with a decreased amount of diagnostic tests being ordered, even if confidence and accuracy were larger decoupled/miscalibrated (Meyer et al, 2013). 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 interpretation of clinical features that favour the initial diagnosis (Leblanc, Brooks & Norman, 2002). Together, these findings have 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.

**Current Work**

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). Clinical experience may also be connected to risk aversion and further information seeking behaviour (Lawton et al, 2019), which offers an important avenue for future medical education. Hence, this research informs medical education of non-technical skills such as diagnostic reasoning, especially around evaluating diagnostic differentials and seeking information during the diagnosis process.

We seek in this work to better understand how information seeking, confidence and differential generation interact within the diagnosis process. We conducted a vignette-based diagnosis study with medical students to inform future work on how diagnostic reasoning is taught to students, especially when it comes to weighing up competing differentials.

**Study 1**

**Methods**

**Participants**

The study was conducted online, with participants able to run the experiment in their browser. The experiment was coded using JSPsych, which is a Javascript plugin used specifically for psychology experiments. 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.

**Materials**

This study involved the usage of patient vignettes, which are patient scenarios that 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. This was done via consultation with three different researchers working with the OxSTaR Centre who were also practising medical staff and students within the NHS. The aim was to consult medical professionals who had completed their clinical education and were at differing experience levels and medical subdisciplines. Their medical roles at the time of the development of this study were as follows: Speciality trainee (ST6) in Anaesthetics, Foundation (F1) Doctor and Gastroenterology Consultant.

As well as translating the language of the cases from US to UK-orientated (e.g. drug names, conditions etc), there were also changes made based on time, given that the original vignettes were developed for a paper published in 2004. Cases made occasional references to specific years in the patient’s history where they had previous medical conditions or hospitalisations. These years were updated to make sense for a contemporary patient (i.e by adding 18 years such that any referenced years were relative to 2022 rather than 2004). Whilst a sizable bank of vignettes were provided by Friedman, certain conditions were considered too rare (either for the current time or for the UK) to be used. Our goal was to test the clinicians’ ability to deal with diagnostic uncertainty, rather than testing their declarative knowledge of obscure medical conditions. In consultation with the aforementioned doctors, cases were therefore chosen based on their underlying medical conditions being those that medical students would be expected to know. We also aimed to choose cases that displayed a variety of both difficulty and of condition type (i.e. involving different anatomical systems).

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.

**Procedure**

A diagram of a patient's flow

Description automatically generated

*Figure 1: Paradigm of Study 1, showing the procedure for a single patient case. For Study 2, the procedure is very similar but excludes the ‘Indicate Differentials’ stage.*

The goal of the task was to determine a diagnosis, or set of diagnoses, for each presented patient. Information on the patient is split into a series of discrete stages so that the researchers are able to control what information the clinicians have access to at any given point in the experiment. We can call each point of new information an “information stage”. Participants are able to seek information freely until they are ready to move on, similar to the paradigm adopted by Kämmer et al (2019).

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. An example of this is: “patient is a 68 year old male presenting with fever and arthralgia”. This remains on screen throughout the entire case. Each case is split into three information stages: Patient History, Physical Examination and Testing. This order of stages is fixed for all cases. At each stage, the participant sees pieces of information or tests that they can request. Participants can view information from a previous stage but cannot see information for a future stage (e.g. if a participant is at the Physical Examination stage, they will be able to see information pertaining to Patient History and Physical Examination, but not information pertaining to Testing). The set of information requests for each stage is the same for all cases. The Patient History stage includes information on “Allergies”, “History of the Presenting Complaint”, “Past Medical History” and “Family History”. The Physical Examination stage includes ‘actions’ that a doctor may take when examining a patient, such as “auscultate the lungs”, “abdomen examination”, “take pulse” and “measure temperature”. Finally, the Testing stage involves information on any bedside tests or tests they may request from another department. This includes “Chest X-Ray”, “Venous Blood Gas”, “Urine Dipstick” and “Clotting Test”. In total, 29 possible tests that can be requested across the three information stages.

When a participant clicks on any of these tests, the screen shows a loading icon for 3 seconds before showing the information for that test on screen. During this loading time, other tests cannot be requested. When any subsequent test is requested, the previous test result is removed from the screen such that participants can only view one piece of information at a time. The time delay for receiving information was added after piloting the study, where the lack of time delay meant that participants were likely to request most information without being selective. It was 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. The information shown for each test is pre-defined as per the medical vignettes and is the same for all participants. Participants are free to request the same piece of information multiple times in order to remind themselves, including information from a previous information stage.

At any point, the participant can choose to stop gathering information for that stage. They are then taken to a new screen where they can report a list of all differential diagnoses that they are considering for that patient at that stage. Participants can report as many diagnoses in their list as they want to. For each differential, participants report a “level of concern” for that differential, which we describe as how concerned the participants would be for that patient if this differential really was the patient’s underlying condition. This is reported on a 4 point scale, with labels of “Low”, “Medium”, “High” and “Emergency”. Participants also reported a likelihood rating for each differential, ranging from 1 (very unlikely) to 10 (certain). When reporting differentials at the first information stage (Patient History), the list of differentials is blank and participants must add at least one differential to proceed. In subsequent stages, the list from the previous stages is available for participants to update concern/likelihood ratings, or to add/remove differentials from the list.

Participants are asked to carefully consider which differentials they have in mind in light of the new set of information they have received. Even at the last information stage, participants can report multiple differentials. Participants are not penalised for reporting a wide set of differentials at any 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. Once all three stages are complete, participants report how difficult they found it to determine a diagnosis for that case, on a scale from 1 (trivial) to 10 (impossible). At the end of all six patient cases, participants are told the true underlying conditions for all the patients.

**Data Analysis**

Responses were coded for correctness manually with help from a medical consultant, who looked at all the information available for each case and determined which diagnoses could be accepted answers. This depended on the nature of the case, as a case may sometimes have a vague set of information such that determining the exact correct diagnosis was considered too challenging. For example, for the TTP case, making a diagnosis of TTP (even with all information requested by the participant) was seen as too difficult given that the information provided was not discriminant enough. Hence, other conditions like ITP and Meningitis were also accepted as correct answers. All lists of differentials were ‘marked’ for correctness manually using the following criteria (the correct condition is followed by the list of accepted diagnoses to be considered correct):

**Temporal Arteritis**: any inflammatory arteritis is accepted

**Ulcerative Colitis**: infectious colitis, ischemic colitis and diverticulitis are also accepted answers.

**Miliary TB**: any TB or lymphoma type is accepted

**Aortic Dissection**: pulmonary embolism and coarctation of the aorta are also accepted answers.

**Guillain-Barre Syndrome**: Cauda Equina Syndrome is also accepted

**TTP**: ITP or Meningitis are also accepted.

For all cases, acronyms or spelling mistakes on the accepted answers were marked as correct.

There are a number of key dependent variables that we are able to derive from our data:

• **Confidence:** the reported confidence at each information stage. 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 and 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. Hence, 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).

• **Perceived Difficulty:** the subjective rating by participants at the end of each case for how difficult they found it to determine a diagnosis for that patient case. This is reported on a scale from 1 (trivial) to 10 (impossible).

• **Accuracy:** For a case to be considered ‘correct’, the participant should have reported the correct condition for that case within their list of differentials regardless of the number of differentials provided. Given that differentials are provided via free text, cases are manually coded as correct or incorrect using the aforementioned criteria. Our main accuracy measure is computed by the taking the likelihood value assigned to the correct differential if it is included in the list of differentials. This means that accuracy ranges from 1-10 when a correct differential is included and has a value of 0 when a correct differentials is not included. If multiple differentials that are considered correct were provided, then the likelihood value of closest differential to the true condition was used.

• **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 a vector of length 29, which is made up of 0s and 1s where for each of the 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. The vectors for all cases for a given participant are combined to produce a 29 x 6 matrix. We calculate the Dice dissimilarity coefficient between each row of the matrix (trial) using R’s dist function (in the proxy package). While several methods exist for calculating distance between, we use the Dice coefficient due to it being well suited specifically for binary data, as well as its increased weighting on discordant pairs (ie a piece of information being sought by one participant but not sought by another). The computation of all pairwise distances produces a 6 x 6 matrix where each trial is given a distance/dissimilarity value relative to every other trial. A lower distance value between two trials indicates that the information sought on those trials are more similar to one another. In order to look at the similarity of information seeking across all six trials, we compute the mean distance of the participant’s cosine distances. A lower mean value indicates that participants seek similar information across the cases whilst a higher value indicates that information seeking is varied more by case.

• **Information Seeking Value:** We sought to compute the average value of sought information for the medical students using these values. To do this, 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. For each group, we compute the proportion of trials where the students included a correct differential, and then take the difference between these two values. A positive value would indicate that students were more likely to identify the correct condition with that information rather than without that information. This difference can be considered that information’s ‘value’. For each of the participants’ cases, we compute this difference for each piece of information that the participant sought (for information they did not seek, the informational ‘value’ would be 0) and then calculate the mean information value for each case. We then sum all cases’ information values for each participant. This gives an overall measure of how useful the information was that participants sought on a case-by-case basis.

**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). We note here that confidence was on average below 50% even at the end of each case, indicating that participants were overall not highly confident to treat the presented patients. 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 |

*Table 1: Showing statistics across participants for each case (leftmost column). Differential Accuracy refers to proportion of cases in which a correct differential is included in the list of differentials, whilst Accuracy refers to the average likelihood (on a 1-10 scale) assigned to a correct differential if included. Both of these measure, as well as Final Confidence, are calculated at the final information stage of each case (i.e. the Testing stage).*

A chart of different levels of performance

Description automatically generated

*Figure 2: Graph showing Accuracy (black), Confidence (green) and Differential Accuracy (red) at each of the three information stages. Differential Accuracy values are divided by 10, and Confidence values are divided by 100 such that all variables are between 0 and 1 for visualisation purposes here.*

Previous work (such as from Meyer et al, 2011) have noted a gap between subjective confidence and objective accuracy. In particular, there has been demonstrated to be a general tendency for less experienced medical trainees to be underconfident and for more experienced medical professionals to be overconfident (Yang and Thompson, 2010). We sought to investigate whether we observe a similar pattern by comparing participants’ final confidence on this task with their objective performance. We operationalise performance in two ways: Differential Accuracy being a similar performance measure to that used in previous studies (i.e. a trial is marked as correct if a correct differential is provided) and Accuracy taking into account the likelihood values assigned to correct differentials. We find that confidence is fairly well calibrated to objective performance except for a slight deviation between the two during the Testing stage. This could imply that past findings on the 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 using 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 case is predicted by training the algorithm on all other cases. By plotting an ROC curve of our classifier, we find an area under the curve (AUC) value of 0.73 (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 0.62. This indicates overall that differences in information seeking are indeed predictive of a difference in participant ability at above chance. Next, we seek to identify and better characterise the specific differences in information seeking that contribute to this relationship with diagnostic ability.

A graph of a number of people

Description automatically generated with medium confidence

*Figure 3: Receiver-Operator Characteristic curve using both GLM (black solid line) and Classification Tree (red dashed line) algorithms to classify individual cases as being performed by either high or low accuracy participants. The models are trained on the raw binary predictor variables for each of the 29 available pieces of information, with 0 indicating that the information was not sought for the case and 1 indicating that the information was sought. Participants were sorted as high and low accuracy based on a median split on their average Accuracy value across the six cases. For the purposes of this analysis, the true condition of the case is not taken into account.*

Interestingly, when conducting a Pearson’s Correlation test, we 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). As previously discussed, participants rarely seem to remove differentials from consideration. Therefore, one can surmise here that higher information seeking is associated with the consideration of more diagnostic differentials.

A graph of a number of differentials

Description automatically generated

*Figure 4: Scatter plot showing the relationship between the number of initial differentials (x-axis) and the proportion of available information sought (y-axis). Each point represents a single student with both variables average across the six cases that each student performs. The x-axis refers to the average number of differentials that participants report in their list at the Patient History stage. The y-axis refers to the average proportion of available information sought, with each cases containing 29 pieces of information across the Patient History, Physical Examination and Testing stages. The line of best fit is plotted using the geom\_smooth function in R with a linear model. The shaded region shows the 95% confidence interval of the correlation.*

We then look at how ability on the task relates to information seeking behaviour, starting with 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).

A graph with black and blue lines and dots

Description automatically generated

Figure 5: *Scatter plot showing the relationship between participant accuracy (x-axis) and information value (y-axis), with both variables calculated across the six patient cases. Accuracy is divided by 10 in this case so that values fall between 0 and 1. Information Value refers to the sum of all mean information values across all 6 cases for a given participant.*

We then find that participants with higher overall accuracy have a lower variance in information seeking. In other words, students with a higher diagnostic ability on this task are found to have varied the information they sought across cases less. This means they sought more similar information for each case when compared to students of a lower diagnostic ability. We find marginal evidence for a negative correlation between accuracy and information seeking variance (r(83) = -0.21, p = .049). We note that for a number of trials (n = 6), participants did not seek any information. We hence also conducted this correlational analysis after excluding these trials when calculating information seeking variance and find a slight decrease in the observed effect (r(83) = -0.21, p = .057). While not strong evidence, this indicates a broad pattern that being more standardised in information seeking across cases is associated with higher diagnostic accuracy.

A graph of a number of black dots

Description automatically generated

*Figure 6: Scatter plot showing the relationship between Information Seeking Variance (x-axis, quantified as the average Dice Distance between all cases for a given participant) and accuracy (y-axis).*

A graph of blue and pink bars

Description automatically generated

*Figure 7: Average Information Seeking Variance for all cases of a given condition (x-axis), with cases median split by participant accuracy, with red indicating high performers and blue indicating low performers.*

Finally, we turn our attention to the amount of information seeking on cases. We find that the proportion of available sought is not shown to correlate with accuracy (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 as it demonstrates that being selective in information seeking is a better marker of performance and giving a lower ability participant all available information does not necessarily translate into accurate diagnoses. In addition, students with a higher diagnostic ability seek better information but also approach each case in a more similar manner. This could indicate a base of information kept constant across cases alongside a more selective set of useful information related to that patient.

A graph of a graph with black dots and red line

Description automatically generated with medium confidence

*Figure 8: Scatter plot showing the average proportion of available information sought across cases for a given participant (x-axis) against the average change in confidence across cases (y-axis, quantified as the difference between reported confidence at the Testing Stage and at the Patient History Stage).*

**Discussion**

We overall find that for medical students on this task, accuracy and confidence are fairly well calibrated unlike previous work. Part of this could stem from the diagnostic uncertainty expressed by students on the task, which they do in two ways. Firstly, students broaden, rather than narrow, their considered diagnostic differentials with more information and still report a broad range of differentials after receiving all available information for a given cases. Secondly, students report fairly low confidence overall to treat patients, with an average confidence of below 50% even after receiving all available information. This may indicate that part of ensuring appropriate confidence or expressions of uncertainty could be related to properly evaluating all possible diagnostic differentials rather than forcing decisions to focus on a single diagnosis, which has been cited previously as a problematic tendency (Redelmeier & Shafir, 2023).

We also find through our analysis of information seeking patterns on this task that accurate diagnoses were associated with both selective and standardised information seeking. 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. There has been increased emphasis on overtesting, such as requesting costly imaging scans when they may not be medically necessary (Carpenter, Raja & Brown, 2015). ‘Overtreatment’ has been estimated to cost the US healthcare system between 158 and 226 billion dollars in 2011 (Berwick & Hackbarth, 2012). Seeking more information during the task made students more confident but not more accurate, which is important to note as it corresponds with previous findings from the cognitive psychology literature (Ko, Feuerriegel, et al, 2022).

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. In our study, participants could remove a differential in the interface by clicking the X button on a differential. One explanation is that the button is not very prominently placed on the screen. However, this feature was explicitly explained in the tutorial to the experiment. This tendency is reflected in the overall pattern of the average number of differentials increasing over the three stages of a case. What this may indicate then is an attachment to hypotheses and unwillingness to remove them from consideration. There is a general adage in healthcare that medical students come across which says that “history is 80% of diagnosis”. It is therefore worth considering whether there is a specific facet of diagnostic decisions whereby students are taught not to disregard diagnostic possibilities easily. We nonetheless sought to investigate this in our follow-up study.

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. 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. We hence also sought in our follow-up study to better understand students’ thought processes when evaluating differentials, especially given our observation on this study that wider sets of initial diagnostic differentials prompted greater amounts of information seeking.

When creating a task that emulates diagnosis, we in a sense conceptualise what diagnosis looks like in a fairly static manner, when really diagnosis is a more fluid and nebulous structure in medicine. For example, a doctor’s approach to a patient is not always going to fit within the idealised structure of taking a patient history, conducting physical examinations and then requesting tests in this order. There are environmental or even patient factors that necessitate information being processed out of order, as well as different diagnostic approaches by doctors. In addition, it was striking that participants in our study rarely removed differentials from their list of suspected conditions despite having the ability to do so. This lack of removing differentials was what drove our observed effect of the number of differentials increasing with more information and it might have been that this was related to the lack of visual prominence in our interface when allowing students to remove differentials.

We hence wanted to see if signs of these results would be evident in the thought process of medical students. Are students seeking information to confirm their existing set of differentials, to rule out differentials or to expand their set of considered possibilities? And are these different approaches interleaving or are they more dependent on individual diagnostic decision making styles? In order to provide more context to the results from study 1, we conducted a follow-up study that utilised a very similar experimental procedure, but instead prompted students to think out loud as they were performing the task. Their utterances were then transcribed and coded to conduct both quantitative and qualitative analysis.

**Study 2**

**Methods**

**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 the same six patient scenarios and were tasked with diagnosing the patient. Participants were free to seek information in three stages: Patient History, Physical Examination and Testing. 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. An initial transcript was generated using Microsoft Office’s transcription feature, but the transcript was checked and modified for accuracy by listening through the audio recordings again. The screen of the experimental interface was also recorded, such that the audio could be linked to specific actions within the task. The focus of this study is on verbal utterances rather than any non-verbal or inferential aspects of the participants’ qualitative data.

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 questions are indicated below, each with a corresponding follow-up question in case they are not answered by responses to the main 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?

**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 define the following codes:

* **Differential Evaluation:** any time that the participant (each of the following is considered a separate subcode):
  + Mentions a new condition that they are considering
  + Rules out or eliminates a condition from consideration
  + Mention of increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition
  + Mention of decreased likelihood of a previously mentioned condition, or that information seems to contradict with a condition
* **Information Seeking Strategies:** any time the participant expresses why they may or may not request a particular piece of information in relation to ruling out or confirming 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 also adopted a think-aloud protocol during 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 pathophysiological 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”).

We first code specific utterances within each case that suggested one of these strategies, and then determined which strategy was most prevalent or influential for cases as a whole such that each case was categorised under one of these strategies. 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. As questions were fairly closely related to diagnostic decision making, we could ensure that participants gave an account of their decisional process.

Although we do not record differentials in the same way as in Study 1 (in a list with corresponding likelihood and severity ratings), we do obtain the other variables from Study 1. Namely, we record confidence at each stage of information seeking and data around the information sought by participants. We also code each case as ‘correct’ if a correct differential is mentioned at some point (using the same marking scheme for each case as utilised in Study 1).

**Results**

First, we look at overall quantitative characteristics of the think aloud utterances. When looking at accuracy (the proportion of cases where a correct differential was mentioned by the participant), participants were 57.3% accurate across all cases. This varied considerably by condition however, with accuracy across participants for each condition being as follows: AD = 62.5%, GBS = 87.5%, MTB = 18.9%, TA = 43.8%, TTP = 68.8%, UC = 62.5%. For utterances coded as Differential Evaluations, participants on average made 5.21 such utterances per case (SD = 2.80). The mean number of Differential Evaluations was relatively constant by condition except for the AD case: AD = 8.18, GBS = 4.63, MTB = 4.81, TA = 4.75, TTP = 4.25, UC = 4.63.

A graph of different colored bars

Description automatically generated

*Figure 9: Accuracy for each case in the think-aloud study across participants. Accuracy is operationalised here as the proportion of participants who mention a differential considered correct during that case.*

As previous mentioned, Differential Evaluations can be further categorised into one of four subcodes: adding/mentioning a new differential, eliminating/removing a differential from consideration, increasing the likelihood of a differential (or mentioning that evidence/information is in support of a previously mentioned differential) and decreasing the likelihood of a differential (or mentioning that evidence/information is counter to a previous mentioned differential). As found in the previous study, there is a general reticence to disregard differentials completely. Participants, on average across cases, produced more utterances of adding differentials (M = 3.13, SD = 0.66) than removing differentials (M = 0.40, SD = 0.30) to a significant degree (t(15) = 15.22, MDiff = 2.73, p < .001). Participants however did also produce more utterances of increasing likelihoods (M = 1.76, SD = 1.06) than of decreasing likelihoods (M = 1.03, SD = 0.60) to a significant degree (t(15) = 2.72, MDiff = 0.73, p = .02).

A graph of different colored squares

Description automatically generated with medium confidence

*Figure 10: Bar graphs showing the average number of each type of differential evaluation across all cases by all participants. The left graph compares the number of utterances of a new differential being considered (green) against the number of utterances of a differential being eliminated or placed not under consideration (purple). The right graphs compares the number of utterances where a differential is decreased in likelihood or is said to have contradicting information (light blue) against the number of utterances where a differential is increased in likelihood or is said to have supporting information (brown).*

Next we look at our coding of reasoning strategies at a case level. As mentioned, our criteria for each code was applied to each individual case based on the transcribed utterances. When looking at reasoning strategies by case, 43% of cases were coded as Hypothetico-Deductive, 29% were coded as Pattern Recognition and 19% were coded as Scheme Inductive (the remainder of cases did not contain enough clear utterances to classify under one of these strategies). Accuracy was higher for cases coded as Hypothetico-Deductive (71%) compared to both Pattern Recognition cases (64%) and Scheme Inductive (39%). It is worth noting here that accuracy was solely based on participants mentioning differentials during their thinking aloud, which is naturally not facilitated by Scheme Inductive reasoning due to its focus on identifying pathophysiological systems acting as sources of patient symptoms rather than specific conditions. This can hence explain the lower ‘accuracy’ for Scheme Inductive cases. We also note that the types of reasoning strategy used varies by condition (see figure below), with the MTB case in particular exhibiting higher usage of Pattern Recognition than others. This could be because this case was considered harder than others and hence participants could not generate a larger set of candidate differentials due to its difficulty.

A graph of multiple colored bars

Description automatically generated

*Figure 11: Proportion of participants who use each type of reasoning strategy for each condition/case, with the overall proportions across all cases shown by the rightmost bars. The strategies shown are: Hypothetico-Deductive (where multiple differentials are considered simultaneously, orange), Pattern Recognition (where a single differential is considered in turn, blue), Scheme-Induced (where participants evaluate pathophysiological systems as causes of patients rather than specific conditions, green) and None (for cases where a clear differential is not mentioned, grey).*

We note, rather unsurprisingly, that we observe a higher number of average Differential Evaluations when cases are correct (M = 5.85, SD = 0.38) compared to when they are incorrect (M = 4.34, SD = 0.39). Given our methodology for defining accuracy, participants are more likely to mention a correct differential if they mention more differentials. The procedure used in the previous study for collecting data on which differentials participants were considering at each information stage was not present here and hence we are not able to operationalise accuracy in the same manner as before. While we look at which differentials are mentioned, we cannot observe how participants weigh up differentials against each other in the same way as in the first study.

Next we discuss two key themes that emerged from the debrief interviews.

**Differential Management**

Participants reported that there are difficulties with managing sets of differentials when making diagnoses (in brackets are anonymised participant ids for the relevant quote):

*“I find it more stressful to have, you know, a list of like a weird way of remembering a list of all potential differentials something else, I'd rather use the information I've given to slowly build up a picture.” (k5376h)*

As we shall explore in the next section, participants reported that they tried to keep an open mind when making diagnoses. Part of this approach involves a general reticence to eliminate differentials from consideration:

*I think I do…rule out definitely some…it's sometimes difficult to rule out completely. (dcjymb)*

*“I often I’ll have like three or four top differentials. But I wouldn't have a very wide collection.” (l3jd8r)*

*“So like a lot of my thinking is like, what like really worrying thing could this be that we need to rule out? And what tests do I need to rule it out? But then also like when I'm thinking or what could this be, I'm also thinking about what investigations would help me to conclusively reach a diagnosis that this is what it is once I think I know what's going on.” (5lvg8j)*

**Attempting to Keep Open Minded**

Participants reported being aware of a tendency to become fixated with a differential prematurely without considering others:

*“I think it's called anchor bias where you have, you can leap onto one thing early on, and then you want other things to fit that. I think we are all vulnerable to it to an extent. And we will look for things that support our initial idea, but I try and keep an open mind.” (k5376h)*

*“I think my brain can sometimes get stuck on an idea. And it's difficult to pull away from that. Or sometimes it can go too broad and it's like, I don't know what's going on.” (gdq7tc)*

Part of this “anchor bias” stems from the relative inexperience that medical students have at their stage of their medical career:

*“I think I try to keep an open mind perhaps because I'm just like, the student and I don't have as much knowledge, as someone who's been training for a long time.” (dcjymb)*

*“But then I do think I, at this stage, I'm quite kind of bias towards what I know more about, if that makes sense. So I think the things which I don't know about, I'm just hoping it’s not that.” (3lkzjq)*

Other participants however reported an opposite concern however: that they may lack the ability to commit to a differential early enough:

*“I don't like to commit…I think being broad can be good so you don't miss things.” (gs6zbl)*

*I think we need to rule out differentials, but I felt like a lot of points, I felt like this is the most likely but I still feel like I want rule everything else out first. (d9b1qf)*

**Discussion**

These results from the think-aloud variant of our online study hint at some interesting avenues for future work to tap into. We observe differences in reasoning strategy that are potentially a function of the case at hand but also of the individual clinician. Students during debrief interviews reported a consideration they tend to have during diagnoses about either becoming fixated too early on a particular differential or about being hesitant to commit to a differential without giving others due consideration first. The tension between these two can hence be resolved depending on reasoning strategy. A diagnostician who is concerned about being anchored to a differential too early may adopt a hypothetico-deductive process to broaden the differentials they consider. Meanwhile, a diagnostician who is concerned about not being able to commit to a differential may adopt a pattern recognition to narrow their thinking toward matching the observed patient’s symptoms to prototypical instances of conditions. It is potentially more likely that a medical student, due to relative inexperience, would be more concerned about the former than the latter.

Coderre et al (2003) found the pattern recognition was utilised more as clinicians increased in experience. On the one hand, this makes sense given that having more experience with seeing patients would improve a diagnostician’s ability to match symptoms to a condition. On the other hand though, as alluded to by students in this student, knowledge and experience brings with it the ability to generate more differentials than a less experienced clinician. One cannot adopt a hypothetico-deductive reasoning process, whereby multiple differentials are considered and then eliminated, if the clinician lacks sufficient knowledge to generate a set of differentials based on the observed patient. This may be where the complexity/difficulty of the case has a bearing on reasoning process too, whereby harder cases are harder because one cannot easily generate differentials for them. However, the inverse could also be true, whereby a set of conflicting symptoms may cast a wider net of potential differentials that are more challenging to narrow down. As we noted in the online study, the number of initial differentials has an impact on information seeking behaviour, but as we explain here, differentials are themselves a result of a particular reasoning strategy. Ascertaining the exact interaction between reasoning strategy, case difficulty and differential evaluation is hence important for us to focus on in the following study, as it informs how diagnosis is characterised as a cognitive process and how cognitive interventions are designed to aid the process.

We note that the use of a think-aloud methodology brings with it a couple of limitations. Firstly, participants may behave differently to how they would otherwise given that they are being observed and recorded by a researcher. Hence, there may be a tendency toward medical students behaving in a manner that they believe to be judged as better by others, such as being thorough in their information seeking and differential evaluation. Relatedly, we found that medical students naturally differed from one another in terms of the amount of verbalising they did during the task. By not explicitly asking students for their diagnostic differentials as we did in Study 1 (and minimising the amount of input that the researcher had during the task), we are constrained to analysing only what students say out loud. Given that some students do not verbalise their thoughts as naturally, we may not aware of the aspects of their thought process that they did not verbalise. As a result, the analysis presented here is relatively exploratory in nature. Future work should utilise more structured methods for eliciting clinicians’ think process during diagnoses.

**Overall Discussion**

There are a few potential limitations with our studies that are worth addressing. 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 participants requested a chest X-ray, they may interpret the X-ray image in different ways. While this difference in perception may be interesting, it adds a potential confound for the purposes of this study. That is why, for this study, if a participant requested a chest X-ray, they instead see a result that reads something like “no abnormalities found”, such that the interpretation of the image has already been standardised for all participants. However, future work could use more naturalistic methodologies to investigate the diagnosis process.

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. Some tests (e.g. a chest X-ray) are not performed by the doctor themselves at the patient’s bedside and require extra staff and technology to produce. 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. In addition, we attempted to make the patient cases as realistic as possible whilst having a moderate degree of difficulty. The original researchers removed certain findings from the cases that may give away the patient’s condition in a fairly obvious manner. In that sense, the patient cases may not replicate the set of information that might be available to clinicians in a similar scenario during medical practice. However, using a paradigm similar to past research does extend and build upon empirical experiments on diagnosis. As previously mentioned, information was chosen in order to be general to all cases and was not very discriminant, such that we can specifically focus on situations of diagnostic uncertainty.

We can however take both studies together to provide a nuanced discussion of the diagnostic process among medical students. We find that information seeking patterns and evaluation of differentials during the diagnosis process contribute to diagnostic accuracy. When students generated a greater number of differentials from a patient history, they sought a greater amount of information. We then observe an association between information seeking and confidence, but not with accuracy. Instead, accuracy was characterised by more standardised and selective information seeking during the diagnostic process. However, the amount of differentials generated is itself dependent on the reasoning strategy adopted by students, be it a hypothetico-deductive process where multiple differentials are generated and then evaluated in turn or whether the students match patient symptoms with a prototypical case of a condition from their medical knowledge. The students’ ability to generate differentials is dependent on how difficult the case is, affecting the medical knowledge available for students to utilise. This is where, in a real medical setting, it is important to express uncertainty as a function of how able clinicians are to generate diagnostic differentials. As is hinted in our qualitative analysis for Study 2, a choice of reasoning strategy can potentially stem from a desire to either avoid fixating too early on a differential or to avoid a situation where there is a lack of commitment to a differential. How this choice is made is currently unclear, but we can hypothesise that this could be related to a specific interaction between case difficulty, expertise and clinical domain (McLaughlin, Rikers & Schmidt, 2008).

There can hence be value in teaching metacognition and uncertainty within medical education (Royce, Hayes, & Schwartzstein, 2019), such as with the use of cognitive aids (Chew, Durning & Van Merriënboer, 2016, Ely, Graber & Croskerry, 2011), especially given that doctors can be reticent to express their uncertainty (Katz, 1984). A more structured aid is needed, as simply looking at a case for a second time may not be sufficient to improve diagnostic accuracy (Monteiro et al, 2015) and current cognitive forcing strategies have not been found to be effective enough (Sherbino et al, 2014). The reason for this might that past distinction between System 1 and System 2 for prompting diagnostic reasoning may be overly simplistic, in that one solution may not fit all possible cases and all clinicians. Future work should hence focus on understanding when certain reasoning styles and hence which cognitive aids may be more useful for a given clinical situation. Past work on cognitive interventions have not tended to focus on prompting appropriate information seeking, and we show here that different facets of information seeking contribute uniquely to both confidence and accuracy. While the most relevant information that should be afforded to clinicians will differ depending on the medical discipline, we note that before any integration of information can take place, interventions can focus on standardising which information is presented to clinicians in the first place. This could not only improve diagnostic accuracy but ensure more appropriate expressions of confidence and uncertainty by reducing a tendency toward overtesting. We emphasise that such recommendations are highly dependent and variable depending on the specific medical context, but this acts an important starting point for medical education around how seeking information relates to reasoning styles and how important these non-technical skills are to integrate into the educational context of medicine.

**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. Berwick, D. M., & Hackbarth, A. D. (2012). Eliminating waste in US health care. Jama, 307(14), 1513-1516.
3. 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.
4. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.
5. 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.
6. 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.
7. 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.
8. 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.
9. Chi, M. T., Glaser, R., & Farr, M. J. (Eds.). (2014). The nature of expertise. Psychology Press.
10. 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.
11. Coderre, S., Mandin, H. H. P. H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. Medical education, 37(8), 695-703.
12. 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.
13. 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.
14. Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. Psychological science, 29(5), 761-778.
15. Ely, J. W., Graber, M. L., & Croskerry, P. (2011). Checklists to reduce diagnostic errors. Academic Medicine, 86(3), 307-313.
16. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological review, 124(1), 91.
17. 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.
18. Frotvedt, T. F., Bondevik, Ø., Seeligmann, V. T., & Sætrevik, B. (2020). Primacy, Congruence and Confidence in Diagnostic Decision-Making.
19. 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.
20. Hall, K. H. (2002). Reviewing intuitive decision‐making and uncertainty: the implications for medical education. Medical education, 36(3), 216-224.
21. 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.
22. 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.
23. Higgs, J., Jones, M. A., Loftus, S., & Christensen, N. (2008). Clinical reasoning in the health professions E-book. Elsevier Health Sciences.
24. 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.
25. 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.
26. 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.
27. 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.
28. Katz, J. (1984). Why doctors don't disclose uncertainty. Hastings Center Report, 35-44.
29. 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.
30. 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.
31. 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.
32. Kovacs, R. J., Lagarde, M., & Cairns, J. (2020). Overconfident health workers provide lower quality healthcare. Journal of Economic Psychology, 76, 102213.
33. Kuipers, B., & Kassirer, J. P. (1984). Causal reasoning in medicine: analysis of a protocol. Cognitive Science, 8(4), 363-385.
34. 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.
35. 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.
36. McLaughlin, K., Rikers, R. M., & Schmidt, H. G. (2008). Is analytic information processing a feature of expertise in medicine?. Advances in Health Sciences Education, 13, 123-128.
37. 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.
38. 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.
39. 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.
40. Morgan, P. J., & Cleave‐Hogg, D. (2002). Comparison between medical students' experience, confidence and competence. Medical education, 36(6), 534-539.
41. Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. Decision making in action: Models and methods, 1, 3-20.
42. Redelmeier, D. A., & Shafir, E. (2023). The Fallacy of a Single Diagnosis. Medical Decision Making, 43(2), 183-190.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12(3), 129-140.
50. Yang, H., & Thompson, C. (2010). Nurses’ risk assessment judgements: A confidence calibration study. Journal of Advanced Nursing, 66(12), 2751-2760.
51. Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. Journal of Experimental Social Psychology, 33(4), 345-366.