Information Seeking and Confidence in Medical Decision Making

Sriraj Aiyer

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

For my family

Acknowledgements

I would firstly like to thank my amazing supervisors, Nick and Helen, for your insights, patience, enthusiasm and boundless knowledge that helped shape this thesis into what it is.

I also would like to thank my mother, father and sister for everything they have done for me, which would require more words to list off than there are contained in this thesis.

I dedicate this to my grandparents.

Sriraj Aiyer
Wolfson College, Oxford
30 September 2024

Abstract

Decisions within healthcare are unique within the wider realm of decision making. They are often made within high-pressure situations and have severe consequences if done so incorrectly. Hence, they require intensive training and a wide knowledge base for clinical staff to draw from. What is remarkable is that despite the intimidating amount of material for medical students to learn and the pressures that can befall them in their everyday line of work, as well as an ever-expanding understanding of medical conditions, treatment methods and technology to maintain, clinicians frequently make swift and accurate decisions that can have a profound impact on patients' lives. When seeking to apply past research within decision making to an applied context, medicine is an interesting domain to study decision making, especially if findings can inform the training of the newer medical students. In particular, 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). Similarly, curricula within medicine place little emphasis on how uncertainty is communicated and approached in medical decision making (Hall, 2002). Hence, this research looks into non-technical skills such as communication of confidence, management of uncertainty and mental model alignment. Over the course of this thesis, we will look at confidence and information seeking in general decision making and then apply insights from cognitive psychology to the realm of medicine.

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Aortic Dissection
GBS	Guillain-Barre Syndrome
MTB	Miliary Tuberculosis
PhEx	Physical Examination
PaHi	Patient History
TA	Temporal Arteritis
Te	Testing
TTP	Thrombotic Thrombocytopenic Purpura
UC	Ulcerative Colitis
VR	Virtual Reality: a form of technology where users wear a headset and are shown an application in a first-person view, where they embody that particular avatar.

Introduction

Imagine a group of doctors within an intensive care unit (or critical care unit, depending on the parlance used in different geographic regions). 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'. These are considered to be some of the most important metrics for tracking the human body's essential capabilities and can include heart rate, body temperature, blood pressure, blood oxygen saturation and respiration rate. In the case of our aforementioned patient, she has been recording a slow decrease in blood pressure and blood oxygen saturation. The doctors have to decide what is the most likely cause of this patient's symptoms and vital signs. 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 cause could be that the patient is suffering 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 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?

Introduction

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 methods for reaching a solution and even produce disagreements over a desirable starting hypothesis and end state (Jonassen, 1997). Individuals involved in clinical decision making have to frequently contend with an uncertain decision making environment, as well as time pressure and personal stresses (Orasanu & Connolly, 1993). Alignment of mental models for the medical staff involved in a patient case is therefore critical that they can formulate very different understandings of a patient's condition and how it would be best to proceed. Medical staff have to align their thoughts in order to align their actions and function as a cohesive team. This is despite the fact that they often have to operate under uncertainty. While clinicians can continue to gather information in order to reduce their uncertainty, there is a further constraint enforced upon them: time. They frequently have to perform their job while observing the deterioration of a patient's health. This means that clinicians have to choose carefully when to commit to a particular working diagnosis in order to guide their future actions for treating the patient, even when they cannot freely choose the point of commitment.

Part of what makes medical decision making particularly challenging is the lack of clear feedback, as there might be in other task contexts. When making a particular diagnosis for a patient, clinicians likely do not receive a lot of feedback about the correctness of their diagnosis. Some may view diagnostic tests (eg blood tests) as a form of feedback: doctors use these test results to either reinforce or re-evaluate their prior beliefs. However, tests are not objective markers of feedback, as they can differing levels of sensitivity and specificity rates, leading to false positives, false negatives or even inconclusive results. The results can be used by clinicians to guide their future beliefs and actions, but they are primarily a form of information gathering that can be used to confirm existing hypotheses or eliminate differential hypotheses. Feedback may only be brought up to the clinician based on how a patient's status changes. Generally, doctors gather information through tests, patient documentation and other means to generate a model of the patient's condition, through which they can surmise 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. Hence, a patient's reaction to treatment, and their rate of recovery, can be seen as a form of feedback on the doctor's initial model of the patient. This in itself is an imperfect form of feedback, as patients can deteriorate or improve due to circumstances outside of the doctor's control or awareness. This makes confidence an interesting notion to study. Confidence is viewed within cognitive psychology as one's subjective probability of their own decisions being correct. In the absence of feedback, confidence can be used as a marker of how likely someone is to be correct. In the case of medicine, a lack of clearly communicable feedback can cause clinicians to proceed as if they have received positive feedback. This means that they do not adequately update their internal model of the patient and hence they increase their confidence inappropriately (Jaspan et al, 2022). In most psychology experiments, the feedback provided to participants is the objective correctness of a decision but in the case of medicine, feedback is more difficult to define given the lack of objective markers of correctness. As we shall discuss, this has implications on the study of confidence calibration. That is, how confident an individual is relative to their true accuracy.

Medicine also has an interesting wrinkle of real-world decision making that makes it different to classic psychology experiments. In a task such as the Information Sampling Task (IST) (reference), participants gather information by revealing squares on a grid that can be one of two colours, with one colour being the majority colour on the grid. Participants reveal colours until they are ready to decide, at which point they report which colour they think is the more prevalent colour. Here, there is a clear delineation between information gathering and decision making. In the healthcare sector, a clinician may decide to prescribe a certain course of treatment to the patient in order to start treating their condition or lessening the symptoms. However, this in itself is a form of information gathering, as the patient's reaction to treatment can itself yield informative information about their underlying condition. This can force clinicians to rethink their model of the patient. Hence, information gathering and decision making is more interleaved in real-world contexts such as healthcare.

Clinicians have to make challenging decisions as part of their occupation, such as administration of medication, allocation of hospital space and delegation of responsibilities to colleagues. However, one such 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. This allows for past findings to be applied within an ecologically valid, real-world setting. Secondly, diagnosis is an important task that has a large impact on a patient's road to recovery. Ensuring that a patient has positive outcomes in their treatment is in large part contingent on an accurate diagnosis of their conditions being made by healthcare professionals.

Firstly, it is worth painting a picture of the wider context of diagnostic errors. Looking into errors more broadly allows healthcare systems to learn from mistakes to improve technical and safety processes for future patients. Understanding the common sources of medical errors and adverse events can be extremely valuable for improving healthcare in the future. For example, Cohen et al (2021) analysed surgical adverse events in California to find that the majority of events were caused by the retention of foreign objects from surgery. However, there is also work looking at errors in diagnosis, which ties into questions of how humans gather information and formulate their confidence that are studied through the lens of cognitive psychology.

Diagnostic Errors

Diagnostic discrepancies are where an initial diagnosis is different to a diagnosis made for a patient upon their discharge from hospital. In other words, this would indicate that the initial diagnosis was incorrect. A report from the US Institute of Medicine (McGlynn, McDonald & Cassel, 2015) concluded that many people will experience a diagnostic error within their lifetime. The Harvard Medical Practice Study found that diagnostic errors were responsible for 17% of adverse events (Leape et al, 1991). The Canadian Adverse Events Study found this value to be 10.5% (Baker, Norton & Flintoft, 2004) whilst a study in New Zealand found this value to be 8% (Davis et al, 2003). When looking at records of new diagnoses for spinal epidural abscess in the US Department of Veteran Affairs, Bhise et al (2017) found that 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). Diagnostic errors have been found to be have downstream consequences, leading to longer hospital stays and even increased patient mortality (Hautz et al, 2019). Even when using the most conservative estimates, this illustrates the large scale of the diagnostic error when extrapolated to the population of patients. Studies have even investigated downstream consequences of diagnostic errors, with unnecessary treatments (or 'overtreatment') estimated to cost the US healthcare system between 158 and 226 billion dollars in 2011 (Berwick & Hackbarth, 2012). There has been increased emphasis on overtesting, such as requesting costly imaging scans when they may not be medically necessary (Carpenter, Raja & Brown, 2015).

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. 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 the causes of diagnostic error can have important implications for future interventions within healthcare settings.

Introduction

One account of diagnostic error is that they can stem from cognitive biases during decision making. Frotvedt et al (2020) looked at primacy (information presented earlier being more influential on judgements than information presented later) and congruence (preferentially seeking information to confirm prior beliefs) biases in mental health diagnoses. Chapman, Bergus & Elstein (1996) found a recency effect (rather than a primacy effect) when presenting physicians with a patient history either at the beginning or end of a patient vignette. Another set of diagnostic pitfalls has been found for physicians when making probability judgements. Arkes, Aberegg and Arpin (2022) showed that a majority of physicians incorrectly estimate the joint probability of two medical outcomes given each of their independent probabilities. Redelmeier and Shafir (2023) found when medical professionals estimated the probability that patient was infected with COVID, their estimates were affected by a test results for an alternative diagnosis even when a patient could have multiple conditions (e.g. both COVID and influenza), especially when the conditions are similar in nature.

Making a diagnosis involves considering a hypothesis as likely because the displayed symptoms seem to correspond with a prototypical case of a particular condition (despite symptoms being presented to the contrary). 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 that attempted to automatically detect uses of heuristics and biases by dermatologists, examples of satisficing bias (premature closure) and anchoring were found, but very few examples of other biases such as availability and representative were found (Crowley et al, 2012). Results of anchoring bias in clinical errors could be driven by a failure to adjust sufficiently based on a self-generated anchor (Epley & Gilovich, 2006) and that the anchoring effect size is affected by the order of information such that information presented later is less influential on decisions (Ellis et al, 1990). This corresponds with work that explains primacy biases as a result of an attentional decrement for successive items of information (Cunnington et al, 1997), which we can investigate as a factor that contributes to premature closure and overconfidence in information seeking. Attempts to lower the likelihood of diagnostic error has involved the use of checklists as cognitive aids (Ely, Graber & Croskerry, 2011, Kämmer et al, 2021) in order to ensure that diagnoses are not missed from the doctor's thought process.

Overall, 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.

Confidence and its Miscalibrations

At this point, we shall revisit the scenario presented in the Preface 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. 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 when faced with high confidence decision makers (Cheng et al, 2021). As we shall explore, confidence is commonly useful as a predictor of another person's accuracy, especially when feedback is not readily available of the true accuracy of the individual. This behaviour has been observed in mock jury trials in which participants hear eyewitness testimonies presented with high confidence and then perceived as more credible than testimonies provided with low confidence (Cutler, Penrod & Dexter, 1989, Roediger, Wixted & DeSoto, 2012). There may be a tacit assumption that others will be metacognitive 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. Highly confident members within a group could also unknowingly reduce the chance of less confident members speaking up about potential errors, which is a problem within healthcare (Hémon et al, 2020).

Confidence can be considered to be an individual's internal probability of a given decision being correct when taking into account the evidence used to make that decision (Fleming & Daw, 2017). Confidence has been proposed as a conscious, introspective property given that it is used in communication with others (Shea et al, 2014). This becomes an especially important point when making group decisions and aligning mental models between members of the group. 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. 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 when faced with high confidence decision makers (Cheng et al, 2021).

Confidence has a number of interesting facets that sheds some light on its cognitive mechanisms. It increases in the face of a larger amount of evidence overall, even if some evidence favours decision alternatives other than the one chosen (Ko, Feurnigel et al, 2022). Confidence also has a relation to decision times, as it increases with viewing time of a stimulus irrespective of decision accuracy (Raush, Hellmann & Zehetteitner, 2018). A faster response time is associated with higher confidence (Audley, 1960), which is a heuristic used not only for an individual introspecting about their own decision but also by observers who are attempting to infer the confidence of others (Patel et al, 2012). Confidence may also be used to predict choosing tasks to perform where the tasks have differing levels of effort involved (Carlebach & Yeung, 2020), with task choice being induced by using extra evidence to boost confidence (Kool et al, 2010) which corresponds with the aforementioned finding that a larger quantity of evidence leads to higher confidence. 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 initial confidence (and hence a lower strength of evidence). Confidence has been thought to tie into the global workspace model of consciousness, whereby confidence is broadcast for other systems (e.g memory, perception, attention) to utilise (Dehaene & Changeaux, 2011). The case for confidence as a conscious introspective property is bolstered when considering its evolutionary origins as a communication tool with other people, especially for group decisions (Shea, 2014).

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, a number of factors may distort subjective confidence such that confidence becomes 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 or uncertain when they are correct. These two cases can be referred to as overconfidence and underconfidence respectively. Miscalibration of confidence is part of a wider corpus of work on deficiencies in self-monitoring, with individuals being far more likely to notice their own execution errors (slips) than their own method errors (mistakes) (Allwood, 1984). These latter errors are where overconfidence can arise from.

Katz (1984) proposed that doctors do not approach uncertainty in the practice of medicine in the same way that they do in theory. This was illustrated in an example where a doctor was keenly aware of the lack of medical consensus (at the time of writing) on the best course of treatment for breast cancer, but the same doctor was highly confident when recommending surgery to the patient.

Evidence for overconfidence has been shown in other clinical contexts too. In a task that involved diagnosing ultrasound scans, it was found that overconfidence was negatively correlated 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) had 103 nursing students and 34 experienced nurses work through risk assessment vignettes and provide confidence judgements. They 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 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).

Linking Confidence and Information Seeking

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). However, this process being a standard to strive for has been argued to increase the risk of confirmation bias by collecting data to fit pre-existing theories rather than crafting theories around collected data (Chi, Glaser & Farr, 2014).

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). Choosing when to stop gathering information has been found to produce a 'boost' in confidence, though this is not the case when participants are forced to stop gathering information at a time point that they do not choose themselves (Wei, 2022).

One of the earliest papers to find evidence of overconfidence and information seeking in clinical settings was by Oskamp (1965). This study focused on clinical psychology and tasked participants with answering questions about a patient who may have been displaying signs of post-traumatic stress disorder. After receiving each set of new information, participants could revise their answers to all questions and report their new confidence. Oskamp found that with each new set of information, participants increased their confidence but did not significantly improve their accuracy. In fact, participants were less likely to change their answers as more information was provided. This showed that confidence could be linked to mere receipt of information and that participants were more confident than they should have been. 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 (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 processing of clinical features that favour the initial diagnosis (Leblanc, Brooks & Norman, 2022).

The relationship between confidence and information seeking is yet to be determined in the context of diagnosis. Hence, one aim is to investigate at information seeking in diagnostic decision making to look at differences in calibrations of confidence.

Expertise

One of the primary implications and practical applications of this research is to inform the training of medical students and novice medical professionals, especially in their communication of confidence and information gathering strategies with making diagnostic decisions. We must therefore look at the differences between novices and experts more broadly and then focus on novices and experts within healthcare specifically.

The differences between experts and novices were explored in the seminal work of Kruger and Dunning (1999). This paper looked at how individuals are aware of their own accuracy, which has been mentioned previously as being called metacognition. What the authors showed was a clear relationship between one's ability and awareness of their own ability. When participants were better at a task (test of humour, grammar and logic), they accurately estimated their own percentile rank amongst the other participants. Participants who performed worse (especially those placed in the bottom quartile) severely overestimated their own accuracy. This finding has been highly influential in researching the differences between experts' and novices' metacognitive abilities. This relationship has been similarly explored in osteopathic medical students asked to classify heart arrhythmias, though the effect was found to be curvilinear such that metacognition was best for participants in the range of 70-85% accuracy (Mann, 1993). While Kurger and Dunning had famously proclaimed that individuals can be 'unskilled and unaware of it', this work from Mann provided evidence for individuals who could also be 'skilled and unaware of it'. The relationship between confidence, accuracy and expertise/experience should hence be explored further in clinicians.

Expertise can be thought of as the progression from a superficial understanding of problems and an effortful problem solving procedure to a principled understanding made evident by automatic pattern recognition (Hoffman, 1998). The nature of expertise is difficult to pin down, namely because it is hard to capture in a meaningful way. A potential indicator could be the length of time that an individual has had experience with a task. However, one's competence at the task likely does not increase monotonically over time and this function may differ from person to person. This is where one might consider that there is some latent trait variable of 'natural aptitude' that affects the change in competence over time. In other words, those who are more 'naturally gifted' at a task may improve more quickly than others over time. However, it is unclear how this latent variable can factor into any given task. For instance, are some people better than others at perceptual tasks? It is difficult to imagine one to be an expert at perception. Hence, expertise may only make sense for tasks where learning is involved, as there is an expected change in competence over time. One could test domain knowledge as a means of testing expertise, but this prompts questions in situations when domain knowledge does not consistently correspond with task competence. Accounts of expertise also do not explain why experts may still differ in their decisions or judgements despite fairly similar levels of experience or domain knowledge.

Medical students have a large body of knowledge to absorb as part of their education. Part of their job with regards to diagnosing patients is to identify patterns in the presented symptoms and match them to the most likely condition that would cause these symptoms. As these students gain more practical experience and see more patients, they obviously become adept at this task. Experts are able to recall information more automatically, as well as having a better organised and integrated knowledge base (Persky & Robinson, 2017). More experienced clinicians also have a higher confidence in medical skills (e.g. administration of intravenous drugs, preparation of equipment for intubation, packed red blood cell transfusion) which correlates with the number of times they report performing these skills (Morgan & Cleave-Hogg, 2002). These findings in themselves are not surprising. However, looking at the differences in confidence and strategies around information seeking that novices and experts use could be useful for understanding what expertise actually means in healthcare. This is especially significant given that perceptions vary greatly between individuals in terms of what expertise actually entails, both among novices and experts (St Pierre & Nyce, 2020). Understanding how expertise translates into the use of information to reduce uncertainty and increase confidence can hence inform the training of newer medical students.

Experts realise that they can, or even are forced to, utilise either mental or organisational shortcuts when caring for a patient. When asked to interpret an electrocardiogram (a recording of electrical activity in the heart, ECG) and give a diagnosis for any abnormalities in the patient's heartbeat, Wood et al (2014) found that experts were quicker and more confident in the task than novices. In addition though, eye gaze data found that experts were quicker in identifying the critical points of interest in the ECG trace. This indicated that experts had more codified strategies for focusing on the most important pieces of visual information. This is backed up by Carrigan et al (2019), who found that expert radiologists were more sensitive to informative visual features of a chest radiograph (ie a lung nodule) that naïve observers did not view as salient. Experts are seemingly able to make inferences of high relevance more easily, which is distinct from their recall ability (Patel & Medley-Mark, 1986). There is also evidence that experts' confidence was sensitive to the consistency of information provided. Tabak, Bar-Tal and Cohen-Mansfield (1996) provided two patient scenarios to experienced nurses and nursing students, with one scenario presenting information about the patient fully consistent with a provided diagnosis and the other containing some information inconsistent with the provided diagnosis. The nurses reported lower confidence and higher subjective difficulty in the case of the inconsistent scenario, whilst the students showed little difference in both measures across the two scenarios. Brannon and Carson (2003) used scenarios adapted from Tabak, Bar-Tal and Cohen-Mansfield (1996) and found a similar effect for confidence across a larger sample size of nurses and nursing students. When considering classical works from psychology, the key properties of expert performance can be summarised as superior pattern recognition (Chase & Simon, 1973) and better forward reasoning from established facts (Larkin, McDermott, Simon & Simon, 1980). In the same way that confidence can lead to a higher influence in group decisions (Zarnoth & Sniezek, 1997), expertise may in itself have a high influence within a group decision. Salem-Schatz, Avorn and Soumerai (1990) found that 61% of surveyed resident doctors had ordered unnecessary transfusions at least once a month due to a suggestion to do so by a more senior physician.

Hence, studying the effects of clinical experience on diagnostic decision making is of interest when it comes to understanding implications for medical education and the relationship between confidence and information seeking may change as doctors become more experienced.

Study 1 - Information Seeking and Confidence in Diagnosis

Methods

Participants

Participants were recruited between July 11th 2022 and April 6th 2023. 85 medical students were recruited for this study, including 32 males, 52 females and 1 participant who self-reported as non-binary. The age ranged between 22-34 (M = 24.2). 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 using a mailing list that those within OxSTAR have access to in order to recruit medical students based in Oxford. In order to recruit from further afield, we were assisted by the Medical Schools Council, who distributed the study to students in other medical schools across the UK. 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).

Materials

Our first study involved the usage of patient vignettes. These are simulated patient cases that have been adapted from actual past cases. We used the bank of patient scenarios from Friedman's (2004) study as a foundation for our scenarios. However, it should be noted that these vignettes could not be used straight away as they were provided. These vignettes were developed by a team of researchers based in the US, meaning that certain medical terms (eg medication, tests etc) had to be 'translated' into the vernacular used by doctors based in the UK. This was done via consultation with researchers working with the OxSTAR Centre who were also practising medical staff and students within the NHS. There also may have been differences in vernacular 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 have experienced previous medical conditions. These years were updated to make sense for a contemporary patient. Whilst a sizable bank of vignettes were kindly provided to us 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. We therefore chose cases that involved medical conditions that medical students would be expected to know.

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 to familiarise the participants with the experimental procedure and the interface.

Procedure

The procedure of a single case (or 'trial') 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, the participant is shown a presenting complaint for a patient, which includes the patient's age and their main symptoms. An example of this is as follows: "patient is a 68 year old male presenting with fever and arthralgia". This information 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 participants. 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". There is a total of 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. We also emphasised during the instructions that participants should only request information that they believe will help them with diagnosing the patient. 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 do so by clicking the "Enter Differentials" button. At this point, they are 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. 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, 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, add differentials or 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 if they do not prune their list down to a single diagnosis. 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. This is different to previous papers as it takes the focus away from merely their confidence that they have the correct answer. 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

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. * *Proportion of Information Requests*: 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 by the number of possible tests available during that stage (which is the same for all cases). * *Number of Differentials*: we take 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). * *Subjective 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 have to be manually coded as correct or incorrect. Spelling errors or alternative names are not penalised. To calculate Accuracy, we first identify the correct differential if provided in the list and find the likelihood rating assigned to that differential. The highest possible value here would be 10 if the participant included the correct condition in their differentials and assigned it the maximum likelihood rating. If a correct differential is not provided, a value of 0 is assigned. 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): - TA: any inflammatory arteritis is accepted - UC: infectious colitis, ischemic colitis or diverticulitis are also accepted answers. - MTB: any TB or lymphoma type is accepted - AD: pulmonary embolism or coarctation of the aorta are also accepted. - GBS: cauda equina syndrome is also accepted - TTP: ITP or Meningitis are also accepted. * *Information Seeking Variance*: We compute a vector of length 29, which is made up of 0s and 1s where for each of the pieces 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 normalised vectors for all cases for a given participant are combined to produce a 29 x 6 matrix. We calculate the Euclidean distance between each row of the matrix (trial) using R's dist function (in the proxy package). The computation of all pairwise distances produces a 6 x 6 matrix where each trial is given a Euclidean distance value relative to every other trial. A lower distance value between two trials indicates that the information sought on those trials are similar to one another. In order to look at the similarity of information seeking across all six trials, we compute the variance (the standard deviation squared) of the participant's Euclidean distances. A lower variance 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 take each of the 29 pieces of information in turn and split all participant trials into two groups: those trials where that information was sought and trials where that information was not sought. For each group, we compute the proportion of trials where participants included a correct differential, and we then take the difference between these two values. A positive value difference would indicate that participants were 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' trials, we calculate the sum of information values for all information that the participant did seek based on these values and we then take the mean of these sums across trials. This gives an overall measure of how useful the information was that participants tended to seek. We avoid circularity in this measure via cross validation. As such, each participant's information values are derived by looking at differences in accuracy for all other participants.

Results

Firstly, we can look at how our dependent variables change over the course of a case by comparing at each of the three stages. We fit a linear model by using the stage as our independent variable and our key dependent variables. A key finding is that the number of differentials increases over the course of the stages. This indicates that students do not tend to narrow their differentials with more information, rather they broaden their differentials. Participants rarely used the option of removing differentials from their consideration, which could be attributed to how they have been taught to make diagnostic decisions. We found that accuracy and confidence increased across the stages in a manner that was well calibrated, unlike previous papers. 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.

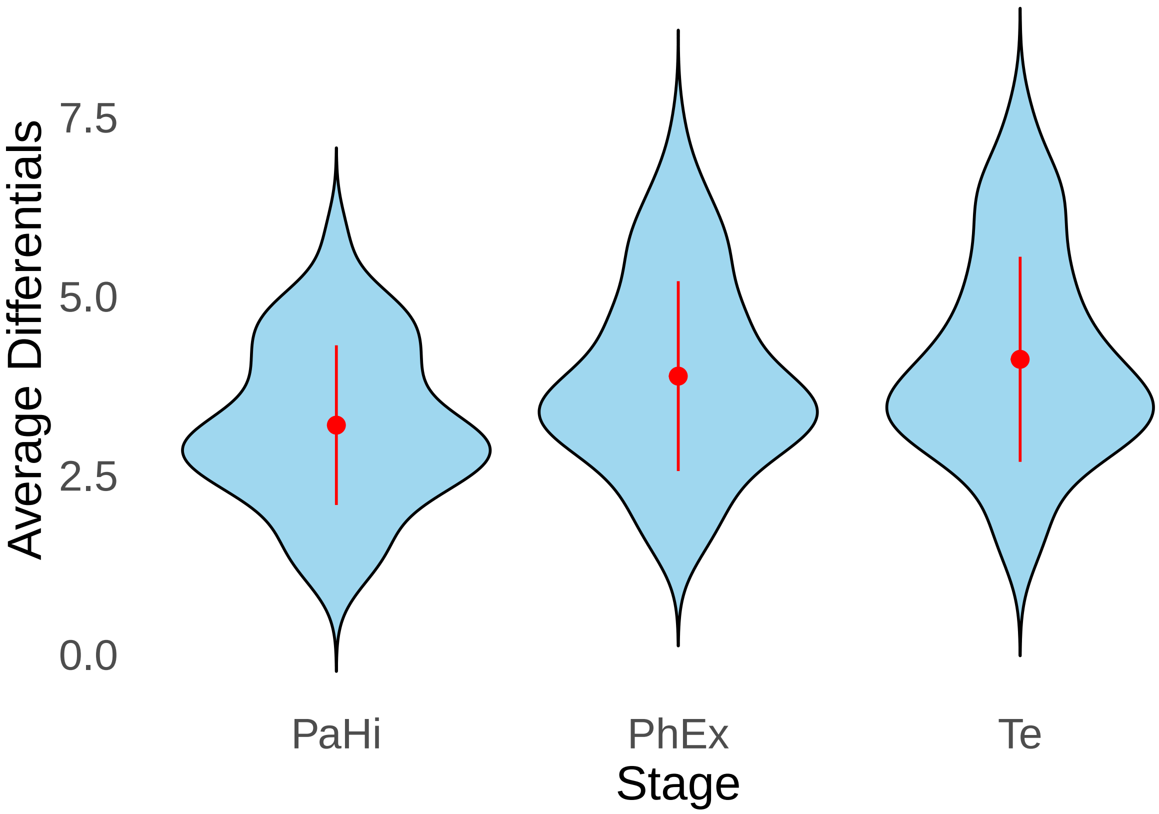
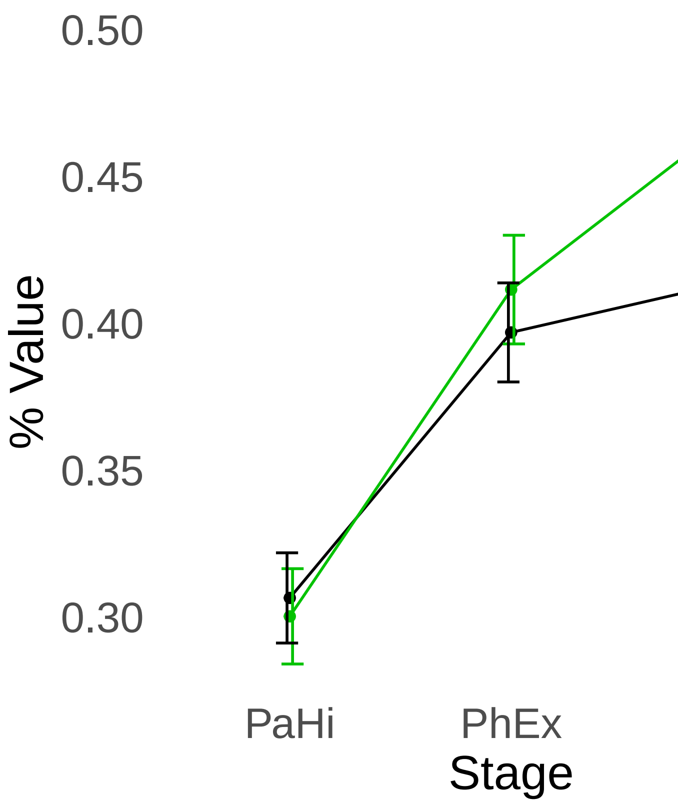


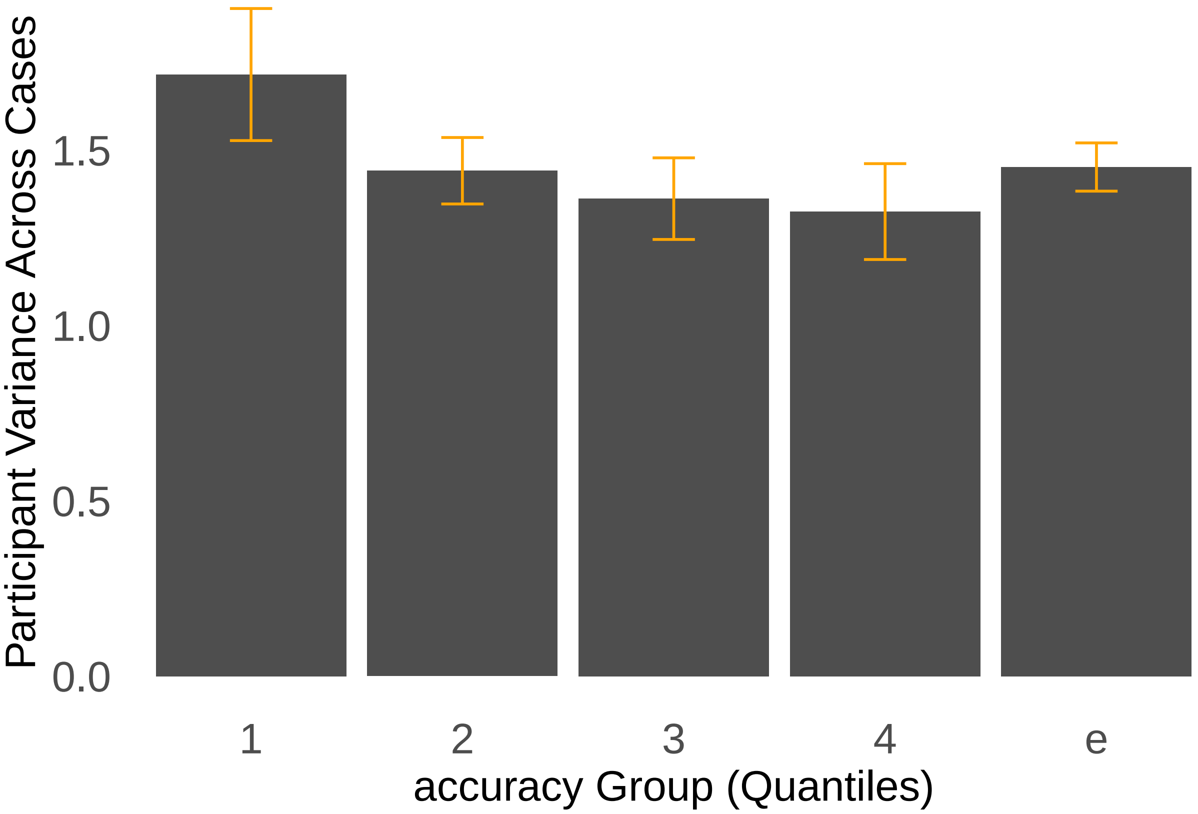
Figure 1: The average number of differentials after each stage of information seeking.

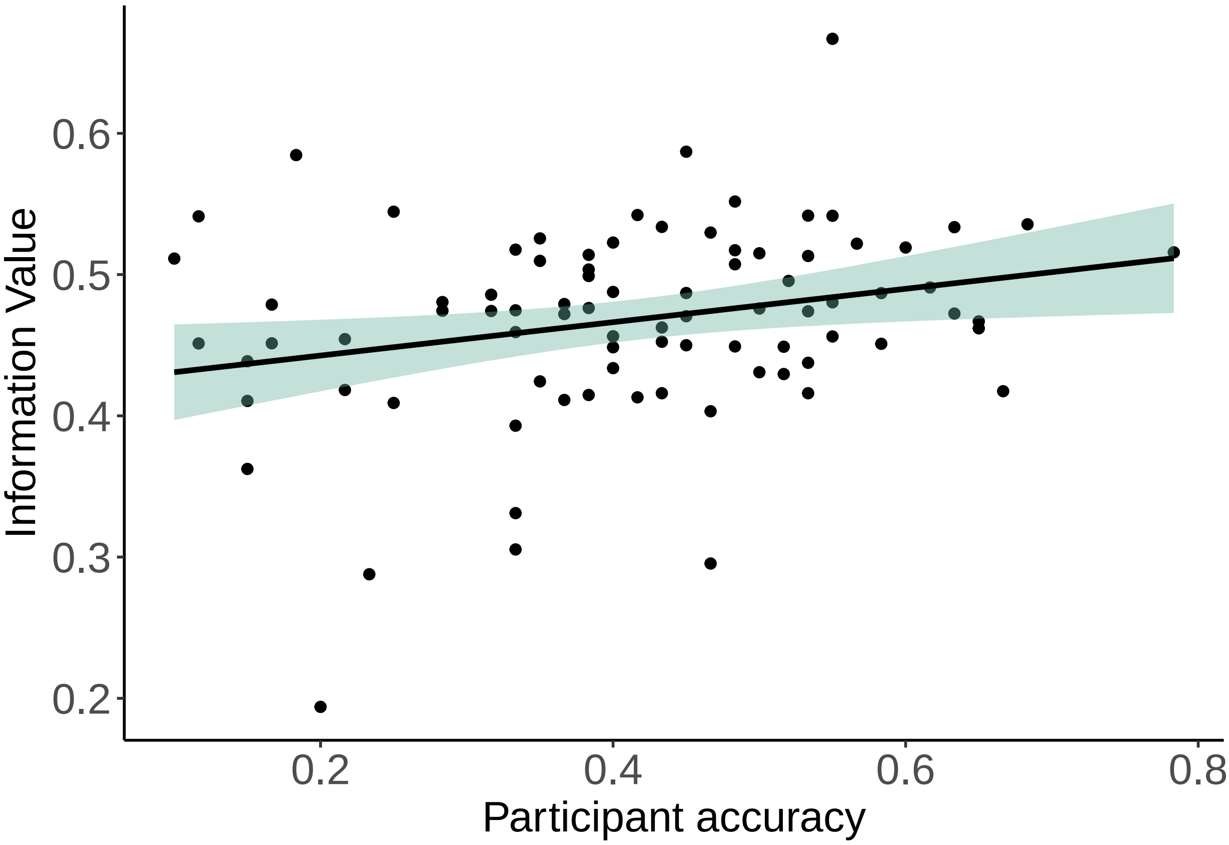
We found that accuracy and confidence increased across the stages in a manner that was well calibrated, unlike previous papers. We do find that confidence and accuracy does deviate during the final stage (Testing) such that confidence is higher on



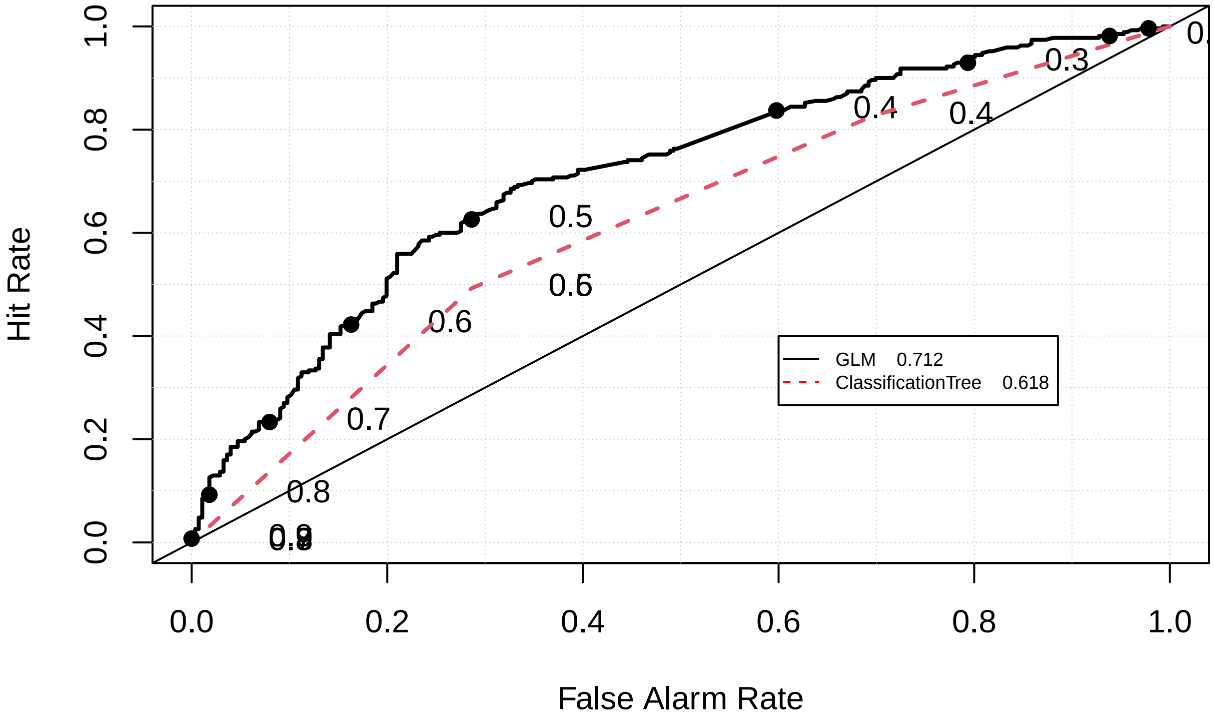
average than the true accuracy of the participants.

We then look at how ability on the task relates to information seeking behaviour. To do this, we calculated the average accuracy for each participant (across cases) and then sorted participants into four groups by quantiled accuracy. We then look at mean information seeking variance for each group of participants. We find that participants with higher overall accuracy have 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, seeking more similar information for each case when compared to students of a lower diagnostic ability. We can also test the same hypothesis by treating participant accuracy as a continuous measure, and we find evidence for a negative correlation between accuracy and information seeking variance (r(83) = -0.23, p = .04). 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). Taken together, 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. Meanwhile, participants with a lower diagnostic ability are not selective with their information seeking and hence do not seem to have a set framework or plan for what information to seek. We do also find that the proportion of available 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 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. 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.





To further investigate the differences in information seeking, we also trained a binary classification algorithm using a generalised logistic regression model. 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. For this, we do not take into account the specific case. 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 0.727 (with $p < .001$ when comparing the ROC curve to AUC = 0.5). This indicates that differences in information are indeed predictive of a difference in participant ability.



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 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 done for the participant.

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 staff and technology from another department. 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.

Within this discussion, it is worth mentioning a few general observations in the data and how they might inform the design of future studies of this nature. 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”. The fact that diagnostic differentials do not change that much between stages is supportive of this. Indeed, accuracy does not improve by a large amount between stages (from 52.2% after Patient History to 65.9% after Testing). It is indeed striking that in over half of all cases, students are able to include the correct condition in their differentials by the patient’s history alone. It is therefore worth considering whether there is a specific facet of diagnostic decisions whereby clinicians are taught not to disregard diagnostic possibilities easily. This also corresponds with participants tending to request most, if not all, information during the Patient History stage (86.1% across all participants) and then becoming more selective in information seeking during later stages. Hence, this indicates a general behaviour to gain the majority of diagnostic differentials from Patient History and to not easily disregard diagnoses.

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.

We should also note the manner in which accuracy was coded manually for each case. 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. This ties into one of the main challenges of designing these vignettes and this study: the set of information available for participants to request were chosen such that they were reasonable to be requested in any of the cases. The participants may have wanted to request more specialised, discriminant tests (e.g. lumbar puncture, biopsy), but including these could clue participants into the nature of the patient's condition. In addition, these types of highly specialised tests that target a specific type of diagnosis tend to take much longer to come back to doctors with results after they request them in a real healthcare setting. Hence, having results available at the touch of a button for these may seem unrealistic unless we alter the design to have patient cases unfold over a longer time period.

Study 2 - Investigating Confidence and Information Seeking in Non-Medical Decisions

In our follow-up study, we wanted to look at a similar style of decision making to Study 1 but in a non-medical context. We wanted to use a task that did not require the same level of specialised domain knowledge that medicine requires to see if similar findings hold. The structure of this task is the same as the medical task of Study 1, in that there are 3 stages of information seeking with 29 pieces of information in total that could be requested, interspersed with reporting of differentials and confidence. The aim of this task, referred henceforth as the Country Task, is for participants to identify a country based on a series of information that they can request about it. None of the information they can request is able to indicate the country on its own (e.g. participants are not told the capital city of the country). Hence, participants have to request and consolidate a decent amount of information in order to determine the country. The aim of this task is to observe whether patterns of information seeking and confidence extend to other domains outside of the specialised field of medicine.

Methods

Participants

The experiment was implemented using the same JSPsych code base as used in Study 1. Participants were recruited from the global general population using the online recruitment platform Prolific. Participant age ranged between 20-66 (M = 32.11, SD = 10.32). In total, 36 participants were recruited for the pilot of the study.

Inclusion criteria for this study stipulated that participants were aged between 18 and 70, were fluent in English and had normal-to-corrected vision. Participants were recruited between the 14th and 19th of October 2022.

We had used R's pwr.anova.test function to conduct a power analysis for a balanced one-way analysis of variance (ANOVA). We intended to compare high and low groups of geographic knowledge (median split by their score on the geography quiz). Hence, our power analysis was computed for two groups to detect an effect size f of 0.4 at 95% power and an alpha value of 0.05. This produced a sample size of 41.59 per group, meaning a total of around 84 across both groups. As a result, for our main study, we recruited 87 participants with ages from 20 to 66 (M = 33.87, SD = 11.79). Of these participants, 35 were female (40.2%). Inclusion criteria for this study stipulated that participants were aged between 18 and 70, were fluent in English and had normal-to-corrected vision. Participants were recruited between the 19th and 30th of January 2023.

Procedure

For this study, participants have to guess a country that is being described by a set of information. This information includes 'the number of colours on the country's flag', 'the country's area', 'native animals' and 'average temperature'. For any numerical values, such as area and population, these are provided as global ranks (e.g. a country might be the 13th largest in the world). This is because absolute values lack any real context on their own (e.g. an area of 30,000km2 is unlikely to be helpful information). The information is split into three stages: Geography, Economy & Politics and People & Culture. Across all three stages, there are 29 pieces of information in total that can be requested (the same amount as in the medical task of study 1). At the start of each trial, participants are told the continent in which the country presides. Participants are then free to gather as much information as they want within each stage. Once they are finished gathering information, they then report a list of all possible countries they are considering based on the current set of information they have seen. With each country, participants report how likely that country is on a scale from 1 to 10. Once this is done, participants rate their confidence that they have the current country included in their list on a scale from 0 to 100. In subsequent stages, the participants' list from the previous stage is shown for them to update. This includes the ability to remove countries from the list if they are no longer under consideration.

This study involved 6 trials, each with a true underlying country. These countries were: South Korea, Mongolia, Colombia, Switzerland, Greece and Botswana. The order in which the countries were presented was randomised for each participant. We also included a practice country (Thailand) to familiarise the participants with the experimental procedure and the interface. When coding responses for accuracy, any spelling mistakes were accepted as correct answers (e.g. 'Columbia' was accepted as a correct answer for the Colombia trial). We also accepted 'Korea' as a correct answer for the South Korea trial, as well as 'Mongol' for the Mongolia trial.

Before beginning the main experiment, participants were asked to complete a geography quiz that featured multiple choice questions. This allowed us to have a measure of expertise (via geographic knowledge) to look at confidence and information seeking against. A separate data collection was run to validate and create this quiz. An initial set of 40 multiple choice questions was first created, each with four possible answers (with only one being the correct answer). This included questions such as "What colour is the star on the flag of Ghana?", "Which volcano lies on the east coast of Sicily?' and "Which of the following countries is landlocked?" The aim was to use questions with clear answers whilst also tapping into a similar bed of knowledge that is used to complete the main task. To determine the validity of these questions, we collected and included responses from 99 participants to all 40 questions presented in a randomised order and with randomised order of multiple choices. This quiz was implemented using Qualtrics and participants were recruited using Prolific. Inclusion criteria for this study were that participants were aged 18 or over, were fluent in English and had normal-to-corrected vision. Each question was timed, such that participants were given only 10 seconds to read and answer the question with one of the four possible options.

If the participant failed to provide the answer in the allotted time, they would be timed out, no response would be recorded for that question and the quiz would move onto the next question. This time limit was to ensure that participants did not spend too long thinking about their answer (instead answering based on intuition) and also to reduce the participant's ability to search the answer to questions online. Amongst the 40 questions were two attention checks (again, randomly placed within the order of questions). One of these said the following: "it is important you pay attention to this study. Please answer Asia for this question" (with the other options between Europe, South America and Africa). The other attention check question said the following "it is important you pay attention to this study. Please do not provide any answer to this question and let the timer run out". For this second attention check, participants would pass the attention check if they did not click on any of the multiple choices provided (all of which were the same continents as the other attention check). If the participant failed either of these attention checks, their participation was immediately ceased and their data was not used.

With the responses to questions on the geography quiz where participants passed the attention checks, we scored each answer based on whether it was the correct answer or not. Based on this data, we can compute the total score for each participant. Out of a possible 40, participants scored between 11 and 34 (M = 21.34, SD = 4.96). We then computed discrimination index for individual questions such that we determine which questions are able to differentiate participants by their overall performance. We sought only to use questions that had a discrimination index of 0.2 or higher . Out of the 40 questions, 17 questions met this criterion. Hence, our geography quiz in the main study were these 17 questions, using the same randomisation and attention checks as in the initial data collection for the quiz. This final set included questions such as "The Angkor Wat is the largest religious monument in the world located in which country?", "What is the capital of Brazil?" and "What is the name of the microstate located between Spain and France?" Each participant included hence achieved a score based on the number of the correct responses in the quiz, with the maximum value being 17. Participants in the main study scored between YA and BLA (M =, SD =).

Recording performance on this geography quiz allows us to obtain a measure of the participants' latent geographic knowledge. We can then investigate whether task knowledge affects overall performance, metacognitive performance and information seeking patterns.

Results

Firstly, we again looked at our key dependent variables changed over the course of the three stages. We found similar patterns to the results from Study 1. When conducting a Welch's ANOVA, we found that participants increased their confidence with the stage on average (F(1.68,145.38) = 21.39, p < .001). We also found that accuracy (by the likelihood of the correct differential if it was present in the participant's list,) increased on average (F(1.40,121.62) = 54.27, p <.001). When measuring the relationship between the two directly, we found a significant correlation between the confidence provided at the final stage and likelihood score at the final stage (r(86) = .55, p < .001). This indicates overall that participants were well calibrated in their confidence judgements when compared with their objective performance. We also find a correlation between geographic knowledge and likelihood score (r(86) = .33, p < .001). In other words, participants with more geographic knowledge before the task performed better on the task. In addition, we found support for a positive association between the amount of information sought across a case and the change in confidence (r(86) = .22, p = .04). We also find evidence for a significant correlation between geographic knowledge and likelihood score (r(86) =.33, p < .001). Whilst there is a large degree of noise at zero (whereby participants who are incorrect on all six trials display a wide range of geographic knowledge), this points to a general trend where having higher geographic knowledge tends to result in better performance on the task. This at least suggests that the measure derived from the geography quiz displays construct validity with regards to the task itself.

We split participants into groups by their score on the geography quiz in two different ways: by using a median split (to create two groups of HIgh and Low knowledge participants) and by quantiles (to create four groups of participants). We also split cases into high and low difficulty by their objective accuracy across participants, with the high difficulty cases being Switzerland, Greece and Botswana.

A hypothesis that we had before the study, which was informed by previous work, is that one marker of expertise is a sensitivity to difficulty that is expressed via subjective confidence. To investigate this, we median split participants by their geographic knowledge and then 2x2 Mixed ANOVA was performed with confidence change (the difference between final and initial confidence) as a dependent variable. We found support for a main effect of case difficulty on confidence change (F(1,86) = 6.14, p = .01) but did not find support for a main effect of knowledge (F(1,86) = 1.76, p = .19). However, we find support for an interaction effect (F(1,86) = 4.81, p = .03). The presence of an interaction effect suggests a sensitivity to difficulty that is expressed via confidence which comes with geographic knowledge, whereby participants with lower knowledge express similar confidence for easy and hard trials.

We also sought to characterise information seeking patterns as a function of geographic knowledge. To do this, we first split participants into quartiles based on their geographic quiz score. Within each quartile, we then compute the variance in Euclidean distances between all trials' information seeking vectors within each quartile. Similar to the medical task, each trial contains 29 pieces of information that can be requested, so each vector is of length 29 where each value is a 1 if that information was requested and 0 if that information was not requested. A lower variance in Euclidean distances within a quartile indicates that participants within that quartile sought information in a more homogenous way (i.e. more similarly to one another). When looking at variance values, we find that a quadratic relationship such that participants with the highest and lowest geographic are more homogenous in their information seeking patterns. When looking at information seeking variance by country/trial, we do not find a consistent pattern for how trial difficulty (where countries are labelled as easy or hard by a median split on objective accuracy across participants) affects information seeking variance.

To investigate how knowledge affects information seeking in a systematic manner, we trained a binary classifier using a neural network algorithm to detect whether participants belong to the high or low geographic knowledge group. The predictors for our classifiers were binary variables on each trial for whether a particular piece of information was requested or not (29 predictors in total). We use all trials across participants for training and testing the classifier, resulting in a total of 528 trials (88 participants performing 6 trials each). To ensure that our classifiers are not overfitted to the data, we split trials into a training and testing set with 80% of trials used for training. We iterate this process of splitting the data and then calculate the average ROC AUC across 100 iterations. In each iteration, after splitting the data into training and testing sets, we again iteratively split the training set into sub-training and sub-test sets to maximise AUC by finding the optimal values for the number of hidden layers and weight decay, which are free tuning parameters in our algorithm. For hidden layers, we iterate between 1 to 10 hidden layers, with the average AUC computed across 5 iterations for each possible value for the number of hidden layers (at a weight decay of 0). For weight decay, we iterate through values of between 0.001 and 0.03 (in increments of 0.001), with the average AUC computed across 5 iterations for each possible value of weight decay (at a hidden layer of the optimal value computed in the previous step). Once the optimal values for hidden layers and weight decay are computed, these parameters are applied to classify the original test trials and an ROC is generated. The AUC of this ROC is recorded on each entire iteration and averaged across the 100 iterations. We find an average AUC of 0.940, which indicates that the classifier has high accuracy in differentiating between low and high geographic knowledge participants based on their information seeking patterns. In other words, geography expertise seems to broadly affect how participants seek information on the task.

To further investigate this finding, we then treat geographic knowledge score as an ordinal variable (taking integer values ranging from 0 to 17) and investigate whether information seeking predicts this score using Principal Components regression with Leave One Out validation. The aim of this process is, given our binary predictor variables, to find a number of linear combinations smaller than the number of total predictors and then use least squares to fit a linear regression model using the principal components as predictors. When computing the number of principal components to use, we visualise the cumulative RMSE across the number of components used. We find multiple 'elbows' in the curve, with the change in RMSE levelling off at 11 components. When testing on the same dataset using a PCR model with 11 components, we find a test RMSE of 4.60 when predicting the geographic knowledge score. Bearing in mind that this score has a range 0-17, this RMSE constitutes 27.08% of the entire range of geographic knowledge scores. This indicates a large degree of error in predicting score, suggesting that information seeking patterns are not precisely predictive of geographic knowledge.

We can also investigate how similar information seeking is across case by knowledge level through the training binary classifiers on the information seeing behaviour on one set of country trials (e.g. South Korea) and test the classifier on another set of country trials (e.g. Mongolia). This was done using the same neural networks method explained above, using the same procedure to derive the optimal parameter values for hidden layers and decay. When training and testing classifiers on every combination of country trials (excluding training and testing on the same country trials), we only find that two classifiers out of thirty produce an AUC of > 0.7 (values 0.74 and 0.71). This indicates that information seeking behaviour is not highly consistent within low and high knowledge participants across trials.

Discussion

It is worth comparing the country task to the diagnosis task to assess the extent to which they are analogous. Structurally, the tasks were designed to be close to each other as possible. In this respect, the number of stages and the amount of information is the same across both tasks. The tasks also both use the same code and interface, as well as the same experimental flow. There are a couple of changes to note from the procedure of Study 1 aside from the obvious change in visual stimuli shown. Firstly, participants only reported a likelihood rating for each differential, whereas medical students had to report both likelihood and severity ratings for each diagnostic differential. We also found from piloting our study that we had to incentivise information gathering. When piloting our medical task from Study 1, we found that medical students had a natural inclination to click on most if not all information given there was no cost to doing so. Hence, we implemented a time delay of 3 seconds in order to prompt participants to consider the information they chose to gather more carefully. When piloting the country task, we found the opposite issue: participants tended not to request information due to this delay. Hence, we shortened the delay between clicking on an information request and receiving the information from 3 seconds to 0.5 seconds. We also removed the 'Ready to Treat' button from the confidence screen, as this was not relevant for this task.

There are notable differences between the tasks. For the medical task, the available pieces of information are all relatively useful such that a participant could realistically request all the information available to make a diagnosis if time constraints were not present. However, for the country task, one could intuit some information to be more useful than others. This is of course subjective on our part in terms of how useful information is, but it is clear that some information is more uniquely discriminant than others. For example, knowing the number of colours in a country's flag is more likely to be useful than knowing the proportion of a country that is covered in forest.

Study 3 - Think-Aloud Study on Diagnostic Decisions

The main results from Study 1 were better diagnosis on our task was characterised by more standardised information seeking and that participants were increasing the number of differentials they were considering with more information. Both of these results were surprising and hence it had to be considered that the results were due to the nature of our specific task. 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. This taken together brings up the question of whether the observed results on reduced variance in information seeking being associated with accuracy was a result of our strict task structure. In addition, it was indeed 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. We wanted to hence see if signs of these results would be evident in the thought process of medical students. Are doctors 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 ran the same study (see that section for more information on the experimental procedure). 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:

“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 but interesting. The audio of the participant’s verbalisations was recorded and then transcribed. The screen of the experimental interface was also recorded, such that the audio could be corresponded to specific actions within the task.

The transcripts were coded for the following utterances:

Differential evaluation: any time that the participant mentions a condition (or set of conditions) that they are considering, ruling out or updating their likelihood of for a patient. 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. Uncertainty Expression: any time the participant expresses being unsure about their diagnosis or surprised by a piece of information (if it doesn’t fit their existing account of the patient’s condition).

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. Due to the richness of the qualitative data (and time required for transcription and coding), we recruited a smaller sample of participants (N = 10). Participants had to be 5th or 6th year medical students based in Oxford in order to participate.

Study 5 - Diagnostic Uncertainty and Information Seeking in Virtual Reality Paediatric Scenarios

A critique with the task used in Studies 1 and 4 are to do with its limits in naturalism. For a start, participants are unable to see the patient, which is important given that the visual state (or distress) of a patient can be informative for a doctor in diagnosing the patient. In addition, the task is static in time, in that the patient does not change over the course of a case (i.e. improving or deteriorating over time). The case also does not include any aspect of treatment of patients, where doctors can start managing the patient's symptoms and even using reactions to their treatment plan in order to change their understanding of the patient. In order to address these shortcomings in realism of our task, we used a virtual reality (VR) paradigm in order to investigate questions of differential evaluation, confidence and information seeking in a more naturalistic manner.

We used VR scenarios implemented by Oxford Medical Simulation (OMS), a company that uses VR for medical education and simulation, in their bespoke software. Participants in this study were medical students based in Oxford who were at the time taking part in VR-based teaching sessions as part of their medical degrees. Students performed the scenarios using Oculus Quest 2 VR headsets. Scenarios were based in paediatrics, meaning that the patients in the scenario were children who were attending the hospital with their legal guardian. Each scenario features a visual 3D implementation of a basic ward room in a hospital. Participants are shown a (child) patient, their guardian and a nurse who can help with certain treatment and testing. Participants are asked to diagnose, begin treatment and escalate the case to a senior with appropriate understanding of the patient. Whilst in the scenario, participants can learn about the patient's medical history, check key parameters (such as temperature, pulse, blood pressure, respiratory rate etc), perform physical exams/tests and begin certain treatment actions (such as administering oxygen or prescribing medication). Compared to the previous studies, participants have more freedom in terms of what information and tests they can request, as well as being able to begin a treatment plan.

Certain data, such as what information is requested, is recorded within the OMS software. However, for our other key data points, participants are asked to answer on paper. After 5 minutes in the scenario (by which point it is expected that participants would have a history of the patient and have started some early assessment of the patient), participants are asked to pause the scenario and fill in a brief questionnaire. Multiple VR participants were performing the scenario simultaneously and were paired with another student who would watch their performance. This other student would aid with administering the questionnaire. The VR participant was asked in the questionnaire to answer the follow (this is considered time point 1):

"Please say all the conditions that you are currently considering or are concerned about for this patient. For each, please rate how likely you think they are on a scale of 1 (low) to 5 (high)." "On a scale of 1-10, how confident are you that you understand the patient's condition?" "How severe do you think the patient's condition is on a scale of 1 to 10?"

The questionnaire was kept relatively short to minimise disruption to the scenario, as well as the questionnaire being administered whilst the participant kept the VR headset on with the scenario paused. This was due to the extra time that could be expended by asking participants to take off and put on the headset again to readjust to VR. Participants were given 20 minutes to complete the scenario, but could end the scenario early if they feel that they have completed the necessary care and tests for the patient. After completing the scenario, participants completed a second questionnaire on a separate sheet (this is considered time point 2). The second questionnaire featured the same three questions as the first questionnaire (see above), as well as the following questions:

“To what extent would you be prepared to leave the patient prior to a senior review” (this question was answered using a visual analogue scale) “Did you complete all the history, examinations and investigations necessary? If not, what else would you do if given more time?”

The dependent variables that we derive are as follows:

Performance: OMS implements a series of objectives for each scenario, which are tasks or actions that the participant is expected to have completed within the allotted time. This can include administering oxygen, prescribing a particular medication or calculating the Patient Early Warning Score (PEWS). The proportion of completed objectives is used as a score of the participant’s performance during the scenario. Latent Knowledge: medical students who participate in the study also complete a series of questions that test their knowledge of other paediatric conditions. Questions are multiple choice, with four possible answers for each. The proportion of correct answers for this questionnaire is used to derive a measure of the participants’ latent medical knowledge that is relevant to the scenarios performed in VR. Confidence Change: the participants’ confidence in their understanding of the patient’s condition is recorded at two time points, with the first being after 5 minutes (out of the 20 minute time limit) and the second being after the participant has finished the scenario. Confidence at each stage is recorded on a 10 point scale (1-10). The difference between the second and the first confidence rating is taken, such that a positive value indicates that the participant has increased their confidence over the course of the scenario. Number of Differentials: participants are asked to record all the diagnostic differentials that they are considering at the two aforementioned time points. Hence, the total number of differentials is recorded at each stage.

Appendices

A
The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

And here's another one from the same chapter, i.e. Chapter ??:

B
The Second Appendix, for Fun

References