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 paatis and my thathas

Acknowledgements

Having a PhD thesis written is a weird position to be in. On one hand, this is very much a new beginning in my career and my professional life. On the other hand, reaching this stage feels like the culmination of many years of work and life experiences. I still have a lot to learn, which is very humbling. But getting to the end of something like this really forces me to take stock of where I am at and where I want to be at in future. I'm grateful to have experienced this kind of personal growth over the course of this DPhil, and will take it with me going forward.

It took a village to make this thesis a reality. I would firstly like to thank my amazing supervisors, Nick and Helen, for your insights, enthusiasm and knowledge that helped shape this thesis into what it is. I have learnt so much from you both, and I will be forever grateful for that. Thank you for your patience and for being willing to help me on this journey and for just being good people. I could not have asked for better, more supportive supervisors.

I also want to thank everyone who was from the OxSTaR group, it was wonderful to be part of such a supportive academic and clinical community. Thank you to Wendy, Rosie, James, Olivia, Will and Aditi for all your support. I want to particularly highlight within the group who directly helped with my research, including Nathan, Kapil, Archie, Anil, Kaushila and Laura. Your inputs, feedback and help are so appreciated! And I would like to thank Mo for his hard work. I have also been lucky to be a part of the ACCLab, and to have met so many brilliant, smart folks. There are a lot of names to mention: Maja, Sarah, Aaron, the two Matts, Rebekka, Jasmine, Joshua. Thank you for your candour, your stimulating conversations and your friendship.

I am so grateful to have had friends in so many places I can rely on, especially for the (very needed) emotional support and for laughs! I could not have got this stage without such amazing support systems around me. In Oxford, I am thankful to have made friends for life. Thank you to The Crisps (Riddhi, Nilgoun and Samuel), Soyoung, Olly, Georgie, Malin. It's incredible to have such brilliant, funny and supportive people around me during this time. Outside of Oxford, I want to thank the Bath Lads (John, Matt, Robert, Gary) for your years of friendship and laughs, you mean a great deal to me.

And to the Foresters and Forester-adjacents! That means you Dom, Alessio, Dan, Ellen, Jesse, Connor, Will, Phil, Edwin, James, Nick and Tim! What a group to be a part of. It may not seem like it, but it's the little things that helped me a lot over the course of this PhD.

Last, and by no mean least, I want to thank my family. My Akka, my Amma and my Appa. I do not have words to express and go through everything that you done for me (that would be another whole thesis to write, and I think that the one thesis is enough). For now, I love you all so very much and you share in this achievement with me, alongside my thathas and paatis, who I am here because of. I miss them every day.

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
DKA	Diabetic Ketoacidosis
GBS	Guillain-Barre Syndrome
HD	Hypothetico-Deductive Reasoning
ICU	Intensive Care Unit
MTB	Miliary Tuberculosis
OMS	Oxford Medical Simulation
OSF	Open Science Framework
PhEx	Physical Examination
PaHi	Patient History
PR	Pattern Recognition
SI	Scheme-Induced Reasoning
TA	Temporal Arteritis
Te	Testing
TTP	Thrombotic Thrombocytopenic Purpura
UC	Ulcerative Colitis
VR	Virtual Reality

Chapter 1 - Introduction

Diagnosis and Error

Imagine a medical consultant within a hospital's intensive/critical care unit. They are asking a colleague for advice about a particular patient. The patient has 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. There has been a slow decrease in her blood pressure and blood oxygen saturation. The consultant is deciding what the most likely causes of this patient's symptoms are and how they may inform her future care/treatment. It is possible that the patient is suffering from pulmonary oedema, 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, when a lung becomes severely compressed. Alternatively, there could be a cardiac cause of the patient's condition. The consultant must integrate the information they have so far, align their individual mental models of the patient's condition (i.e. the root causes and contributing factors of a patient's symptoms and risks associated with them) with colleagues (e.g. nurses, specialists) and decide the following:

1. Do they have enough information to diagnose the patient's condition?
2. If not, what extra information do they need? Are there further tests that need to be performed?
3. What actions should they start taking to treat the patient given the most likely diagnosis?

One of the difficulties within this scenario is that the symptoms are indicative of multiple underlying conditions, which is a common occurrence in medical decision making. This example is illustrative of why many medical decisions are ‘ill-structured’ problems, in that they present several possible courses of action, and can produce disagreements between clinicians over both the current hypothesis for the patient’s condition and desired end goal (be they or short-term or long-term) for that patient’s care (**jonassen_instructional_1997**). During this thesis, we will investigate and aim to elucidate the cognitive mechanisms of medical diagnoses. Diagnosis is a core aspect of several medical subdisciplines and we choose it as an area of study for a few reasons. Firstly, accurate diagnosis is crucial to a patient’s care and treatment. Secondly, from a psychological standpoint, it allows for an extension of previous research on the relationship between information gathering and confidence to an ecologically valid, real-world setting. Finally, past work looking at diagnosis has not yet provided clarity on the causes of diagnostic errors (**van_den_berge_cognitive_2013**; **norman_etiology_2014**; **restrepo_annals_2020**).

Diagnosis in medicine has been defined as “the science…to distinguish one disease from another and trace symptoms to causes from which they spring” (**fenwick_medical_1891**). Identifying the condition that a patient has is important for guiding subsequent treatment. Accurate medical diagnosis is crucial to safe, high quality patient care and is a core part of a doctor’s job. Research on diagnosis has then been grounded in the incidence of errors in order to better ensure safer patient care. Looking into medical errors allows healthcare systems to learn from past mistakes and improve both technical and safety processes for future patients. For instance, the Healthcare Safety Investigation Branch (HSIB) has reviewed patient case studies to guide future diagnosis of aortic dissection (**hsib_investigation_2021**) and lung cancer (**hsib_investigation_2022**) among others, due to these case studies exemplifying delayed recognition of these conditions and their negative consequences for patients.

Past work has attempted to quantify the scale of diagnostic errors within healthcare. A report from the US Institute of Medicine 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_errors_2017**) found that up to 55.5% of patients experienced a diagnostic error. Other papers have estimated a lower incidence of diagnostic error: the Harvard Medical Practice Study found that diagnostic errors were responsible for 17% of adverse events (injuries/harm that were caused by medical management, rather than the underlying disease) (**kohn_errors_2000**), whilst the Canadian Adverse Events Study found this value to be 10.5% (**baker_canadian_2004**). The Quality in Australian Health Care Study found that 20% of adverse events were due to delayed diagnosis (**wilson_analysis_1999**). Even when using the most conservative of these estimates, the scale of the diagnostic error is substantial when extrapolated to the population of patients. Past work has hence attempted to find the sources of diagnostic error in order to identify how to reduce their prevalence. All in all, understanding the common sources of medical errors and adverse events can be extremely valuable for improving healthcare in the future.

With this scale of diagnostic errors in mind, another subset of past work has connected diagnostic errors to clinicians' diagnostic thinking, including how they weigh up diagnostic hypotheses (also known as differentials) against each other and how they seek information to support or rule out these hypotheses. Around 32% of clinical errors have been found to be caused by clinician suboptimal assessment, particularly the clinician's failure to weigh up competing diagnoses (**schiff_diagnostic_2009**). Another study estimated that 75% of diagnostic failures could be attributed to failures in clinicians' diagnostic thinking (**thammasitboon_diagnostic_2013**), including having inadequate knowledge, faulty gathering of information, and not thoroughly verifying information. In terms of their downstream consequences, diagnostic errors have also been found to lead to longer hospital stays and increased patient mortality (**hautz_diagnostic_2019**). In addition to longer stays in hospital, errors also manifest in differences to treatment for patients. Unnecessary treatment (or ‘overtreatment’) was estimated to cost the US healthcare system 158-226 billion dollars in 2011 (**berwick_eliminating_2012**). There has been increasing emphasis in the research literature on overtesting, such as requesting costly imaging scans when they may not be medically necessary (**carpenter_overtesting_2015**). **salem-schatz_influ** 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. Taken together, this set of literature illustrates the broad impact of diagnostic error and the amount of work that has gone into understanding its impact on patients.

Diagnostic error is by no means the sole cause of medical incidents. There are several factors tied to the wider work environment, culture and technology that can contribute to incidents and errors. Individuals involved in clinical decision making must frequently contend with an uncertain decision making environment, as well as time pressure and personal stresses (**yates_physician_2020**). However, by understanding the individual psychological factors that affect the diagnostic process, we better understand how sociotechnical and environmental factors may interact with and amplify individual contributing factors to diagnostic error. Gaining a greater understanding of the causes of diagnostic error can have important implications for future interventions within healthcare settings and improvement of patient care. Studying diagnosis also has added insights for the field of cognitive psychology, within which decision making has been studied in the past. Such insights can then be applied to the real-world context of medicine with the aim of improving diagnostic decisions. In the next two sections, we describe the extant psychology literature on decision making and how it has been previously applied to medical diagnoses.

Cognitive Biases in Diagnoses

One potential account of diagnostic errors is that they stem from cognitive biases during the diagnostic decision making process. Cognitive biases have been investigated substantially in past psychology research on judgement and decision making. Studies of biases have aimed to elucidate the ways in which the decisions we make are reliant on heuristics that can often contribute to suboptimal or erroneous decisions. These heuristics are believed to be used as 'shortcuts' to make quicker and more automatic decisions. This was originally posited by **kahneman_thinking_2011** as part of the dual-system theory of thinking, such that decisions are either made on a fast, intuitive level (System 1) or on a slower, systematic basis (System 2). Biases can include weighting information differently depending on whether they arrive earlier (primacy bias, **saetrevik_anchoring_2020**) or later (recency bias, **chapman_order_1996**) in a decision making process. They can also include judging a decision as better when it results in a success rather than failure ceteris paribus (outcome bias, (**baron_outcome_1988**; **aiyer_outcomes_2023**)) and judging members of an outgroup less favourably when compared to an ingroup (ingroup bias, (**tarrant_social_2012**)). The study of decision making biases has been broad and considered a large number of ways in which decision making deviates from what we might consider as optimal or rational.

The dual system theory of reasoning has been proposed as being applicable for diagnostic decisions (**croskerry_clinical_2009**). According to one review, 77% of studies in the allied health professions found evidence of a cognitive bias (**featherston_decision_2**
Whilst the biases in the previous paragraph have tended to be implicated within System 1 thinking, there are also error-prone approaches that arise from the more analytical, deliberate mode of System 2 thinking: reasoning from a false premise, having inadequate contextual knowledge and being paralysed with indecision are a few notable examples (**croskerry_deciding_2014**). Making a simple 'debiasing' intervention to aid accurate diagnoses is not a trivial task given that in many cases, de-

cision makers are not aware of their biases (**croskerry_mindless_2013**). It has been proposed that integrating education on cognitive biases within medical education would reduce diagnostic errors and improve patient safety (**royce_teaching_2019**). What complicates this picture however is the argument that the use of heuristics, that may seem biased or suboptimal, can actually be beneficial in many situations where environmental and cognitive constraints are placed upon the decision maker (**gigerenzer_why_2008**), which is common within healthcare. For example, a clinician who has less time with a patient due to other commitments may employ certain heuristics to expedite their decision making process. Hence, it is important to understand the situations in which medical decisions are biased in a manner that increases the risk of errors and situations in which these biases instead reflect good decision making.

While it seems intuitive that classical decision making biases affect those in healthcare too (**restrepo_annals_2020**), the empirical evidence of the impact for medical decision making is scant (**van_den_berge_cognitive_2013**). One example from dermatology found examples of satisficing bias (also known as premature closure, whereby clinicians arrive at an incorrect diagnosis too early and subsequently stop considering alternatives or seeking other information) and anchoring bias (whereby individuals are highly influenced by a reference point that other options for a decision are judged against), but few examples of other biases, such as availability bias (overweighting information or symptoms that correspond with a recent or memorable case from a clinician's experience) (**crowley_automated_2013**). These biases have been considered to be most applicable to medical decision making based on case studies and compelling anecdotal evidence (**groopman_how_2010**). For many of these biases, it can be challenging to establish a causal empirical relationship such that errors can be shown to be a result of a bias. For example, investigating availability bias requires researchers to simulate a patient situation that clinicians have experienced recently and then show that relating the current situation to that recent case is what caused a diagnostic error (if indeed the clinician does make an error).

The literature on cognitive biases is vast, with a lot of different biases named and defined. In an attempt to condense the large number of defined biases, **oeberst_toward_2023** categorised several biases from psychology research under a number of groups of beliefs about how individuals process information. For instance, a belief that one's opinion is shared by many others (false consensus effect, (**nickerson_how_1999**)) and the tendency to judge others as similar to oneself (social projection, (**robbins_social_2005**)) are both heuristics/biases driven by a common underlining belief: that one's own experience is a reasonable reference to extrapolate others. Similarly, tendencies to believe that successes come from one's self rather than external factors (self-serving bias, (**mullen_self-serving_1988**)) and that one's performance is better than others (better-than-average effect, (**alicke_better-than-average_2005**)) are both underscored by a belief that we are good or high performing at what we do. This latter belief is of particular interest to our present work, as it is related to the estimation of one's own ability. In other words, it can be considered as one's subjective 'confidence'. Confidence is important to study given that it can often not be matched to objective performance and underscores a large amount of medical practice, as we shall explore during the remainder of this section.

Confidence is important to consider within healthcare given the scarcity of clear feedback available to clinicians. In addition, in line with the implications of cognitive biases in diagnoses, one type of bias that has more consistently been revealed in experimental findings within medical decision making (when compared to other aforementioned decision making biases) is overconfidence (**berner_overconfidence_2008**; **meyer_physicians_2013**). These findings revealed clinicians to report themselves as being more confident in their decisions than they should be given their objective accuracy. When making a diagnosis for a patient, clinicians likely do not receive a lot of feedback about the correctness of their diagnosis. 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 cause of a patient's symptoms. 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 have differing levels of sensitivity and specificity rates, leading to false positives, false negatives or even inconclusive results. Clearer feedback may be available to the clinician based on how a patient's condition changes. A patient's reaction to treatment, and their rate of recovery, can be seen as a form of feedback. This in itself is imperfect however, as patients can deteriorate or improve due to circumstances outside of the doctor's control or awareness. This also requires a patient's condition to develop over time and cannot be known in the moment of a diagnosis. In the absence of clear feedback then, confidence is used as one of the only markers available for how likely someone is to be correct. We shall now explore how confidence has been studied within cognitive psychology and why it is very relevant to medical diagnoses.

Confidence and Miscalibration

Confidence can be defined as one's "subjective probability of their decision being correct" (**fleming_self-evaluation_2017**), and has been viewed as a way for humans to communicate their thought process to others (as per Kahneman's System 2 mode of thinking) (**shea_supra-personal_2014**). Individuals have been shown to be able to evaluate their own decisions without any objective feedback via judgements of confidence (**henmon_relation_1911**) and evaluations of their own accuracy (**rabbitt_three_1968**) that correlated well with objective performance. Confident individuals tend to be more influential with others in a group (**zarnoth_social_1997**) and can even causally increase the confidence of other observers (**cheng_social_2021**). This effect has been observed in mock jury trials, during which participants heard eyewitness testimonies presented with high confidence and then perceived those testimonies as more credible than testimonies provided with lower confidence (**cutler_eyewitness_1989**; **roediger_iii_curious_2012**). Confidence is a commonly used predictor of another person's accuracy, especially when feedback is not readily available on an individual's true accuracy. Confidence also varies across individuals with what may be considered a 'subjective fingerprint' (**ais_individual_2016**), meaning that individuals may be 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_effects_1982**). After a decision is made, we continue to process evidence (i.e. we continue to think about a decision after it has been made). Having 'second thoughts' or changes of mind are more likely with a lower level of confidence (**charles_dynamic_2019**). If an individual is systematically overconfident, they would be less likely to change their mind in the face of evidence that is contradictory to their beliefs (i.e. confirmation bias, **nickerson_confirmation_1998**).

What can be surmised here is that confidence is important for decision making, as it allows for an indicator of how likely a decision is to be correct. This is especially useful in contexts where objective feedback is scarce, such as in medicine. In the absence of objective feedback, confidence can be used as a marker of how likely someone is to be correct (**price_intuitive_2004**). Confidence is also important for deciding when a individual commits to a decision or whether more information is needed first before committing. Building on the current research landscape of confidence is then important within medical decision making. If there is an assumption that others will calibrate their confidence to their true accuracy, this would mean that heeding high confidence advice/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 specialty experience, a clinician's confidence is one of the only markers available for other clinicians and for patients during key medical decisions.

We refer to confidence as being 'calibrated' if it closely predicts objective accuracy (i.e., such that the individual is neither overconfident nor underconfident, see Figure 1.1 below). In experimental studies, confidence sometimes exhibits impressive calibration to objective accuracy (**boldt_shared_2015**), which is thought to reflect people's ability to evaluate the quality of evidence on which they base their decisions (**xue_challenging_2023**). But calibration is rarely perfect because confidence also depends on factors that do not directly correlate with accuracy, such as the time spent deliberating and the total amount of evidence considered (independent of the quality or consistency of this evidence) (**kiani_choice_2014**), as well as the mood (**rouault_psychiatric_2018**), personality (**schaefer_overconfidence_2004**), gender (**syzmanowicz_gender_2011**) and status (**see_detrimental_2011**) of the decision maker. Miscalibration of confidence come from a lack of deliberation over one's decisions and confidence, leading to an overreliance on intuitive decision making (**mata_metacognitive_2013**). Overconfidence has also been associated with insufficient consideration of reasons to choose alternative options/decisions (**koriat_reasons_1980**; **scherer_trust_2015**). The resulting under- and overconfidence matters: overconfident decision makers leap to premature conclusions and ignore useful information or advice, whilst underconfident decision makers waste time collecting evidence that will not improve their decisions (**desender_subjective_2018**). Especially in the absence of feedback, decision makers may develop systematically incorrect evaluations of their general performance (**rouault_forming_2019**) and their performance in comparison to their peers (**kruger_unskilled_1999**). Individual miscalibrations of confidence can also become amplified within groups. Effective decision making in groups depends on team members sharing calibrated information about their uncertainty: Confident team members tend to be listened to more, which can lead others astray if they are overconfident (**zarnoth_social_1997**). Conversely, underconfident team members may be ignored or may fail to share potentially useful information (**silver_wise_2021**).

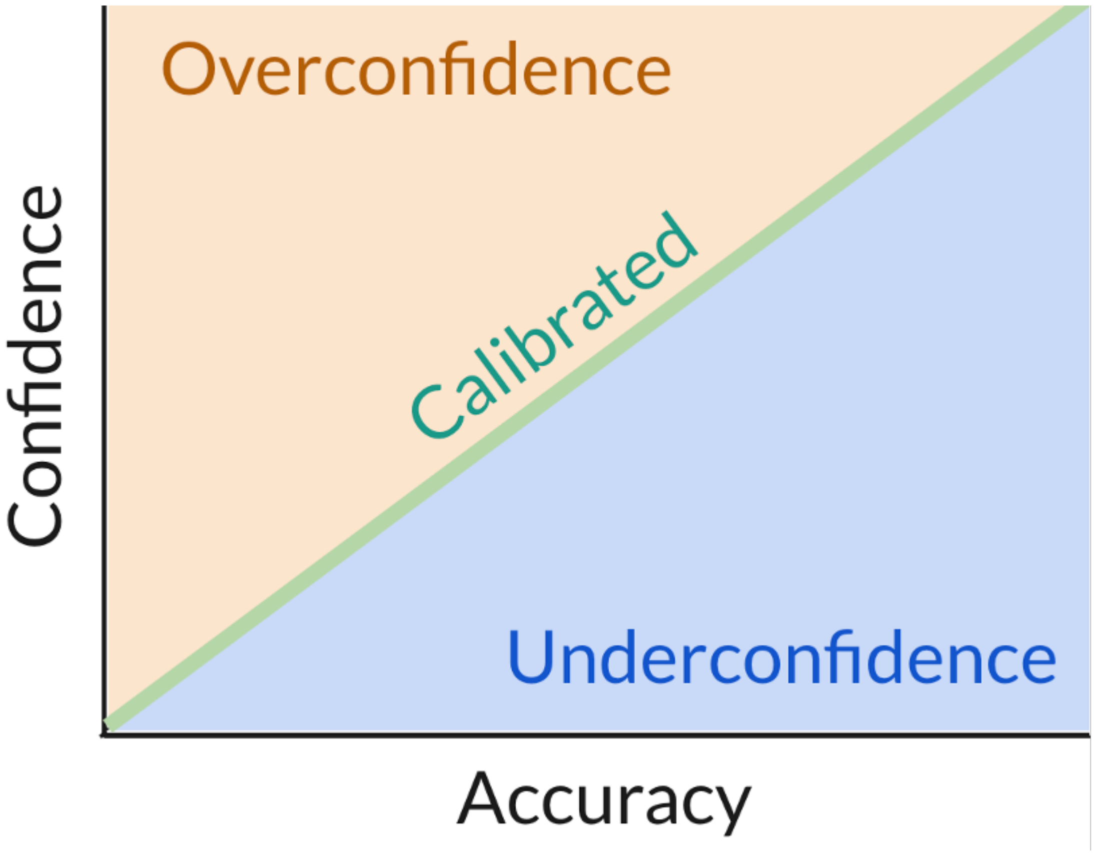


Figure 1.1: Visual representation of confidence calibration when comparing objective accuracy (x-axis) to subjective confidence (y-axis). Confidence is said to be calibrated when the two are relative equivalent (green line). Individuals are considered underconfident when their confidence is lower than their true accuracy (blue area) and overconfident when their confidence is higher than their true accuracy (orange area).

These findings of miscalibrated confidence are important to highlight specifically within healthcare, as overconfidence can lead to insufficient consideration of diagnostic alternatives and inadequate care in terms of seeking appropriate tests and treatment for patients (**kovacs_overconfident_2020**). In medicine, a lack of clearly communicated feedback can cause clinicians to proceed as if they have received positive feedback (also known as diagnostic momentum bias, **aron_diagnostic_2024**). Without clear feedback on whether their decisions are correct, clinicians may not adequately update their internal model of the patient and then increase their confidence inappropriately, whether working individually or in teams (**jaspan_improving_2022**). As we shall explore in the next chapter, the link between confidence and eventual patient care/treatment has been explored in past work, demonstrating the importance of confidence calibration of studying in medical decisions.

Some past work has explored instances of miscalibrated confidence specifically within the context of diagnoses. (**meyer_physicians_2013**) found that instances of overconfidence in physicians, even with the receipt of further information, were especially stark for difficult cases when objective accuracy was very low. In a task that involved diagnosing ultrasound scans, it was found that overconfidence was inversely associated with the amount of clinical experience that the clinicians/participants had (**schoenherr_subjective_2018**). However, it has also been found that underconfidence can be more prevalent than overconfidence, especially when comparing medical students to residents (**schoenherr_subjective_2018**). Similarly, (**yang_nurses_2010**) found that experienced nurses exhibited similar performance to nursing students, but were more confident in their judgements, resulting in differences in confidence calibration across experience levels. However, (**brezis_does_2019**) found that compared to students, experienced physicians were both more confident and less accurate at making a diagnosis for a paediatric case. Similarly, (**friedman_physicians_2005**) found that residents were overconfident in their diagnoses on 41% of cases, whilst students were overconfident on 25% of cases. As can be observed from this set of research, past work has tended to focus on drawing out the link between experience and confidence calibration. The overall finding that additional experience as a clinician does not lessen (and could even exacerbate) miscalibration of confidence shows that studying the mechanisms of diagnostic confidence would have benefits for clinicians across all levels of experience. This is especially pertinent in healthcare environments where more experienced clinicians tend to be listened to more. Highly confident members within a group could unknowingly reduce the chance of less confident (or less experienced) members speaking up about potential errors (**hemon_speaking_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_overconfident_2020**).

To summarise, confidence can be thought of as a readout of the evidence/information received in favour of a particular decision relative to the evidence against that decision. Past work has not only shown evidence for miscalibration of confidence (i.e. overconfidence or underconfidence), but also that such miscalibration has an impact on patient treatment. In sum, one can infer that the decoupling between confidence and accuracy is linked to the way in which evidence/information is sought or received. This is pertinent within medicine where doctors must synthesise a vast array of patient information (e.g. documents, test results, examinations etc.). In addition, the fact the papers covered in this section span different medical subdisiciplines and experience levels indicate the broad relevance of and interest in confidence calibration across the field of medical decision making. In the next section, we review the extant literature on the relationship between information seeking and confidence and how it may provide additional insight into how confidence can become miscalibrated during diagnoses.

Information Seeking and its link to Confidence

The way that individuals seek information is important, as it affects what information they are using to make their decisions. Information seeking is also thought to signal the importance of a task and quality of the source of the information among factors (**xu_who_2006**). Information seeking is an aspect of real-life medicine that should be considered: two clinicians confronted with the same patient case are likely to not use the same information to make a diagnosis if they seek different investigations/examinations. In addition, considering information seeking allow us to conceptualise decision making as an active, ongoing process where information is sought in response to previously seen information. For instance, an individual may be more likely to seek further information when they receive information that is contrary to, as opposed to being supportive of, their prior beliefs (**adams_reduction_1961**). Similarly, individuals with staunchly held beliefs have been found to be less likely to seek new information to refine their beliefs (**schulz_dogmatism_2020**). Information seeking patterns are also a signal of certainty or uncertainty, especially when a clinician has to decide whether the information they have is adequate (**gehlbach_illusion_2024**) to make a diagnosis or whether they need more information before coming to a decision. Seeking confirmatory information has been thought of as indicative of calibrated judgements of confidence when information is not processed in a biased way (**rollwage_confidence_2020**), such as weighting confirmatory information higher than corrective information (**schulz-hardt_biased_2000**). Taken together, information seeking as a research area from cognitive psychology has potential applications for a deeper study within medical diagnoses.

The link between confidence and information seeking has been previously investigated in cognitive psychology research. This association can be studied in two directions, looking at either how the evidence collected subsequently informs confidence or how confidence informs subsequent information seeking. In relation to the former of these, confidence is primarily determined by the accumulated strength of evidence in favour of a particular decision alternative relative to others (**vickers_effects_1982**). However, it has also been shown that the mere quantity of information, even if some information favours the non-preferred option, increases confidence in of itself (**ko_divergent_2022**). One potential reason that has been suggested for this is that individuals experience ‘decision inertia’, whereby they tend to maintain their previous choices regardless of the evidence presented against those choices (**akaishi_autonomous_2014**). Alternatively, past decisions made by an individual may bias any subsequent information sampling (i.e. ‘confirmation bias’, (**nickerson_confirmation_1998**)). The utilisation of confirmatory information has been found to be especially higher when presenting information sequentially rather than simultaneously (**jonas_confirmation_2001**), which can be the case in healthcare when it comes to requesting tests or examining the patient. Individuals have been shown to have a tendency to sample information that corresponds with a previous choice, with confidence increasing the extent to which information sampling is biased (**kaanders_humans_2022**). The qualities of the information received matters for confidence too however, with more variability in information being associated with lower confidence (**desender_subjective_2018**). Choosing when to stop gathering information has also been found to produce a 'boost' in confidence when compared to being forced to stop gathering information at a certain time (**wei_confidence_2021**). In summary, the process of seeking evidence is implicated in the confidence that individuals have in their decisions.

In relation to the latter directionality, such that confidence informs subsequent information seeking, confidence has been found to affect the accumulation of perceptual evidence **balsdon_confidence_2020**, as well as the amount of time spent viewing (**rausch_cognitive_2020**) and deliberating over evidence (**kiani_choice_2014**). Confidence also affects the type of evidence that is sought, as it has been found to be associated with a tendency to seek confirmatory evidence (**rollwage_confidence_2020**) and how likely incoming evidence is to change one's mind (**pescetelli_confidence_2021**). 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 relative strength of evidence). When taken together, confidence and information seeking are deeply connected during the decisional process. Given that the confidence that individuals have in their decision is based on the evidence evaluated either in favour or against that decision, systematic differences in information seeking and evaluation could be responsible for differences in confidence calibration. Hence, studying information seeking in the context of medical decisions can elucidate how clinicians' confidence can become decoupled from their objective accuracy. We now look at research on information seeking and confidence during medical decisions.

One of the earliest papers to investigate the link between overconfidence and information seeking in clinical settings was by **oskamp_overconfidence_1965**. This study focused specifically on clinical psychology and tasked participants with answering questions about a patient who had been displaying signs of post-traumatic stress disorder caused by the patient's army service. Participants received some information about this former soldier named Joseph Kidd and were asked 25 multiple choice questions about Kidd's past and predicted future behaviour. They finally reported their 'confidence' by estimating the percentage of questions they answered correctly, ranging from 20% (at chance) to 100% (all correct). Participants then received more information about Kidd in three subsequent stages, focusing on Kidd's childhood, his time in school and his time in the army. 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 (see Figure 1.2 below). This demonstrated that confidence could be linked to mere receipt of information and that participants were more confident than they should have been given their objective accuracy.

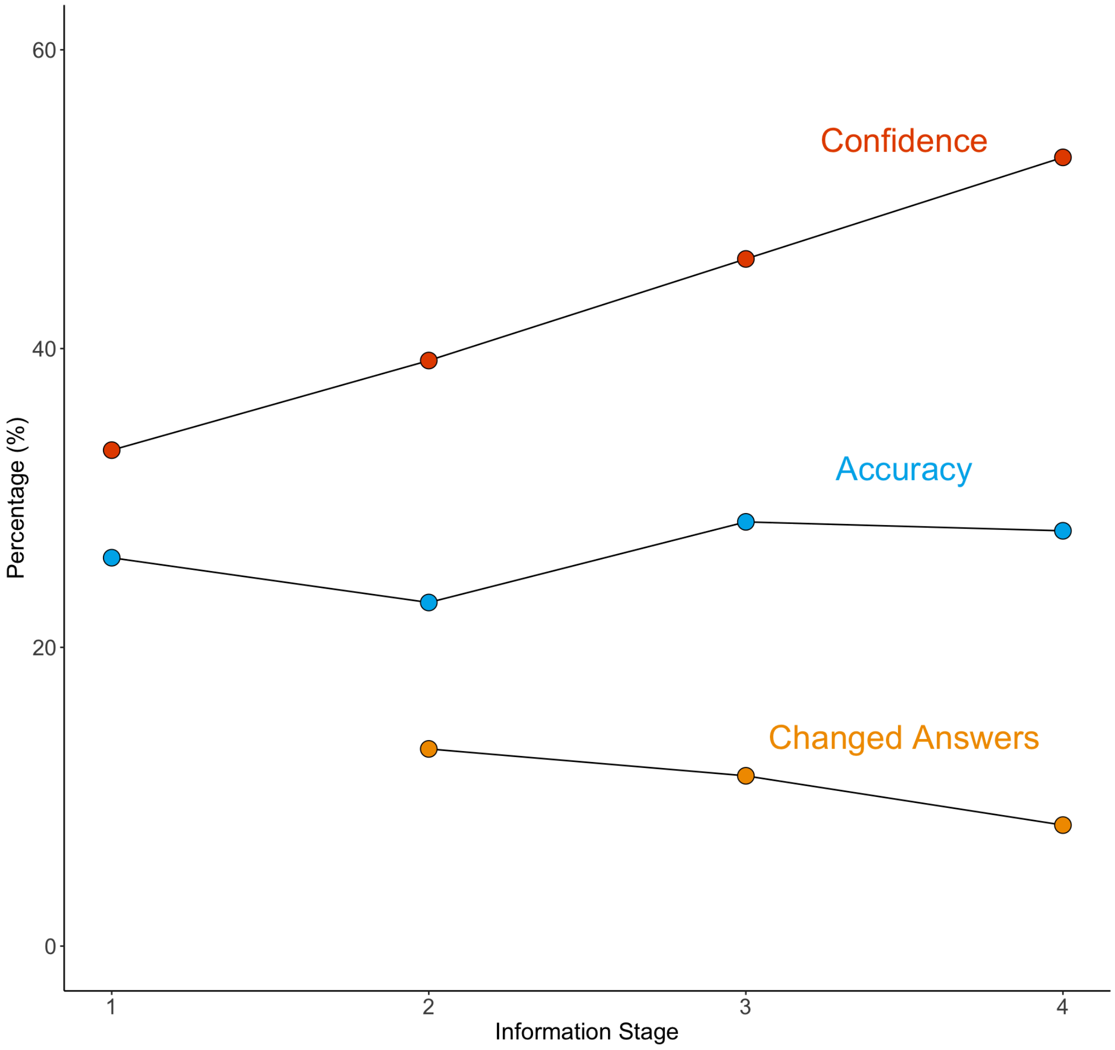


Figure 1.2: Graph representing the results from **oskamp__overconfidence__1965***, which was plotted using the reported data (the original paper did not have such a figure). The graph shows at each stage of information being provided about the patient (x-axis), the average confidence (y-axis, red), accuracy (y-axis, blue) and proportion of answers that were changed from the answers provided during the previous stage (y-axis, orange). This graph, showing that confidence increased with more information whilst accuracy remained at a similar level, is representative of findings from other papers (e.g.* **meyer__physicians__2013***) of overconfidence with the receipt of further information.*

Studying the interaction between confidence and information seeking in medical diagnosis has revealed similar interactions and imperfections. Notably, **gruppen_information_1991** found that clinicians were less confident when they had to seek relevant information for themselves compared to when all information was already provided, indicating that information seeking as a task is contributory to formulating diagnostic confidence. While this shows the relationship in one direction, past work has also viewed confidence as contributory to further information seeking. Pathologists with more calibrated confidence were found to request more information, such as second opinions or ancillary tests, when unconfident (and hence less accurate) in their judgements (**clayton_are_2023**). In a sample of 118 physicians who were presented with patient vignettes, it was found that higher confidence was associated with a decreased number of diagnostic tests being ordered, even if confidence and accuracy were largely decoupled/miscalibrated (**meyer_physicians_2013**).

It has also been observed previously that physicians may 'distort' neutral or inconclusive evidence to be interpreted as supporting prior beliefs (**kostopoulou_information_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 this initial diagnosis (**leblanc_believing_2002**). Together, these findings have implications for how clinicians may seek and integrate evidence when making diagnostic decisions and how patterns of receiving information could affect decision confidence and in turn confidence calibration.

Evaluation of Diagnostic Hypotheses

Information seeking has a clear goal during medical diagnoses: weighing up possible diagnoses. When making a diagnosis, clinicians generate hypotheses and then gather information to evaluate the likelihood of these hypotheses. They should ideally eliminate hypotheses from consideration only when it makes sense given the incoming evidence. By the same token, they should also not be attached to a hypothesis when there is overwhelming evidence to the contrary. One conclusion of **wason_failure_1960** was that individuals struggle to remove a hypothesis from consideration even if they receive evidence against it. Individuals may look to sample positive evidence for their generated hypotheses and then reject alternative hypotheses that were relatively undersampled (i.e. such that they did not seek as much information to support them) (**hunt_approach-induced_2016**). Past work on positive evidence bias predict confidence to be biased towards confirmatory evidence in favour of a chosen decision (**peters_perceptual_2017**). When taken together with the previous papers on information seeking, hypotheses are potentially evaluated based on the quantity of evidence, rather than the balance of evidence for and against. Understanding how individuals generally reason about a possible space of hypotheses is important for understanding how the reasoning process works differently for novices and experts, especially in a specialised domain such as medicine. We would argue then that the information seeking process feeds into hypothesis generation during diagnoses.

There are interesting questions here around how individuals consider and maintain multiple hypotheses at once. Past work looking at hypothesis consideration (**robinson_revision_1985**; **van_wallendael_tracing_1990**) has tended to show that when participants assign probabilities to each hypothesis in their 'list' that they are considering, probabilities are considered for each hypothesis in turn. This leads to situations where the sum of probabilities of all hypotheses exceeds 1, indicating a poor grasp of probability theory but perhaps a more realistic depiction of how individuals consider a set of hypotheses. Individuals may not be maintaining a set of hypotheses in their mind that they modify as they receive information, but instead focus on one at a time. Hypothesis generation and information seeking are linked together, as has been found in one study, in which individuals were more willing to integrate information that conflicted with a prior belief when they had already considered an alternative hypothesis that could explain said information (**vallee-tourangeau_role_2000**).

Understanding how individuals generally reason (and vary from each other in their approach) about a possible space of hypotheses is interesting for understanding how the clinical reasoning process works. We are particularly interested in how one's 'process of elimination' (if clinicians do practice this) affects confidence. Contrary to experiments that prompt a set of two or three alternatives (**meyer_physicians_2013**; **chartan_isolating_2019**; **kuper_mitigating_2024**), a lot of real-world decisions have a large set of potential options (which depends on the individual's task-specific knowledge in order to generate plausible hypotheses). In theory, individuals gather information in order to reduce the initial set of potential alternatives to a more manageable set (or in some cases, deciding on a single option). On the one hand, individuals with more domain knowledge will be able to generate a larger set of plausible alternatives (including more 'obscure' or lesser known options), but their knowledge also allows them to eliminate hypotheses earlier in a decision process based on less information. A question here is how the amount of hypothesis elimination is related to information seeking patterns and confidence. If an individual has a larger set of initial hypotheses, this means that the problem space is more complex and potentially harder to whittle down.

Diagnostic decisions have been thought of as 'ideal' when using the hypothetico-deductive process (**kuipers_causal_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_clinical_2019**). This account was challenged by **coderre_diagnostic_2003**, who found, via analysis of clinicians' verbal explanations as they worked through diagnostic cases, that more accurate diagnoses were based more on pattern recognition (matching observed information/symptoms to the most likely diagnosis), especially for more experienced clinicians. (**gilhooly_cognitive_1990**) proposed that both novice and expert clinicians used a mixture of reasoning strategies to arrive at a diagnosis. Either way, the bridge between confidence and information seeking could be considered as a function of the reasoning strategy utilised by clinicians. The reasoning strategy used impacts how many diagnoses are considered at once, how much information is needed to increase or decrease the number of diagnoses being considered and, in turn, how confidence changes as a function of the information received. For example, a clinician may consider many diagnoses to start with, require more information to eliminate the diagnoses being considered to form a more manageable set, and by eliminating more diagnoses, they increase their confidence to signal how much the information has refined their thinking. Diagnostic reasoning is currently taught using cognitive frameworks such as the surgical sieve (which prompts clinicians to individually consider each pathophysiological system in turn as a possible cause of the patient's condition) and the ABCDE mneumonic for patient assessment (Airway, Breathing, Circulation, Disability, Exposure). However, current education does not teach differences in reasoning strategies, whether strategies may vary meaningfully by case or by clinician and how these strategies have a downstream influence on the diagnostic process in terms of seeking information, generating differentials and formulating confidence. Making clinicians and medical students more aware of their own reasoning processes can be useful for addressing cognitive errors in seeking and integrating information (**nendaz_diagnostic_2012**).

Current Work

Based on the literature previously presented, we can surmise that there has been considerable work on understanding the sources of diagnostic error. This has led to key insights, such as cognitive biases being implicated in diagnostic decisions and the role of information seeking in diagnostic errors. However, past research has often used simplified diagnostic tasks that do not capture the complex interplay between information seeking and decision making that are evident in real-world decisions, both medical and otherwise. There is evidence from both psychology and medicine that this rich interplay is important to study. As we have explained, lab experiments within the field of psychology have found imperfections and biases in the link between information seeking and confidence, and there is evidence of similar biases in clinicians. There is also evidence that clinicians vary in the strategies they use to seek and integrate information within their consideration of diagnostic hypotheses. Taken together, this implies gaps in our understanding of how clinicians make diagnostic decisions on a cognitive level, which has implications for the development of cognitive interventions or educational/training resources on how to make accurate diagnostic decisions. We aim to fill these gaps with research that better represents the complexities of real diagnostic decisions, whilst also studying the cognitive mechanisms of how the decision making evolves over time and as more information is afforded to the clinician. In particular, we aim to study diagnosis not as a single decision but as a more continuous 'practice' of information interpretation, hypothesis evaluation and choosing when to commit to a particular course of action (**alby_diagnostic_2015**).

There is a need for the teaching and assessment of non-technical skills and human factors in healthcare (**higham_observer-based_2019**), which is currently not addressed in a widespread standardised manner in medical curricula (**greig_lack_2015**). Curricula within medicine also place little emphasis on how uncertainty is communicated and approached in medical decision making (Hall, 2002). This is especially important to note given that doctors can be reticent to express their uncertainty (**katz_why_1984**). Clinical experience may also be connected to risk aversion and further information seeking behaviour (**lawton_are_2019**), which offers an important avenue for future medical education. Uncertainty within medicine can stem from many sources, be they cognitive, emotional or behavioural (**han_varieties_2011**; **lee_towards_2021**), and the ongoing tolerance of uncertainty is an important skill that medical students develop but also is a source of stress for them (**hancock_tolerance_2020**; **stephens_exploring_2021**). If we can understand the mechanisms by which uncertainty arises during medical decision making, this can be beneficial both for patients and for clinicians.

This research aims to inform medical education on non-technical skills such as diagnostic reasoning, especially around evaluating diagnostic differentials and seeking information during the diagnosis process. This work also allows for the application of cognitive psychology to important real-world decisions, testing the ecological validity of the field's findings and using the field to better understand diagnosis as a complex decision making process. This has been considered as a necessary avenue for research:

"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_diagnosing_2014**)

Over the course of thesis, we aim to elucidate the cognitive mechanisms that underpin medical diagnoses. In particular, we aim to gain a better understanding of how clinicians seek and interpret the information they receive pertaining to patients and translate their understanding into a set of diagnoses and their subjective confidence. Clinical reasoning is taught as a non-technical skill to medical students, but the field is currently lacking the input from cognitive sciences on how to foster accurate decision making and appropriate expressions of uncertainty. By emulating the diagnostic decision process, we aim to formulate some practical implications for medical education and clinical practice. We also aim to evaluate the methods from cognitive psychology and their applicability to everyday decisions that have significant real-world impact.

Thesis Structure

In this chapter, we presented an overview of past work studying medical diagnoses. In particular, we underscored the need to study their cognitive mechanisms due to the larger incidence of diagnostic errors and their impact on patient care. We laid out research that sought to draw a causal link between cognitive biases and diagnostic errors, with a particular focus on overconfidence/underconfidence. We outlined the importance of confidence from a cognitive psychology perspective, particularly in the field of medicine where objective feedback that could correct biases in confidence is scarce. Finally, we explored the link between information seeking and confidence, which has been investigated in cognitive psychology but only to a limited degree in the context of medical diagnoses. To this end, we orient our current work on further elucidating this link in diagnostic decisions. The overarching hypothesis of our work is that miscalibrations of confidence (when compared to objective accuracy) are caused by clinicians' suboptimal information seeking. For instance, a clinician may cease seeking information and decide on a diagnosis too early (known as 'premature closure'). Alternatively, a clinician may preferentially seek more information to confirm their beliefs, inflating their subjective confidence. We also aim to understand how information seeking relates to the breadth of diagnoses being considered by clinicians. By developing our understanding of how information seeking relates to diagnostic uncertainty, we can present implications for future work on how to prompt appropriate information seeking, and in turn appropriate diagnostic confidence.

We now present the structure for the subsequent chapters in this thesis.

In Chapter 2, we present a systematic scoping review of the medical and psychological literature in which confidence or certainty has been studied within diagnostic decisions. Whilst some of the extant literature has been presented here, we use this review to describe the existing literature more thoroughly. This is especially pertinent given the broad relevance of diagnostic confidence across medicine alongside the siloed nature of literature within the various medical specialties, necessitating a more systematic approach to capturing the relevant literature. The main aim of this review is to look at whether past work has found evidence for calibration of confidence judgements provided during diagnostic decisions. We also look at how confidence predicts aspects of the patients' care pathway. Alongside these research aims, the review is used to map out the themes of the extant literature and types of methodologies used. We use the reviewed literature to propose a conceptual model for how diagnostic decisions are affected by decisional, clinician-based and contextual factors.

In Chapter 3, we present an online behavioural study where participants (medical students) freely sought information and provided diagnostic differentials at different stages during a series of patient vignettes. This study allows us to look at how diagnostic differentials and confidence are affected by patterns of information seeking. In particular, we introduce and investigate different aspects of information seeking, namely how much, how valuable and how variable the information that medical students seek is, with these aspects differentially informing diagnostic confidence and accuracy.

In Chapter 4, we present an in-person study using a similar vignette-based paradigm where medical students verbalised their thought process as they were making these diagnoses, with the aim to use these think-aloud utterances to classify different diagnostic reasoning strategies. We use this paradigm to investigate how reasoning strategies affect confidence and information seeking. This study also allows for a qualitative analysis of medical students' thought process whilst they are making diagnoses, including their main considerations regarding the evaluation of diagnostic hypotheses.

In Chapter 5, we present the third empirical study, where we investigate diagnostic decisions in a more naturalistic manner by using virtual reality paediatric scenarios to investigate differences in information seeking and confidence. This paradigm allows for the study of information seeking in an open-ended manner that is more akin to real medical practice. Given the higher fidelity of this paradigm, we also study how the ongoing treatment of patients is informed by confidence and information seeking.

In Chapter 6, we present a reflective chapter based on observations in both Adult Intensive Care and Emergency Medicine, whereby the findings from this DPhil are contextualised within the decisions made during actual medical practice. This chapter is used to better characterise the strengths and limitations of the studies within this DPhil by evaluating how well they relate to aspects of everyday medical decision making, as well as discussing real examples of where diagnostic uncertainty arises. This section acts as a precursor to our General Discussion section, such that evaluations of this DPhil's ecological validity and generalisability can be grounded in everyday medical practice.

In Chapter 7, we present an overall discussion of the studies conducted in the context of the current literature on diagnostic decision making. We aim to demonstrate how the cognitive mechanisms of confidence and information seeking impact medical diagnoses. We discuss the implications of this work for both the field of cognitive psychology and medicine (in particular, how our findings can inform future medical education given our focus on medical students). We finally evaluate the generalisability of our work and suggest directions for future research (within both cognitive psychology and medical decision making) and medical practice.

Chapter 2 - Systematic Scoping Review on Confidence and Certainty in Diagnoses

Introduction

There is extensive evidence of diagnostic error in most healthcare specialities. As we discussed in the Introduction section, overconfidence is an important source of medical error. It has been suggested that cognitive biases, such as overconfidence, are causally linked with these errors. To our knowledge, there has not been existing work to synthesise past research on diagnostic confidence. Such a review would enable us to understand the factors that contribute to clinicians' confidence in their diagnoses, as well as how diagnostic confidence affects treatment and patient care.

In this chapter, we present a systematic scoping review to collate and synthesise the existing literature studying diagnosis as a cognitive process. To our knowledge, this is the first scoping review with such a remit to include studies of confidence across medical subdisciplines, given its broad importance across medicine. The use of scoping review is suitable given this remit. We aimed to identify key determinants of confidence and characterise how judgements of confidence affect the wider medical decision making process. For the purposes of this review, and in common with practice in the reviewed articles, we treat "confidence" and "certainty" as interchangeable terms, while noting that the psychological literature discusses whether they are subtly different concepts (**pouget_confidence_2016**).

We systematically searched SCOPUS, MEDLINE, PsycINFO and Global Health.

Scoping Review

Articles were categorised according to methodology and clinical speciality. Findings were analysed thematically. Our review methodology adheres to the JBI's PRISMA-ScR Checklist for Scoping Reviews. We then performed citation tracking within these papers' references to identify additional articles. Papers were included if they reported quantitative results from an empirical study in which participants reported their confidence or certainty during a diagnostic decision. Studies comprised several medical subdisciplines. 79 articles met the inclusion criteria.

We organised articles under two main themes: the determinants of confidence and the uses of confidence during the patient's care pathway. Firstly, we find little evidence that clinicians' confidence is aligned to their diagnostic accuracy, even when using certain cognitive interventions or aids. Confidence is found to be affected by several factors including case complexity, early diagnostic differentials, and the healthcare environment. Factors that affect confidence, but not accuracy, demonstrate how the two can become decoupled, resulting in overconfidence/underconfidence. Secondly, confidence is predictive of actions in many parts of the patient care process, such as ordering investigations, referrals to specialists or prescribing, which may be suboptimal if confidence is miscalibrated.

Based on the literature from this review, we propose a theoretical model of factors that affect diagnostic confidence/certainty and accuracy throughout various stages of the patient care process. The proposed conceptual model highlights our current understanding of diagnostic confidence and how future research might focus on underexplored areas, particularly on group decisions, individual differences in confidence, and on the link between information seeking and confidence. Such a model can inform future work on how appropriate diagnostic confidence can be prompted and communicated amongst clinicians. Improving the calibration of confidence should hence be a priority for medical education and clinical practice (e.g., via decision aids), with this model acting as a starting point for future work to target specific factors that contribute to misclibrated confidence.

Methods

Search Strategy

Our review protocol was preregistered on the Open Science Framework: https://osf.io/wz5se. We conducted a systematic scoping review of empirical studies on confidence and certainty in medical diagnosis using JBI's PRISMA-ScR Checklist for Scoping Reviews (**peters_scoping_2024**). The search strategy was designed in cooperation with a subject specialist librarian at the University of Oxford's Bodleian Libraries group. The search string comprised keywords that captured the intersection of four elements: confidence/certainty, medical diagnoses, decision making and a study population of medical staff/students (i.e., clinicians, physicians, doctors and medics). The full search terms were as follows:

(clinicians OR physicians OR doctors OR medics) AND (confiden* OR uncertain* OR certain*) *AND* (diagnosis AND medical) *AND (decision OR (decision AND making) OR decision-making)*

The databases SCOPUS, MEDLINE, PsycINFO and Global Health were searched during February 2024. Finally, we hand-searched the citations of the included articles from these databases for further relevant articles via backward and forward scanning (**webster_analyzing_2002**; **tranfield_towards_2003**).

Study Selection

The inclusion criteria for screened studies were as follows:

1. Studies must report original empirical work with quantitative results.
2. Studies must be written in the English language.
3. Studies must use an experimental paradigm with medical diagnostic decisions
4. Confidence or certainty in diagnoses must be measured as a dependent variable

5. Studies can be from any medical discipline.
6. Editorials, review papers and opinion papers are all excluded

Identified articles were uploaded onto Rayyan (https://rayyan.ai/) to detect duplicate papers for manual checking and removal. This is the only part of the review process in which an automation tool is used. Data from the papers was collected using manual review. Research characteristics were derived iteratively and can be found on OSF (https://osf.io/4g6s3/).

Research Synthesis

Papers selected for review were first categorised by their broad research methodology (e.g., patient vignettes, in situ questionnaires, etc.) and their medical population of study (e.g., medical students, general practitioners/hospital physicians etc.). We reviewed the experimental procedures to extract their key manipulations and independent variables (e.g., case complexity, use of a cognitive intervention, level of medical experience). We also extracted dependent variables as they pertain to confidence/certainty and, where relevant, recording of both diagnostic differentials and information seeking. Each of the paper's key findings were summarised and then all findings were categorised under recurring themes.

Results

Findings of Scoping Review

The initial search returned a total of 3,332 articles. Applying the inclusion criteria identified 50 eligible articles. 439 further articles were retrieved for review from the included articles' citations. After applying both exclusions of duplicates and our inclusion criteria, 29 further articles were identified. This produced a total of 79 articles for inclusion and synthesis (see Figure 2.1 below for PRISMA diagram). The full set of papers can be found in Table S1 of the Supplemental Materials.

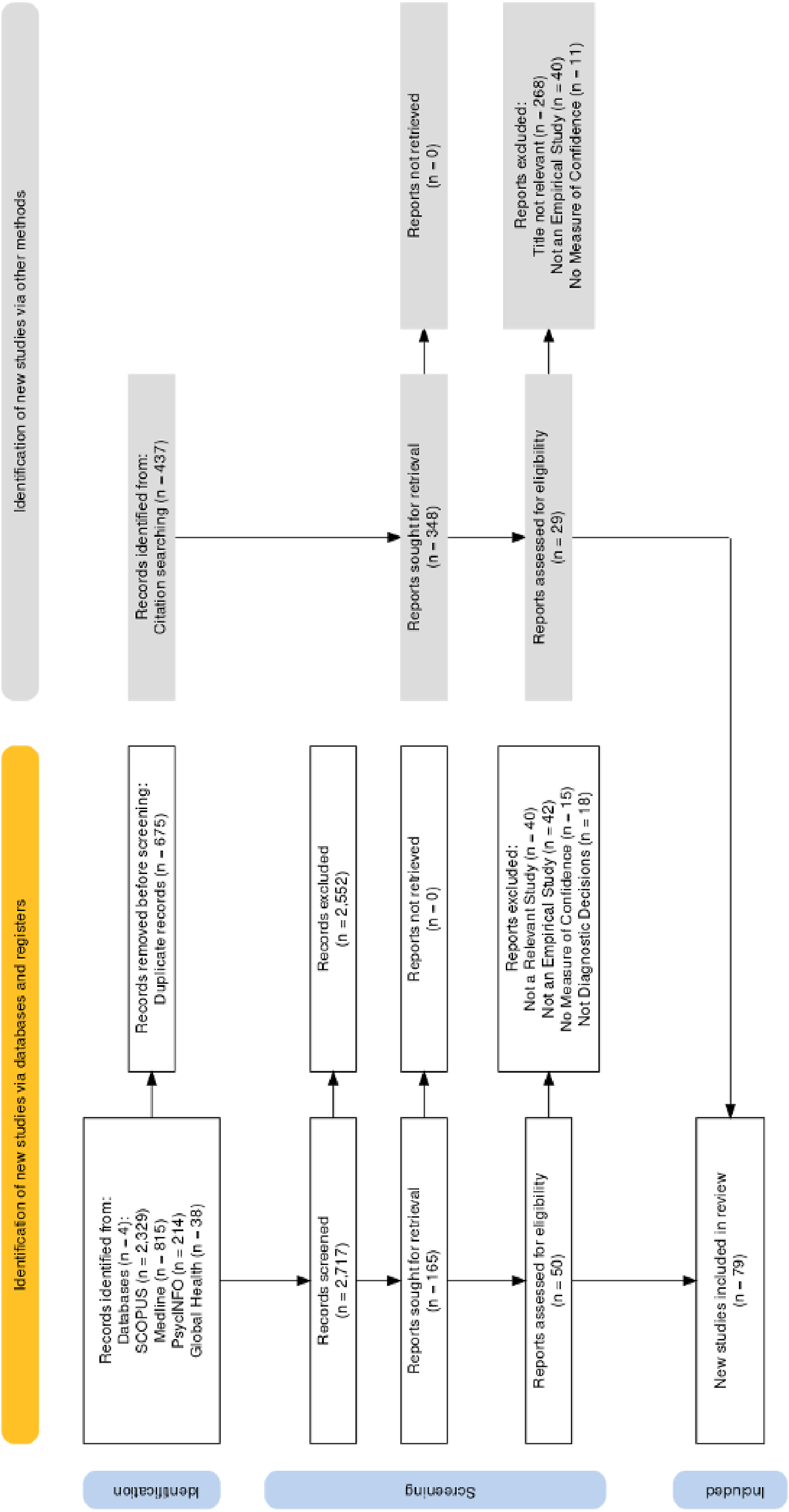
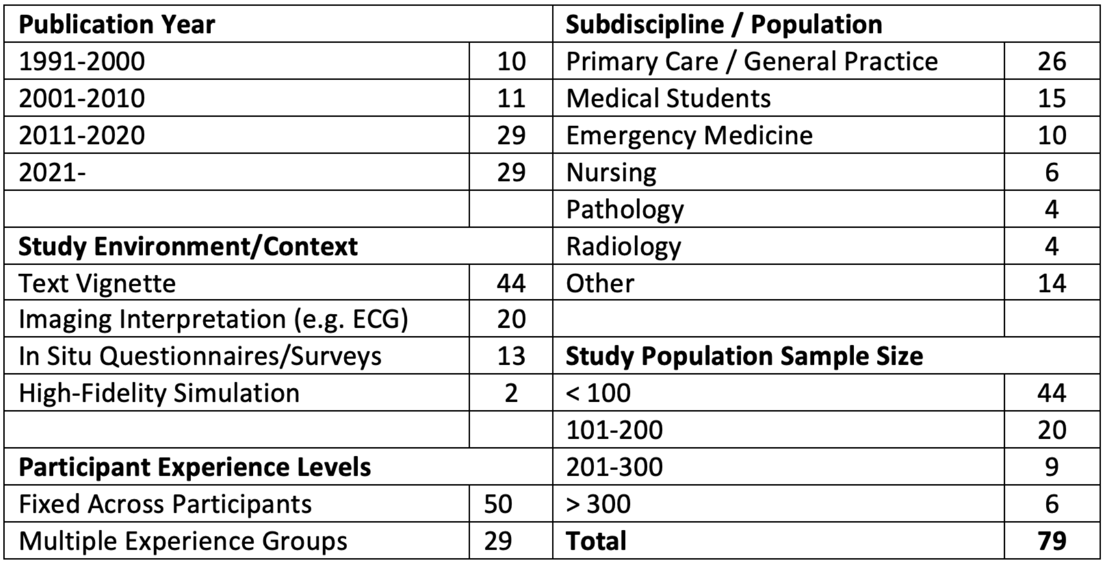

Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Records identified from:
Databases (n = 4):
SCOPUS (n = 2,329)
Medline (n = 815)
PsycINFO (n = 214)
Global Health (n = 38)
Records removed before screening:
Duplicate records (n = 675)
Records screened
(n = 2,717)
Records excluded
(n = 2,552)
Reports sought for retrieval
(n = 165)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 50)
Reports excluded:
Not a Relevant Study (n = 40)
Not an Empirical Study (n = 42)
No Measure of Confidence (n = 15)
Not Diagnostic Decisions (n = 18)
New studies included in review
(n = 79)
Records identified from:
Citation searching (n = 437)
Reports sought for retrieval
(n = 348)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 29)
Reports excluded:
Title not relevant (n = 268)
Not an Empirical Study (n = 40)
No Measure of Confidence (n = 11)

Figure 2.1: PRISMA Diagram of Literature Review.

Study Characteristics

Table 1.1 below summarises study characteristics and Figure 2.2 shows that 36 of the 79 studies (46%) were published since 2019, indicating a recent surge of research interest in this field and the timeliness of this scoping review. The studies appeared in 59 different publications, including both medical and psychological journals, with medical education journals being most common (19 studies). Clinical areas most represented were Primary Care/General Practice, Emergency Medicine and Nursing.



Publication Year		**Subdiscipline / Population**	
1991-2000	10	Primary Care / General Practice	26
2001-2010	11	Medical Students	15
2011-2020	29	Emergency Medicine	10
2021-	29	Nursing	6
		Pathology	4
Study Environment/Context		Radiology	4
Text Vignette	44	Other	14
Imaging Interpretation (e.g. ECG)	20		
In Situ Questionnaires/Surveys	13	**Study Population Sample Size**	
High-Fidelity Simulation	2	< 100	44
		101-200	20
Participant Experience Levels		201-300	9
Fixed Across Participants	50	> 300	6
Multiple Experience Groups	29	**Total**	**79**

Table 1.1: Characteristics of Included Studies, including year of publication, study environment used and medical population (recruiting single or multiple levels of participant experience, medical subdiscipline, sample size). A full list of all included papers can be found in the Supplemental Materials (Table S1).

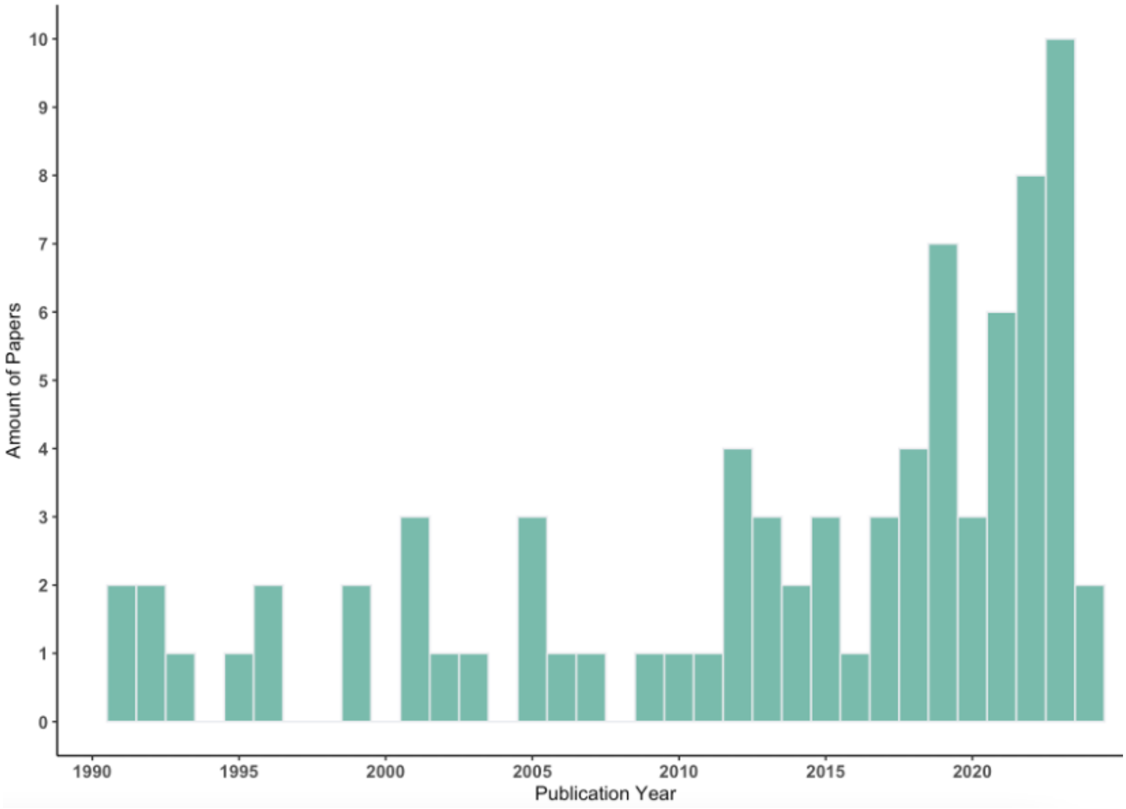


Figure 2.2: Distribution of Papers by Publication Year.

Study designs were split roughly evenly between focusing on how confidence varies across individuals (35 studies) and on how confidence varies according to features of the patient case (31 studies), with remaining (13 studies) studying both. Thirty-four studies (43%) looked at the level of medical experience or training's effect on confidence, either measured as a dependent variable or by recruiting participants in 'novice' versus 'experienced' group. Nineteen studies (24%) manipulated the complexity or difficulty of the patient case. Finally, ten studies (13%) investigated how diagnostic confidence varies with the information presented or the opportunity to seek information.

Most of the studies (44 (56%)) used clinical patient text vignettes. For vignettes, there is an established ground truth in each case (which may not be possible for in situ studies involving real patients) to compare the participants' confidence to their true accuracy in order to gauge calibration. Because vignettes are quick and simple to administer, participants can complete several diagnoses during a single study such that both their confidence and accuracy can be averaged across cases. Other experimental methodologies include the use of imaging (e.g., ECG, X-Rays, MRI) for diagnosis, high-fidelity simulations (either using extended reality tools or a patient mannequin), or questionnaires administered in situ to measure confidence during real patient cases as they are happening (Table 1). The preponderance of vignette studies is noteworthy given the finding from one study that nurses were both less accurate and less confident in a high-fidelity simulation compared to a paper-based vignette (**yang_effect_2012**), suggesting the need for caution when generalising experimental findings to how medical professionals behave in their everyday practice.

Studies varied in how confidence and diagnostic accuracy were assessed. Studies mostly used a self-reported scale for confidence (usually 1-10 or 1-100) as opposed to verbal expressions of confidence (e.g., “not sure” to “certain”) or visual analogue scales. The use of self-report numerical scales is common within cognitive psychology, where measured confidence values predict other behavioural indices of uncertainty, such as the tendency to seek further information or to opt out of making a decision (**gherman_neural_2015**). Twenty-four studies (30%) allowed participants to input multiple diagnostic differentials rather than a single diagnosis. Confidence is then either measured for each differential or in the set of differentials as a whole.

In terms of accuracy, most studies prompt clinicians for a single diagnosis that is marked as correct or incorrect. However, clinicians may consider multiple possible diagnoses in their everyday practice. Hence, 24 studies (30%) allowed participants to record multiple differentials in their diagnosis. This complicates scoring accuracy and confidence: If accuracy is operationalised as whether a correct diagnosis is included in this set of multiple differentials, clinicians are more likely to be correct with more differentials, and it may remain unclear how clinicians weigh up competing differentials. Hence, the operationalised calibration of confidence judgements is heavily contingent on how diagnoses are recorded.

Emerging Research Themes

Miscalibration of Confidence and Certainty Judgements to Objective Accuracy Calibration is assessed by comparing confidence ratings with objective diagnostic accuracy: When clinicians rate 100% (or 50% or 60%, etc.) certainty in their diagnosis, are they in fact correct 100% (or 50%, 60%, etc.) of the time? Calibration is then an indirect measure that is calculated by comparing two other observed measures: confidence and accuracy. In our study sample, there was limited evidence of calibrated confidence judgements, with some studies reporting underconfidence (**mann_relationship_1993**; **yang_effect_2012**; **brezis_does_2019**)and others overconfidence (**friedman_are_2001**; **fernandez-aguilar_use_** **garbayo_metacognitive_2023**). To examine these findings in more detail, we considered factors that impact and promote calibration in diagnoses.

The Impact of Experience on Calibration The first major theme of interest is whether calibration improves with experience. This was not always observed in the results (**yang_nurses_2010**; **clayton_are_2023**). However, experienced clinicians seem better able to identify when a case is more complex and adjust their confidence accordingly (**tabak_clinical_1996**; **brannon_nursing_2003**). Looking at the link between calibration and experience alone may be too simplistic, and there are other aspects of experience that influence diagnoses. Experienced clinicians were found to be less likely to 'distort' neutral information to be in support of their reported diagnoses, indicating a lower tendency toward confirmation bias (**kostopoulou_information_2012**). Past work has also suggested a distinction between experience (operationalised as years of experience or role seniority) and knowledge (measured using standardised tests of medical knowledge). In medical students, the calibration of confidence judgements were found to improve with years of education but not with medical knowledge (**hautz_diagnostic_2019**). Information 'distortion' was found to affect novice clinicians more (**kostopoulou_information_2012**) and lower knowledge was found to be related to higher susceptibility to irrelevant, distracting features of a patient (**mamede_role_2024**), However, the latter study from **mamede_role_2024** found that medical knowledge on the part of resident physicians was not directly associated with calibration.

The Impact of Contextual Factors on Calibration The second major theme relates to contextual and environmental factors. Studies have found that calibration is affected by the complexity or difficulty of the presented case (**meyer_physicians_2013**; **hausmann_sensitivity_2019**; **li_relationship_2023**). When confidence judgements are not sensitive to the difficulty or complexity of the case, confidence stays fairly constant for difficult cases whilst accuracy decreases, leading to increased overconfidence (and decreased calibration) (**meyer_physicians_2013**). In past studies, complexity is manipulated by either presenting patient cases with more comorbid conditions (**hausmann_sensitivity_2019**) or by showing conflicting information about the patient to indicate multiple possible conditions (**yang_nurses_2010**). Calibration can be improved by the presence of feedback during a training period (**kuhn_improving_2022**; **staal_impact_2024**).

Contextual factors that pertain to the situated medical environment can also affect overall levels of confidence (rather than calibration), as found using naturalistic paradigms. For example, clinicians may be constantly interrupted by other tasks (**soares_accuracy_2019**), especially during busier shifts where they have to manage more patients (**gupta_associations_2023**) and may not be present for the sharing of information during handovers (**bergl_factors_2024**). Studies that simulated these situations found they resulted in lower diagnostic confidence. However, such studies cannot assess the effect of contextual factors on confidence calibration with diagnostic accuracy because they were conducted in-situ, meaning that researchers do not yet have a ground truth of the patient's condition. At this stage, we can only determine how these contextual factors affect confidence, rather than calibration.

Imaging and Decision Support Systems A subset of papers found evidence for an increase in confidence when providing clinicians with specialised imaging for a patient to assist diagnoses, be they MRI scans (**mackenzie_magnetic_1996**; **albrechtsen_impact_2022**), CT scans (**abujudeh_abdominopelvic_2011**), evacuation proctography (**harvey_evacuation_1999**) or photos of wounds (**sanger_diagnosing_** Another subset of papers used various forms of computer-aided decision support systems with the goal of increasing confidence (**hillson_effects_1995**; **berner_influence_1999**; **dreiseitl_physicians_2005**; **neugebauer_clinical_2020**). **hillson_effects_1995** found that the adoption of diagnoses that were recommended by a computer-aided decision support system was not associated with an increase in confidence. **neugebauer_clinical_2020** did find evidence for such an association, however, with use of decision support leading to both increased confidence and increased diagnostic accuracy when compared to diagnoses made without using the system. On the other hand, both **berner_influence_1999** and **dreiseitl_physicians_2005** found that usage of decision support recommendations were associated with lower confidence when compared with decisions in which such recommendations were not utilised. Taken as a whole, whilst useful imaging increases confidence when available to clinicians, the efficacy of decision support systems at increasing confidence is likely dependent on other factors that require future work to elucidate.

Interventions at the Point of Generating Differentials The process of generating diagnostic differentials has been subject to experimental manipulations and interventions (such as early diagnostic suggestions) to investigate their effect on accuracy and confidence. This work is applicable, for instance, to understanding how a clinician transitions care of a patient to another clinician and gives a handover of relevant information. A general theme of this work is that there is a tendency toward higher weighting of early information. Early diagnostic suggestions have been found to be highly influential in the subsequent decision process where clinicians find these suggestions difficult to ignore and have more confidence in them (**kammer_differential_2021**; **kourtidis_influences_2022**). This also affects the breadth of differentials considered, with fewer differentials considered when provided with early suggestions (**kourtidis_influences_2022**; **staal_does_2022**) and an underweighting of differentials if they were considered later in the diagnostic process (**eva_influence_2001**). Interventions aimed at mitigating this tendency by asking clinicians to explicitly consider alternatives, increased their accuracy and calibration (**feyzi-behnagh_metacognitive_2014**), or prompting the consideration of the patient's 'red flags' in diagnoses, which increased confidence on simpler cases but not accuracy (**chartan_isolating_2019**). These interventions seem to require explicit instructions: Simply asking clinicians to reflect on their decision without guidance (**lambe_guided_2018**; **costa_filho_effects_2019**) or participate in an educational training course (**benvenuto-andrade_level_2006**; **kuhn_learning_2023**) does not seem to improve diagnostic accuracy and calibration.

Studies have also investigated how confidence is affected by the manner in which information is presented to clinicians during the diagnostic process. Higher confidence was found when clinicians were presented with additional patient information even when this did not carry diagnostic value (**heller_heuristics_1992**) and when given all available patient information rather than having to gather information themselves (**gruppen_information_1991**). Clinicians were also found to be more confident and more accurate when presented with an Electronic Health Record of the patient alongside other information (**ben-assuli_improving_2015**) and when presented with the patient history first rather than out of order (**tio_effect_2022**). This finding indicates that complete patient history available early on has a positive impact on confidence. However, an erroneous patient history has also been found to cue both novice and experienced clinicians to incorrect diagnoses whilst confidence remained relatively high, resulting in overconfidence (**fawver_seeing_2020**).

Uses of Confidence With more naturalistic studies, it is possible to isolate ways in which confidence is utilised within the wider diagnostic process, especially where healthcare involves transitions of care between multiple clinicians and departments. Past work has attempted to establish a link between confidence and further seeking of patient information and tests, with mixed results. US hospitalists (medical staff who provide care for patients specifically within US hospitals) with lower confidence were found to order more tests (**gupta_associations_2023**) whilst pathologists who were better calibrated (i.e., who tended to report confidence judgements that were closer to their true accuracy) were found to be more likely to request further tests when they were unsure (**clayton_are_2023**). Confidence has also been linked to prescribing medication, though overtreatment with unnecessary medications has been linked to both underconfidence (**levin_antimicrobial_2012**) and overconfidence (**garbayo_metacognitive_2023**). Higher confidence has also been linked to more referrals to specialists in other departments (**calman_variability_1992**) and to a lower willingness to admit mistakes (**brezis_does_2019**). One study found that whilst experienced clinicians were not more accurate in their initial diagnoses, they were more willing to change diagnoses and request more information (**krupat_avoiding_2017**). Lower confidence has been found to result in less specific diagnoses for patients in situ (**hageman_surgeon_2013**). Although psychology research on confidence has examined its role within groups (as discussed in the Introduction), only one article looked at confidence in group decisions in medicine. This study found that a multidisciplinary panel was more confident and better calibrated than a single clinician (**thorlacius-ussing_comparing_2021**).

Conceptual Model for Diagnostic Decisions

We synthesised the reviewed findings into a theoretical model (Figure 2.3) that illustrates how various factors distinctly impact diagnostic confidence and accuracy. This model aims to clarify existing research and identify directions for future work. The model starts by mapping out the stages of the diagnostic process (Figure 2.3, bottom panel). Based on initial patient presentation, clinicians gather and interpret patient information (e.g. history, examinations, tests) to inform their diagnosis of the patient's condition. The clinician's confidence in their diagnosis guides their judgment on when they have enough information to begin treatment versus whether further tests or additional information are needed. Once a diagnosis is reached, this guides patient treatment and care, the success of which is evident in the outcome for the patient.

The middle panel of Figure 2.3 characterises the cognitive processes of the clinician that determine the accuracy of the diagnosis and confidence with which the diagnosis is made. A key feature is that many factors have dissociable effects on accuracy vs. confidence. Diagnostic accuracy depends more on the level of medical knowledge and the quality of information gathering and interpretation; confidence depends more on years of medical experience and the quantity of information gathered (**kostopoulou__information__2012**; **hautz__accuracy__2019**; **mamede__role__2024**). Knowledge is improved through feedback on how a patient case was handled and its outcome, which in turn improves future diagnostic accuracy (though this feedback loop's impact on later confidence is yet to be explored).

The top panel of Figure 2.3 highlights factors pertaining to the medical environment/context. Separately from the patient case, confidence is reduced by time pressures (**yang__effect__2012**), interruptions to work (**soares__accuracy__2019**), busy shifts (**gupta__associations__2023**) and complex patient cases (either due to conflicting information or comorbidities) (**meyer__physicians__2013**; **hausmann__sensitivity__20** **li__relationship__2023**).

Figure 2.3 highlights three primary directions for future research. First, given the focus of research to date on diagnosis by individual clinicians, we recommend that future work also studies diagnoses in groups, given that diagnoses are often made by teams rather than individuals, particularly in secondary care settings. This is especially pertinent given the social influence that experience/seniority can have within a group: junior clinicians may be less likely to speak up about potential errors in the presence of more experienced clinicians (**hemon__speaking__2020**).

Second, future work should study individual differences on the part of clinicians to characterise how personality and trait level factors impact diagnostic confidence. In particular, factors such as personality (**schaefer_overconfidence_2004**), gender (**syzmanowicz_gender_2011**) and status (**see_detrimental_2011**) may impact a clinician's confidence in their diagnoses. Finally, we recommend future work investigate the association between the ongoing receipt of information and confidence. We recommend future work on the implications of diagnosis as a dynamic process where confidence and information seeking interact. Past work has tended to frame information seeking as a further action after diagnosis, rather than information seeking as a process that forms the diagnosis in the first place. Future work should prioritise examining how to prompt appropriate information seeking (i.e., neither overtesting nor undertesting) via educational tools or cognitive interventions.

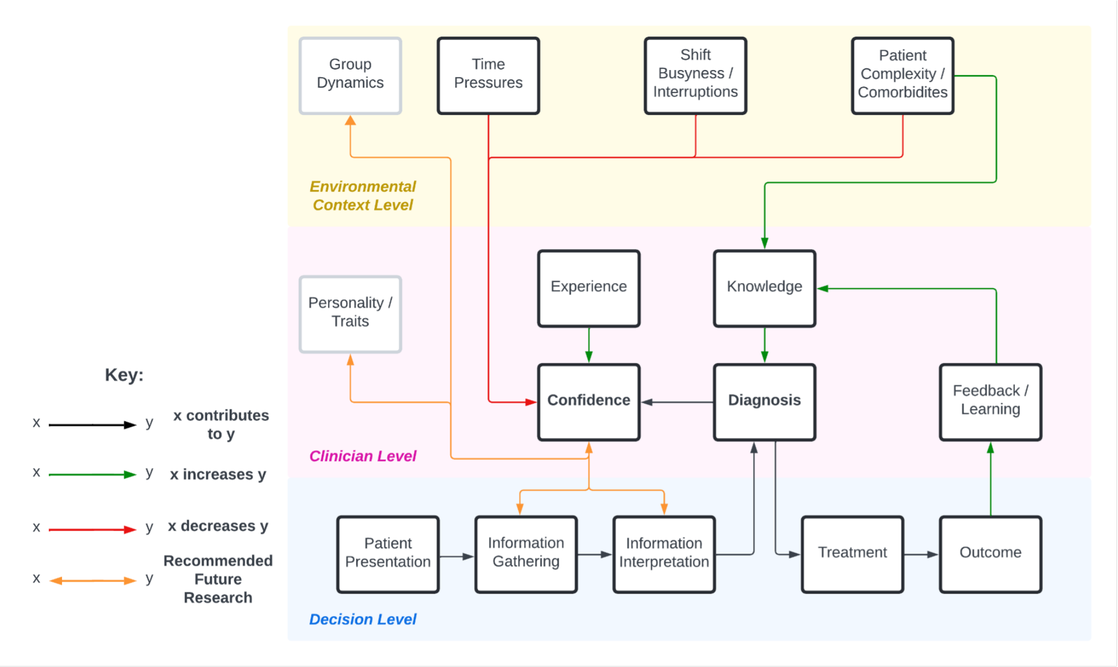


Figure 2.3: Conceptual model that depicts the various factors that impact the course of a diagnostic process, with links established between concepts based on findings from this systematic scoping review. Factors are categorised in three levels: the level of the diagnostic decision process (bottom box, blue, where the course of the

decision proceeds from left to right), the level of the clinician (middle box, pink) and the level of the environmental context within which the clinician operates (top box, yellow). Black arrows represent a progression from one concept to another. Green arrows indicate positive impacts between concepts; red arrows indicate the opposite (i.e. a negative relationship). Orange arrows represent links between concepts that are areas for future research. Light grey boxes represent factors that are known to affect decisions and confidence within the psychology literature but are currently less understood in the context of medical decisions.

Discussion

Review Findings

The present work comprehensively maps out the literature on confidence in medical diagnoses, thus extending previous work exploring how cognitive biases contribute to medical error (**kostopoulou_diagnostic_2008**; **graber_cognitive_2012**; **saposnik_cognitive_2016**) and characterising medical uncertainty more broadly (**hall_reviewing_2002**; **bhise_defining_2018**). This scoping review shows the importance of, and the recent surge in interest in, diagnostic confidence. Although confidence has been linked to diagnostic error in the past (**berner_overconfidence_2008**), full understanding will benefit from leveraging insights from cognitive psychology to inform medical education and practice (**wears_diagnosing_2014**). How clinicians evaluate their decisions contributes to their effectiveness: An overconfident clinician may overlook diagnostic possibilities, delay treatment or ignore crucial information. Conversely, an underconfident clinician may be less likely to speak up in a group about potential errors (**hemon_speaking_2020**).

Our review finds that confidence and accuracy are rarely aligned during diagnoses. Notably, miscalibration of confidence is not only a function of social and environmental factors: Miscalibration is consistently observed in vignette studies performed by individual participants, where decision makers tend to be overconfident particularly when dealing with complex cases. Nevertheless, social and environmental factors may amplify systematic tendencies toward misaligned confidence/certainty. Overconfidence is associated with overlooking differentials, ignoring important patient information and being less willing to admit mistakes. Hence, mitigating overconfidence is an important direction for future research. Underconfidence has received less attention, but is observed in medical trainees (**mann_relationship_1993**; **yang_effect_2012**; **brezis_does_2019**) and can lead to negative outcomes such as delayed treatment (**thompson_nurses_2004**) and ordering of unnecessary tests (**gupta_associations_2023**). Interventions have been tested to improve confidence calibration (such as considering alternative diagnoses and guided reflection), but these have not proven fully successful (**lambe_guided_2018**; **costa_filho_effects_2019**). More work is therefore needed to design interventions to improve calibration, as one can surmise here a link between miscalibrated confidence and suboptimal patient care. Findings from metacognition are already being used to inform educational practices outside medicine to improve students' memory retention (**dunlosky_strengthening_2013**; **putnam_optimizing_2016**). Although cognitive interventions such as considering alternative diagnoses and guided reflections have been tested, there is yet to be a standardised cognitive framework to teach non-technical skills such as constructive confrontation or expressions of uncertainty.

Our conceptual model of the diagnostic decision process reflects how different levels of factors (related to clinician and environmental context) differently impact accuracy and confidence. This model demonstrates the importance of both behavioural and work system factors within healthcare and how environmental aspects can inform an individual's decision process. Considering the work environment is important given our findings of lower confidence due to environmental factors such as shift busyness and time pressures. This corresponds with other findings of stress being associated with decreased confidence for intermediate levels of uncertainty (**heereman_stress_2011**) and this stress could be contributed to by the healthcare environment that the clinician operates in. Future interventions on diagnosis can refer to this model to understand the part and context of the decision process at which they are administering the intervention.

Beyond these key research themes, two further points emerge from the scope and variety of work identified in our review. First, our review highlights the broad relevance of confidence across different medical subdisciplines, suggesting the value of focusing on confidence calibration within medical education as a generally applicable approach to improve diagnostic decisions. Second, in terms of methodology, we find that confidence and certainty have been studied in a variety of ways (e.g. using 'assessments' or 'interpretations' as well as diagnostic decisions), but can sometimes be seen a primary outcome measure to improve. Increasing the confidence of clinicians without considering their objective accuracy may exacerbate instances of overconfidence. Future work should focus instead on prompting calibrated rather than increased confidence given the aforementioned impacts on patient care. As objective feedback is often unattainable in medical settings, confidence could be studied using methodologies like virtual reality and high-fidelity simulations that better emulate real medical settings when compared to vignette methodologies whilst also having markers of objective accuracy.

Implications for Future Clinical Research

The current healthcare context presents additional challenges to clinicians with substantial increases in clinical workload in the aftermath of the pandemic, and workforce stress and burnout at their highest in the NHS since recording began (**noauthor_state_2024**). Support to enhance clinical decision making through improved confidence/accuracy calibration could help to relieve pressure on the frontline.

Through our conceptual model of the diagnostic process, we identify three primary areas for future work that have been relatively underexplored in the extant literature. First, the majority of the studies we reviewed have studied diagnosis as a linear process in which information is presented sequentially prior to a final diagnosis, at which point confidence is assessed. This is a helpful idealisation of a process that is often more complex and dynamic in practice, with diagnosis evolving via back-and-forth between seeking information and evaluating that information in the context of currently considered diagnostic possibilities. Moreover, patients' unexpected responses to initial treatment can lead to revision of an initial diagnosis and prompt further information seeking. Effects of (miscalibrated) confidence are likely to be amplified within these dynamics, such as an overconfident clinician paying too much attention to evidence supporting their diagnosis and neglecting opposing evidence, resulting in greater overconfidence. These dynamics could be studied in more naturalistic, in situ methodologies that are closer to everyday medical practice. Interrupting clinicians in real time to report their diagnostic thinking can be a distraction and potentially a patient safety risk. However, some methodologies permit capturing diagnostic reasoning as it evolves with time and the receipt of new information, such as asking clinicians to think aloud as they make diagnoses (**arocha_novice_1995**; **coderre_diagnostic_2003**) or using a visual representation of clinicians' thought processes to capture paths and sources of diagnoses (**feyzi-behnagh_metacognitive_2014**). The use of high-fidelity or virtual reality simulations may also be useful for emulating the pressure and work environment of the clinician (which may affect decision making) (**schmidt_simulation_2013**; **jans_examining_2023**), as well as providing an actual 'patient' to observe (unlike in textual vignettes). Use of such paradigms would also improve the generalisability of results.

Secondly, the vast majority of studies have investigated confidence from the perspective of individual clinicians. However, diagnosis and treatment decisions are often made by teams rather than individuals, particularly in secondary care settings. Evidence from organisational psychology indicates that group decisions depend critically on communicated confidence and uncertainty (**silver_wise_2021**): Overconfident team members can anchor a group on an incorrect decision (**mahmoodi_equality_2015**) Conversely, underconfident team members may fail to share important information that is unknown to the rest of the group, exacerbating the problems of 'hidden information' and 'shared information bias' (**stasser_pooling_1985**). In addition, clinicians may modify how they communicate certainty with others, especially given the collaborative nature of healthcare and the social benefits of communicating opinions with confidence in order to be listened to in a group (**brezis_does_2019**). Situational awareness (SA) is also important in a group, and higher stress may be associated with overconfidence in SA (**price_acute_2016**). Taken together, group medical decisions are clearly an important and naturalistic area for future study.

Finally, the papers in this review have also not looked at individual differences in expressions of confidence, where past work from cognitive psychology has found individual systematic tendencies toward higher or lower confidence (**ais_individual_2016**; **navajas_idiosyncratic_2017**). Hence, individual clinicians may have a consistent tendency toward underconfidence or overconfidence that impacts their clinical practice and that training or cognitive aids could address.

Empirical Studies of the Current Work

In terms of empirical work within this DPhil, we focus on the foremost of these underexplored avenues within the current research landscape. Namely, with the experimental studies presented in the remainder of this thesis, we aim to elucidate how diagnosis evolves over time, particularly with the back-and-forth between seeking information and evaluating that information in the context of currently considered diagnostic possibilities. Throughout our studies, we record diagnostic hypotheses whilst allowing participants to seek information they find most useful for determining a diagnosis for a patient and recording their confidence in their diagnosis. Clinicians have been found to be less confident in their diagnoses when they have to seek information themselves compared to when they are provided with all available patient information straightaway (**gruppen_information_1991**). Hence, the information seeking process during diagnostic decisions impacts confidence. Understanding in more detail how information seeking patterns impact confidence can help the future design of cognitive interventions for appropriately calibrated confidence. For example, clinicians could reflect on their confidence by considering what information has been made available to them on the patient and what further information they may need to determine a diagnosis. In addition to studying the interplay between information seeking and confidence, we also use a variety of methodologies to study the cognitive mechanisms of medical diagnoses. Across the remainder of the thesis, we increase the naturalism of our methods with each chapter in order to address our point on the lack of high-fidelity methodologies in the extant literature. We start with a dynamic version of previous vignette studies that allow us to study information seeking patterns whilst maintaining the experimental control afforded to us by the use of vignette paradigms. We then use a think-aloud methodology to record the thought processes of clinicians as they make diagnoses. This is followed by the use of a virtual reality paradigm that allows for the simulation of more realistic clinical situations whilst still allowing for the recording of information seeking patterns. Finally, we use a rapid ethnography approach within real clinical settings to record instances of diagnostic uncertainty and information seeking within medical practice. Taken together, these approaches allow us to triangulate our findings and ground them in real medical practice. In the next chapter, we start with the first of our studies that uses a dynamic version of previous vignette paradigms where diagnostic hypotheses are recorded and updated over the course of a patient case, with participants able to freely seek information to inform their diagnoses.

Chapter 3 - Information Seeking and Confidence During Medical Diagnoses

Introduction

In the previous chapter, we presented a systematic scoping review of the extant literature on certainty and confidence during medical diagnoses. One output from this review was that we identified one gap in the literature in that past work has not studied the association between the ongoing receipt of information and confidence. Past work has tended to frame information seeking as a single action/choice taken after diagnosis, rather than an ongoing activity that causes regular reconsideration of a diagnosis and course of treatment. In this chapter, we aim to fill this gap by presenting results from an empirical study that investigates the interaction between confidence and information seeking during medical diagnoses.

In this study, we aim to retain the control and simplicity of vignette-based approaches while incorporating some of the complexities that characterise real diagnostic decision making. By doing this, we aim to study diagnostic confidence and accuracy as it develops over the course of a diagnostic decision. In this chapter, we first introduce our paradigm and its flexibility in allowing free information seeking and updating of diagnostic differentials over time. We then introduce the research questions that such a paradigm allows us to investigate.

Our systematic scoping review on confidence during diagnoses revealed two main findings. Firstly, past work that measured confidence and accuracy during diagnostic decisions found that confidence was rarely calibrated to objective accuracy, leading to overconfidence (**friedman__physicians__2005**; **fernandez-aguilar__use__2022**; **garbayo__metacognitive__2023**) or underconfidence (**mann__relationship__1993**; **yang__effect__2012**; **brezis__does__2019**). Secondly, confidence was associated with many aspects of the patient care process, such as prescriptions (**levin__antimicrobial__2012**; **garbayo__metacognitive__2023**), referrals (**calman__variability__1992**) and requesting investigations (**tabak__clinical__1996**; **gupta__associations__2023**). The lattermost of these factors is of particular relevance to our research interests. As we identify in our scoping review, the link between information seeking and confidence is a current gap in the extant literature. This link currently made between confidence and information seeking in the literature is that confidence is a subjective judgement that then guides a clinician's subsequent testing and requests for information (**tabak__clinical__1996**; **gupta__associations__2023**).

Crucially, this past work studies information seeking by asking participants/clinicians a single question of whether they would (hypothetically) seek further information or not. However, the link between confidence and information seeking can be expanded upon in three ways. Firstly, we can study how information seeking prior to the point at which confidence is reported affects this confidence. Secondly, we can look at specific aspects of information seeking that are linked to confidence aside from merely the intention to seek or not seek. This includes the amount of information sought, how relevant the information is to the patient case, and the degree to which clinicians vary their information seeking on a case-by-case basis. Finally, confidence and information seeking are likely to influence each other over time. Past work has tended to study diagnosis by asking clinicians to provide a single diagnosis/condition after delivering all available information. Whilst this is a useful simplification for the sake of empirical study, it leaves open the key aspects of how the diagnostic process unfolds in real clinical work. In everyday practice, clinicians engage with diagnosis as an active, ongoing decision process that develops with more time and as more information about the patient becomes available. With a more open-ended paradigm, we investigate how diagnoses evolve over time. For instance, does a clinician reach an initial diagnosis and then change their mind when they received unexpected information (e.g. a test result)? And does a clinician tend to have a single diagnosis in mind or do they tend to keep an open mind by having considering several diagnostic possibilities at once? Then, with these questions in mind, how does a clinician seek information to further validate their diagnosis?

With these points in mind, we aimed to design a paradigm that better reflects the evolving nature of diagnosis and allows to us to study aspects of the information seeking process. Our paradigm is then a step towards more realistic diagnostic decisions, as it retains the simplicity and control of vignette-based diagnosis whilst allowing more flexibility in information seeking and committing to a diagnosis (or set of diagnoses). This allows us to investigate more fine-grained aspects of information seeking and how they impact diagnoses. Specifically, is clinician confidence informed by the quantity and quality of information sought during the diagnostic process?

For this study, we designed and implemented a novel vignette-based experimental paradigm where participants are asked to provide a list of all diagnostic differentials they are considering based on the information they have received. We ask clinicians to update this list and their confidence at each of a series of stages related to the information sought about the patient: Patient History, Physical Examinations and Testing. We then ask participants to update this list in light of new information by adding or removing differentials. This allows us to more comprehensively capture their thought process in terms of how differentials are being weighed up against each other. Participants report how severe and likely each of their differentials are to draw a more nuanced distinction between differentials. Whereas past work has tended to provide a preset amount of information to clinicians, we instead prompt participants to actively seek out information that they feel is useful for diagnosing the patient they are presented with. This is more analogous to real medical practice where all the required information is not immediately available to clinicians when presented with a patient. We can then look at information seeking patterns within participants to study how they impact confidence.

Past work from cognitive psychology has shown a link to the quantity of information received and confidence, even if the information is disconfirmatory of one's beliefs (**ko_divergent_2022**). We can hence investigate in this experiment if this holds during medical diagnoses; if so, we would observe that higher amounts of information seeking would be associated with higher confidence. Information seeking could also be a marker of accuracy in addition to confidence though, as we can study whether clinicians who make more accurate diagnoses seek more appropriate information for the patient. This is important to study as some tests/information are less relevant than others for helping to reach a diagnosis for a patient, resulting in instances of overtesting. With all this in mind, allowing clinicians to freely seek information was then an important tenet for designing this experimental paradigm.

Another aspect of past work we aimed to expand on was on generating differentials (a term used in medicine to refer to hypotheses for diagnoses that a patient could have). Past work has tended to frame diagnosis as a single decision where a clinician responds either to a single diagnosis (**redelmeier_fallacy_2023**) or a limited number of conditions that a patient could have (**meyer_physicians_2013**). In the latter case, clinicians may report multiple differentials when prompted to consider alternative differentials via a cognitive intervention that encourages clinicians not to miss other diagnoses (**feyzi-behnagh_metacognitive_2014**). These experimental approaches do not necessarily represent the manner in which clinicians make diagnoses in their everyday medical practice. While clinicians may focus on a single differential at a time, they may also generate multiple diagnostic possibilities that past experimental paradigms do not capture. For instance, a clinician usually has to weigh up differentials (**schiff_diagnostic_2009**), based on their likelihood (taking into account the base rate of medical conditions within a given patient population) and severity (which may be less likely for a given patient, but would be more dangerous if not considered by the clinician as a possibility). In this sense, a clinician may have, at least, a primary diagnosis that is most likely for the patient and a more serious diagnosis that is less likely but can be dangerous if missed. Our paradigm should then allow clinicians to report multiple differentials at a time without constraints, in order to capture both the primary differentials being considered and the differentials that clinicians keep 'in the back of their mind'. We can then use the breadth of differentials considered by clinicians as another marker of uncertainty that may guide their subsequent information seeking. By allowing participants to record a list of all differentials they are considering at each stage, we can capture their thought process as it pertains to the information they have received prior to that point.

As our paradigm is designed to capture the diagnosis process as evolving over time, we can also study confidence differently to past work. Rather than seeing confidence as a static quantity, confidence may shift to reflect the current relative strength of evidence in favour of a decision alternative (**vickers_effects_1982**). Our paradigm then records confidence alongside the participants' list of differentials as it is being updated. We can not only use this facet to link confidence to the breadth of diagnoses considered but also to examine how confidence changes over the course of a case. For instance, a clinician may receive a surprising or inconclusive test result for a patient, causing them to reduce their confidence and seek more information as a result to increase their confidence. Our measure of confidence is also distinct from measures used in past work as we aim to capture the diagnostic process as it pertains to subsequent treatment of patients. An ideal diagnostic process would involve a clinician seeking information to formulate a diagnosis of a patient and, in the process, create a treatment plan to address this diagnosis. We then capture confidence in this study specifically to measure how ready the clinician is to treat the patient, as opposed to past studies that have tended to ask clinicians how confident they are that their diagnosis is the correct one.

There are multiple ways we can define how calibrated participants' confidence is. To recap, measuring calibration requires a subjective judgement of confidence and an objective measure of accuracy to compare this confidence judgement against. For past work where a single differential is provided by clinicians when they are asked to make a diagnosis, accuracy is relatively easy to measure, as it simply requires marking the provided differential as either correct or incorrect. In our paradigm however, participants not only provide all possible differentials that they are considering but also provide assessments of how likely each differential is. We must then consider how to assess each set of differentials as being accurate or not. A lenient definition of accuracy is to simply mark a set of differentials as accurate if it includes a correct differential. Henceforth, we refer to this measure as Differential Accuracy. However, this measure does not take into account the likelihoods assigned to differentials, so it does not consider how clinicians weigh up differentials against each other. Participants are also more likely to be correct by simply including more differentials in their list. A stricter definition of accuracy would be to look at whether the most likely differential (as rated by the participant) is correct and use the likelihood value assigned to this. Henceforth, we refer to this measure as Highest Likelihood Accuracy. However, this penalises participants who consider the correct differential as likely but not as their primary diagnosis. We therefore use the following measure of accuracy as our primary measure: we look at the likelihood rating assigned to the correct differential if it is present in the participant's list. This provides a more nuanced measure of accuracy that takes into account how differentials are weighed up against each other, which marks a difference from accuracy as it is defined in past work. We should note however that assessing the calibration of participants' confidence judgements is potentially contingent on the accuracy measure used. We therefore measure calibration using our primary measure of accuracy (the likelihood of the correct diagnosis), but also provide results using the other two measures mentioned here (Differential Accuracy and Highest Likelihood Accuracy).

For our studies, we chose to focus on medical students who were relatively advanced in terms of their medical education but were still early in their clinical experience. Medical students are yet to settle on a particular medical subdiscipline to specialise in, which allows our vignettes to cover a variety of medical conditions and pathologies. We also focus on students as findings from our work could have implications for future medical education in terms of how clinical reasoning and cognitive psychology is taught. Finally, recruiting students allows us to collect a relatively large sample to facilitate detailed analysis of information seeking patterns.

Research Questions

With this study, we investigated the following research questions:

- **Is confidence calibrated to accuracy within medical students?** - Whilst past work has found disassociations between diagnostic confidence and accuracy, these were found in the context of simple tasks with limited flexibility in terms of information seeking and recording multiple diagnostic differentials. We therefore investigate if similar miscalibrations of confidence occur within a more flexible experimental paradigm.
- **How do medical students weigh up competing differentials during the diagnostic process** - Past work has considered that clinicians may have multiple differentials in mind when diagnosing a patient, but such research has not studied how the differentials being considered changes with the receipt of new information. Specifically, do medical students tend to narrow the differentials over time (i.e. akin to a process of elimination) or do they tend to broaden their thinking as new information on the patient is received?
- **How do confidence and information seeking interact in the diagnostic process?** - We expect that confidence would predict information seeking, such that confidence in diagnoses is predicted by both the quality and quantity of information sought.

- **Do differences in confidence and information seeking predict differences in diagnostic accuracy?** - We expect accuracy to associated with the quality/suitability of information seeking but not the quantity of information sought or by confidence (as per the aforementioned miscalibrations of confidence).

Methods

This study was designed to understand how information seeking, confidence and differential generation interact within the diagnosis process. Specifically, we investigated whether information seeking patterns were associated with diagnostic accuracy and confidence. We conducted a vignette-based diagnosis study with medical students to characterise their diagnostic process and potentially to inform future on how diagnostic reasoning is taught to students, especially when it comes to weighing up competing differentials. Data is openly available on OSF: https://osf.io/kb54u/.

Participants

We recruited final year medical students within the UK. 85 medical students completed the study, including 32 males, 52 females and 1 participant who identified as non-binary. Their ages ranged between 22-34 years (M = 24.2). Participants were recruited between July 11th 2022 and April 6th 2023 via emails sent to all UK medical students within a UK Medical Schools Council 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 study was conducted online, with participants able to run the experiment in a browser on a desktop computer or laptop (and not a phone or tablet) in a location of their choice. The experiment was coded using the JSPsych Javascript plugin. The code is publicly available on Github: https:

//github.com/raj925/DiagnosisParadigm. Ethical approval was granted by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE001.

Materials

This study involved patient vignettes that we adapted from anonymised past cases developed by **friedman_are_2001**. Six cases were chosen, each designed to indicate a specific underlying condition the patient had: 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. Cases were chosen to reflect a variety of affected pathophysiological systems and to test medical students on medical conditions that they were expected to know given their level of education/training.

A panel of 3 subject matter experts (practising doctors and researchers within the NHS and the OxSTaR centre: www.oxstar.ox.ac.uk) were recruited to design the vignettes used in this study. These medical professionals were at differing experience levels, with their medical roles at the time of this study as follows: Speciality trainee (ST7) in Anaesthetics, Foundation (F1) Doctor and Gastroenterology Consultant. The panel assisted with translating terms (e.g., medication names, tests etc.) from US to UK doctors' vernacular, updated patient details to be more current and provided input on the choice and complexity of the cases chosen.

Procedure

The goal of the task was to determine a diagnosis, or diagnoses, for each presented patient (see procedure in Figure 3.1 below). Information on the patient was split into a series of discrete stages to control what information the participants had access to at any given point in the experiment. Each point of new information was termed an "information stage". Participants were able to seek information freely until they were ready to move on.

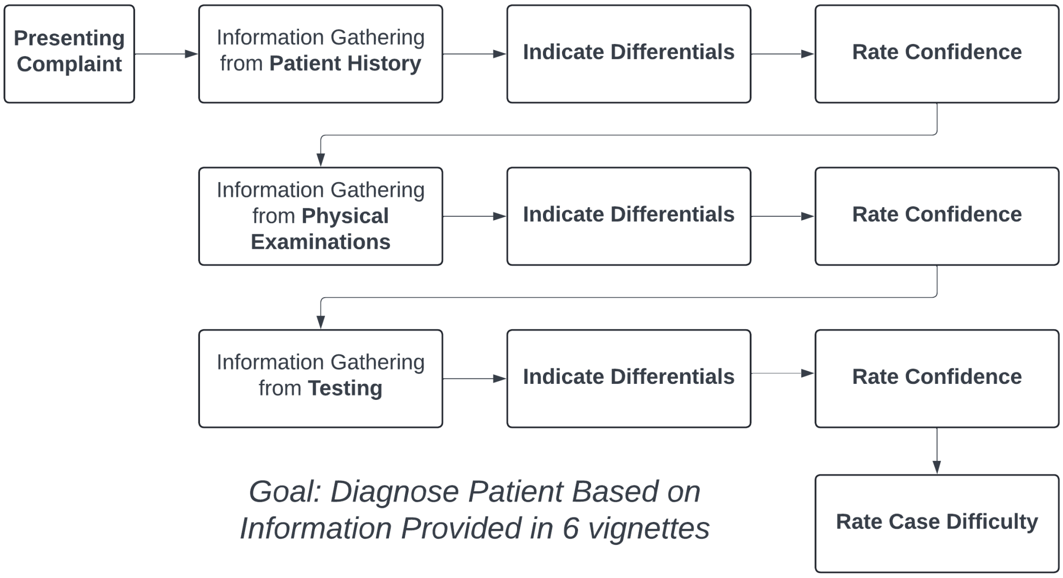
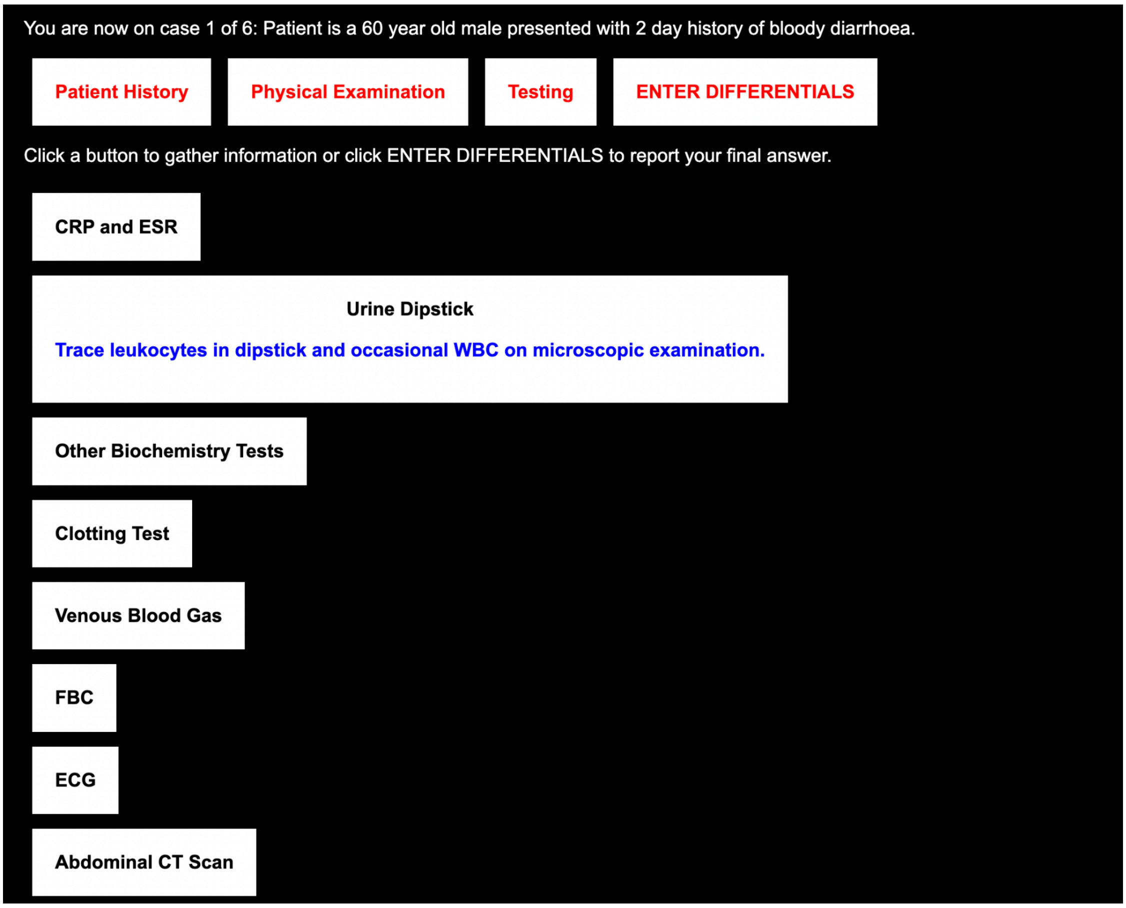

Presenting Complaint
Information Gathering from Patient History
Indicate Differentials
Rate Confidence
Information Gathering from Physical Examinations
Indicate Differentials
Rate Confidence
Information Gathering from Testing
Indicate Differentials
Rate Confidence
Rate Case Difficulty
Goal: Diagnose Patient Based on Information Provided in 6 vignettes

Figure 3.1: Paradigm of the online vignette study, showing the procedure for a single patient case.



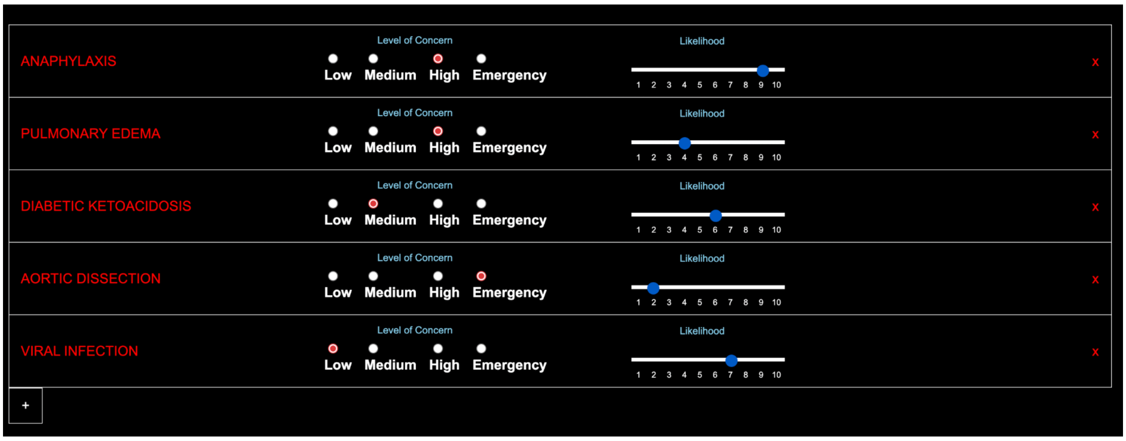
You are now on case 1 of 6: Patient is a 60 year old male presented with 2 day history of bloody diarrhoea.

Click a button to gather information or click ENTER DIFFERENTIALS to report your final answer.

Trace leukocytes in dipstick and occasional WBC on microscopic examination.

Figure 3.2: Screenshot of the interface. Shown here is the screen in which the participants seek information during the Testing stage.

The procedure of a single case is as follows. The participant was 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 was shown a description of a patient, which includes the patient's gender, age and their presenting complaint. An example of this was: "Patient is a 68 year old male presenting with fever and arthralgia". Each case is split into three information stages: Patient History, Physical Examination and Testing (in this order). This structure has been utilised in past work as being prototypical for a diagnosis (**hampton_relative_1975**; **peterson_contributions_1992**). The Patient History stage included information on "Allergies", "History of the Presenting Complaint", "Past Medical History" and "Family History". The Physical Examination stage included 'actions' that a doctor may take when examining a patient, such as “auscultate the lungs”, “abdominal examination”, “take pulse” and “measure temperature”. Finally, the Testing stage involved 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, there were 29 possible information requests across the three stages, with the available set of information being the same for all patient cases.



ANAPHYLAXIS — Level of Concern: ○ Low ○ Medium [✓ High] ○ Emergency — Likelihood: 1 2 3 4 5 6 7 8 [✓ 9] 10 — x

PULMONARY EDEMA — Level of Concern: ○ Low ○ Medium [✓ High] ○ Emergency — Likelihood: 1 2 3 [✓ 4] 5 6 7 8 9 10 — x

DIABETIC KETOACIDOSIS — Level of Concern: ○ Low [✓ Medium] ○ High ○ Emergency — Likelihood: 1 2 3 4 5 [✓ 6] 7 8 9 10 — x

AORTIC DISSECTION — Level of Concern: ○ Low ○ Medium ○ High [✓ Emergency] — Likelihood: 1 [✓ 2] 3 4 5 6 7 8 9 10 — x

VIRAL INFECTION — Level of Concern: [✓ Low] ○ Medium ○ High ○ Emergency — Likelihood: 1 2 3 4 5 6 [✓ 7] 8 9 10 — x

Figure 3.3: Screenshot of the interface. This is the screen in which participants report their current list of differentials, including the name of each condition as well as the severity and likelihood ratings for each condition. Participants remove conditions by clicking the red cross on the right hand side of each differential. Participants add a new differential by clicking the plus icon below the list.

When a participant clicked on any of these requests, the information for that request was shown on screen after a 3 second delay. This delay was added after pilot testing (with 10 participants) revealed that participants tended to select most, if not all, of the information available to them. We mitigated this tendency by adding this delay and by emphasising to participants during the task instructions that they should only request information that they believe will help them with diagnosing the patient for that specific case. Participants were free to request the same piece of information multiple times, including information from a previous stage. At any point, they could choose to stop gathering information for that stage. They were then taken to a new screen where they reported a list of all differential diagnoses that they were considering for that patient at that stage. For each differential, participants reported a likelihood rating, ranging from 1 (very unlikely) to 10 (certain), and a "level of concern" (which was how concerned they would be for that patient if this differential really was the patient's underlying condition) on a 4 point scale (labels of "Low", "Medium", "High" and "Emergency"). In subsequent stages, the list from the previous stages was available for participants to update concern/likelihood ratings, or to add/remove differentials from the list. Even at the last information stage, participants could report multiple differentials.

After recording their differentials, participants were then asked to report their confidence that they were "ready to start treating the patient" on a 100 point scale, ranging from not at all confident to fully confident. Participants also indicated using a checkbox whether they are ready to start treating the patient, at which point a text box appeared 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. This allowed participants to express what actions they would take that were not covered by our set of available information requests. Once all three stages were completed, participants reported 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 were told the 'true' conditions for all the patients. The session took approximately 40-60 minutes to complete.

Data Analysis

During analysis, no sought information was recorded for three cases across participants (i.e. at all three stages during a case, the participant did not appear to seek any information). These cases were excluded from analysis. We now describe the key dependent variables for this study. The first set of the measures (Case-Wise Measures) are calculated at each of the three information stages (except for Perceived Difficulty). When averaging these variables within a participant, we use the values obtained at the final stage (i.e. Testing). The second set of measures (Derived Information Seeking Measures) are based on information seeking by participants on each case across all three information stages.

Case-Wise Measures

- *Correct Differential Included*: This measure captures whether participants consider a correct diagnostic differential. 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 valid answers. Each case is marked as correct if the list of differentials provided includes the correct condition or a differential considered correct as per our marking scheme in Table A1 of the Appendices. Otherwise, the case is considered incorrect if a 'correct' differential is not included.

- *Accuracy*: Our main measure of diagnostic accuracy is computed as the likelihood value assigned to the correct differential for the case (and scored as 0 if this differential is not listed). 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. Likelihoods range from 1-10 when a correct differential is included and has a value of 0 when a correct differential is not included. The value is then rescaled to range from 0 to 1, where 1 corresponds to a correct differential assigned maximum likelihood. If multiple differentials that are considered correct were provided, then the likelihood value of the closest differential (as per our marking criteria with help from a medical consultant) to the true condition was used.

- *Highest Likelihood Accuracy*: This stricter measure accuracy is computed as the likelihood value assigned to the differential with the highest likelihood (in comparison to other differentials provided in the participant's list) if this differential is considered correct. If not, a value of 0 is assigned. Again, likelihoods range from 1-10 for correct differentials, so this is rescaled to range from 0 and 1.

- *Confidence*: Participants reported their confidence that they are ready to start treatment 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). As with accuracy, confidence is rescaled to fall between 0 and 1 to allow for direct comparison between the two variables. 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*: This measure captures the breadth of diagnoses considered by participants. The number of items in the list of differentials was recorded 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).

- *Change in Differentials*: This measure captures how much participants change the differentials they consider over the course of the case. In other words, we can look at how much participants have narrowed or broadened their list of differentials as they receive more information. This is calculated by taking the absolute value of the difference between the number of Initial Differentials and the number of Final Differentials.

- *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 subjectively by each participant on a scale from 1 (trivial) to 10 (impossible).

Derived Information Seeking Measures

- *Amount of Information Seeking*: This measure captures the amount of information that participants seek on cases relative to how much they could have sought if seeking all available information. We take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage and excluding repeat tests) and divide this by the number of possible tests available.

- *Information Value*: We calculate a measure of information value to capture how appropriate the information sought for a case is for the patient's condition. We compute the average value of sought information across cases. To do this, we take each of the 29 pieces of information in turn by case and split all cases completed across participants into two groups: cases where that information was sought at any stage and cases 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'. We then calculate the sum of all information values for each case. This gives an overall measure of, on average, how useful the information was that participants sought on each case.

- *Information Seeking Variability*: We calculate a measure of how much, for a given set of cases, information seeking varies across cases. This is operationalised as the average dissimilarity between cases' information seeking (by

> taking the average of all pairwise comparisons) using each piece of information as a binary variable (i.e. whether it was sought or not). This measure is calculated both within participants, to tell us how much each participant varied the information they sought across their cases, and between participants, to tell us how dissimilar participants are to each other in terms of the information sought for a given condition. We calculate this value using the Dice coefficient (**dice_measures_1945**), 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 in one case but not the other). A higher value between two cases indicates that the information sought on those trials are more dissimilar to each other.

We used statistical analyses to consider differences in confidence, accuracy and information seeking. When looking at how our variables change over the three information stages, we used Analysis of Variance models with Bonfferoni-corrected pairwise T-tests on all pairwise comparisons. We test if there is a relationship between confidence and information seeking (Amount, Value Variability) and between accuracy and information seeking using Pearson's product moment correlation tests (an alpha value of less than 0.05 was regarded as statistically significant). These help us answer how confidence and information seeking interact during the diagnostic process and whether differences in diagnostic are predicted by information seeking and confidence. Our sample of 85 participants is calculated as having 80.4% power to detect a medium effect size of r = 0.3 (using an approximate arctangh transformation correlation power calculation). In addition to correlations for variables averaged (across cases, per participant), we also fit linear mixed effects models to predict information seeking, changes in differentials and changes in confidence (between the Patient History and Testing stages). In order to test if information seeking patterns are predictive of differences in accuracy, we used generalised logistic regression to classify cases as being performed by high or low accuracy participants (via a median split). To test if information seeking patterns are predictive of the case (i.e., whether participants tailor their information seeking to each patient case), we use penalised multinomial regression to classify cases by their patient condition. Both models were trained on the information requests as binary variables (with a 1 signifying that the information was sought for that case and 0 when the information was not sought). We used Leave One Out Cross Validation for both models, such that each case is predicted by training the algorithm on all other cases.

Results

Overall Performance and Calibration

We first look at our research question as to whether confidence is calibrated within medical students. When comparing Accuracy (taking into account the likelihood assigned to correct differentials) to Confidence, we find, across stages, participants' Confidence was aligned to their Accuracy (see Figure 3.4 below). To determine whether there is any systematic discrepancy between subjective confidence and objective accuracy across stages, we compute a paired t-test between average Confidence and average Accuracy (across cases) at each stage. There was no evidence of a difference between the two at the Patient History (t(84) = 0.29, MDiff = 0.01, p = 0.77) and Physical Examination stages (t(84) = 0.74, MDiff = 0.01, p = 0.46), but there was a statistically significant difference between the two at the Testing stage (t(84) = 2.35, MDiff = 0.05, p = 0.02). This indicated well-calibrated confidence after Patient History and Physical Examination, but a slight overconfidence across participants after Testing.

To investigate the dynamics of confidence and accuracy further, we look at how both variables change over the course of the information seeking stages. Across cases, accuracy increased with each stage of information gathering as per our Accuracy measure, (F(2, 252) = 21.6, η^2G = 0.15, p < .001). Participants had lower accuracy at the Patient History stage (M = 0.31, SD = 0.14) than during the Physical Examination (M = 0.04, SD = 0.15) and Testing stages (M = 0.41, SD = 0.15). Pairwise comparisons between the History stage and each of the other two stages are significant (ps < .001). Table 3.2 shows overall accuracy (at the Testing stage) by case, indicating that there was variability in performance between cases.

Confidence also increased as participants received more information (F(2, 252) = 21.6, η^2G = 0.15, p < .001). Participants reported lower confidence during the Patient History stage (M = 0.3, SD = 0.15) than during the Physical Examination (M = 0.41, SD = 0.17) and Testing stages (M = 0.47, SD = 0.47). Pairwise comparisons between History and each of the other two stages are significant (ps < .001). We note here that confidence was on average below 50% even at the end of each case, which indicates that participants were not highly confident to start treatment. This is reflected in participants expressing their readiness to treat the patient in the vignette, which allows them to enter a treatment plan for the patient. In 38% of cases, participants reported they were ready to treat the patient and entered a treatment plan.

Case	Differential Accuracy	Accuracy	Highest Likelihood Accuracy	Final Confidence	Difficulty	Information Seeking
AD	0.60	0.28	0.12	0.49	5.9	0.59
GBS	0.75	0.41	0.30	0.37	6.9	0.63
MTB	0.43	0.24	0.10	0.45	6.7	0.64
TA	0.74	0.50	0.45	0.49	6.2	0.62
TTP	0.61	0.34	0.20	0.41	6.8	0.66
UC	1.00	0.73	0.69	0.62	5.2	0.55

Table 3.1: Average statistics across participants for each case (leftmost column, AD = Aortic Dissection, GBS = Guillain Barré Syndrome, MTB = Miliary Tuberculosis, TA = Temporal Arteritis, TTP = Thrombotic Thrombocytopenia Purpura,

UC = Ulcerative Colitis). Differential Accuracy (0-1) refers to the proportion of participants who correctly included the correct condition or a condition considered correct for that case based on our marking criteria. Highest Likelihood Accuracy refers to the likelihood assigned to the differential with the highest likelihood if it is correct (1-10), otherwise the value for a given case is 0 if this differential is incorrect. This value is then rescaled to range between 0-1. Accuracy refers to the average likelihood (on a 1-10 scale, rescaled to range between 0-1) assigned to a correct differential if included. Confidence refers to the confidence provided by participants on their readiness to treat the patient at the Testing stage (on a scale of 0-100, rescaled to fall between 0-1). All these measures are calculated based on values observed at the final information stage of each case (i.e. the Testing stage). Difficulty refers to the subjective rating provided at the end of each case of how difficult participants found the case to be in terms of determining a diagnosis (on a scale of 1-10).

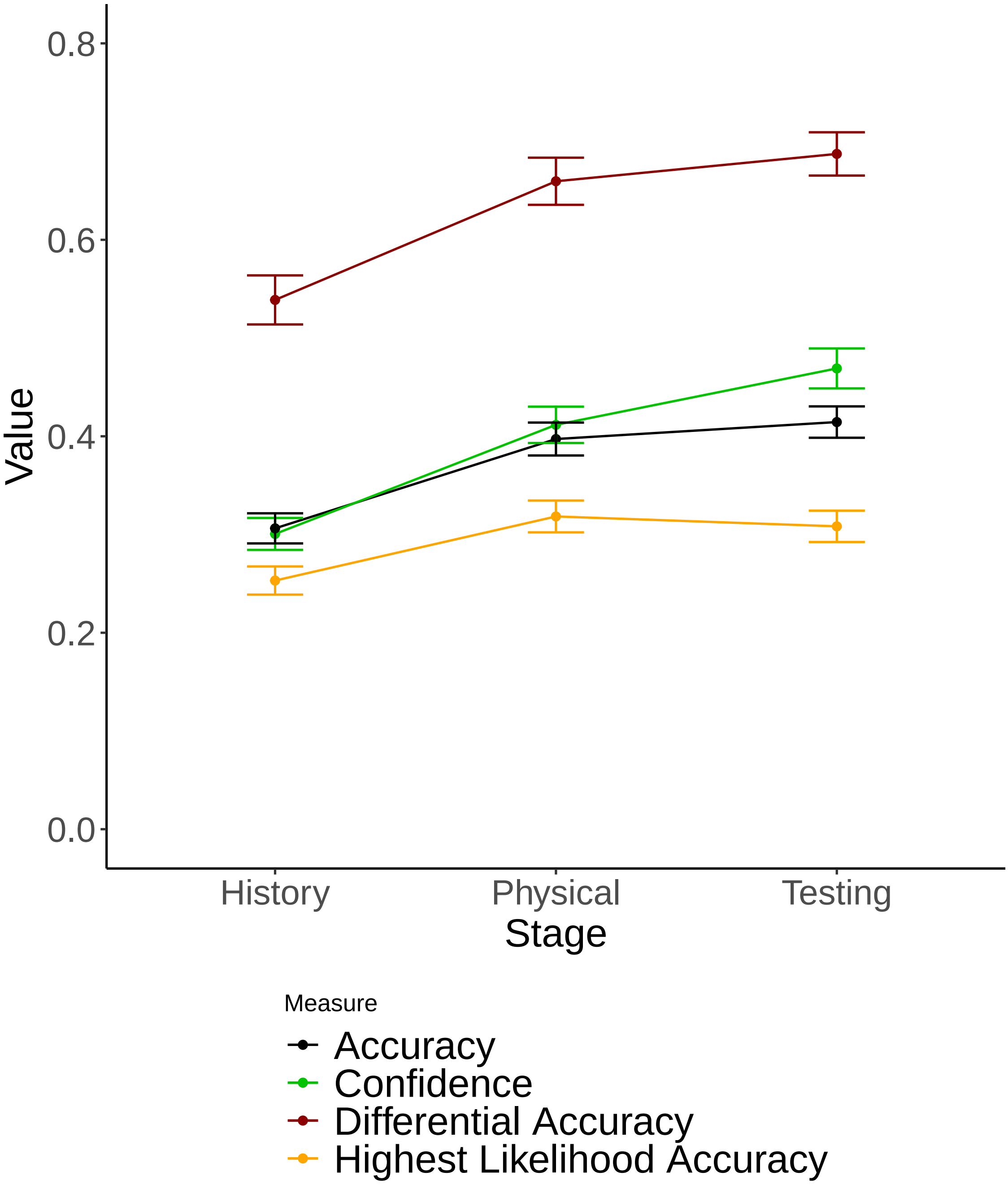


Figure 3.4: Graph showing Confidence (green) at each of the three information stages (History = Patient History, Physical = Physical Examinations, Testing = Testing) in comparison to our main accuracy measure (black, likelihood value assigned to the correct diagnosis), the more lenient measure of the proportion of trials where a correct differential was included (dark red) and the stricter measure of the value assigned to the highest likelihood differential if it is correct (orange). Values

shown are averaged across participants and cases, with the error bars representing standard error.

In order to examine the observed overconfidence in more granularity, we compare confidence and our primary accuracy measure by case (the mean values of which can be found in Table 3.1). We conducted paired t-tests for each condition's cases by comparing accuracy and confidence values (at the final Testing stage) to observe if they significantly differ from each other. A p value of less than .05 is interpreted as evidence for overconfidence or underconfidence (depending on the direction of the effect). We observed overconfidence for the AD case (t(84) = 4.71, MDiff = 0.21, p = < .001) and for the MTB case (t(83) = 4.31, MDiff = 0.21, p = < .001). We observe underconfidence for the UC case (t(82) = -3.51, MDiff = -0.12, p = < .001). The remaining cases did not yield a significant effect, indicating calibrated confidence judgements across participants. The overall overconfidence after Testing that we observe in Figure 3.4 is then driven by the AD and MTB cases, for which accuracy was lowest compared to other cases and confidence was not sufficiently adjusted to reflect this.

Differentials

We analysed the number of differentials to provide insights into the diagnostic decision process across stages, specifically the degree to which it follows a process of deductive narrowing (decreasing differentials) or open-minded broadening (increasing differentials). Analysis of the number of differentials considered by participants at each stage provides little evidence for an overall strategy of deductive narrowing towards a single differential. Instead, participants overall increased the number of the differentials they reported as they received more information (F(2, 252) = 11.66, η^2G = 0.08, p < .001). Participants reported fewer differentials during the Patient History stage (M = 3.2, SD = 1.12) than during the Physical Examination (M = 3.89, SD = 1.32) and Testing stages (M = 4.13, SD = 1.43). Pairwise comparisons between the History stage and each of the other two stages are significant (ps

$< .05$). The majority of participants (74/85) did not decrease the number of differentials between Patient History and Testing on any case, indicating a tendency to widen rather than narrow the set of considered diagnoses through the evolving decision process (even while, on average, growing increasingly certain of the correct diagnosis). As can be observed in Figure 3.5 below, there is general consistency in terms of participants broadening their differentials with more information despite some inter-participant variability, with a small minority of participants narrowing their differentials on average.

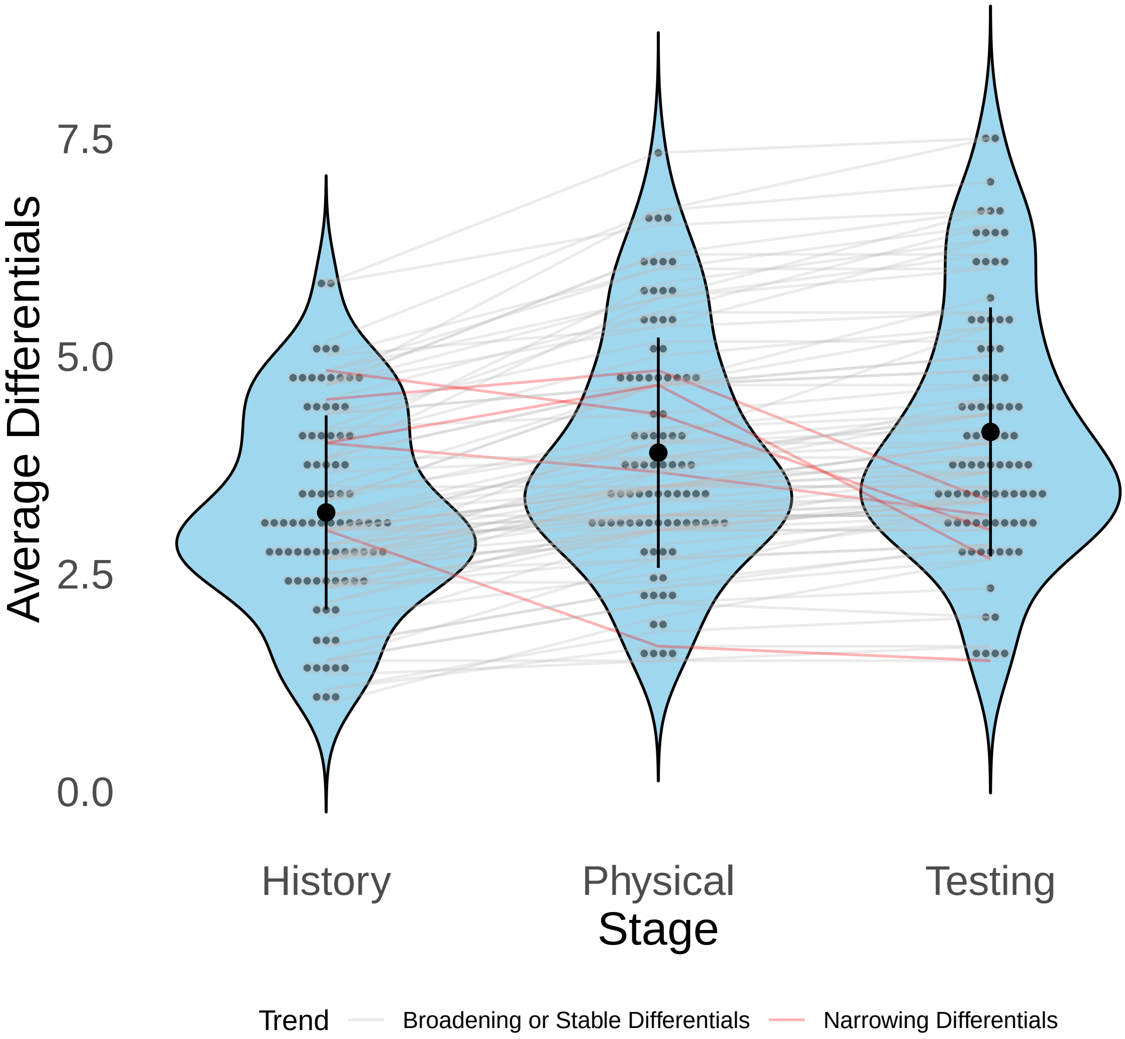


Figure 3.5: The average number of differentials after each stage of information seeking (x-axis, History = Patient History, Physical = Physical Examinations, Testing = Testing). The width of the blue area corresponds to the amount of data points that fall within that part of the y-axis, with a wider area meaning a higher concentration of data points. The larger black dots indicate the mean values, whilst the larger black vertical lines indicate standard deviations. The grey dots show individual values at each stage, with lines connecting the dots at each stage to represent individual participants' trend across the information seeking stages. The participants who show a narrowing of differentials (i.e. recording fewer differentials at the Testing stage compared to the Patient History stage) are marked with a red line, whilst the remainder of participants are marked with a grey line.

As a first probe of the dynamics of the diagnostic process, we analysed whether participants who generated more differentials early in the diagnostic process go on to seek more information by conducting a Pearson's Correlation test on individual differences. We find a positive correlation (see Figure 3.6) between the average number of differentials generated from the Patient History and the average amount of information sought during cases (r(83) = 0.3, 95% CI = [0.09, 0.48], p = 0.005, Figure 3.6a). 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. We also find evidence for a positive association between the number of initial differentials and the change in confidence (i.e. the difference in confidence reported during the Patient History stage and the Testing stage) (r(83) = 0.23, 95% CI = [0.02, 0.42], p = 0.04, Figure 3.6b).

Given that we observe an broadening (increasing number) of differentials across participants, we ask how this change in differentials related to information seeking and changes in confidence. As well the initial diagnostic breadth of participants, we are also interested in whether information seeking and changes in confidence relate to how much participants change the number of differentials they consider over the course of the case. This allows us to capture how much their diagnostic differentials have changed based on the information received. We find the amount of Differential Change was associated with both the amount of information sought (r(83) = 0.3, 95% CI = [0.09, 0.48], p = 0.005, Figure 3.6c) and change in confidence (r(83) = 0.39, 95% CI = [0.19, 0.56], p = < .001, Figure 3.6d). These results indicate that participants who tended to increase differentials also tended to seek more information and increase their confidence to a greater extent. If broadening of differentials was a reflection of diagnostic uncertainty, we may have expected a decrease in confidence, but this does not appear to be the case. We also fit a linear mixed model to predict Differential Change on individual cases, using the amount of information seeking (after Patient History), the number of initial differentials and Confidence Change as fixed effects, as well as both the patient case/condition and participant as random effects. We find that the amount of information seeking was positively predictive of Differential Change ($\beta = 0.86$, SE = 0.23 t = 3.76, p < .001), whilst the fixed effects were not found to be significant. This indicates that higher changes in the number of differentials being considered (i.e. how much a participant was able to broaden or narrow their differentials) was associated with higher information seeking.

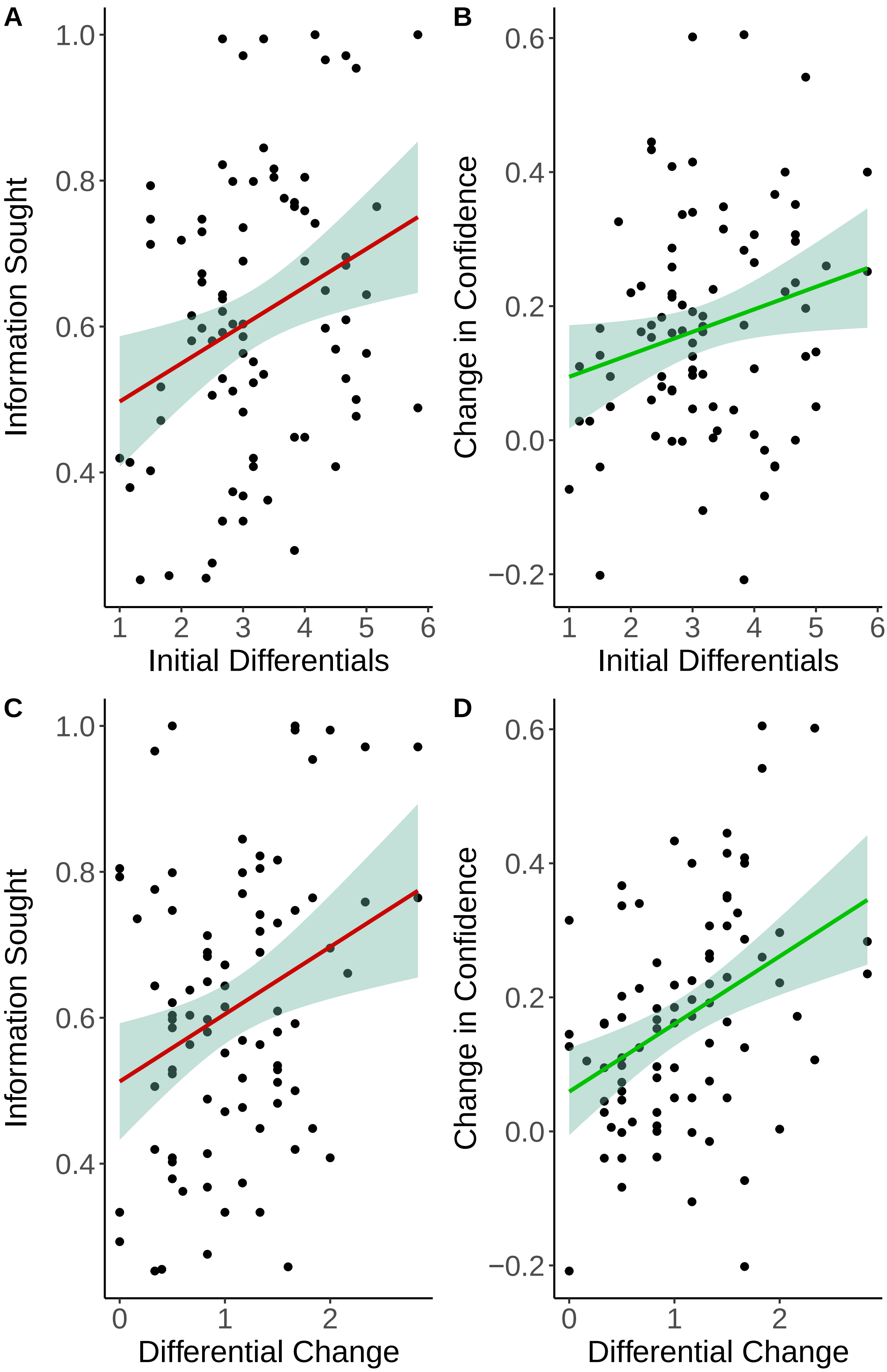

A
Information Sought
1.0
0.8
0.6
0.4
1
2
3
4
5
6
Initial Differentials
B
Change in Confidence
0.6
0.4
0.2
0.0
−0.2
1
2
3
4
5
6
Initial Differentials
C
Information Sought
1.0
0.8
0.6
0.4
0
1
2
Differential Change
D
Change in Confidence
0.6
0.4
0.2
0.0
−0.2
0
1
2
Differential Change

Figure 3.6: Scatter plot showing the relationship between the number of initial differentials reported at the Patient History stage (x-axis, figures 3.6A and 3.6B) and the change in differentials (x-axis, figures 3.6C and 3.6D) against both the proportion of available information sought (y-axis, figures 2.6A & 3.6C) and change in confidence (y-axis, figures 3.6B & 3.6D). Each point represents a single participant with all three variables averaged across the six cases that each participant performs. Initial Differentials refers to the average number of differentials that participants report in their list at the Patient History stage. Differential Change refers to the absolute difference in the number of Initial Differentials (at the Patient History Stage) and the number of Final Differentials (at the Testing Stage). Information Sought refers to the average proportion of available information sought, with each case containing 29 pieces of information across the Patient History, Physical Examination and Testing stages. Change in Confidence refers to the difference in reported confidence at the Patient History and Testing stages, such that a positive represents that the participant on average increased in their confidence over the course of the cases. 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.

Information Seeking

To investigate our research questions of how both confidence and accuracy interact with information seeking during the diagnostic process, we first look at broad characteristics of information seeking and then ask if they are predictive of differences in confidence and accuracy. When investigating whether participants became more selective in their information seeking over the course of cases, we find that the Proportion of Information Seeking decreased with each information stage (F(2, 252) = 57.26, η^2G = 0.31, p < .001). Participants sought more of the available information during the Patient History stage (M = 0.85, SD = 0.19 than during both during the Physical Examination (M = 0.59, SD = 0.24) and Testing stages (M = 0.5, SD = 0.22). All pairwise comparisons are significant (ps < .05). This selectivity in information seeking does not seem to reflect participants being less certain about their diagnoses, which the general pattern of broadening differentials may have indicated.

Given the design of our task, we ask if seeking all available information is in fact a helpful strategy for increasing diagnostic accuracy by testing for a correlation between the two. We do not find that participants who sought more information across cases were also more accurate in their diagnoses (r(83) = 0.16, 95% CI = [-0.05, 0.36], p = 0.13, Figure 3.7A). However, participants who sought more information tended to have increased their confidence more during cases (r(83) = 0.24, 95% CI = [0.02, 0.43], p = 0.03, Figure 3.7C). While seeking more information may imbue students with a greater level of confidence, we do not find evidence that this translates consistently into more accurate diagnoses. This finding links to the results presented in Figure 3.4, in which confidence and accuracy were related to one another but imperfectly (especially during the Testing stage, during which subjective confidence was higher than objective accuracy across participants). We also fit a linear mixed model to predict Information Seeking (after Patient History) on individual cases, using the number of initial differentials and Initial Confidence as fixed effects, as well as both the patient case/condition and participant as random effects. We find that the number of initial differentials was positively predictive of the amount of information seeking (β = 0.01, SE = 0.01 t = 2.44, p = 0.02) and that Initial Confidence was negatively predictive (β = 0, SE = 0 t = -3.81, p = 0). This indicates that higher information seeking was predicted by lower confidence and higher initial diagnostic breadth.

In order to examine more specifically what differences in information seeking are driving differences in both accuracy and confidence, we look at their relationship with informational value. We assess the degree to which each participant's accuracy is predicted by the quality of the information they sought and find evidence for a positive relationship between accuracy and information value (r(83) = 0.22, 95% CI = [0, 0.41], p = 0.05, Figure 3.7B), as well as between confidence and information value (r(83) = 0.29, 95% CI = [0.08, 0.47], p = 0.01, Figure 3.7D). When comparing the correlations between both information amount and information value to accuracy via a Fisher's z-Test of dependent correlations, we find they are not significantly different from one another (z = 1.08, p = 0.28). This means that we cannot make a valid comparison between the correlations with information amount and information value with respect to accuracy.

We also fit a linear mixed model to predict Change in Confidence on individual cases, using the number of initial differentials, the amount of information seeking (after Patient History) and change in differentials as fixed effects, as well as both the patient case/condition and participant as random effects. We find no significant fixed effects (ps > .1).

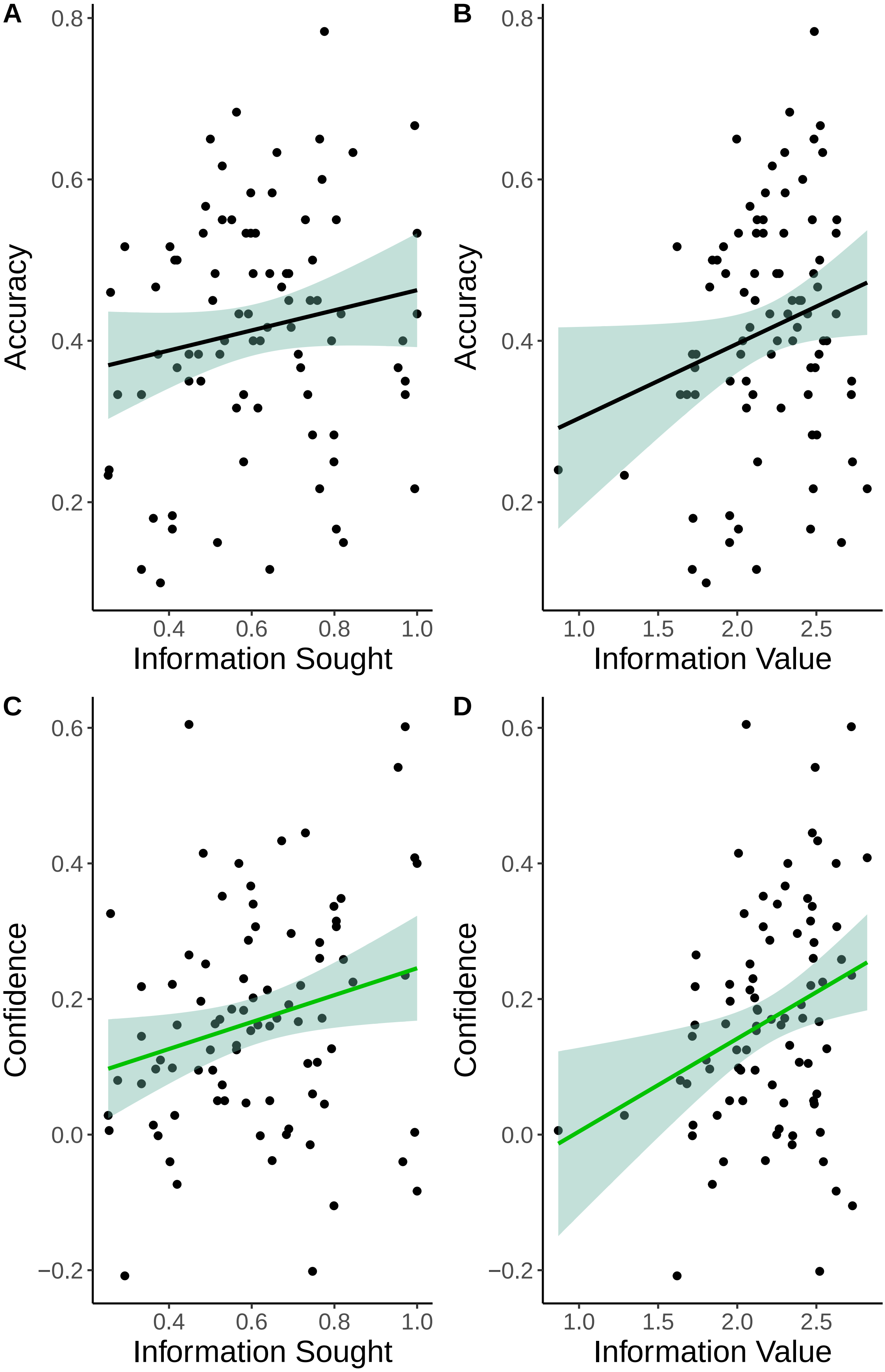

A
B
C
D
Accuracy
Confidence
Information Sought
Information Value
0.8
0.6
0.4
0.2
0.0
−0.2
1.0
1.5
2.0
2.5

Figure 3.7: Scatter plots showing our information seeking variables (amount in figures 3.7A & 3.7C and value in 3.7B & 3.7D) against our key dependent variables of accuracy (the likelihood assigned to a correct differential if provided, figures 3.7A & 3.7B) and change in confidence (difference between final confidence and initial confidence, figures 3.7C & 3.7D). Information Sought refers to the proportion of available information sought across cases. Information Value refers to the sum of all mean information values across all 6 cases for a given participant. All data points are for a single participant where variables are averaged across all 6 cases they completed.

Whilst we do not find evidence that the amount of information sought is predictive of accuracy, it may be that there are identifiable 'fingerprints' reflected in information seeking patterns that differentiate between high and low accuracy diagnosticians. If this is the case, participants who are high and low accuracy participants could be predicted based on their information seeking patterns.

In order to test this, we investigate whether information seeking is predictive of participants who are higher or lower in their diagnostic accuracy using binary classification and receiver operating characteristic (ROC) analysis. ROC is a form of analysis that assesses how well a model performs at predicting a binary outcome (in this case, whether a case was performed by a high or low performing participant). We trained a binary classification algorithm using a generalised logistic regression (GLM) model with Leave One Out Cross-Validation (LOOCV) to identify if participants exhibited high or low accuracy based on the information they sought. LOOCV is where our classifier is trained on all data except one case to ask if, based on the learnt patterns from this data, the classifier is able to predict the participant's accuracy (high or low) on the remaining case. This process is then repeated with each case being left out of training and used as this 'test' case. We first split all cases into two groups by whether they were performed by a high and low Accuracy participant. This was done using a median split by participants' average Accuracy across the six cases. By doing this, we can look at whether participants who perform better at diagnoses seek information in a markedly different way to participants who performed worse.

When plotting an ROC curve, the area under the curve (AUC) is indicative of how well a model performs at correctly categorising cases. An AUC of 0.5 would signify that our model is performing at chance and is not able to predict participant accuracy in any meaningful way. By plotting an ROC curve for our model, we find an AUC value of 0.72 (plotted in Figure 3.8). When conducting a DeLong test, to test the null hypothesis that the AUC is equal is 0.5 (i.e. that the classifier is unable to differentiate between high and low accuracy participants), we find p < .001, indicating that the AUC differs significantly from 0.5 and that the classifier is able to reliably predict high and low accuracy participants.

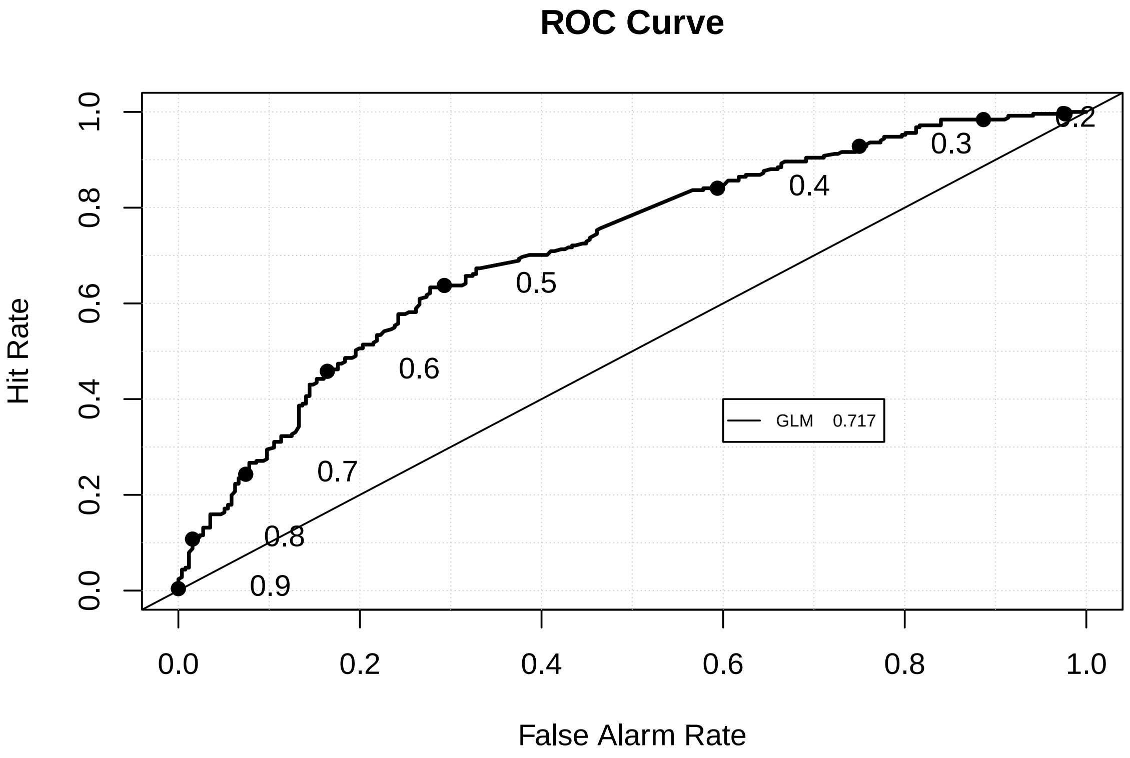


Figure 3.8: Receiver-Operator Characteristic (ROC) curve using a Generalised Linear Model to classify individual cases as being performed by either high or

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 or low accuracy based on a median split on their average Accuracy value across the six cases.

This result indicates overall that differences in information seeking are indeed predictive of a difference in participant ability at above chance, in terms of high and low accuracy participants seeking different sets of information. Essentially, information seeking patterns are different between high and low accuracy participants. This analysis alone, however, does not tell us what aspects of information seeking in particular are predictive of accuracy. We know from figure 3.7D that seeking more valuable information is associated with higher accuracy. We next seek to characterise the specific differences in information seeking that contribute to higher diagnostic performance.

By looking at the extent to which participants vary the information they seek by case, we can ask the following: is diagnostic accuracy characterised more carefully tailoring information seeking to each individual case, or is it characterised by adopting a more consistent information seeking approach regardless of the patient case? With our measure of how much participants vary in their information seeking across cases, we can see if the variability in information seeking is associated with higher diagnostic accuracy. If higher variability is associated with higher accuracy, this would indicate the former approach being more beneficial (tailored information seeking). If lower variability is associated with higher accuracy, this would indicate the latter approach being more beneficial (consistent information seeking).

We find marginal evidence for a negative association between Information Seeking Variability and Accuracy (r(83) = -0.22, 95% CI = [-0.42, -0.01], p = 0.04). This data is plotted below in Figure 3.9. We can also look at variability between groups of participants for each case to ask: are higher performers (in terms of accuracy) more alike in their information seeking than lower performers? To do this, we median split participants into high and low overall accuracy across cases (similar to the ROC analysis in Figure 3.8). We then look at variability in information seeking between participants for each case. If variability is higher, this would indicate that for a given case, participants adopt information seeking approaches that are more different from one another. A plot of variability by case is shown in Figure 3.10. When performing a t-test across conditions, we find that higher performers are more alike in their information seeking (i.e. exhibit lower variability) (t(10) = 2.64, MDiff = 0.36, 0.31, p = 0.02). As can be seen in Figure 3.10, better performing participants show less variability in their information seeking patterns for 5 out of 6 cases, indicating that higher diagnostic accuracy is associated with a consistent 'optimal' information seeking strategy.

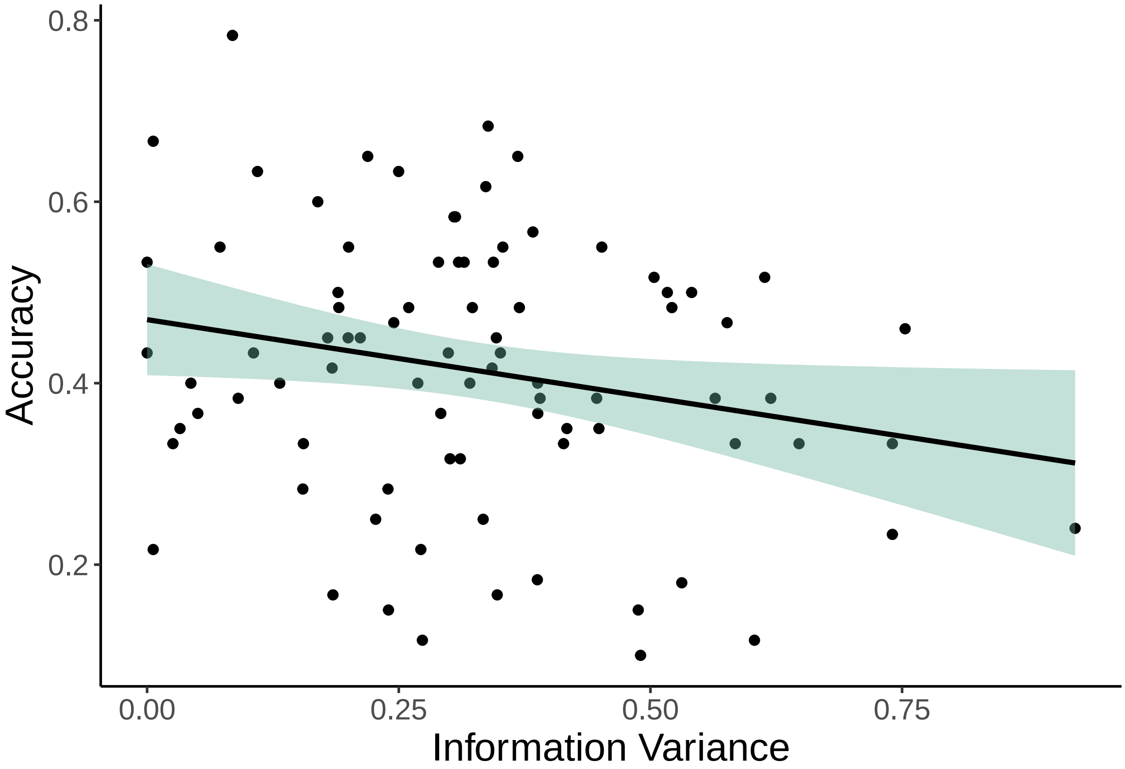


Figure 3.9: Scatter plot showing the relationship between Information Seeking Variability (x-axis, quantified as the average Dice Distance between all pairwise

comparisons of cases for a given participant) and Accuracy (y-axis). Each data point represents a single participant.

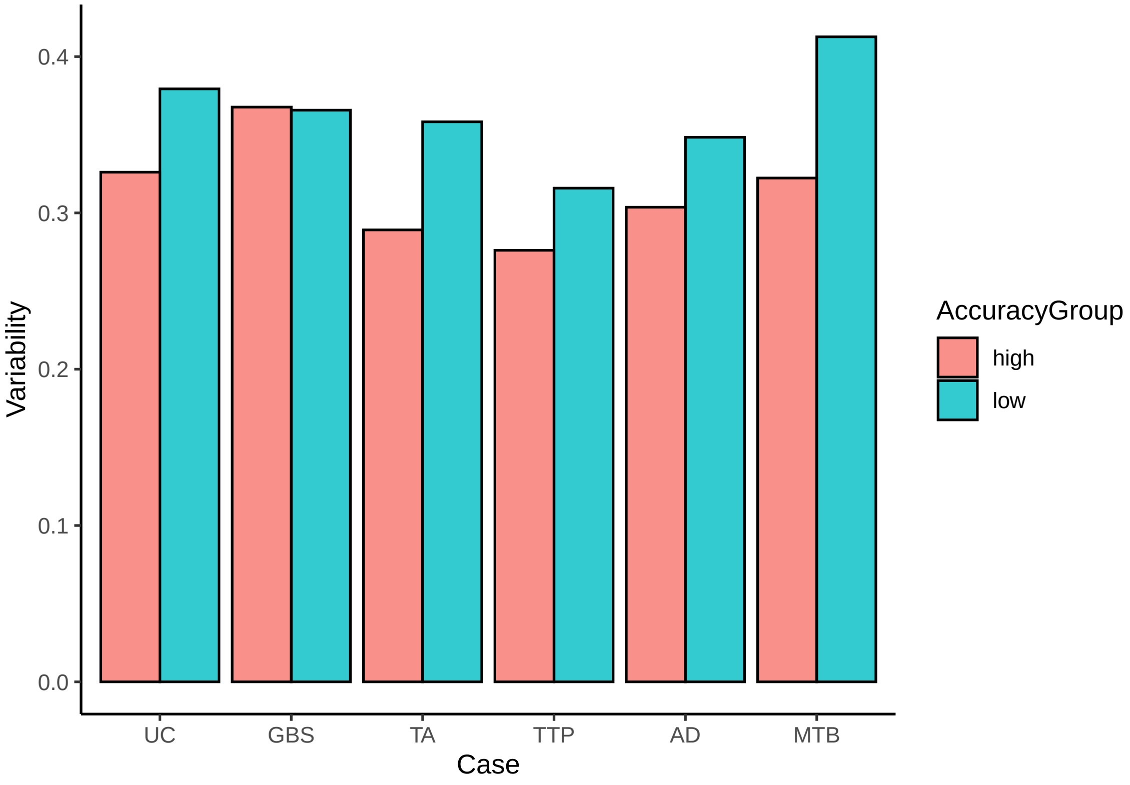


Figure 3.10: Information Seeking Variability (y-axis) for all cases of a given condition (x-axis), with cases median split by participant accuracy. Red bars indicate high performers and blue indicating lower performers. Cases are in descending order (UC = Ulcerative Colitis, GBS = Guillain Barré Syndrome, TA = Temporal Arteritis, TTP = Thrombotic Thrombocytopenic Purpura, AD = Aortic Dissection, MTB = Miliary Tuberculosis) by their average accuracy across participants. Higher variability values signify that participants were less alike one another in terms of the information they sought on a given case.

Given that information seeking variability has a weak negative association with accuracy on our task, we next ask if information seeking is also specific to patient conditions. If so, we would expect the information sought to be predictive of which case the participant is performing. To investigate this, we train a classifier using Penalized Multinomial Regression and Leave One Out Cross Validation (i.e. we train the classifier on all data except one case, and ask if, based on learnt patterns in the remaining data, whether we are able to predict which case/condition it is based on information seeking patterns). Our input parameters are the available information requests as binary predictors (i.e. to denote whether they were sought on each case or not). The outcome variable of the classifier is the patient condition. We generate model predictions and then look at whether they correctly match the actual condition for that case. Across 510 cases (85 participants performing 6 cases each), the accuracy of the classifier was 57%, which is higher than the chance level of 16.6%. When breaking down accuracy of our classifier by condition, we find accuracy to be above chance across all conditions (see Table 3.2 below).

Condition	Prediction Accuracy
AD	0.65
GBS	0.46
MTB	0.48
TA	0.47
TTP	0.56
UC	0.78

Table 3.2: The accuracy of our multinomial classifier that predicts patient condition for each case based on the information sought/not sought as binary predictors. We then test the accuracy of the classifier by comparing the predicted condition from the model against the actual patient condition for each case. We then split cases by condition to look at accuracy on a case-by-case level. Given that participants perform 6 cases each, accuracy would be 1/6 (16.6%) when at chance.

Taking these findings together, keeping information seeking more constant (i.e. requesting similar high-value information) across cases was found to have an association with accuracy whilst there also being some information that is useful for clinicians to know for patients with specific conditions. To reconcile these, we derive which information requests were most weighted in our classifier models to find which were considered markers of accuracy (by being sought across cases) and which were considered markers of identifying specific cases. We extract coefficients from the logistic classifier of accuracy (the ROC curve for which was shown in Figure 3.8) and the multinomial classifier (the accuracy of which was depicted above in Table 3.2). We identify the highest weighted information requests as input parameters for each model. The five highest weighted information requests for each model are shown below in Table 3.3. We also show how often each piece of information was sought for each of the cases in Figure 3.11 below. Viewing this figure shows individual tests that are useful for specific cases. For example, an ECG is sought by most participants for the AD (a heart condition) case.

	Accuracy			**Condition**		
Rank	Test Name	Coefficient	Odds Ratio	Test Name	Coefficient	Odds Ratio
1	FBC	0.45	1.57	Neurologic Exam Record	0.22	1.25
2	Venous Blood Gas	-0.44	0.65	Measure Blood Pressure	0.22	1.25
3	Urine Dipstick	-0.39	0.67	UREA and Electrolytes	0.2	1.22
4	Assess Extremities	0.38	1.47	Rectal Examination	0.17	1.19
5	Other Biochemistry Tests	-0.36	0.7	Urine Dipstick	0.16	1.17

Table 3.3: The five highest weighted parameters (by the absolute value of the coefficient values) for our logistic classifier of participant accuracy (under the "Accuracy" heading above) and our multinomial classifier of patient condition (under the "Condition" heading above). We also show coefficient values and odds ratio values for each parameter.

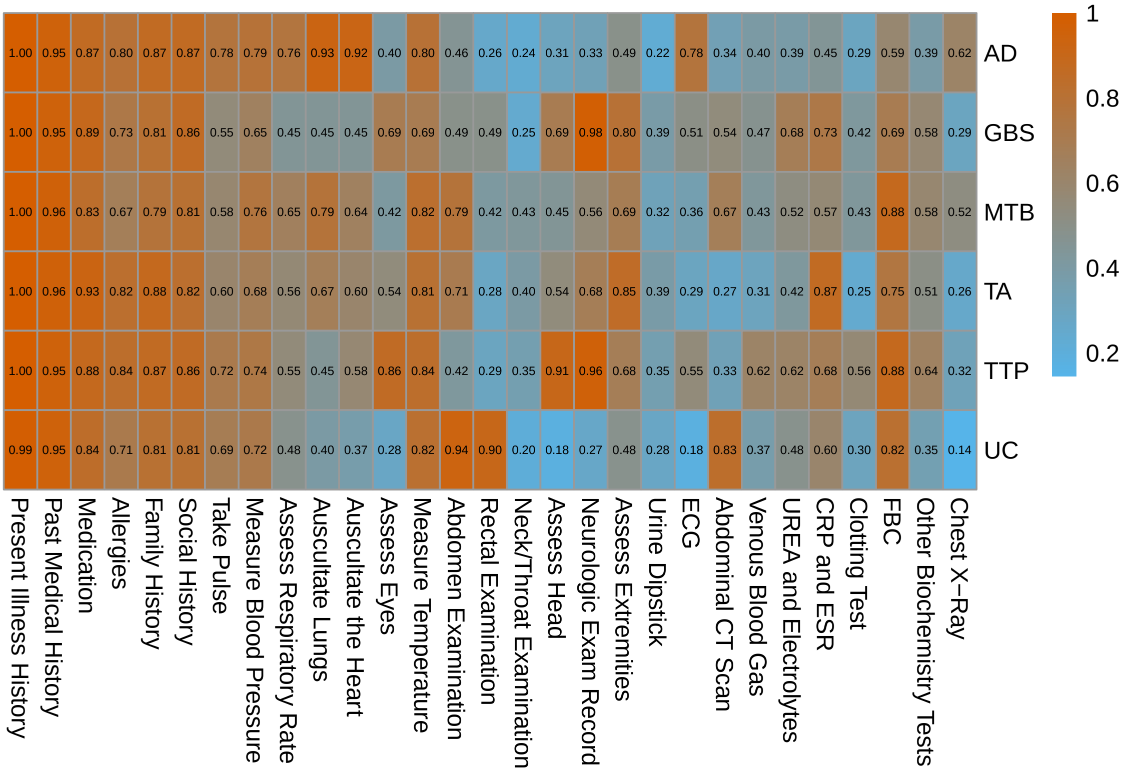


Figure 3.11: Visualisation of the proportion of participants who sought each available piece of information (columns, x-axis) broken down by case (rows, y-axis). Lighter blue colours indicate that fewer participants sought that information for a given case (i.e. towards 0%), whilst lighter orange colours indicate more participants sought that information for a given case (i.e. towards 100%)

Discussion

This study of medical students explored the interplay between confidence, accuracy and information seeking in a novel medical diagnosis task. Using an online interface, we explored how medical students work through diagnostic scenarios, freely seeking information to develop and test sets of possible differentials. Our aim was to look at how different aspects of information seeking impacts both diagnostic confidence and accuracy. The main strength of this study's paradigm is in allowing us to investigate the diagnostic process as it evolves over time and with more information, rather than as a single decision at a single point in time. By tracking how both confidence and the diagnoses considered by participants changes over time, we gain a better understanding of how the manner in which information sought is key to the diagnostic process and to clinicians' subjective confidence.

Calibration of Confidence

On the question of whether medical students provided confidence judgements that were calibrated to their objective accuracy, we found that students become more accurate across successive stages of information seeking as well more confident. However, cases varied in difficulty as reflected in participant accuracy. In particular, the AD and MTB cases exhibited lower observed accuracy across participants. We observed overconfidence for these two cases, and underconfidence for the UC case (for which accuracy was highest). This indicates a classical hard-easy effect of confidence (**lichtenstein_calibration_1977**), whereby individuals have a greater tendency to be overconfident for more difficult decisions when compared to easier decision (**merkle_disutility_2009**). Confidence also increased as participants received more information. However, students reported 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_fallacy_2023**). This may also be a function of undertaking the diagnostic process in isolation (i.e. without being able to discuss with colleagues, as would be the case in naturalistic medical environments). Such a reduction in confidence when making a decision alone rather than in group would be justified from a calibration perspective, as combining medical students' diagnoses has been found to improve accuracy (**kammer_potential_2017**).

Previous work (e.g. **meyer_physicians_2013**) has revealed a gap between subjective confidence and objective accuracy. In particular, a general tendency has been demonstrated for less experienced medical trainees to be underconfident and for more experienced medical professionals to be overconfident (**yang__nurses__2010**). Part of the discrepancy between our findings and past findings could stem from the way that diagnostic uncertainty is expressed by students in this study. Using our primary measure of accuracy, which is obtained by using the likelihood values assigned to correct differentials (if included), we find that accuracy tracks confidence quite closely at each information stage. We note however that our finding of calibrated confidence is highly contingent on the measure of accuracy used. When using a more lenient measure, the proportion of cases where a correct differential was reported (as used in previous papers, (**friedman__physicians__2005**; **meyer__physicians__2013**; **lambe__guided__2018**; **kuper__mitigating__2024**)), participants were found to be underconfident. When using a stricter measure, the likelihood value assigned to the most likely differential if it is correct, participants are found to be overconfident. Calibration also varied across cases, with participants sometimes showing overconfidence and sometimes showing underconfidence. While we therefore temper our finding of calibration, this has implications for further research that looks at calibration during diagnoses, given that accuracy can be defined in multiple ways when participants record multiple differentials. In addition, our confidence measure is related to the participants' subjective readiness to treat the patient, rather than confidence in the set of differentials. Such a measure of confidence is novel to our study and has not been used in previous studies of diagnostic confidence. This limits the extent to which we can compare accuracy and confidence directly. However, rather than confidence being a subjective judgement, we connect to it to clinical action that would be taken by participants if the patient presented were real. This is similar to one paper in which confidence was measured as the subjective likelihood of seeking assistance to reach a diagnosis (**friedman__physicians__2005**), with the authors finding that medical students had a lower tendency toward confidence than both medical residents and faculty. When considering this result alongside our own finding of low confidence across medical students, it is possible that tempering overconfidence may stem from tying judgements to specific clinical actions. Given that medical students lack the experience of more senior clinicians, they may generally be less confident as a result: the lower reported confidence is partly a reflection of their general aptitude/experience with the clinical action being prompted during the confidence judgement. Future work could then measure how confidence relates to specific aspects of the patient care pathway and differences in calibration.

Broadening of Differentials

On the question of whether participants tend to broaden or narrow their differentials with new information, participants exhibited a general pattern of broadening the range of differentials they were considering across successive information seeking stages. In addition, we observed that participants did not tend to remove differentials from consideration despite having the option to do so. This marks a novel finding when situated within past research, which has not studied how the differentials being considered evolves over time. We can interpret this as students being careful not to miss differentials from consideration, indicating a focus on being comprehensive in their generation of differentials rather than a focus on narrowing in on a single diagnosis. It is therefore worth considering whether students are explicitly taught not to disregard diagnoses completely, instead focusing on remaining open-minded to new possibilities for differentials. (**joseph_domain_1990**) found that clinicians with lower domain knowledge generated accurate hypotheses but were unable to differentiate eliminate hypotheses when receiving more information, unlike clinicians with higher domain knowledge who were able to confirm and eliminate hypotheses using the information received. This may help explain the broadening pattern of medical students, as their relative inexperience meant that they were not able to easily eliminate hypotheses.

We also found that the initial breadth of diagnoses considered from the patients' history was predictive of the amount of subsequent information seeking and changes in confidence. We also find that how much participants change the number of differentials they are considering is predictive of information seeking and changes in confidence. Relatedly, information seeking and confidence was associated, such that participants who sought more information tended to increase their confidence more over the course of the diagnoses. However, the amount of information sought was not predictive of diagnostic accuracy, with accuracy instead being associated with seeking more valuable/appropriate information for a given patient condition. When taken together, these findings give an interesting picture of the diagnostic process as we capture it within our task. Our account of how participants approach this task can be summarised as follows (note that this account requires follow-up study to elucidate further):

- Medical students generate an initial set of differentials from the patient history and use this to guide their information seeking.
- With more differentials to consider, students seek more information to 'test' each of these hypotheses.
- Seeking more information increases the likelihood that new differentials are brought to mind, resulting in more differentials being added as under consideration.
- When participants have more information and have considered a wider range of differentials, they are likely to increase their confidence due to being more comprehensive (i.e. considering more differentials) during their thought process.
- With more differentials being considered, participants are more likely to consider a 'correct' differential. However, considering a larger set of differentials makes it more difficult to focus on finding a differential that is most likely.

Aspects of Information Seeking

Given the flexibility afforded by our paradigm, we are able to monitor fine-grained aspects of how participants seek information. We find that the accuracy and confidence gained over the course of cases was related to the quality of the information sought. We also find that higher accuracy was associated with less variability in information seeking (i.e. seeking a similar set of information regardless of the patient case). Higher accuracy participants were found to be more alike in their information seeking compared to lower accuracy participants. Putting these findings together, we can surmise that each patient condition has associated valuable pieces of information that are worth seeking, but that there is a consistent set of information that accurate participants tend to seek across cases. When combined, each case can be seen to have an 'optimal' set of information that participants should seek. In addition, while seeking more information may increase confidence, having more information may be problematic for weighing up differentials against each other. This is because it can be harder to synthesise more information into a cohesive account of the patient. While past work has called for greater standardisation within healthcare (**wears_standardisation_2015**), what seems to constitute accurate diagnoses in our task is a degree of standardisation with certain selectivity of information given the patient condition. As depicted in Table 3.3, certain information is useful regardless of patient condition whilst others are useful for specific medical conditions. While we show certain tests/examinations as being most useful across patient cases or for specific cases, we recommend caution in interpreting these as representative of all diagnostic decisions outside of this task. These specific information requests were found to be useful for our task, but may not generalise to other patient conditions or diagnostic decisions.

Follow-up Think-Aloud Study

Given these results, we know what information is sought by medical students, but only have a limited insight into why they sought certain information and how it directly affected the diagnosis they provided. For one, are all students using a similar decision making process when making diagnoses? As of now, we are inferring the participants' thought processes from data of their differentials and information seeking without context of how they are thinking about the task. One possibility is that there are differences in how medical students approach diagnoses that stem from differing reasoning strategies, which we cannot infer from this current dataset. In order to ascertain this, we would need to record the students' thought processes as they are doing the task. To this end, we conduct a follow-up study using a similar diagnostic paradigm conducted in-person where students think out loud as they make diagnoses.

coderre_diagnostic_2003 used a think-aloud paradigm to characterise distinct diagnostic reasoning strategies: a "hypothetico-deductive" strategy that is closest to the idealised process of elimination that is the typical characterisation of diagnosis, a "pattern matching" strategy where clinicians draw similarities between the current patient and either a past patient or prototypical case of a particular condition, and a "scheme-inductive" strategy in which clinicians follow a structured framework for diagnoses (e.g. a surgical sieve, that considers each pathophysiological system in turn). Of interest to our work was whether we would observe similar variation in reasoning strategies in our medical trainees and, if so, how these strategies related to patterns of information seeking and confidence We hypothesise not only that we can detect reasoning strategies based on the verbalisations of participants' thought process, but that different reasoning strategies for generating differentials are useful for some cases more than others. We also hypothesise that information seeking and changes in confidence vary as a function of the reasoning strategy employed.

Given the recording of qualitative data during this task, we can understand both how medical students are thinking about diagnoses as they are making them but also how they reflect on their thought process outside of the task. This detection of reasoning strategies, if successful, can then subsequently be used to detect the same reasoning strategies in this online study dataset (where we do not have access to the participants' thought process) based on the information sought. Given the higher sample size afforded by the online study, we can more robustly look at differences between reasoning strategies and whether they can tell us about what makes more accurate and more confident diagnoses.

Chapter 4 - Characterising Diagnostic Reasoning Strategies via a Think-Aloud Paradigm

Introduction

The study presented in the previous chapter was online vignette study that investigated confidence and information seeking within evolving diagnostic decisions. Overall, we found that medical students provided confidence judgements that were well-calibrated with their objective accuracy. We found that higher information seeking was associated with larger increases in confidence over the course of diagnostic decisions. Accuracy was associated with selectivity in information seeking, such that certain information was beneficial to students regardless of the patient's condition. We also found that medical students tended to broaden the range of differentials they were considering with more information, as they rarely removed differentials from their consideration. Students who considered more differentials early on tended to seek more information and increase their confidence to a greater extent. We also found that there were case-specific differences in confidence and accuracy, which affected calibration for individual cases.

Given that we find that certain information seeking patterns were associated with accuracy, and that students were overall well-calibrated in their confidence judgements, we now aim to better understand how students approach the task in terms of their decision making strategies. As the previous study was conducted online, we are not able to ascertain why students sought certain information or why they considered certain differentials (as likely or unlikely). To address this, we present in this chapter a study using a modified version of the previous study's vignette methodology where students verbalised their thought processes as they were performing their diagnoses.

In this mixed-methods study, we aimed to gain insight on the types of reasoning strategies used by medical students and how these strategies influence both their information seeking patterns and changes in confidence over the course of their diagnostic decisions. We also investigated why reasoning strategies may vary on a case-by-case basis. We utilise a very similar experimental procedure to our previous study (using the same patient vignettes), but rather than explicitly asking students to report the differentials they were considering, we instead prompted students to think out loud as they were performing the diagnostic task. Everything said by participants was audio-recorded, transcribed and then coded for both quantitative and qualitative analysis. We used this method to derive a richer understanding of the diagnostic process as it pertains to medical students' evolving thought processes.

With this study, we aim to understand how differences in information seeking patterns arise. One possibility is that these differences stem from reasoning strategies adopted when making diagnoses. Clinical reasoning is a key skill that is taught, either explicitly or implicitly, within medical education. Medical decisions are frequently made in uncertainty, with clinical reasoning taught as a skill to navigate this uncertainty (with perhaps an intention to reduce the uncertainty perceived by clinicians). However, clinical reasoning has a broad remit and covers multiple different approaches to making medical decisions. Doctors may have different considerations when diagnosing a patient and may draw on different approaches to making medical decisions accordingly. As we noted in the design of our previous study, doctors may have to consider what conditions are likely for a patient and what conditions are too severe to not miss. A doctor's reasoning strategy may then reflect this dichotomy, such that the focus is either on determining what is most likely or being thorough such as to consider all possible diagnoses.

Diagnostic decisions have traditionally been thought of as ‘ideal’ when using the hypothetico-deductive process (**kuipers_causal_1984**), whereby hypotheses are initially formulated based on patient symptoms and established criteria for diagnoses. Further information is then gathered to test these hypotheses (**higgs_clinical_2019**) or eliminate others. One can think of this approach as akin to a ‘process of elimination’: by starting broad, clinicians then seek information to reduce the potential set of diagnoses to a more manageable set. This theory, that hypothetico-deductive processes are the gold standard for diagnostic decisions, was challenged by the results of **coderre_diagnostic_2003**. The authors found that reasoning strategies differed between novice and expert clinicians and that, crucially, a pattern recognition approach (rather than hypothethico-deductive) was associated with higher diagnostic accuracy. Pattern recognition was used by expert clinicians on a majority of cases (with novice clinicians using a hypothethico-deductive approach most often), with expert clinicians being more accurate in their diagnoses. A pattern recognition approach involves considering fewer diagnostic hypotheses and instead matching the symptoms to prototypical cases of a patient condition. Hypothetico-deductive reasoning represents a more structured, systematic approach to make diagnoses, in which clinicians maintain a more open mind to different diagnostic possibilities, whereas pattern recognition is more automatic and driven by intuition. The difference between these approaches can be related to past work (**norman_etiology_2014**) that has connected medical decision making to the dual-system model of thought (**kahneman_thinking_2011**). System 1 thinking represents an automatic, intuitive mode for making decisions (akin to pattern recognition), whilst System 2 represents a more deliberative, rational mode for decision making (with hypothethico-deductive reasoning as an example).

In their paper, **coderre_diagnostic_2003** asked novice and expert clinicians to provide diagnoses based on patient vignettes and asked them to verbalise their thought processes whilst performing the diagnostic task. Using these verbalisations, the researchers categorised the clinicians' reasoning strategy on each case. They defined three reasoning strategies as follows (paraphrased from their paper):

- *Hypothetico-deductive (HD) strategy*: prior to selecting the most likely diagnosis, the clinician analysed, one by one, each alternative diagnosis.
- *Scheme-inductive (SI) strategy*: This strategy consisted of key predetermined propositions that linked categories and thus provided evidence for chunking (i.e. scheme use). These propositions were presented as structuring diagnoses by different pathophysiological systems/processes (e.g. small bowel vs. Large Bowel, gastrointestinal (GI) vs. non-GI causes).
- *Pattern Recognition (PR) strategy*: The clinician directly reached a single diagnosis with only perfunctory attention to alternative diagnoses.

These reasoning strategies may have been explicitly taught during a clinician's medical education or implicitly developed with experience. With these differing strategies, we can consider whether there are normatively 'better' strategies to use in certain clinical situations. As highlighted with the previous set of conflicting literature, there is currently no consensus within medicine as to which strategy is ideal for diagnostic accuracy. In their study, Coderre et al. found that novice clinicians tended to adopt a HD strategy more often, whereas experts tended to use a PR strategy. In addition, using a PR strategy was associated with higher diagnostic accuracy, which in turn was used to explain why experts were more accurate than novices. In addition to explaining differences in accuracy, these strategies point to different approaches in how diagnostic differentials are generated and considered.

The different diagnostic strategies (PR, SI and HD) reflect different approaches for reaching a diagnosis, whether it be a more intuitive approach (PR) or a more thorough, systematic consideration of possible differentials (SI and HD). The work of **coderre_diagnostic_2003** did not reveal whether these strategies result in differences in information seeking and confidence. We would expect that with a more thorough SI or HD reasoning strategy, participants would seek more information in order to consider a larger number of diagnostic possibilities. By contrast, we would expect that a PR strategy would result in selective (but less) information seeking in order to gather evidence in support of the single diagnosis being considered. By better understanding how different reasoning strategies manifest in differences to diagnostic decisions, we aim to inform future work within medical education and cognitive intervention design to consider when each strategy is appropriate and prompt clinicians accordingly to follow that strategy's decision process. This can also shift our understanding of clinical reasoning as having a 'one size fits all' approach where one method for making decisions is 'ideal' for all possible medical decisions.

In order to pick apart these different reasoning strategies within our vignette-based paradigm, we adopted a think-aloud methodology similar to **coderre_diagnostic_2003**, whereby participants verbalise their thought process as they are doing the vignette-based diagnosis task. Think-aloud methodologies are useful for directly accessing ongoing thought processes during decisions (**someren_think_1994**). The use of thinking aloud (or 'verbal protocols') in research is useful for being able to access the information that is attended to participants in short term memory (**payne_thinking_1994**) and can be treated as the ongoing behavioural state of a participant's knowledge (**newell_human_1972**). Think-aloud protocols have historically been used to study problem solving, particularly for comparing how novices and experts solve problems such as finding the best move in chess (**groot_thought_1978**). Diagnosis is a decisional process that develops over time and allowing participants to think aloud reflects this by providing a time-ordered sequence of how thought processes develop (**payne_thinking_1994**). This is especially well-suited to our task where the information available to participants is controlled at discrete time points, allowing us to investigate how diagnostic thinking develops with more information. We also can connect the verbalisations in our task to the exact information participants received to prompt that thought. In order to understand how medical students view their own thought process, we administer a post-experiment interview to ask students about how they tend to make diagnostic decisions and what their main considerations are.

In order to bolster our findings from this study, we aim to use the reasoning strategies determined from this study to reanalyse the cases in the online study. Whilst we record information seeking, confidence and differential/hypothesis generation behaviour during the previous study, we do not have an understanding of how participants are approaching each case from a reasoning strategy perspective. One way in which we can to some extent infer this is via the information seeking patterns that participants adopt. If our hypothesis is correct that reasoning strategies result in different patterns of information seeking, we should then we able to predict what strategy a participant is using solely from their information seeking. If we can also establish that these predictions are reliable, we can then study the properties of these reasoning strategies with the larger sample size afforded to us in the online study. This work would then improve our understanding of how these reasoning strategies not only affect diagnostic accuracy, but how they contribute to information seeking, confidence and differential generation.

As well as addressing new questions about strategies in diagnosis, the present think-aloud study also provided an opportunity to replicate some key findings from the online study. In our previous study, we observed a general tendency for participants to broaden the set of differentials they considered as they received more information. This was reflected in the average number of differentials reported by the end of the case being higher than the average number of initial differentials (based on the patient history). This effect was driven by participants (74 out of 85 participants, 87%) never reporting fewer differentials at the Testing stage compared to the Patient History stage. This is a surprising result, as we may have expected participants to use a 'process of elimination', which would manifest in decreasing the number of differentials considered as participants receive more information. What this speaks to however is a general reticence to remove differentials from consideration. One of our focuses for this study then is to replicate this finding by examining if students' thought processes reflect this tendency to focus on broadening rather than narrowing differentials being considered. This is to ensure that the finding in our previous study is not merely a quirk of our experimental interface and design, as it is possible that participants are not sufficiently encouraged to disregard differentials within our online paradigm. Replicating this finding would potentially reveal this general reticence to remove differentials as a potential driver of diagnostic uncertainty.

Research Questions

In this study, we investigate the following research questions as they are reflected in the verbalised thought processes of medical students as they perform our diagnosis task:

- Do students report ruling out differentials as they seek information on patients during diagnoses (as we observed from the online study presented in Chapter 2)?
- Can we define different reasoning strategies based on the think-aloud utterances of medical students?
- If so, what reasoning strategies are medical students using when making diagnoses and weighing up differentials?
- How do differences in reasoning strategy manifest in terms of information seeking, both in terms of the quality and quantity of information sought?
- Are differences in reasoning strategy related to the individual or are they dependent on the case at hand?
- Do more accurate medical students utilise specific reasoning strategies?

- What considerations do medical students report having whilst they are making diagnoses? And how do these differ from how medical students reflect on their thought processes after performing the task?

Methods

Participants

16 participants were recruited for this mixed-methods study. Participants were 5th or 6th year medical students at the University of Oxford (including 2nd year Graduate Entry Medical students) recruited via physical posters at Oxford's John Radcliffe Hospital and via a mailing list for students managed by the Medical Sciences Division at the University of Oxford. The study was conducted onsite at the John Radcliffe hospital. Participants were recruited between July 5th 2023 and December 1st 2023. This study was reviewed and granted ethical approval as an amendment to our existing protocol to allow for audio recordings by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE004.

Materials

The same set of vignettes and a similar computer interface to the online experiment were used for this study, with the exception that participants no longer explicitly recorded their differentials at the end of each information gathering stage. Instead, participants' differentials were recorded in a more naturalistic way as participants verbalised out loud their thought process as they worked through each diagnostic case. 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. Informed consent was obtained anonymously using an online electronic information sheet and consent form. Information, including experimental data and audio recordings, collected during the study were stored under anonymised

IDs with no linkages to participants. Data was kept on a password-protected computer and hard drive.

Procedure

The general procedure was very similar to that of the online experiment, except that participants were given the following standardised 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 (Aiyer) occasionally prompted participants with content-neutral probes: *"can you tell me what you are thinking?"* in cases of periods of long silence, and *"can you tell me more?"* when the participant said something vague that may warrant further detail. We emphasise that these are non-leading questions. The audio of the participants' verbalisations was recorded and then transcribed. An initial transcript was generated using Microsoft Office's transcription feature, then the transcript was checked and modified for accuracy by listening through the audio recordings. 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. Given that participants were encouraged to verbalise their considered differentials as they were performing the task, we did not show participants the screen where they explicitly listed the differentials they were considering. At the end of the experiment, the researcher administered a semi-structured interview to better understand participants' views of their own diagnostic reasoning approach. This interview allows us to detect common themes across students on how they tend to make diagnostic decisions from a reasoning perspective, as well as coding each student's subjectively preferred reasoning strategy. The interview questions are provided in the Appendices.

Aside from these differences, participants performed the same six patient vignettes (in a randomised order) from the online vignette study using the same interface that allowed them to seek information that they think is useful for that particular case. We also used the same practice case (Colon Cancer) from the online study to familiarise participants with the experimental interface. Participants were able to practice thinking aloud during this practice case, but utterances during this case were not included in any analyses.

Data Analysis

Our data analysis process for this mixed-methods study is split into a few parts. We first describe the main quantitative variables and analysis for this study. We then detail our coding process for detecting reasoning strategies based on participants' think-aloud utterances, followed by quantitative analysis that we perform based on these coded reasoning strategies. We then describe the qualitative analysis performed based on the recorded debrief interviews with participants. Finally, we detail the process by which we code for reasoning strategies in the previous experiment's dataset.

Descriptive Quantitative Analysis

The variables defined for this study are similar to those in the online experiment, as we utilised the same interface and vignettes. Specifically, the variables for confidence, subjective difficulty and information seeking are the same as in the online experiment. We note some key differences however. Firstly, given that participants did not explicitly report the differentials they were considering at each information stage, we are not able to record the number of differentials in the same way at each stage. Secondly, we also define accuracy differently due to the lack of this differential reporting screen:

- *Accuracy*: Each case is defined as correct if a differential that is considered correct (as per our marking criteria in the Appendices) is mentioned by the participant at some point during the case.

We also code all utterances related to differential/hypothesis generation. We define instances of Differential Evaluation, which is a main code that comprises a number of subcodes that we apply to think-aloud utterances. These are defined as follows:

- **Differential Evaluation:** any time that the participant (each of the following is considered a separate subcode):
  - *Differential Added:* - Mentions a new condition that they are considering
  - *Differential Removed:* - Rules out or eliminates a condition from consideration
  - *Likelihood Increased:* - Mention of increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition
  - *Likelihood Decreased:* - Mention of decreased likelihood of a previously mentioned condition, or that information seems to rule out a particular condition

Based on this, in lieu of the Number of Differentials variable from our previous study, we define a new variable to look at the number of instances in which participants evaluate or reevaluate the differentials they are considering:

- **Number of Differential Evaluations**: The number of instances of the above subcodes belonging to the main Differential Evaluation code. The number of such utterances are defined for each individual case. The higher this number, the more participants are 'updating' their thinking around what differentials they were considering as likely/unlikely for the patient.

Coding of Reasoning Strategies

We aimed to detect which reasoning strategies are used by students on each case. To code for reasoning strategies, we adopt a similar approach to **coderre_diagnostic_2003**. Codes were initially chosen by two independent coders, and then conflicts were resolved between the two. We defined coding criteria that indicate three different diagnostic reasoning strategies: hypothetico-deductive reasoning, scheme-inductive reasoning and pattern recognition **coderre_diagnostic_2003**. These were defined as follows:

- **Hypothetico-Deductive Reasoning (HD)** - 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 (SI)** - 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 (PR)** - 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").

- **None** - cases are defined as not having a clear reasoning strategy if there are insufficient utterances to make an inference that a participant is using a particular reasoning strategy (as agreed by both coders).

We first code specific statements 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. Coding of utterances and case-wise reasoning strategies were conducted with a second independent coder from a non-medical background who was not involved with the project prior to coding (in order to have an impartial perspective). For reasoning strategies, initial interrater reliability was low, with both coders agreeing on 58.3% of cases. When resolving these initial conflicts, changes were made to the coding criteria to prioritise strategies used early on in a case, as some participants were noted to utilise multiple strategies within a single case. For example, a participant may use a SI approach to focus on different pathophysiological systems and then adopt a PR approach to identify an appropriate diagnosis within a specific pathophysiological system. The coding criteria was also changed to allow cases to be coded as not having a clear strategy due to a lack of utterances. The coding criteria provided above are after these changes were made. Cases were then independently coded for a second time with these updated criteria. Both coders agreed on 78% of cases when coding for correctness, with conflicts resolved in consultation with a member of expert panel used to develop the vignettes (as described in the online experiment).

We hypothesise not only that this think-aloud methodology can be used to detect different reasoning strategies but also that usage of these reasoning strategies will vary across medical students and as a function of the specific case/patient condition being treated. For the former, the reasoning strategy may be more related to the individual medical student, in that each student will have their own preferred strategy that they tend to use regardless of the patient. For the latter, there would be properties of specific patient conditions that prompt usage of certain reasoning strategies. We investigate both of these theories within this study, using deductive analysis as well as inductive analysis (i.e. we are prepared to reveal additional insights whilst analysing data that does not feed into these two theories). To look at individual-level strategy, we ask participants about their diagnostic reasoning process during the debrief interview. This includes questions such as *"What's your general approach to making diagnoses?"* and *"Do you tend to keep a broad set of differentials in mind?"* (full list of questions available in the Appendices). Based on their responses, participants were each categorised as belonging to one of the three reasoning strategies, with each coder independently coding each participant, with then conflicts resolved. This is considered their 'subjective strategy'. To look at condition-level strategies, after classifying each case using independent coders, we find the most commonly used strategy for each condition. This is considered the condition's 'dominant strategy'. Once both of these are defined, we compare the strategies coded for each case against both the subjective and dominant strategies. By comparing the cases' reasoning strategies with both the subjective and dominant strategies (using Binomial Exact Tests), we can investigate whether it is the individual medical student or the patient's medical condition that is responsible for the choice of reasoning strategy on a given case.

We then compare our dependent variables (Accuracy, Confidence, Information Seeking, Differential Evaluations) as a function of reasoning strategy. As we compare observations on a case-by-case basis, observations are not independent, as student each record multiple cases and there are multiple cases for each condition. As such, we conduct generalised mixed effect modelling to analyse the effect of reasoning strategy on these variables. For Accuracy, we fit a binomial logistic model for correctness as a binary outcome variable. For the other variables, because we are using non-normally distributed continuous outcome variables, we compare model fit metrics (Aikake Information Criterion and Bayesian Information Criterion) for generalised models (with the *glmer* function in R's *lme4* package) using inverse Gaussian and Gamma distributions, reporting results for the better fitting model (as outputted by the *anova* function in R's *stats* package for model comparison).

Qualitative Thematic Analysis

The aim of this thematic analysis was to identify the reasoning strategies that medical students in this study reflectively report using in their medical practice, as well as understanding considerations made by students when making diagnoses. Similar to the think-aloud utterances by participants during the vignette task, we record and transcribe the responses given by participants during the debrief interviews (administered after the vignette task). With these interviews, we aimed to understand how participants report making diagnostic decisions, including how they seek information, weigh up differentials against each other and whether they use any cognitive frameworks during diagnoses. These interviews were important for providing an account of how participants would characterise the general approach they adopt during diagnoses, which may or may not correspond with the approach they actually adopt (as coded from their think-aloud utterances). Based on these transcribed responses, we conducted a theory-driven semantic thematic analysis (as per definitions detailed by (**braun_using_2006**)) to code utterances under specific categories. This kind of thematic analysis is suitable given that our qualitative data is from a semi-structured interview with predefined research questions of interest (whilst allowing participants to deviate slightly based on their responses), rather than a dataset with a looser structure. Similar to the coding of reasoning strategies outlined in the previous section, thematic analysis was conducted using two independent coders. Each individual coder analysed the data from the debrief interviews and independently coded emerging themes using NVivo. After this point, themes that were common to both coders were included for further analysis, with these themes being discussed in the Results section of this chapter and related to our quantitative analyses for triangulation.

Recoding of Reasoning Strategies in Online Vignette Study

As laid out in the introduction, we ask with this whether we are able to detect each participant's reasoning strategy on a case-by-case basis in the online experiment. We used the think-aloud data to identify strategies that our participants used, with the aim of then exploring the use of these strategies in the online experiment dataset. We did this in two ways: one was a fine-grained analysis of information seeking behaviour as 'fingerprints' of the various reasoning strategies being used. The other was simpler, in terms of identifying the dominant strategy associated with each of the different cases.

For the former of these analyses, we assume that the reasoning strategies we have defined are differentiated from each other by their information seeking. We establish whether the predictions of reasoning strategy are able to reliably correspond with the 'ground truth' labels of reasoning strategy via our independent interrater coding process. If these predictions of reasoning strategy are reliable within the think-aloud study, we can then apply the same method for predicting based on information seeking in the online study. We can then determine, given medical students' reasoning strategies, whether these strategies are associated with differences in information seeking and confidence (with the increased robustness of the online study's higher sample size). In order to apply reasoning strategies to the data from the online experiment, we train a classifier using penalised multinomial logistic regression to classify cases as HD, PR or SI using the cases from the think aloud study (with Leave One Out Cross Validation). Cases that are not coded with a reasoning strategy due to insufficient utterances to infer are excluded from this model training. The input parameters for the classifier are the 29 pieces of information as binary predictors (similar to the approach depicted in Figure 3.8 of the previous study) and the cases' condition. In other words, the cases from the think-aloud study make up the training data for the classifier whilst the cases from the larger online study make up the test dataset. The classifier was implemented using R's caret package with the train() and glmnet() functions. The testing data is then labelled with predicted strategies using R's predict function. We report accuracy values for the classifier when applied to the think-aloud study (i.e. how well the classifier is able to reproduce the coded reasoning strategies).

For the latter analysis, we use the think-aloud methodology to determine the dominant strategy for a given case by observing which reasoning strategy was used by a majority of participants. We then look across cases for a given dominant strategy, how our key dependent variables are affected by the type of case. We group cases in the online dataseet by the dominant strategy used in Study 3 for each condition. For example, if a majority of participants use a HD strategy in the think-aloud study for the Ulcerative Colitis (UC) case, we then consider UC to be HD-dominant.

Results

Across the 16 participants, there were no audio issues that impeded transcription to the point where participants were no longer audible. As a result, all participants were included in the subsequent analyses. This constituted over 21 hours worth of recorded audio that was transcribed, with participants taking between 50 and 114 minutes to complete both the experiment and the debrief interview. As there was minimal guidance given to participants as they performed the task and verbalised their thought process, they varied widely in how much they spoke during the study, uttering 1038-7730 words (M = 4194) across the scenarios. Part of this range is driven by some participants repeating information they see during the task, but participants also varied in terms of how much they externalised their thought processes.

Descriptive Quantitative Results

Overall Performance and Calibration

Accuracy (the proportion of cases where a correct differential was mentioned by the participant) was 0.57 across all cases. This differed by case, as can be seen below in Table 4.1. When compared to the online experiment, confidence judgments were similar to those provided in this study. Accuracy, on the other hand, was lower particularly for three cases (MTB, TA, UC) during this study when compared to the online experiment. Similar to the previous study, we ask whether participants provided confidence judgements that were, on the whole, calibrated to their objective accuracy. In order to investigate this for this study, we compare the final confidence reported by participants on cases when they were objectively correct and when they were objectively incorrect. Participants would be considered calibrated if we found evidence of higher confidence when correct compared to when participants were incorrect. Participants did indeed report higher confidence when correct (M = 50, SD = 24) compared to when they were incorrect (M = 42, SD = 27). Given that the samples from each of these groups are not independent, we test for a difference between these groups using a binomial mixed effects model that predicts correctness using final confidence as a fixed effect and both the individual participant and condition as random effects. We find evidence that the case correctness was predicted by final confidence (β = 0.03, SE = 0.01, z = 2.26, p = 0.02). This provides evidence that participants' confidence judgements that were well calibrated to their objective accuracy, as confidence judgements were sensitive to whether participants were correct or incorrect across cases.

Case	Accuracy	Final Confidence	Difficulty	Information Seeking
AD	0.62	0.43	6.8	0.77
GBS	0.88	0.37	7.4	0.76
MTB	0.19	0.52	5.9	0.82
TA	0.44	0.48	6.3	0.79
TTP	0.69	0.40	7.1	0.80
UC	0.62	0.62	5.3	0.73

Table 4.1: Table of data from this think-aloud study showing, by case (AD = Aortic Dissection, GBS = Guillain Barré Syndrome, MTB = Miliary Tuberculosis, TA = Temporal Arteritis, TTP = Thrombotic Thrombocytopenic Purpura, UC =

Ulcerative Colitis), from left to right, average values across participants for Accuracy (the proportion of participants who mentioned a correct differential during the case), Final Confidence (reported at the Testing stage), Difficulty (as rated by participants at the end of the case on a scale of 1-10) and Information Seeking (the proportion of available information sought).

Differential Evaluations

Case	Differential Added	Differential Removed	Increased Likelihood	Decreased Likelihood	Total Differential Evaluations
AD	4.69	0.50	2.06	1.12	8.88
GBS	2.50	0.50	0.44	1.44	4.94
MTB	3.19	0.12	0.94	0.44	4.75
TA	2.12	0.25	0.69	0.62	3.75
TTP	2.81	0.12	0.56	1.12	4.69
UC	3.50	0.12	0.88	1.19	5.88

Table 4.2: Descriptive Statistics for subcodes within the Differential Evaluation main code as detailed above in the Data Analysis section. Shown above are mean values for the number of instances/utterances for each of the following subcodes (from left to right): a new differential being considered, a differential being removed from consideration, a differential being seen as more likely given a piece of information, a differential being seen as less likely given a piece of information, the average total of these subcodes.

For Differential Evaluations, participants on average made 5.5 such utterances per case. The mean number of Differential Evaluations was similar across case except for the AD case, for which we observed a higher number of Differential Evaluations (see Table 4.2 above). As previously mentioned, Differential Evaluations can be further categorised into one of four subcodes: Differential Added, Differential Removed, Likelihood Increased and Likelihood Decreased (average numbers of each are visualised below in Figure 4.1). As found in the previous study, there is a general reticence to disregard differentials completely. Participants expressed significantly more statements adding differentials (M = 3.14, SD = 1.66) than removing differentials (M = 0.27, SD = 0.53) (t(15) = 14.14, MDiff = 2.86, p < .001). Out of the 16 participants, 6 participants never recorded an utterance where they removed a differential from consideration. Participants expressed slightly more statements of decreasing likelihoods (M = 0.99, SD = 1.09) rather than increasing likelihoods (M = 0.93, SD = 1.09) but this difference did not reach statistical significance (t(15) = 0.34, MDiff = 0.06, p = 0.73). Taken together, this indicates that participants tended to keep differentials in mind, increasing their likelihood when receiving confirmatory evidence and decreasing their likelihood when receiving contradictory evidence (rather than ruling differentials out entirely).

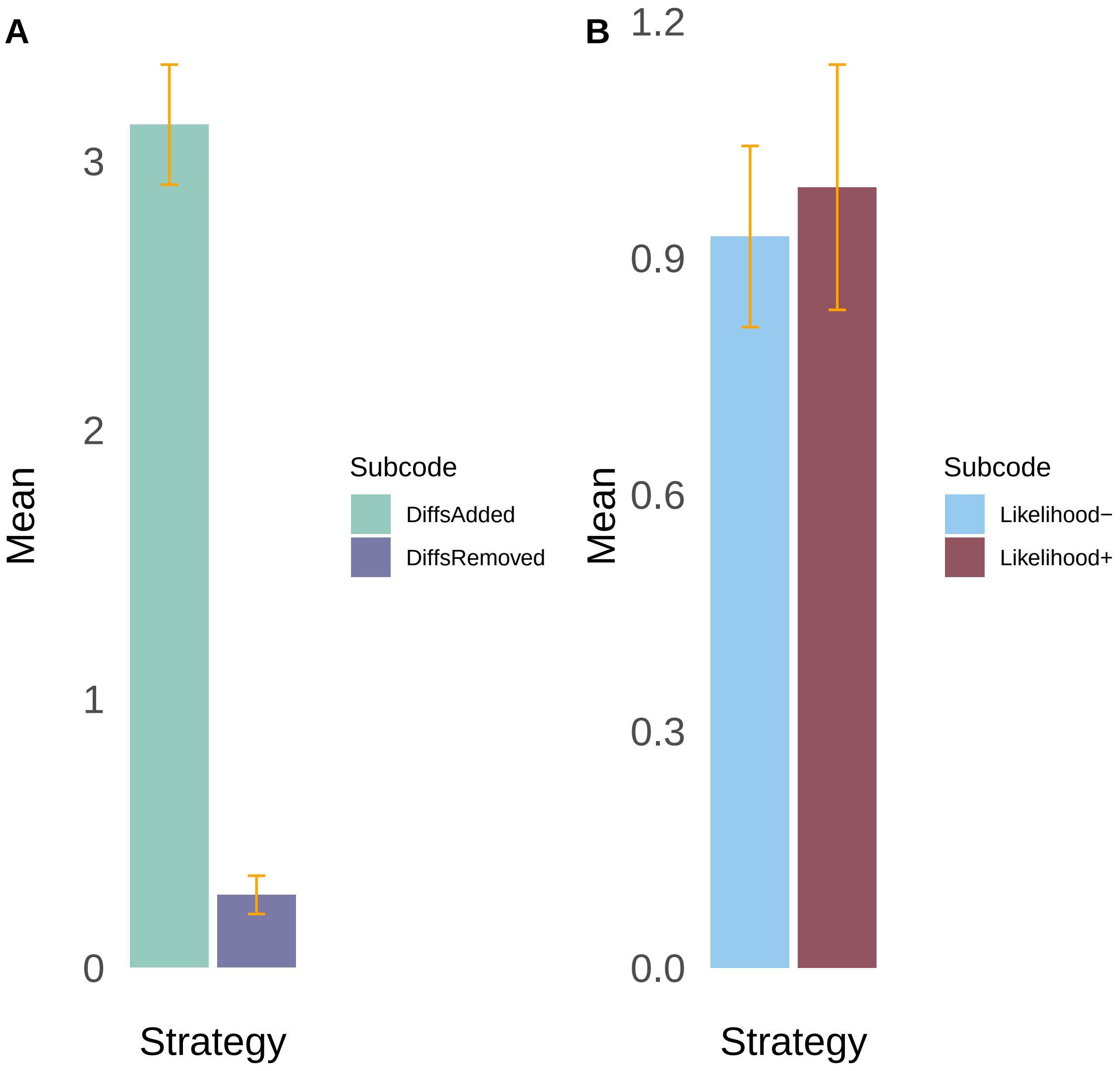


Figure 4.1: *Bar graphs comparing incidences of each of the four subcodes within Differential Evaluations. We compare instances of differentials being added (green) and removed (purple) from consideration (Figure 4.1A) and compare instances of differentials decreasing (blue) and increasing (red) in likelihood (Figure 4.1B).*

Reasoning Strategies

Incidence of Strategies

In Table 4.3, we show all 96 cases from the think-aloud strategy and the strategy coded to each after resolving all interrater conflicts. Of these cases, 6 were coded as not having a clear reasoning strategy due to both an insufficient number of think-aloud utterances and no diagnostic differentials being mentioned. 50 cases were coded as having a HD strategy, 35 cases were assigned a PR strategy and 21 cases were coded as SI. In Figure 4.2 below, we plot the proportion of cases for each patient condition that were categorised under each of the reasoning strategies. We note that the types of reasoning strategy used varies by condition (see Figure 4.2 below), with the MTB and TTP cases in particular exhibiting higher usage of PR than others, whilst HD was used by the majority of participants for the UC and AD cases in particular. Whilst data is sorted by case difficulty, this does not reveal any systematic pattern of reasoning strategy usage according to difficulty. Table 4.4 gives illustrative examples of quotes that resulted in coding of reasoning strategies.

ID	AD	GBS	MTB	TA	TTP	UC	Self Reported Strategy
p1	HD	SI	PR	SI	HD	HD	HD
p2	HD	HD	HD	SI	PR	HD	PR
p3	PR	PR	PR	HD	HD	HD	PR
p4	HD	HD	HD	SI	HD	HD	HD
p5	HD	HD	HD	HD	HD	HD	SI
p6	HD	HD	HD	PR	PR	HD	PR
p7	SI	SI	PR	SI	SI	HD	HD
p8	SI	HD	PR	PR	NONE	PR	PR
p9	SI	HD	SI	PR	PR	PR	PR
p10	HD	SI	PR	PR	PR	PR	HD
p11	HD	SI	NONE	HD	PR	HD	HD
p12	HD	HD	PR	SI	PR	HD	HD
p13	HD	SI	PR	PR	PR	PR	SI
p14	NONE	NONE	NONE	PR	PR	NONE	HD
p15	HD	HD	HD	HD	HD	HD	SI
p16	PR	SI	PR	SI	SI	HD	PR

Table 4.3: Strategy coded for each case by participant (rows) and by patient condition (column) after resolving conflicts between both independent coders. The

rightmost column shows the subjectively reported strategy that participants felt they tended to use. Anonymised participant IDs are used.

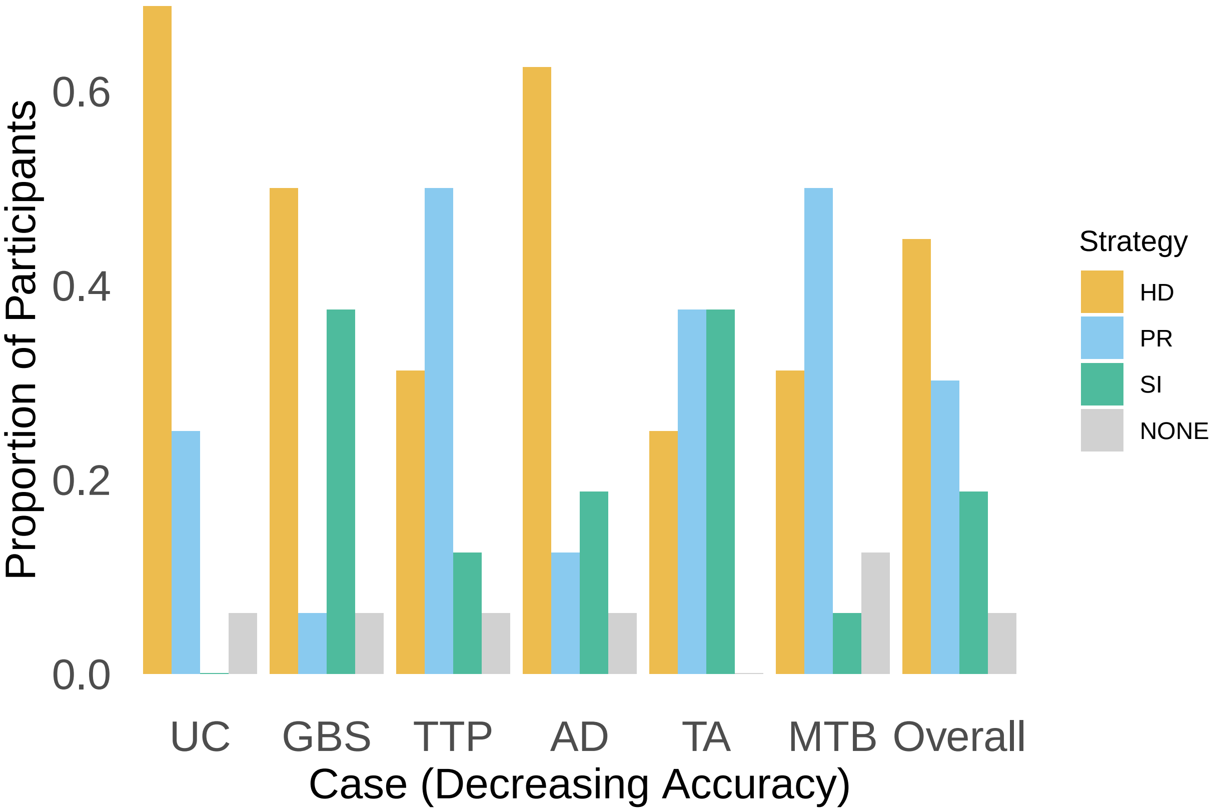


Figure 4.2: Proportion of participants who use each type of reasoning strategy for each case, with the overall proportions across all cases shown by the rightmost ('Overall') bars. Cases are sorted, from left to right, in descending order of Accuracy across participants for each case. The strategies shown are: Hypothetico-Deductive (where multiple differentials are considered simultaneously, orange), Pattern Recognition (where a single differential is considered in turn, blue), Scheme-Inductive (where participants evaluate pathophysiological systems as causes of patients' conditions rather than specific conditions, green) and None (for cases where a clear differential is not mentioned or if there are not enough utterances to infer a clear strategy, grey).

Participant (Case)	Quote	Coded Strategy	Interpretation
p1 (AD)	”A long time ago she had a total hysterectomy, doesn’t really tell me much. Although... why did she have that? Was that for cancer? Has that come back? She must have been quite young when that happened. So yeah, it would be worth investigating further...so I would need more investigations before knowing how to treat her in this case. So I’m not very confident at all. If this is an infection, if this is cardiac related, is it a PE? Yeah, it could be many things or is this malignancy? Yeah. So hopefully the exam will tell us more.”	HD	Considering multiple diagnostic possibilities before starting physical examinations, naming specific conditions (e.g. PE) rather than simply focusing on pathophysiological systems
p5 (UC)	”So I think diverticulitis is probably at the top of my differential, IBD would be up there. Infective gastroenteritis. I would also just be worried generally about his hydration status because it sounds like he’s losing a lot of water.”	HD	Explicit mention of a ’list’ of differentials, with one at the top of consideration
p3 (GBS)	”Reduced tone, reduced power, reduced coordination, reduced reflexes. Upper limbs normal. So it’s looking like lower motor neuron given this reduced tone. And so could this lower motor neurone be something like Guillain Barre? What would cause that, would be sudden onset. Okay, but yeah, given that it’s not upper motor neuron stroke.”	PR	Use ’looks like’ to denote pattern matching, points to a specific condition, mentions ruling out a previous differential before moving on to this current one.

(continued)

Participant (Case)	Quote	Coded Strategy	Interpretation
p13 (MTB)	”Crackles on the right and left lower zones. So bilateral crackles could be congestive heart failure and then sounds like there’s maybe a pneumonia there with the crackles in the right middle zone.”	PR	Uses ’sounds like’ to denote pattern matching, provides two concurrent conditions as a single diagnosis rather than separate possibilities
p8 (TTP)	”So the slurred speech, headache and weakness in the arms sounds neurological in origin. And his high blood pressure is concerning for such a young man...I don’t have a definitive diagnosis. Now could be something infective or it could be something cardiac in origin.”	SI	Mentions multiple pathophysiological systems that could be implicated, explicitly states that they do not have a definitive diagnosis
p16 (TA)	”Wonder if this is an underlying rheumatology thing and this fever is something separate so let’s just auscultate his lungs. Lungs are clear, get the heart. Okay neurological exam, normal, eyes, want to do abdo as well. Ok, fine so I think it’s more of an infective picture but I don’t know where the source of infection is...So testing, do CRP and ESR. CRP is raised, ESR is raised. So again could be rheumatological or infective.”	SI	Weighing up two different groups of conditions that could be causing the patient’s symptoms, but not mentioning specific conditions

Table 4.4: Examples of cases that were coded for a particular reasoning strategy and key quotes that suggest the coded reasoning strategy. We provide the quotes from each case, as well as how we interpret the quotes in line with the particular reasoning strategy.

Characterising Reasoning Strategies

Strategy	N	Accuracy	Differential Evaluations	Differentials Added	Information Seeking	Information Value	Confidence Change
HD	43	0.74	7.33	3.98	0.77	2.77	0.19
PR	29	0.48	4.28	2.48	0.79	3.38	0.20
SI	18	0.50	4.11	2.67	0.80	2.55	0.25
NONE	6	0.00	2.17	1.67	0.71	2.99	0.18

Table 4.5: Mean values for dependent variables broken down by the reasoning strategy coded after resolving conflicts between the two independent coders. From left to right, Accuracy refers to the proportion of cases where a correct differential was mentioned. Differential Evaluations refers to the number of coded utterances under one of the subcodes (Differential Added, Differential Removed, Increased Likelihood, Decreased Likelihood). Information Seeking refers to the proportion of available information sought across cases. Information Value refers to how useful the information sought by participants was by calculating the difference in accuracy between participants who did and those who did not seek each piece of information for a given case (see Data Analysis section for the online experiment). Confidence Change refers to difference between initial confidence and final confidence.

Next we look at our coding of reasoning strategies at a case level (see Table 5 above). Accuracy was numerically higher for cases coded as Hypothetico-Deductive (0.74) compared to both Pattern Recognition cases (0.48) and Scheme Inductive (0.5). When running a binomial mixed effects model with Accuracy as the outcome variable and reasoning strategy as a predictor, the effect of reasoning strategy on accuracy was not statistically significant (F = 2.28, p = 0.1). On cases with a SI strategy, participants gained more confidence over the case (0.25) when compared to PR (0.2) and HD cases (0.19) but the effect of reasoning strategy on confidence was not statistically significant (F = 0.03, p = 0.97). Participants evaluated differentials more often during HD cases (7.33) when compared to other strategies, and we find evidence of an effect of reasoning strategy on the number of differential evaluations (F = 5.84, p = 0.003). Participants also added more differentials for consideration when using an HD strategy (3.98) when compared to other strategies, and we find evidence of an effect of reasoning strategy on the number of differentials added (F = 4.05, p = 0.02). We do not find evidence for a significant effect of reasoning strategy on information seeking or informational value (ps > .1).

Dominant Reasoning Strategies

Given that each case attracted different strategies, but some were more common than others for particular cases (e.g. HD being used most for the AD case, PR being used most for the TTP case, as shown above in Table 4.3), we asked whether participants using the 'dominant' strategy for each case were more accurate than participants using a less common reasoning strategy. We first categorise each of the 6 cases as having a 'dominant' reasoning strategy based on which was utilised the most across participants. Through this process, we categorise three conditions as HD (AD, GBS, UC), and three conditions as PR (MTB, TTP, TA). We note that there was an equal number of PR and SI cases for TA condition, but we use PR as its dominant strategy to easily compare HD and PR directly. HD was assigned to 62% of AD cases, 50% of GBS cases, and 69% of UC cases. PR was assigned to 50% of MTB cases, 50% of TTP cases, and 38% of TA cases. Accuracy was numerically higher for cases when participants matched the condition's dominant strategy (0.67) compared to when they did not (0.54). However, this difference was not found to be significant via a mixed effects logistic regression (on accuracy as a binary outcome measure with both condition and participant as random effects) (β = 0.6, SE = 0.5 t = 1.19, p = 0.23). Overall, participants matched the dominant strategy on 51 cases (56.7% of cases, excluding those cases without a clear reasoning strategy). Participants then use a range of strategies, with variability observed here for a given case.

Dominant Strategy	Matching Dominant Strategy	N	Accuracy	Differential Evaluations	Information Seeking	Confidence Change
HD	No	16	0.62	5.12	0.78	0.21
HD	Yes	29	0.83	7.83	0.75	0.20
PR	No	23	0.48	5.17	0.81	0.20
PR	Yes	22	0.45	3.86	0.79	0.21

Table 4.6: Table showing average accuracy values by cases where the participants used or did not use the dominant reasoning strategy for that case. Dominant strategies are decided based on which of the reasoning strategies was utilised by the majority of participants in the think-aloud study. Cases without a coded reasoning strategy are excluded from this table. The first column refers to the dominant strategy for that condition, whilst the second column refers to whether the cases' coded strategy matches the condition's dominant strategy.

Subjective Reasoning Strategies

Not only do we find reasoning strategy usage varying on a given case, we also show in this section evidence that there is variability for a given participant. In addition to reasoning strategies being coded based on the participants' think-aloud utterances, we also asked participants about their diagnostic process during the debrief interviews that can be used to infer the reasoning strategies participants think they use in their regular medical practice. We use this to determine if participants were more accurate when using their subjectively preferred reasoning strategy. In Table 4.7 below, we categorise participants based on their subjective reflection of their diagnostic process. In Table 4.8 below, we provide quotes from the debrief interview that exemplify each participant's subjectively preferred reasoning strategy in order to better show how these strategies were determined. Through this process, we categorise 7 participants under a HD reasoning strategy, 6 participants as PR and 3 participants as SI. Given these categorisations of reasoning strategy based on subjective reflection by participants, we compare these participant-level strategies to the case-level strategies assigned by our independent coders. Accuracy was numerically higher for cases when participants matched their subjective strategy (0.68) compared to when they did not (0.58). However, this difference was not found to be significant via a mixed effects logistic regression (with accuracy as a binary outcome measure and both condition and participant as random effects) (β = 0.84, SE = 0.68 z = 1.25, p = 0.21). We can also ask whether participants tend to use the strategy that they subjectively prefer, or that they tend to use according to own perception. If coded reasoning strategies tends to match the subjective preferred strategy of participants, we would surmise that participants have some insight into their own decision making process. We find that there are 31 cases (34%, excluding cases without a coded reasoning strategy) where participants match the reasoning strategy during the case to their subjectively defined strategy that they tend to use for diagnostic decisions. As this is around chance of 33% (given that there are three possible reasoning strategies), we overall find that participants do not tend to match their subjectively preferred strategy. When computing a Binomial Test, we find evidence that participants do not match their subjective strategy significantly more than chance (number of cases with strategies matching subjective preference = 31, probability = 0.34, 95% CI = [0 , 0.44], p = 0.002). To summarise, participants either do not seem to have insight into what reasoning strategies they tend to use or the nature of the patient cases presented caused participants to deviate from their usual process for making diagnoses. Either way, we find there is not a close match between diagnostic strategies reported by participants and how they seem to actually work through diagnoses as derived by think-aloud utterances during the diagnoses.

Matched to Subjective Strategy	N	Accuracy	Differential Evaluations	Information Seeking	Confidence Change
No	59	0.58	5.56	0.80	0.21
Yes	31	0.68	5.97	0.74	0.19

Table 4.7: Dependent variables by cases where the reasoning strategy used (as categorised by the independent coders) matches the subjective strategy coded for that participant (as per responses to the debrief interview, see Table 4.8 below).

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p1	"I try to go based on kind of...the stock, that kind of differentials I have for the different presentations and then work down, kind of trying to narrow it down with further information and kind of rule out certain things in my head."	HD	I generate differentials based on the patient presentation and then narrow those down	Process of elimination

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p4	"So going through like a system of starting with the history and sort of gathering as much information as I can there and thinking already what I think might be happening, and then examining them and seeing if that sort of changed anything, but then sort of getting investigations. And yeah, basically, I like to sort of, yeah, piece everything together, sort of, as I get information, and then sort of try to reach a diagnosis that I think might be going on sort of at the at the end of that...I think probably, especially as a medical students, we get taught to rule out like red flag stuff. 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."	HD	Seek as much information as possible to find what differentials are likely but also what differentials need to be ruled out. Based on what is ruled out, I can decide what is most likely.	Focus on ruling out differentials to find what is likely and what is potentially serious

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p7	"So after taking a history, as long as the patient's not like acutely unwell, taking history and accurate family history, and their past medical history, creating a sort of list of most likely differentials in my head, and then seeing what would be useful in ruling some of those in and out."	HD	Generate a list of differentials first based on patient history and then decide what information is needed to either support or rule them out	Work from an initial set of differentials to guide information seeking
p10	"I guess, trying to go from their main symptom and then trying to bring up general, maybe not specific conditions, but kind of what can go wrong that would lead to this picture. And then trying to rule in or out things that fit with that...Like, if someone's coming in with a presentation that could be quite urgent and serious, then obviously you want to rule out like a stroke, you want to rule that out quite quickly."	HD	From the symptom, generate a set of differentials that could create this picture in the patient, then rule in and out from this set	Initial set of differentials to rule in or out
p11	"I guess you've always got to think about two things. What's common? And what do we have to rule out?... even just to rule out a few more things that perhaps then you don't even have to test for and consider later, if you've ruled them out in history."	HD	Generate differentials based on what's common and what's severe if missed, rule some differentials out during history taking so you don't have to test for them later	From an initial set of differentials, work on ruling some out as early as possible

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p12	“Take a history, like detailed history, formulate my top differentials. And then basically, look at investigations and examinations to confirm or rule out these differentials or any that are sufficiently different? Consider different differentials.”	HD	Generate a set of differentials based on a detailed history, seek information to rule them in or out	Work from an initial set of differentials to guide information seeking
p14	“Just sort of work through from the most like, basic things like history and examination within like least invasive tests, and then try to work up from there. trying to rule out the most serious things...Like some things, you just have to get the differentials broad because there’s so many things that could cause it.”	HD	Rule out serious differentials, important to remain as broad as possible	Focus on remaining open minded and broad to rule differentials out
p2	“Yeah, I think I would like to say that I would think about different systems and stuff, but I think I’m a bit more frantic...I’m not sure if I have a specific approach, I think I just wait for something into pop into my head, really, which is quite bad.”	PR	I would like to be more structured but I tend to wait for a diagnosis to come to mind based on the information I have	Matching information with most likely diagnosis
p3	“But yeah, I definitely, like, often have an idea from the start. I think I probably do think about that the whole way through, which probably can be beneficial, but can also sometimes hold me back from looking at other options.”	PR	Early differential used to guide rest of the diagnostic process, not much consideration of alternative diagnoses	Seeking information in favour of a single diagnosis

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p6	"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 to rule everything else out first...But also for the last one I especially, I have a very strong feeling it was GCA. Therefore I wasn't as keen to broaden my differentials."	PR	Ruling out differentials is important, but there tends to be a primary diagnosis that comes to mind	Focal diagnosis tends to come to mind, though some ruling out of differentials is also necessary
p8	"Trying to go from their main symptom and then trying to bring up general, maybe not specific conditions, but kind of what can go wrong that would lead to this picture...I think my brain can sometimes get stuck on an idea. And it's difficult to pull away from that."	PR	Think about what comes to mind based on the initial presentation, can be influenced by an early idea of what the patient has, less focused on specific conditions to rule out	Generate a general diagnosis based on what comes to mind from early patient presentation
p9	"Either spot diagnosis if it's a really typical case, like otherwise, basically, if I'm not sure I'll work through a surgical sieve."	PR	Come up with an idea for a diagnosis on the spot or go through a structured process if not sure	Default is to match the current patient to what comes to mind as a prototypical case

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p16	"So probably getting history, I look at initial observations first...and then I look at ECG as well, then I want to get initial bloods being guided by what I think could be going on and then think about potential images. If it's quite obvious early on what's going on. And for example, it's some kind of infection, I'd want to start them on antibiotics early, but otherwise, I might get advice or wait until I get all my information, before starting treatments."	PR	Based on intial history and tests, see if it is obvious what is happening with the patient and if not, seek additional advice and information	Seek information until a clear diagnosis comes to mind matching the information
p5	"But I think most of the time I try and keep quite a broad set of ideas. And then narrow it down and try and think, like, system wise...like the one that was aortic dissection, that I was sort of like, this sounds like they're coming with abdo pain, but it could be something cardiac, it could be something like...trying to keep ruling out systems."	SI	Keeping broad set of ideas based on the possible systems that could be implicated in the presenting complaint and then ruling systems out	Structuring diagnosis by what patho-physiological systems can be ruled out

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
p13	"I try and think about the, like the different systems that might be involved before thinking about...so for example, if someone presents with chest pain, there's lots of things in the chest that could lead to chest pain. So try and think about the systems more generally and then focus on the specific symptoms that align with this specific system if that makes sense."	SI	Based on the presenting symptoms, I think about what pathophysiological systems might be causing them and then go from there	Primary focus on determining the system involved before thinking about specific conditions
p15	"But I think generally the big things are is there a system under which it generally falls? Primarily, like, does it sound haemotological, cardiac, whatever."	SI	Trying to match symptoms to a particular pathophysiological system	Structures diagnosis by pathophysiological systems

Table 4.8: Categorisation of participants under one of three possible reasoning strategies based on their responses during the debrief interview. We capture here the subjective reasoning strategy for each participant based on how they reflect on how they tend to make diagnostic decisions. In the second column are key highlighted quotes related to each of the participants' diagnostic decision making processes. In the fourth column, we provide our summary of the quote and then in the fifth column, our interpretation of the quote that explains the choice of reasoning strategy for that participant. Participants have been sorted/ordered by their coded strategy.

To combine the two previous sections, we consider whether participants are more accurate when in fact their reasoning strategy matches both the case's dominant strategy and the participant's subjectively preferred strategy. For briefness, we refer to this as the "fully matched strategy" going forward. Participants used a fully matched strategy on 21 cases (23.33%), compared to 40 cases when strategy matched either the case's dominant strategy or the participant's subjectively preferred strategy. Participants were more accurate with a fully matched strategy (0.71) than for cases when strategy matched either the case's dominant strategy or the individual's subjectively preferred strategy 0.625 and cases when neither matched (0.52). We do not find evidence of an effect of matching strategies via a mixed effects logistic regression (with accuracy as a binary outcome measure and both condition and participant as random effects) (β = 0.39, SE = NA t = NA, p = NA).

Thematic Analysis from Debrief Interviews

Building on the previous section and going beyond identifying reasoning strategies, we can utilise the rich data from debrief interviews to understand how doctors think about their diagnostic processes. In this section, we present key themes from the thematic analysis of participant responses to the debrief interviews The interview questions were designed to ask participants how they think they tend to make diagnostic decisions and what their main considerations are during the decisional process. We provide quotes from participants exemplifying each of these themes. Participants are referred to by their anonymised identifiers.

Heuristics when Making Diagnostic Decisions

Participants reflected a few general tendencies (or rules of thumb) when making diagnoses regardless of the patient. Firstly, seven participants mentioned that they prioritised any serious/emergency differentials early on when making diagnoses, which would affect the urgency with which they would approach ruling these differentials out. This suggests that some focus would be taken away from determining likely diagnoses and instead, ruling out more serious diagnoses that would require more immediate medical attention. The need to consider serious diagnoses offers a potential reason for why participants do not simply utilise their preferred reasoning strategy on all cases, as some patients may display symptoms that prompt consideration of diagnoses that are harmful if missed.

"I do have an approach, the first (thing) I always want to think is if I miss something, is this patient gonna be bad? So, like, thinking about emergency stuff." (p2)

"I think probably, especially as a medical student, we get taught to rule out red flag stuff...a lot of my thinking is like, what really worrying thing could this be that we need to rule out? And what tests do I need to rule it out?" (p4)

"If it's...an acute versus a non-acute thing, I think that would change the pace I approach it." (p7)

"Like, if someone's coming in with a presentation that could be quite urgent and serious, then obviously you want to rule out like a stroke, you want to rule that out quite quickly." (p8)

"If I think something's remotely possible, that's really like say, so like for GBS, I'm even thinking about Cauda Equina syndrome, like, regardless of how high my index of suspicion for it is, even if it's pretty low. If it's an urgent diagnosis, I'll just do it anyway." (p9)

"I'm trying to rule out the most serious things." (p14)

"And essentially if there's any serious conditions, I make sure to rule those out...and then go from there. I probably should go through each one and weigh each one individually. Because that would avoid being as biased. But it's not something I do as much as I should...The other big things, are there any of the red flag symptoms that are really important that should influence what I'm thinking? Like fevers especially, that sort of thing...So generally, acute situations are where I narrow a little bit." (p15)

Another tendency was for participants to report a form of progressive investigation that stems from the patient's history. In this sense, participants report a decision process that quite closely matches our experimental procedure of gradually seeking information based on patient's medical history to build up a picture of them. This also corresponds with past work that found that history taking is responsible for around 80% of final diagnoses after examinations and testing (Hampton et al., 1975, Sandler, 1980, Peterson et al., 1992), which has been adopted as a general adage in medicine. This illustrates the importance of a comprehensive medical history for the patient being available and how much it guides medical students' decisional process. We can also ascertain from this theme that the initial diagnostic differentials generated from the patient's history has a large influence on the subsequent diagnostic process:

"Definitely start like history...I think to go from there and like, kind of think about that in the context of the patient. Yeah, I feel I've definitely been taught in terms like that methodical, like do it in that order." (p3)

"I guess going through like a system of starting with the history and sort of gathering as much information as I can there and thinking already what I think might be happening. And then examining them and seeing if that sort of changed anything, but then sort of getting investigations." (p4)

"You can get a lot from the history. So I think sort of, I guess my general approach is like, take your history, and then from the history, have a little, it's not like, if you wrote it down, it'd be like a little bubble, like brainstorming thing as...the key big differentials I'm considering." (p5)

"But if it was a patient...who had sort of not very clear symptoms, but wanted to be a bit more thorough, like take a history first and then looking at any test they've had, starting with like more basic tests like observations blood tests, and then, depending on the cause, or the symptoms, doing more invasive tests, perhaps." (p7)

"Take a history, like detailed history, formulate my top differentials. And then basically, look at investigations and examinations to confirm or rule out these differentials." (p12)

"I think, start, think systematically. So start with a thorough history, asking kind of about what's happening currently, and then going through the kind of past medical history and focusing on that, asking what the patient thinks might be going

on. And then focus on a thorough examination which sometimes for the interest of time is focused on the, the kind of symptom at hand, but you should do a kind of formal full checkthrough as well…see if there's anything that points you towards a diagnosis. I think it (the experiment) was set up in the way that I go about things in the way that you do the history first, you do the examination, you do the investigations." (p13)

"Just sort of work through from the most like, basic things like history and examination, least invasive tests, and then try to work up from there." (p14)

"So probably getting history, I look at initial observations first…And then I look at ECG as well, then I want to get initial bloods being guided by what I think could be going on and then think about potential images." (p16)

Within this process of progressive investigation, seven participants (including the quote above from participant *p16*) noted that there are pieces of information or tests that they would seek for all patients (regardless of their condition) as part of a routine diagnostic approach. This indicates that some aspects of the diagnostic process are seen as fairly standardised by medical students:

"I would always want to do like full blood count, VBG [Venous Blood Gas]… Probably, as I said, I think like most people in the emergency department get a chest X ray" (p2)

"With the examination, I think, normally, I would like…auscultating the heart and feeling the heart, abdomen, etc. These are things I think I would do in any patient, irrespective." (p4)

"And then for investigations…I'll take all the bloods, do an ECG, chest X Ray, just in case. Yeah, yeah. So I am a bit more like, on the side of caution." (p6)

"I think if you went to your senior and you said I'm really concerned about this patient, but I've not done an FBC [Full Blood Count], a U&E [Urea & Electrolytes], an ECG, VBG…They'd be like, what are you on about? So there's a few that

you would do anyway, that are largely non invasive, in terms of...higher degree investigations, very much depends on anatomically, what you're seeing, what your differential is." (p11)

"I wasn't sure (during the experiment) whether I should try and be very focused to the presentation at hand...because for example, when I was going through the examinations, in reality, I would do a full exam on someone, even if they presented with something that was very specific, like a very specific symptom just so that you can have a full kind of clerking assessment." (p13)

"And then if the patient's unwell, I do an A to E assessment of them. Try to take history from them if I can, or collateral history from staff or family members. And while I'm doing the A to E assessment, also, I don't have exact logic for it. But I'd run like blood tests, FBC, Us and Es, LFTs, CRP [C-Reactive Protein blood Test] and then adding other things like troponin, if I think it's like cardio related or D dimer [blood clot test], if I think it's PE [pulmonary embolism]." (p10)

Along similar lines to the above quote from participant *p10*, four other participants noted the use of a standardised framework for structuring their diagnostic process, with the two mentioned (as covered during the students' medical education) being the ABCDE assessment tool (Airway, Breathing, Circulation, Disability, Exposure, Smith & Bowden, 2017) and the surgical sieve (which guides students and clinicians through different pathphysiological systems, Chai, Evans & Hughes, 2016):

"So I feel like the A to E has very drilled into us at medical school. So I think I still rely on that. And even when all the things were jumbled up, I try and like, pick them out in that order." (p5)

"If it was an acute patient, I want to do like an A to E assessment, like airway, breathing circulation..." (p7)

"Basically, if I'm not sure I'll work through a surgical sieve." (p9)

Think Aloud

"I think the other way I can do it is the sort of surgical sieve idea to make sure you've ruled out" (p15)

When taken as a whole, students tended to view diagnosis as quite a standardised process. The extent to which it is based around intuition or a more logical, structured process is what may contribute to differences in how diagnostic decisions are made in terms of what information is sought and how clinicians/students generate diagnostic hypotheses. In particular, structured frameworks such as A to E and surgical sieve tended to be viewed as approaches to fall back on when medical students were unsure. This indicates a potential relationship between confidence and the reasoning process using by students.

Avoidance of Anchoring

To focus on a key consideration of diagnoses, as mentioned by six participants, we discovered a theme of anchor bias within the qualitative data. This was explained by one of these participants as follows:

"I'm quite aware that there's, I've tried to remember what it's called, 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." (p11)

These participants showed awareness of this phenomenon, whereby clinicians may focus too early on a particular diagnosis and then seek information to confirm this existing belief (a form of confirmation bias). This can then prevent participants from considering alternative diagnoses early on in their decisional process. Given their awareness of this bias and its pitfalls, we can then infer that participants approach their diagnoses in such a way as to avoid this bias. Other participants cited this as a consideration of theirs when making diagnoses:

"I'm quite rubbish, I often get fixated like 'I think this is this' ' '...but I'm not that good at thinking, 'oh, what else could it be' into like, 'I've got something that's proved to me it's not.' " (p2)

"I try to, but I think my brain can sometimes get stuck on an idea. And it's difficult to pull away from that." (p8)

"I think I probably do think about that the whole way through, which probably can be beneficial, but can also sometimes hold me back from looking at other options" (p3)

One reason provided by students for such a bias is that they are relatively early into their medical experience. As a result, they have not developed as much medical knowledge as experienced clinicians and may focus on diagnoses that they have more familiarity with. Medical students also reported making a conscious effort to keep an open mind with regards to alternative differentials given their inexperience:

"But then I do think I, at this stage, I'm quite kind of biased 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." (p1)

"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." (p7)

"My knowledge isn't broad enough...to remember all the differentials for everything." (p14)

This is important to note for three reasons. Firstly, medical students take their relative inexperience into account as a factor when making diagnoses. This could then mean that as medical students become intermediate/experienced clinicians, their decision making style may change to reflect their increased medical knowledge. Secondly, medical students may be more likely to express uncertainty if there are more diagnoses/conditions that they are unfamiliar with due to their lack of knowledge. Thirdly, the awareness shown for the relative inexperience of medical students indicates that students would be less likely to tend toward overconfidence, given this sense of 'humility' about what knowledge they have and do not have. Taken together, medical students are likely to approach medical decisions very differently from experienced clinicians mainly because they have different perceptions of their own medical knowledge.

Challenges Related to the Vignette Paradigm

Participants cited a number of challenges related to our diagnostic task. These were related to ways in which our study did not emulate real-world aspects of the diagnostic process, but also reveal insights into how diagnostic decisions tend to be made in healthcare settings. Firstly, three participants noted that it was difficult to retain all of the information they needed during the task:

"I kept thinking that I couldn't like hold onto all of the information." (p2)

"I'm quite a visual person. So…reading on a screen is quite different to I don't know, actually having seen a patient, seeing the exam findings, or even looking at the scans myself. I feel like I find that easier to retain the information. Whereas when it's, I find it, it's kind of hard to take it in when it's like just written down" (p3)

"I think just thinking on the spot and coming up with the diagnosis quite quickly, it's quite hard remembering the management afterwards as well." (p14)

Four participants noted that, in their real medical practice, they would be consulting other doctors, frameworks or online resources, which made our task difficult given that these were not available to participants:

Think Aloud

"I'd find as much information as possible and to ask for help...I think I'm someone who looks up stuff a lot. Like I rely a lot on looking up things. And that gives me a lot of comfort. I feel like when I don't have those tools, yeah, I feel a bit shaky...And I often kind of just take my phone out and look at that, and even just glancing at them kind of helps me structure my thoughts. So, yeah, I would wish it would be kind of more, more of an organic process. But at the moment, I think I rely quite a lot on prompts and things like that, or guidelines, even if I don't read them thoroughly, I'd need kind of reminders, especially when I feel like there's so much that it could be and I lose myself a little bit in the possibilities." (p1)

"I guess in real life, I also have, like, Google. So at points where I forgot the disease, or like, what is the first line, I would have like checked before, before typing up my management plan...when I'm confused, I'm definitely going to approach the senior. So I wouldn't be the one making the diagnostic decision. So I think it's a bit harder to do it alone." (p6)

"When I'm not sure, I'll definitely be running my train of thought past my consultants." (p12)

"I think it's weird doing it in isolation. Because I guess in an actual clinical setting, you kind of bounce ideas off someone else." (p16)

Reasoning Strategies in the Online Vignette Study

Our insights from detecting reasoning strategies in this think-aloud study can be linked to the statistical power gained from the larger sample size in the online experiment to then pick up on such strategies in the online dataset. Our first approach was based on training a multinomial logistic classifier to identify reasoning strategy in this think-aloud study, with each of the information requests as binary predictors (i.e. whether they were sought or not on each case). We then compare the predicted strategies from the classifier to the objective 'ground truth' strategies in order to assess the model's accuracy. Before we can interpret the reasoning strategies predicted for the online experiment dataset, we first need to determine that the classifier reliably reproduces the reasoning strategies coded in the Study

3 dataset based on think-aloud utterances. In order to train the classifier, we account for imbalance in the training data (due to the larger number of HD cases and the lower number of SI cases) using downsampling and limited regularisation. The accuracy of our classifier is 0.34, which is significantly lower than the No Information Rate of 0.48 (i.e. the accuracy of the classifier if HD was predicted on all cases). We compare our classifier accuracy to a bootstrapped null distribution where the ground truth labels are repeatedly shuffled (i.e. such that the labels of reasoning strategies are meaningless). Based on 10 shuffles, we find an average accuracy of 0.36. We also compare Balanced Accuracy values (i.e. the accuracy within each reasoning strategy, in order to account for the data imbalance) to those of the null classifier. We find that the Balanced Accuracy of our trained classifier (HD = 0.46, PR = 0.61, SI = 0.49) does not exceed that of the null classifier (HD = 0.49, PR = 0.48, SI = 0.5). To summarise, we are not able to predict reasoning strategy reliably based on information seeking alone using our think-aloud dataset. This could be due to a combination of data imbalance and too low a sample size. Hence, we do not perform further analysis based on the predicted reasoning strategies in the online study's dataset.

As per our previous analysis, we are not able to reliably predict reasoning strategy on a case-by-case basis. We instead look at behaviour as a function of the dominant reasoning strategy for the patient conditions based on the coded reasoning strategies in the think-aloud study. Our aim to is to look at whether cases result in different diagnostic behaviour assuming that each patient condition is associated with a particular reasoning strategy. We find that three of the six conditions (UC, GBS, AD) were approached using a HD strategy by a majority of participants. The remaining cases tended to be performed using a PR strategy (except for the TA case, where there were equal numbers of PR and SI coded participants). We consider these cases (TTP, TA, MTB) as PR-dominant cases for ease of comparison with HD-dominant cases. We calculate the mean values across conditions within each group of cases (HD-dominant or PR-dominant) in the online study dataset. This allows us to investigate broadly if reasoning strategy on a case-wise level affects diagnostic behaviour.

First, we look at whether our key variables vary as a function of the dominant strategy for cases in order to determine if strategy is associated with a difference in behaviour. As we have two groups of cases, we use paired Wilcoxon Signed Rank to compare median values between these groups (averaged across conditions), due to Accuracy not being normally distributed and violating the assumption of normality (Shapiro-Wilk Test $p = 0.008$) needed to perform a t-test. We observe higher accuracy for HD-dominant cases (Mdn = 0.47) when compared to PR-dominant cases (MDn = 0.4) (V = 2512.5, pseudomedian difference = 0.12, 95% CI = [0.05, 0.17], $p < .001$). This psuedomedian difference is greater than the sample median difference of 0.07, indicating a skewness in the distribution of the Accuracy value. We also observe more initial diagnoses being considered during HD-dominant cases (MDn = 3.3) when compared to PR-dominant cases (MDn = 3) (V = 2171.5, pseudomedian difference = 0.5, 95% CI = [0.33, 0.83], $p < .001$). When looking at differences in information seeking, we observed that more information was sought on PR-dominant cases (MDn = 0.63) when compared to HD-dominant cases (MDn = 0.6) (V = 719, pseudomedian difference = 0.04, 95% CI = [0.06, 0.02], $p < .001$). We also observed that informational value was higher for PR-dominant cases (MDn = 2.9) when compared to HD-dominant cases (MDn = 1.53) (V = 4, pseudomedian difference = 1.23, 95% CI = [1.31, 1.14], $p < .001$). We did not observe a significant difference between groups of cases in terms of changes in confidence.

Finally, we look at predictors of diagnostic accuracy, hypothesising that the dominant reasoning strategy for a case interacts with the number of initial diagnoses when predicting accuracy. Given that HD has been previously associated with more diagnoses being considered early on, we expect that diagnostic accuracy is increased with the initial diagnostic breadth for cases with a HD strategy. In order to investigate this hypothesis, we fit a linear mixed effects model that predicts diagnostic accuracy with an interaction between the number of initial diagnoses and case-dominant strategy, with participant as a random effect. As in the previous analysis, variables are averaged per participant across all conditions within each case group (i.e. for HD-dominant cases, accuracy and initial diagnoses are averaged across the UC, GBS and AD cases for each participant). We find main effects for the number of initial diagnoses (F(1,125.95) = 10.19, p = .002) and case-dominant reasoning strategy (F(1,89.48) = 11.74, p = .001). We also find evidence for an interaction between the number of initial diagnoses and case-dominant reasoning strategy (F(1,109.24) = 10.69, p = .001). As depicted in Figure 4.3 below, accuracy does not seem to depend on the number of initial differentials for HD cases, whereas there is a strong positive relationship between the two for PR cases.

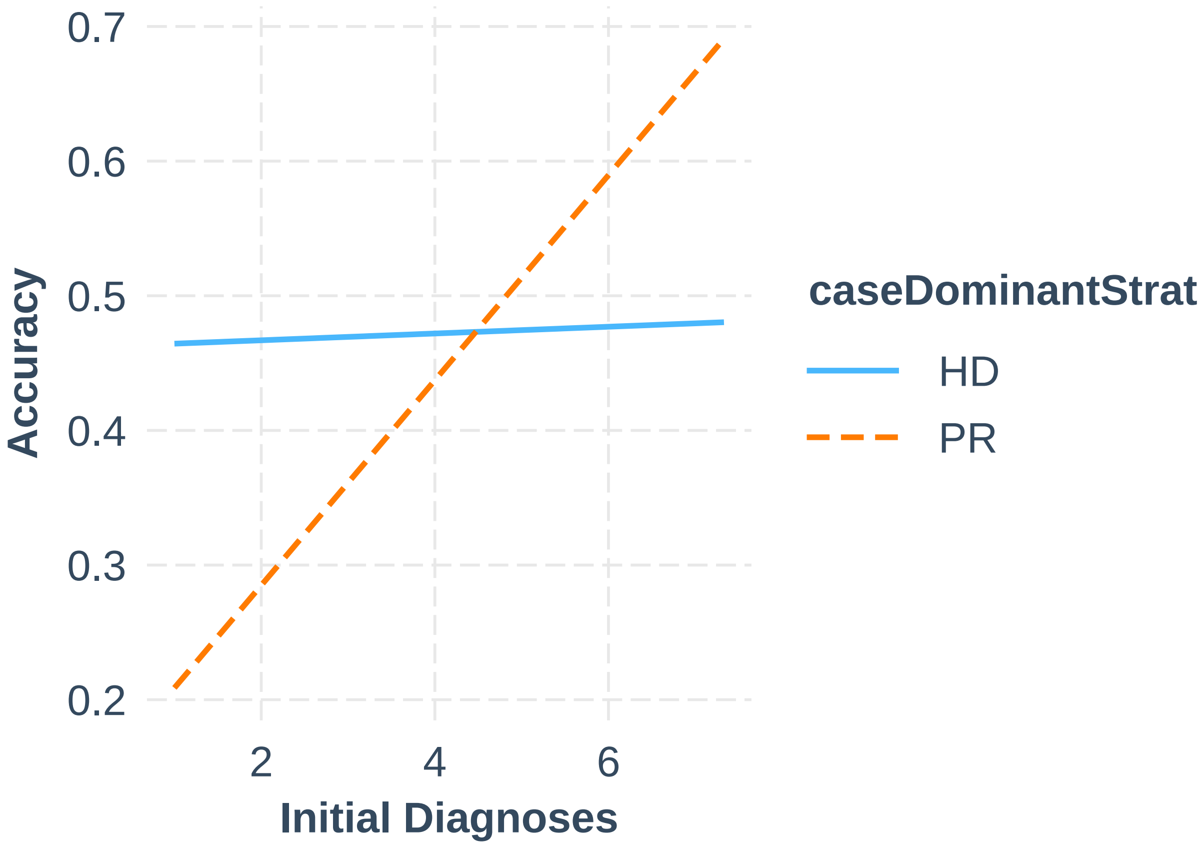


Figure 4.3: Fitted regression line for a linear mixed effects model that predicts accuracy (y-axis) with an interaction between the number of initial diag-

noses/differentials (x-axis) and dominant reasoning strategy (HD in blue averaged across the UC, GBS and AD cases, PR in orange averaged across the MTB, TA and TTP cases). Data shown here is from the online experiment dataset of Study 2.

Discussion

This study of 16 medical students explored the usage of a think-aloud methodology to understand thought processes during medical diagnoses. Using our online interface and recorded verbalisations by students, we aimed to detect clinical reasoning strategies based on criteria adapted from (**coderre_diagnostic_2003**). The strength of this paradigm is in qualitatively recording medical students' thought processes as they evolve with information as per our flexible, evolving vignette-based interface design. By recording how participants consider different diagnoses in real time, we are able to understand the reasoning approach students are applying for each case and how this affects their information seeking and confidence behaviour. We are also able to investigate if these reasoning strategies affect diagnostic accuracy, both in the context of this current study and in the previous online study. In this section, we summarise the main findings from this mixed-methods study.

Broadening of Differentials

Similar to the online vignette study, we find that medical students are reticent to remove differentials from consideration. In this study, participants report low occurrences of disregarding/removing differentials from consideration The reticence to remove differentials seems to be a clear tendency for medical students, rather than a quirk of our interface from the previous study (where we explicitly prompted participants to record their differentials). By rarely removing differentials from consideration, students are then observed to broaden their differentials with more information, as observed more directly in the online vignette study.

One way that this broadening of differentials can be explained is in terms of 'decision inertia', whereby individuals tend to report their past choices regardless of the evidence presented against those choices (**akaishi_autonomous_2014**). However, this account seems unlikely given that we found medical students to be sensitive to information that was either confirmatory or disconfirmatory of their diagnostic hypotheses. This is because we found that medical students were equally as likely to increase or decrease the likelihoods of differentials via coding of such utterances. We can surmise from this finding that medical students do not solely think about positive/confirmatory evidence for their beliefs (cf. **kaanders_humans_2022**). The 'positive test account' put forward by (**klayman_confirmation_1987**) claimed that individuals formulate beliefs and then seek information to support these beliefs, rather than against them (as also explained by **hunt_approach-induced_2016**). In our task, we find that medical students integrated information to both support and oppose their differentials despite the sequential nature of the task's information seeking (cf. **jonas_confirmation_2001**). It is however possible that with more freedom in information seeking, they may be more strategic in choosing information that is more likely to yield supporting information. This has been investigated in some previous studies as 'diagnostic momentum', in which individuals seek positive test results to resolve uncertainty/ambiguity caused by inconclusive tests (**aron_diagnostic_2024**). Future work then requires a more fine-grained understanding of information seeking within medicine and whether certain tests are chosen based on their likelihood of producing a positive test result. While the receipt of information seems to increase diagnostic confidence (as found in our online study), the choice of information itself may be an important factor to consider and a focus for future interventions/study. However, this informational account would not explain our findings in this study that students tended to broaden their differentials.

This tendency to broaden diagnostic thinking could instead correspond with medical students remaining open minded in their diagnoses in order to conscious avoid biased diagnoses, which was also reflected in our qualitative data. A number of participants mentioned that they aware of an 'anchoring bias' and that they should avoid it. The conscious attempts to avoid narrowing on a diagnosis too early could explain our findings of differentials rarely being removed consideration. There has been past work studying how to reduce such instances of 'premature closure' (**voytovich_premature_1985**; **eva_difficulty_2006**; **krupat_avoiding_2017**), whereby clinicians focus in on a diagnosis and disregard other possibilities too early in the decision process. The foremost of these papers in particular found that experienced clinicians were more susceptible to settling on a diagnosis too early than those who were less experienced. It has also been found that medical residents with greater knowledge of discriminating features between diseases were less susceptible to anchoring bias (**mamede_role_2024**). It is then reasonable for medical students to be more susceptible to such a bias due to their relative inexperience compared to other clinicians (as reflected by participants during debrief interviews). As this work was conducted using controlled experimental paradigms, it does not include environmental factors that may exacerbate tendencies toward anchoring/premature closure (e.g. work stress, busyness in managing multiple patients at once) (**gupta_associations_2023**). Prior to conducting both of these studies, we may have expected a 'process of elimination' to be used by students but this does not appear to be the case across both of these studies. Rather than frequently disregarding differentials altogether, medical students instead incorporate new information to adjust how likely their considered differentials are, whilst also being able to incorporate new differentials for consideration that the information suggests.

Reasoning Strategies and Their Relationship with Accuracy

On reasoning strategies, we were able in this study to use think-aloud utterances to detect reasoning strategies on the part of the medical students, corresponding with **coderre_diagnostic_2003** in their finding of different strategies being utilised during diagnoses. We considered three different strategies: Hypothetico-Deductive

(HD), Pattern Recognition (PR) and Scheme-Inductive (SI). These strategies represent different approaches to diagnosis, either seeking to be comprehensive in both the information sought and differentials considered or focusing in on a single diagnosis. We extend previous work to show variability in strategies that is not linked to individual medical students or to specific cases. Specifically, we found that reasoning strategies were not primarily determined by either an individual's general decision making approach (as per their subjectively preferred strategy during debrief interviews) or by specific patient conditions.

This could be because of the practicalities of patient cases meant that students were forced to make decisions in ways that they were not used to. This begs the question of what the properties are of a patient case that determine the choice of reasoning strategy on a given case. One account is that reasoning strategy is determined by how much experience/familiarity the student/clinician has with that type of patient presentation. If they had seen a similar patient before (during their education or practice), they may be more likely to use pattern recognition to identify the patient's condition (**nendaz_diagnostic_2012**). **arocha_novice_1995** found that intermediate medical students displayed a deterioration in performance (compared to advanced students) as they were able to generate plausible differentials but did not have sufficient knowledge yet to use incoming information to narrow their differentials (which could also explain the finding from the previous section on students tending to broaden their differentials on average). Hence, the use of reasoning strategies may be dependent on experience and knowledge, with clinicians/medical students using their assessment of their own knowledge to guide their reasoning strategy (even if it is not their preferred method for making diagnostic decisions). Future work can validate this account by looking at reasoning strategies as a function of clinicians' experience with similar patient conditions/symptoms.

However, this study did emulate conditions within healthcare environments (e.g. time pressures as per **gupta_associations_2023**, interruptions as per **soares_accuracy_20** that may force clinicians to make decisions in ways that do not prefer, making this account less likely. A more likely account stems from a key difference between our study's methodology and **coderre_diagnostic_2003** : in the latter study, participants were asked to retrospectively think aloud about how they arrived at a diagnosis, whilst our study asked participants to think aloud as they were performing a diagnosis. Given our finding that participants do not tend to use their preferred strategy, it seems that clinicians are often not able to accurately reflect on their own decision making process retrospectively due to inaccurate self-perceptions. This could relate closely to miscalibrations of confidence, whereby individuals have false self-beliefs (**oeberst_toward_2023**) about their own performance/efficacy (**alicke_better-than-average_2005**) and can then include false beliefs about how they make decisions too (for instance, believing that others are more biased that one's self when processing information, **pronin_bias_2002**).

Unlike **coderre_diagnostic_2003** however, we do not find the same association between reasoning strategy and accuracy. **coderre_diagnostic_2003** had found that pattern recognition was associated with higher accuracy and was utilised more by experienced clinicians. We investigated this relationship in two ways: by coding reasoning strategies in the think-aloud study based on participants' utterances and using the dominant reasoning strategy for each case in the think-aloud study to determine how this affects behaviour in the online vignette study (utilising its larger sample size). Both of these methods provide some concordant results: whilst we observed an association between PR and beneficial information seeking (i.e. informational value), HD was associated with higher initial diagnostic breadth and higher diagnostic accuracy (for HD-dominant cases). A HD reasoning strategy being associated with greater diagnostic breadth corresponds with the nature of HD being that of considering a broad set of differentials to either add to or subtract from. We find through modelling that accuracy was highest when participants had higher initial diagnostic breadth on PR-dominant cases (whilst increasing diagnostic breadth had less of an impact on accuracy for HD-dominant cases). This deviates from the assumption in past literature that HD, as a process, is the optimal one for medical diagnoses (**kuipers_causal_1984**; **higgs_clinical_2019**).

Given that we averaged across individual cases/conditions however, we advise caution in interpreting these findings, as the difference in information seeking could also be a result of these being different cases as opposed to involving reasoning strategies. Our differing results to **coderre_diagnostic_2003** hint that effective strategy usage is predicated on the experience/knowledge available to the clinician. For example, pattern recognition is more effective when a clinician has more past cases to draw from and relate to any given patient. This would explain why we find that medical students are more accurate when using a hypothetico-deductive process instead.

Though we find evidence of a link between reasoning strategy and initial diagnostic breadth, the causal direction between the two is not yet fully ascertained. It seems unlikely that the reasoning strategy is consciously applied by clinicians to guide their subsequent reasoning process and differential generation. This is because of our finding that participants have limited insight into their own reasoning process, meaning that participant do not seem to be consciously applying a reasoning strategy from the outset that then manifests in a number of differentials being generated. Our account would then be that based on the initial presentation of the patient and the early differentials that come to mind; clinicians then reason within this space of differentials using the strategy that they perceive as being most appropriate in terms of broadening or narrowing from this initial set of differentials. Future work would be needed to elucidate this account by assessing the awareness that clinicians have of their own decision process whilst they are making diagnoses, with think-aloud methodologies providing a promising avenue for this research.

Progressive Investigations and History Taking

Whilst finding that there is variance in how medical students make diagnostic decisions, there was some agreement on certain tendencies to follow that emerged by our qualitative thematic analysis. In debrief interviews, several participants reported progressively investigating patient symptoms based on the patient's history and their associated set of initial diagnostic differentials. This corresponds with our finding in the previous study that the number of initial differentials considered based on the patient history was predictive of information seeking and changes in confidence. This supports evidence for the large weighting on early information received by clinicians, especially to do with history taking, because early information is responsible for the initial set of diagnoses that then guides subsequent information seeking. It has also been found that a patient's case history that is suggestive of a particular diagnosis prompts selective processing of clinical features that favour said diagnosis (**leblanc_believing_2002**), with quality of working hypotheses influencing the relevance of information sought on a patient (**brooks_difficulty_2000**). The influence of early differentials is also found in the extant literature, with differentials generated early on being harder to disregard later on (**kourtidis_influences_2022**; **redelmeier_fallacy_2023**). Given the influence of early differentials on the diagnostic process, it follows that history taking is an important skill to teach medical students that has been researched extensively in the past (**keifenheim_teaching_2015**).

We also find a qualitative theme that participants report certain information being standard to seek regardless of the patient case. This corresponds with the finding from the online study that lower information seeking variability was associated with higher accuracy. Variability in information seeking across cases has not been previously studied during diagnostic decisions to our knowledge, but previous work has called for a degree of standardisation within medical decisions (**wears_standardisation_2015**), such as checklists (**ely_checklists_2011**). When taking findings from both our coding of reasoning strategies and thematic analysis together, these findings portray that diagnosis is a decision process where there is considered to be 'optimal' information to seek for any given patient, with a particular focus on history taking to inform the rest of the decisional process. Whilst reasoning processes varied as a function of the patient condition, an optimal strategy for diagnostic accuracy across all cases seems to be generating a larger set of initial differentials from the patient history and then selectively choosing from this set the diagnosis that closely resembles the patient's symptoms and observations. In other words, medical students performed best by starting broad and then narrowing their differentials. This is predicated on students and clinicians being able to identify plausible diagnoses early on, whilst remaining open minded to other possibilities. This speaks to a strength of our methodology, whereby we are able to observe triangulation between quantitative analysis and qualitative themes.

Implications and Limitations

From our coding of reasoning strategies, we found that medical students utilise a variety of reasoning strategies and that we were able to pick up on these strategies from medical students' verbalisations as they thought out loud during diagnoses. However, students did not show clear insight into their own decision making process, as they tended to deviate from their subjectively preferred strategies. In addition, reasoning strategies were not reliably explained by the case/patient being treated. This brings an unresolved question from this work of the factors that inform the diagnostic reasoning processes used by medical students or clinicians. Clinical reasoning can be described and taught in a variety of ways (**royce_teaching_2019**), such as teaching the strategies coded in these studies (e.g. hypothethico-deductive, pattern recognition) and other reasoning processes (e.g. Bayesian reasoning, inductive reasoning). With this range of reasoning approaches however, there is currently not a clear consensus on which approach, or approaches, are most useful for diagnoses and the situations in which a clinician should adopt a certain approach. Our study reveals this as an open question for future work: understanding the factors related to individual clinicians and to patient cases that necessitate the use of certain reasoning strategies. A real strength of this study is its use of a think-aloud protocol, which has only been used in a few past papers for diagnoses (**arocha_novice_1995**; **coderre_diagnostic_2003**). We recommend the use of this methodology or similar qualitative means (e.g. focus group discussions) in future work given its ability to pick out nuances of diagnostic reasoning process that would not otherwise be apparent based on the diagnostic decisions alone (without the context of how the decision was arrived at). By gaining a greater understanding of when certain reasoning strategies as being most appropriate, it shifts the use of cognitive interventions away from a single framework being appropriate for all decisions (as adopted in previous interventions, (**graber_cognitive_2012**; **lambe_dual-process_2016**)). Instead, one can imagine that clinicians are explicit taught a variety of reasoning techniques and then a framework for deciding which reasoning strategy to utilise for a given case when we develop a better understanding of when certain strategies are more effective for diagnoses. For instance, we discussed that effective strategy use may be predicated on past experience, and clinicians could be prompted to reflect on how much experience they have with similar patients to guide their choice of reasoning strategy (with a more structured reasoning process to fall back on when clinicians are unconfident).

We now note some limitations with our experimental studies thus far. A limitation with our vignette studies is the assumption we make that the dominant reasoning strategy for each patient's conditions is the same in the think-aloud study as it was in the online vignette study. It should be noted that, as our multinomial classifier was unable to accurately predict strategies for individual cases in the online study, we are unable to verify this assumption. This is likely because of the relatively small and unbalanced think-aloud dataset used to train the classifier. This could be rectified by future work that uses think-aloud methodologies for larger scale data collection for quantitative analysis such as this. Disentangling reasoning strategy from the medical conditions being treated is useful when looking at individual reasoning strategies on each case is useful, though we would require a much larger sample size in our think-aloud study to investigate this. We were unable to reliably predict reasoning strategies on an individual-case level for the online study where we did not have access to think-aloud utterances, which was likely a result of our think-aloud study dataset being both underpowered and imbalanced in terms of the incidences of reasoning strategies. Looking at reasoning strategies on a case-by-case level with larger samples (and different levels of expertise) would be incredibly useful and we recommend future work to adopt such a methodology to do so, as the think-aloud methodology has been found to provide useful insights into the diagnostic decision process. We would also recommend the use of other techniques such as focus group discussions to explore reasoning strategies in depth, particularly in groups of clinicians rather than simply individuals (which has been relatively underexplored in previous research).

We should also consider the generalisability and ecological validity of these studies. This study is more naturalistic than our online study, as it allows medical students to verbalise their thought process as they might to other clinicians that they are working with. By using a vignette-based paradigm however, participants do not actually interact with, observe and treat a patient. Participants noted during debrief interviews that there were elements of the task that were not analogous to real life (e.g. being able to see a patient, getting more specific information about patient history). We are also limited in terms of the information that is available for clinicians to seek. In addition, participants completed the studies in relatively controlled environments, outside of their usual medical context. Hence, our next study hence aims to look at the link between information seeking and confidence, but with a more naturalistic paradigm. This limitation is one we address more directly in an experimental study presented in the next chapter. To address this limitation, we require a paradigm that allows for more open-ended information seeking, visual observation and treatment of a patient and the use of a clinical environment akin to the one in which clinicians operate. As previously explored in our systematic review, the use of in-situ research lacks objective markers of accuracy that we utilise. To this end, we use virtual reality (VR) in our next study to simulate a realistic medical environment, as well as the patients themselves. This allows for a realistic, interactive paradigm where participants observe a (virtual) patient in real-time and can administer treatment (and observe reactions to this treatment in the patient). There is also more openness in terms of the information that can be sought and clinical actions taken, making its use more analogous to real medical contexts.

References