Information Seeking and Confidence in Medical Decision Making

Sriraj Aiyer

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

For my family, paatis and thathas

Acknowledgements

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

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

I dedicate this to my grandparents.

Sriraj Aiyer
Wolfson College, Oxford
30 September 2024

Abstract

Decisions within healthcare are unique within the wider realm of decision making. They are often made within high-pressure situations and have severe consequences if done so incorrectly. Hence, they require intensive training and a wide knowledge base for clinical staff to draw from. What is remarkable is that despite the intimidating amount of material for medical students to learn and the pressures that can befall them in their everyday line of work, as well as an ever-expanding understanding of medical conditions, treatment methods and technology to maintain, clinicians frequently make swift and accurate decisions that can have a profound impact on patients' lives. When seeking to apply past research within decision making to an applied context, medicine is an interesting domain to study decision making, especially if findings can inform the training of the newer medical students. In particular, there is a need for the teaching and assessment of non-technical skills and human factors in healthcare (Higham et al, 2019), which is currently not addressed in a widespread standardised manner in speciality curricula (Grieg, Higham & Vaux, 2015). Similarly, curricula within medicine place little emphasis on how uncertainty is communicated and approached in medical decision making (Hall, 2002). Hence, this research looks into non-technical skills such as communication of confidence, management of uncertainty and mental model alignment. Over the course of this thesis, we will look at confidence and information seeking in general decision making and then apply insights from cognitive psychology to the realm of medicine.

Contents

List of Figures

List of Tables

List of Abbreviations

AD . . . . . Aortic Dissection

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

Introduction

Diagnosis

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

Imagine a group of doctors within a hospital's intensive/critical care unit. They are engaged in a collective discussion about a particular patient. The patient has presented with a series of symptoms, including dizziness, breathing difficulties and eventual chest pain. She has been placed under continuous monitoring of her 'vital signs', including heart rate, body temperature, blood pressure, blood oxygen saturation and respiration rate. There has been a slow decrease in her blood pressure and blood oxygen saturation. The doctors are deciding what is the most likely cause of this patient's symptoms and how this may inform 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 collapses. Alternatively, there could be a cardiac cause of the patient's condition. The doctors must integrate the information they have so far, align their individual mental models of the patient and decide the following:

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

Diagnosis is a core aspect of a doctor's job and is important for a number of reasons. Firstly, accurate diagnosis is crucial to a patient's treatment. Secondly, from a psychological standpoint, it allows for an extension of previous research on 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.

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

Diagnostic error is by no means the sole cause of medical incidents. There are a number of factors tied to the wider work environment, culture and technology that can contribute to incidents and errors. A lot of these factors are challenging to isolate and emulate in an experimental setting. By understanding the individual psychological factors of the diagnostic process however, we better understand how sociotechnical and environmental factors interact with and amplify individual contributors to diagnostic error. Gaining a greater understanding of the causes of diagnostic error can have important implications for future interventions within healthcare settings.

Cognitive Biases and Overconfidence in Diagnoses

Diagnostic error can stem from cognitive biases during the diagnostic decision making process, such as primacy (Frotvedt et al., 2020) or recency (Chapman, Bergus & Elstein, 1996) biases. While it seems intuitive that classical decision making biases affect those in healthcare too (Restrepo et al., 2020), the empirical evidence of impact for medical decision making is scant, (van den Berge & Mamede, 2013). One example from dermatology looked found examples of satisficing bias (premature closure) and anchoring were found, but few examples of others such as availability and representative biases (Crowley et al., 2012). One type of bias that has manifested in more experimental findings is overconfidence (Berner & Graber, 2008, Meyer et al.., 2013).

At this point, we shall revisit the scenario presented at the start of this section. In summary, a patient is presenting with a set of symptoms, requiring doctors to assign a diagnosis to guide future treatment. One of the doctors confidently presents their opinion that the patient has suffered a pneumothorax.. The certainty with which the diagnosis is suggested makes it more difficult for others to disagree with, especially if the doctor is a consultant/attending such that there is a disparity in seniority.

Introduction

Confidence can be viewed as one's "subjective probability of a decision being correct" (Fleming & Daw, 2017). Confident individuals tend to be more influential on others in a group (Zarnoth & Sniezek, 1997) and can even causally increase the confidence of other observers (Cheng et al., 2021). This behaviour has been observed in mock jury trials, during which participants hear eyewitness testimonies presented with high confidence and then perceive those testimonies as more credible than testimonies provided with low confidence (Cutler, Penrod & Dexter, 1989, Roediger, Wixted & DeSoto, 2012). Confidence is a commonly used predictor of another person's accuracy, especially when feedback is not readily available of an individual's true accuracy. Confidence also varies across individuals with what may be considered a 'subjective fingerprint' (Ais et al., 2016), and 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 & Packer, 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 and having 'second thoughts' or changes of mind are more likely with a lower level of confidence (Resulaj et al., 2009).

Individuals are 'well-calibrated' with regards to confidence if their internal likelihood of being correct is predictive of their true accuracy. However, confidence can become decoupled from true accuracy. This decoupling is known as 'miscalibration'. One would show miscalibration of confidence if they tended to be more confident than they are correct (overconfidence) or more uncertain than they are correct (underconfidence).

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, Waechter & Millington, 2018). However, it has also been found that underconfidence can be more prevalent than overconfidence, especially when comparing medical students to residents (Friedman et al., 2005). Similarly, Yang and Thompson (2010) found that experienced nurses exhibited similar performance to nursing students, but were more confident in their

judgements, showing differences in confidence calibration across experience levels. More broadly, highly confident members within a group could unknowingly reduce the chance of less confident members speaking up about potential errors, which is a common problem within healthcare (Hémon et al., 2020). Overconfidence has also been linked to a lower likelihood of sufficient patient management and clinical effort as per a field study in Senegal (Kovacs, Lagarde & Cairns, 2019).

We would argue that building on the current research landscape of diagnostic confidence is important. If there is an assumption that others will calibrate their confidence to their true accuracy, this would mean that heeding high confidence advice or judgements would be an optimal strategy for maximising accuracy. However, this can be a serious issue when high confidence errors lead others astray. This is important, as in addition to seniority and speciality experience, a clinician's confidence is one of the only markers available for other clinicians and for patients when making key medical decisions. One underexplored avenue in current research is the role that information seeking during the diagnostic process affects confidence.

Information Seeking

Clinicians generate hypotheses and then gather information to reduce the space of 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 continue attaching themselves to a hypothesis when there is overwhelming evidence to the contrary. One conclusion of Wason (1960) was that individuals struggle to remove a hypothesis from consideration even if they receive evidence against it. Understanding how individuals generally reason about a possible space of hypotheses is interesting for understanding how the reasoning process works differentially for novices and experts, especially in a specialised domain such as medicine. One question that is worth investigating is how the 'process of elimination' affects confidence.

Introduction

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

There is also evidence to assume that information seeking is important within medical diagnoses too. Notably, Gruppen, Wolf & Billi (1991) found that clinicians were less confident when they had to seek relevant information for themselves compared to 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 in their judgements (Clayton et al., 2022). In a sample of 118 physicians presented with patient vignettes, it was found that higher confidence was associated with a decreased amount of diagnostic tests being ordered, even if confidence and accuracy were larger decoupled/miscalibrated (Meyer et al., 2013). It has also been observed previously that physicians may 'distort' neutral or inconclusive evidence to be interpreted as supporting prior beliefs (Kostopolou et al., 2012). Similarly, it has been found that a patient's case history that suggests a particular diagnosis prompts selective interpretation of clinical features that favour the initial diagnosis (Leblanc, Brooks & Norman, 2002). Together, these findings have implications for how clinicians may seek and integrate evidence when making decisions and how patterns of receiving information could affect decision confidence and in turn confidence calibration.

Diagnostic decisions have been thought of as 'ideal' when using the hypothetico-deductive process (Kuipers & Kassirer, 1984), whereby hypotheses are formulated

based on specific features of a patient and are then linked to established criteria for a diagnosis, with further information gathering to test these hypotheses (Higgs et al., 2008) or eliminate others. This account was challenged by Coderre et al. (2003), who found that diagnosis can based more on pattern recognition, especially for more experienced clinicians. Either way, the bridge between confidence and information seeking is the reasoning strategy utilised by clinicians. Diagnostic reasoning is currently taught using cognitive frameworks such as the surgical sieve and the ABCDE pneumonic. However, current education does not account for differences in reasoning strategies, whether strategies may meaningful vary by case and by clinician and how these strategies have a downstream influence on the diagnostic process in terms of seeking information, generating differentials and formulating confidence.

Current Work

There is a need for the teaching and assessment of non-technical skills and human factors in healthcare (Higham et al., 2019), which is currently not addressed in a widespread standardised manner in speciality curricula (Grieg, Higham & Vaux, 2015). Curricula within medicine also place little emphasis on how uncertainty is communicated and approached in medical decision making (Hall, 2002). In addition, there is little work that informs how information seeking is taught within medical reasoning other the use of cognitive frameworks (such as the 'surgical sieve') and pneumonics (such as Airway, Breathing, Circulation, Disability, Exposure). Clinical experience may also be connected to risk aversion and further information seeking behaviour (Lawton et al., 2019), which offers an important avenue for future medical education. Hence, this research informs medical education of non-technical skills such as diagnostic reasoning, especially around evaluating diagnostic differentials and seeking information during the diagnosis process.

The following sections are structured as follows. Firstly, I will present a scoping review of the medical and psychological literature in which confidence or certainty has been studied within diagnostic studies. This review will map out

the broad findings from this extant literature and identify gaps in our current understanding, some of which this DPhil aims to fill. Next, I will present an online behavioural study with medical students where participants freely sought information and provided diagnostic differentials at different stages during a series of patient vignettes. This study allows us to study how diagnostic differentials and confidence are affected by patterns of information seeking. The following section then details an in-person using a similar paradigm where medical think aloud as they are making these diagnoses, with the aim to use these think aloud utterances to classify different diagnostic reasoning strategies. These different strategies can be used to reanalyse the online study to investigate how reasoning strategies affect confidence and information seeking. The third empirical study seeks to look at diagnostic decisions in a more naturalistic manner by using virtual scenario paediatric scenarios to investigate differences in information seeking and confidence. Finally, I present a reflective chapter based on observations in Intensive Care, whereby the findings from this DPhil are contextualised within the decisions made during actual medical practice.

Study 1 - Scoping Review of Literature on Confidence and Certainty in Diagnoses

Study 2 - Information Seeking and Confidence in Diagnosis

Methods

Participants

Participants were recruited between July 11th 2022 and April 6th 2023. 85 medical students were recruited for this study, including 32 males, 52 females and 1 participant who self-reported as non-binary. The age ranged between 22-34 (M = 24.2). The study was conducted online, with participants able to run the experiment in their browser. The experiment was coded using JSPsych, which is a Javascript plugin used specifically for psychology experiments. We recruited fifth or sixth (foundation) year medical students using a mailing list that those within OxSTAR have access to in order to recruit medical students based in Oxford. In order to recruit from further afield, we were assisted by the Medical Schools Council, who distributed the study to students in other medical schools across the UK. Participants were emailed with a study information sheet and a link to access the experiment, where they first provided consent via an anonymous online form. After doing so, the participant provided demographic information (age, gender and years of medical experience).

Materials

Our first study involved the usage of patient vignettes. These are simulated patient cases that have been adapted from actual past cases. We used the bank of patient scenarios from Friedman's (2004) study as a foundation for our scenarios. However, it should be noted that these vignettes could not be used straight away as they

were provided. These vignettes were developed by a team of researchers based in the US, meaning that certain medical terms (eg medication, tests etc) had to be 'translated' into the vernacular used by doctors based in the UK. This was done via consultation with researchers working with the OxSTAR Centre who were also practising medical staff and students within the NHS. There also may have been differences in vernacular based on time, given that the original vignettes were developed for a paper published in 2004. Cases made occasional references to specific years in the patient's history where they have experienced previous medical conditions. These years were updated to make sense for a contemporary patient. Whilst a sizable bank of vignettes were kindly provided to us by Friedman, certain conditions were considered too rare (either for the current time or for the UK) to be used. Our goal was to test the clinicians' ability to deal with diagnostic uncertainty, rather than testing their declarative knowledge of obscure medical conditions. We therefore chose cases that involved medical conditions that medical students would be expected to know.

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

Procedure

The procedure of a single case (or 'trial') is as follows. The participant is asked to imagine that they are working in a busy district hospital and they encounter patients in a similar way to how they would in their real medical practice. At the start of each, the participant is shown a presenting complaint for a patient, which includes the patient's age and their main symptoms. An example of this is as follows: "patient is a 68 year old male presenting with fever and arthralgia". This information remains on screen throughout the entire case. Each case is

split into three information stages: Patient History, Physical Examination and Testing. This order of stages is fixed for all participants. At each stage, the participant sees pieces of information or tests that they can request. Participants can view information from a previous stage but cannot see information for a future stage (e.g. if a participant is at the Physical Examination stage, they will be able to see information pertaining to Patient History and Physical Examination, but not information pertaining to Testing). The set of information requests for each stage is the same for all cases. The Patient History stage includes information on "Allergies", "History of the Presenting Complaint", "Past Medical History" and "Family History". The Physical Examination stage includes 'actions' that a doctor may take when examining a patient, such as "auscultate the lungs", "abdomen examination", "take pulse" and "measure temperature". Finally, the Testing stage involves information on any bedside tests or tests they may request from another department. This includes "Chest X-Ray", "Venous Blood Gas", "Urine Dipstick" and "Clotting Test". There is a total of 29 possible tests that can be requested across the three information stages. When a participant clicks on any of these tests, the screen shows a loading icon for 3 seconds before showing the information for that test on screen. During this loading time, other tests cannot be requested. When any subsequent test is requested, the previous test result is removed from the screen such that participants can only view one piece of information at a time. The time delay for receiving information was added after piloting the study, where the lack of time delay meant that participants were likely to request most information without being selective. We also emphasised during the instructions that participants should only request information that they believe will help them with diagnosing the patient. The information shown for each test is pre-defined as per the medical vignettes and is the same for all participants. Participants are free to request the same piece of information multiple times in order to remind themselves, including information from a previous information stage.

At any point, the participant can choose to stop gathering information for that stage. They do so by clicking the "Enter Differentials" button. At this point,

they are taken to a new screen where they can report a list of all differential diagnoses that they are considering for that patient at that stage. Participants can report as many diagnoses in their list as they want. For each differential, participants report a "level of concern" for that differential, which we describe as how concerned the participants would be for that patient if this differential really was the patient's underlying condition. This is reported on a 4 point scale, with labels of "Low", "Medium", "High" and "Emergency". Participants also reported a likelihood rating for each differential, ranging from 1 (very unlikely) to 10 (certain). When reporting differentials at the first information stage, the list of differentials is blank and participants must add at least one differential to proceed. In subsequent stages, the list from the previous stages is available for participants to update concern/likelihood ratings, add differentials or remove differentials from the list. Participants are asked to carefully consider which differentials they have in mind in light of the new set of information they have received. Even at the last information stage, participants can report multiple differentials if they do not prune their list down to a single diagnosis. Participants are not penalised for reporting a wide set of differentials at any stage.

After recording their differentials, participants are then asked to report their confidence that they are "ready to start treating the patient" on a 100 point scale, ranging from fully unconfident to fully confident. This is different to previous papers as it takes the focus away from merely their confidence that they have the correct answer. Participants are also able to indicate using a checkbox that they are ready to start treating the patient, at which a text box appears for them to report what further tests they would perform, any escalations they would make to other medical staff and treatments they would start administering for the patient. Once all three stages are complete, participants report how difficult they found it to determine a diagnosis for that case, on a scale from 1 (trivial) to 10 (impossible). At the end of all six patient cases, participants are told the true underlying conditions for all the patients.

Data Analysis

There are a number of key dependent variables that we are able to derive from our data: * *Confidence*: the reported confidence at each information stage. Initial Confidence refers to the reported confidence after the first stage of information seeking (Patient History), whilst Final Confidence refers to the reported confidence after the third and last stage of information seeking (Testing). We can then use these two variables to calculate Confidence Change, by subtracting the participants' Initial Confidence from their Final Confidence. Hence, a positive value for Confidence Change means that the participant has gained confidence over the course of the patient case. * *Proportion of Information Requests*: we take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage or including tests that had been requested before during that stage) and divide by the number of possible tests available during that stage (which is the same for all cases). * *Number of Differentials*: we take the number of items in the list of differentials at each stage. Initial Differentials refer to the number of differentials after the first stage of information seeking (Patient History), whilst Final Differentials refer to the number of differentials after the third and last stage of information seeking (Testing). * *Subjective Difficulty*: the subjective rating by participants at the end of each case for how difficult they found it to determine a diagnosis for that patient case. This is reported on a scale from 1 (trivial) to 10 (impossible). * *Accuracy*: For a case to be considered 'correct', the participant should have reported the correct condition for that case within their list of differentials regardless of the number of differentials provided. Given that differentials are provided via free text, cases have to be manually coded as correct or incorrect. Spelling errors or alternative names are not penalised. To calculate Accuracy, we first identify the correct differential if provided in the list and find the likelihood rating assigned to that differential. The highest possible value here would be 10 if the participant included the correct condition in their differentials and assigned it the maximum likelihood rating. If a correct differential is not provided, a value of 0 is assigned. Lists of differentials were 'marked' for correctness

manually using the following criteria (the correct condition is followed by the list of accepted diagnoses to be considered correct): - TA: any inflammatory arteritis is accepted - UC: infectious colitis, ischemic colitis or diverticulitis are also accepted answers. - MTB: any TB or lymphoma type is accepted - AD: pulmonary embolism or coarctation of the aorta are also accepted. - GBS: cauda equina syndrome is also accepted - TTP: ITP or Meningitis are also accepted. * *Information Seeking Variance*: We compute a vector of length 29 (as this is the maximum number of unique pieces of information that one can request during each case), which is made up of 0s and 1s where for each of the pieces of information available for a case, a value of 1 is assigned if that information is requested and 0 is assigned if that information is not requested during the case. The normalised vectors for all cases for a given participant are combined to produce a 29 x 6 matrix. We calculate the Euclidean distance between each row of the matrix (trial) using R's dist function (in the proxy package). The computation of all pairwise distances produces a 6 x 6 matrix where each trial is given a Euclidean distance value relative to every other trial. A lower distance value between two trials indicates that the information sought on those trials are similar to one another. In order to look at the similarity of information seeking across all six trials, we compute the variance (the standard deviation squared) of the participant's Euclidean distances. A lower variance value indicates that participants seek similar information across the cases whilst a higher value indicates that information seeking is varied more by case. * *Information Seeking Value*: We take each of the 29 pieces of information in turn and split all participant trials into two groups: those trials where that information was sought and trials where that information was not sought. For each group, we compute the proportion of trials where participants included a correct differential, and we then take the difference between these two values. A positive value difference would indicate that participants were likely to identify the correct condition with that information rather than without that information. This difference can be considered that information's value. For each of the participants' trials, we calculate the sum of information values for all information that the

participant did seek based on these values and we then take the mean of these sums across trials. This gives an overall measure of how useful the information was that participants tended to seek. We avoid circularity in this measure via cross validation. As such, each participant's information values are derived by looking at differences in accuracy for all other participants.

Results

Firstly, we can look at how our dependent variables change over the course of a case by comparing at each of the three stages. We fit a linear model by using the stage as our independent variable and our key dependent variables. A key finding is that the number of differentials increases over the course of the stages. This indicates that students do not tend to narrow their differentials with more information, rather they broaden their differentials. Participants rarely used the option of removing differentials from their consideration, which could be attributed to how they have been taught to make diagnostic decisions. We found that accuracy and confidence increased across the stages in a manner that was well calibrated, unlike previous papers. When conducting a Pearson's Correlation test, we found evidence for a positive correlation between the change in confidence (the difference in confidence in the first and last stages) and the proportion of information sought ($r(83) = 0.24$, $p = .03$), such that seeking more information was associated with higher gains in confidence.

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

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

We then look at how ability on the task relates to information seeking behaviour. To do this, we calculated the average accuracy for each participant (across cases)

and then sorted participants into four groups by quantiled accuracy. We then look at mean information seeking variance for each group of participants. We find that participants with higher overall accuracy have a lower variance in information seeking. In other words, students with a higher diagnostic ability are found to have varied the information they sought across cases less, seeking more similar information for each case when compared to students of a lower diagnostic ability. We can also test the same hypothesis by treating participant accuracy as a continuous measure, and we find evidence for a negative correlation between accuracy and information seeking variance ($r(83) = -0.23$, $p = .04$). We apply a similar analysis to look at how information value varies as a function of participant ability. We find evidence for a positive relationship between accuracy and information value ($r(83) = 0.25$, $p = .02$). Taken together, students with a higher diagnostic ability seek better information but also approach each case in a more similar manner. This could indicate a base of information kept constant across cases alongside a more selective set of useful information related to that patient. Meanwhile, participants with a lower diagnostic ability are not selective with their information seeking and hence do not seem to have a set framework or plan for what information to seek. We do also find that the proportion of available sought is not shown to correlate with accuracy on at the final stage ($r(83) = 0.17$, $p = .11$) but does correlate with the participants' change in confidence, which is the difference in confidence between the first and final stages ($r(83) = 0.24$, $p = .03$). While seeking more information may imbue students with a greater level of confidence, it does not necessarily translate into more accurate diagnoses. This is important to note as it demonstrates that being selective in information seeking is a better marker of performance and giving a lower ability participant all available information does not necessarily translate into accurate diagnoses. This has interesting implications for medical practice, as the ordering of unneeded tests or patient examinations may not contribute to better decisions. Given the constraints within most hospitals and healthcare to obtain certain tests, being selective with information seeking is already a frequent

necessity and results from this study seem to show evidence that it is also a good marker of diagnostic performance.

To further investigate the differences in information seeking, we also trained a binary classification algorithm using a generalised logistic regression model. To do this, we first split all trials into high and low ability participant trials with a median split of participants by their average accuracy across the six cases. We train the classifier by treating the 29 binary variables for each information as predictors (with a 1 signifying that the information was sought for that case and 0 when the information was not sought) to predict the binary outcome of whether the participant is a low or high accuracy participant. For this, we do not take into account the specific case. We used Leave One Out Cross Validation, such that each trial is predicted by training the algorithm on all other trials. By plotting an ROC curve of our classifier, we find an area under the curve (AUC) value of 0.727 (with $p < .001$ when comparing the ROC curve to AUC = 0.5). This indicates that differences in information are indeed predictive of a difference in participant ability.

Discussion

There are a few limitations with our study. We did not use more naturalistic stimuli, such as images of scans/test results or audio cues (such as the sound of lung auscultation) and instead used solely textual results for all tests. While this may make the experiment more ecologically valid, it takes away the interpretation of complex stimuli which could affect information seeking. For example, if two participants requested a chest X-ray, they may interpret the X-ray image in different ways. While this difference in perception may be interesting, it adds a potential confound for the purposes of this study. That is why, for this study, if a participant requested a chest X-ray, they instead see a result that reads something like "no abnormalities found", such that the interpretation of the image has already been done for the participant.

Our experiment also assumed that all tests were equal in terms of how long they take for results to be shown. If the tests were analogous to real medical practice, certain tests would take longer to produce results after being requested. Some tests (e.g. a chest X-ray) are not performed by the doctor themselves at the patient's bedside and require staff and technology from another department. We should also note that our experiment was run via an internet browser, meaning that study participants were taken out of the setting within which they would usually make these decisions. This means that participants may act differently than they might do in their regular medical practice. In addition, we attempted to make the patient cases as realistic as possible whilst having a moderate degree of difficulty. The original researchers removed certain findings from the cases that may give away the patient's condition in a fairly obvious manner. In that sense, the patient cases may not replicate the set of information that might be available to clinicians in a similar scenario during medical practice. However, using a paradigm similar to past research does extend and build upon empirical experiments on diagnosis. As previously mentioned, information was chosen in order to be general to all cases and was not very discriminant.

Within this discussion, it is worth mentioning a few general observations in the data and how they might inform the design of future studies of this nature. Firstly, participants did not tend to use the ability to remove differentials from their list. In our study, participants could remove a differential in the interface by clicking the X button on a differential. One explanation is that the button is not very prominently placed on the screen. However, this feature was explicitly explained in the tutorial to the experiment. This tendency is reflected in the overall pattern of the average number of differentials increasing over the three stages of a case. What this may indicate then is an attachment to hypotheses and unwillingness to remove them from consideration. There is a general adage in healthcare that medical students come across which says that "history is 80% of diagnosis". The fact that diagnostic differentials do not change that much between stages is supportive of this. Indeed, accuracy does not improve by a large amount between stages (from 52.2% after

Patient History to 65.9% after Testing). It is indeed striking that in over half of all cases, students are able to include the correct condition in their differentials by the patient's history alone. It is therefore worth considering whether there is a specific facet of diagnostic decisions whereby clinicians are taught not to disregard diagnostic possibilities easily. This also corresponds with participants tending to request most, if not all, information during the Patient History stage (86.1% across all participants) and then becoming more selective in information seeking during later stages. Hence, this indicates a general behaviour to gain the majority of diagnostic differentials from Patient History and to not easily disregard diagnoses.

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

We should also note the manner in which accuracy was coded manually for each case. This depended on the nature of the case, as a case may sometimes have a vague set of information such that determining the exact correct diagnosis was considered too challenging. For example, for the TTP case, making a diagnosis of TTP (even with all information requested by the participant) was seen as too difficult given that the information provided was not discriminant enough. This ties into one of the main challenges of designing these vignettes and this study: the set of information available for participants to request were chosen such that

they were reasonable to be requested in any of the cases. The participants may have wanted to request more specialised, discriminant tests (e.g. lumbar puncture, biopsy), but including these could clue participants into the nature of the patient's condition. In addition, these types of highly specialised tests that target a specific type of diagnosis tend to take much longer to come back to doctors with results after they request them in a real healthcare setting. Hence, having results available at the touch of a button for these may seem unrealistic unless we alter the design to have patient cases unfold over a longer time period.

Study 3 - Diagnostic Reasoning Strategies via a Think-Aloud Paradigm

We aimed to replicate the finding of considered differentials increasing with more information when the method by which these differentials were reported. Are students seeking information to confirm their existing set of differentials, to rule out differentials or to expand their set of considered possibilities? And are these different approaches interleaving or are they more dependent on individual diagnostic decision making styles? In order to provide more context to the results from study 1, we conducted a follow-up study that utilised a very similar experimental procedure, but instead prompted students to think out loud as they were performing the task. and the transcripts were coded to conduct both quantitative and qualitative analysis.

Think-aloud methodologies are useful for directly accessing ongoing thought processes during decisions (van Someren, Barnard & Sandberg, 1994). The use of thinking aloud (or verbal protocols) in research is useful for being able to access the information attended to participants in short term memory (Payne, 1994) and can be treated as the ongoing behavioural state of a participant's knowledge (Newell & Simon, 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 (de Groot, 1946, Bilali, McLeod & Gobet, 2008). 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, 1994). This is especially well-suited to our task where the information available to participants is controlled with time, allowing us to investigate how diagnostic thinking develops with more information. A think-aloud

methodology has previously been used to study the differences between novice and expert clinicians during diagnostic reasoning (Coderre et al., 2003). This study found a general trend that experts tend to use a 'pattern recognition' approach to diagnosis more than novices, who tended to use a 'hypothetico-deductive' process (which is aforementioned to be the 'textbook' description of the diagnostic process), but this was highly dependent on the case presented. We build on the work of Coderre et al. (2003) here to further investigate how reasoning strategies contribute to accuracy and why certain cases result in differing strategies.

Methods

Participants

16 participants were recruited for this study. Participants were 5th or 6th year medical students at Oxford University (including 2nd year Oxford University Graduate Entry Medical students) recruited via posters in the John Radcliffe Hospital in Oxford and via a mailing list for students managed by the Medical Sciences Division at the University of Oxford. The study was conducted onsite at John Radcliffe hospital. Participants were recruited between July 5th 2023 and December 1st 2023. Data was reviewed on an ongoing basis to cease recruitment when emerging themes were exhausted. 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 cases and a similar computer interface from Study 1 were used for this study, with the exception that participants no longer recorded their differentials in a specific screen at the end of each information gathering stage. Instead, participants' differentials were recorded as a more naturalistic part of their diagnostic process as they reported aloud their thoughts 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 Study 1, except that participants were given the following instructions at the start of the study:

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

The experimenter 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, but the transcript was checked and modified for accuracy by listening through the audio recordings again. The screen of the experimental interface was also recorded, such that the audio could be linked to specific actions within the task. The focus of this study is on verbal utterances rather than any non-verbal or inferential

aspects of the participants' qualitative data. At the end of the experiment, the researcher administered a semi-structured interview to better understand what the participants feel their diagnostic reasoning approach tends to be. These questions are provided in the Supplemental Materials.

Data Analysis

We conducted a theory-driven semantic thematic analysis (as per definitions detailed by Braun and Clarke, 2006) to code utterances under specific categories. This kind of thematic analysis is suitable given that our qualitative data is from a structured experiment, rather than a dataset with a looser structure (e.g. interview recordings). As a result, we apply deductive analysis using predetermined codes for think-aloud utterances and for a debrief interview where we administer a semi-structured interview with specific questions of interest.

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

- Differential Evaluation: any time that the participant (each of the following is considered a separate subcode):
  - 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 contradict with a condition

Think Aloud

We also define a group of codes that indicate three different diagnostic reasoning strategies: hypothetico-deductive reasoning, scheme-inductive reasoning and pattern recognition (Coderre et al.., 2003). These were defined as follows:

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

We first code specific 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. In addition to coding each case under one of these strategies, we also code participants on an overall level based on their subjective perception of how they make diagnostic decisions. This is based on responses provided during the debrief interview (as described in the Procedure section). Hence, reasoning strategy codes are at the case level and also at the participant level.

Coding of utterances and case-wise reasoning strategies were conducted with a second independent coder. For reasoning strategies, initial interrater reliability was low, with both coders agreeing on 58.3% of cases. Conflict resolution led to changes made to the coding criteria by prioritising strategies used early in a case, as some participants were noted to utilise multiple strategies within a single case, as well as allowing some cases to be coded as not having a clear strategy

due to a lack of utterances. Conflicts were then resolved 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 mentioned in Study 1).

Although we do not record differentials in the same way as in Study 1 (in a list with corresponding likelihood and severity ratings), we do obtain the other variables from Study 1. Namely, we record confidence at each stage of information seeking and data around the information sought by participants. As we do not explicitly record differentials in the same manner as in Study 1, accuracy is operationalised differently. We code each case as 'correct' if a correct differential is mentioned at some point by the participant (using the same marking scheme in table S1).

Results

First, we look at overall quantitative characteristics of the think aloud statements. When looking at accuracy (the proportion of cases where a correct differential was mentioned by the participant), accuracy was 0.57 across all cases. This varied considerably by condition however, with accuracy across participants for each condition being as follows: AD = 0.63, GBS = 0.88, MTB = 0.19, TA = 0.44, TTP = 0.69, UC = 0.63. For utterances coded as Differential Evaluations, participants on average made 5.21 such utterances per case (SD = 2.80). The mean number of Differential Evaluations was relatively constant by condition except for the AD case: AD = 8.18, GBS = 4.63, MTB = 4.81, TA = 4.75, TTP = 4.25, UC = 4.63. Participants varied in how much they spoke during the study, uttering 1038-7730 words (M = 4194) across the scenarios. Part of this range is driven by participants repeating information they see during the task, but participants also varied in terms of how much they externalised their thought process.

As previously mentioned, Differential Evaluations can be further categorised into one of four subcodes: Differential Added, Differential Removed, Likelihood Increased and Likelihood Decreased. 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 = 0.89) than removing differentials (M = 0.27, SD = 0.28) ($t(15) = 14.14$, MDiff = 2.86, $p < .001$). Participants expressed more statements of decreasing likelihoods (M = 0.99, SD = 0.62) rather than increasing likelihoods (M = 0.93, SD = 0.46) but we did not find evidence of a significant difference ($t(15) = 0.34$, MDiff = 0.06, $p = .73$).

Reasoning Strategies

Next we look at our coding of reasoning strategies at a case level. As mentioned, our criteria for each code was applied to each individual case based on the transcribed utterances. When looking at reasoning strategies by case, 43 cases were coded as Hypothetico-Deductive, 29 were coded as Pattern Recognition and 18 were coded as Scheme Inductive (the remainder of cases did not contain enough clear utterances to classify under one of these strategies). Accuracy was higher for cases coded as Hypothetico-Deductive (71%) compared to both Pattern Recognition cases (64%) and Scheme Inductive (39%). It is worth noting here that accuracy was solely based on participants mentioning differentials during their thinking aloud, which is naturally not facilitated by Scheme Inductive reasoning due to its focus on identifying pathophysiological systems acting as sources of patient symptoms rather than specific conditions. This can hence explain the lower 'accuracy' for Scheme Inductive cases. We also note that the types of reasoning strategy used varies by condition (see Figure 13 below), with the MTB and TTP cases in particular exhibiting higher usage of Pattern Recognition than others. This could be because this case was considered harder than others and hence participants could not generate a larger set of candidate differentials due to its difficulty.

We note, rather unsurprisingly, that we observe a higher number of average Differential Evaluations when cases are correct (M = 5.85, SD = 0.38) compared to when they are incorrect (M = 4.34, SD = 0.39). Given our methodology for defining accuracy, participants are more likely to mention a correct differential if they mention more differentials. The procedure used in the previous study

for collecting data on which differentials participants were considering at each information stage was not present here and hence we are not able to operationalise accuracy in the same manner as before. While we look at which differentials are mentioned, we cannot observe how participants weigh up differentials against each other in the same way as in the first study.

To connect the results of this study to those of Study 1, we break down the same dependent variables (as operationalised in that study) by reasoning strategy. We do not apply statistics to this study due to the lower sample size. 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, UC, GBS), three conditions as PR (MTB, TTP, TA). The proportions of participants who use each reasoning strategy for each condition can be viewed in Figure 10. We then compare the individual case classifications of strategy to this reasoning strategy that is most commonly used for that medical condition. Table 2 shows how dependent variables are affected by reasoning strategy. We find that the amount of information seeking was fairly consistent across reasoning strategy, but that PR cases were associated with higher value in information seeking. In order to derive informational value, we used the same values of each piece of information for each case that were derived in Study 1. This higher informational value does not translate into higher accuracy for PR cases, though we should note that the manner in which accuracy was defined for this study limits the analysis only to statements made out loud of specific conditions rather than formally recorded differentials as we did in Study 1. In order to formally replicate this finding with the larger dataset, we use the cases from this study and the coding of strategies to apply the same coding to our online dataset from Study 1.

Reasoning Strategies in Study 1 Dataset

In order to apply reasoning strategies to the data from Study 1, we train a classifier using penalised multinomial regression to classify cases as HD, PR or SI using the cases from the think aloud study (with Leave One Out Cross Validation). The input

parameters for the classifier are the 29 pieces of information as binary predictors (similar to the approach depicted in Figure 7) 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 is the test dataset. The classifier was implemented using R's nnet package (version 7.3-19). The testing data is then labelled with predicted testing strategies using R's predict function. We note that the training data was initially labelled with reasoning strategies using the think-aloud utterances and thus is separated from the information sought during the case.

We show a breakdown of cases by their coded reasoning strategy in Table 4. We now look to compare our key dependent variables by strategy, in particular comparing PR and HD cases. In line with our expectations based on the definitions of HD and PR reasoning approaches, we find that HD cases are associated with more differentials being considered (M = 3.37, SD = 1.64) average when compared to PR cases (M = 2.84, SD = 1.58) and find evidence of a difference between the two via a Welch Two Sample t-test (t = 2.89, MDiff = 0.53, p = .004). We find that PR cases are associated with higher informational value (M = 2.35, SD = 1.07) when compared to HD cases (M = 2.15, SD = 1.32) (t = 1.48, MDiff = 0.20, p = .14). However we do find evidence of higher amounts of information seeking for HD cases (M = 0.63, SD = 0.21) when compared to PR cases (M = 0.50, SD = 0.21), (t = 5.28, MDiff = 0.13, p < .001). Overall, this indicates that PR reasoning were associated with lower but more selective information seeking when compared to HD reasoning.

We hypothesised that an interaction with reasoning strategy is associated with accuracy on the task. This is because a single reasoning strategy is considered unlikely to be more accurate for all cases. As indicated by Figure 10, different patient conditions seem to result in varying reasoning strategies being utilised, which begs the question of what properties of a condition contribute to changes in strategy and in accuracy. One possibility is that reasoning strategy interacts with the diagnostic uncertainty of a case (i.e. the breadth of conditions that a

patient could have given their current symptoms and history, with some conditions presenting in a more apparent way than others), as captured by the number of initial differentials reported by participants. To test this hypothesis, we fit a linear model to predict accuracy with an interaction between the number of initial diagnoses and reasoning strategy. The interaction regression lines are plotted below in Figure 11.

Discussion

In the quantitative results shown in Table 4, we firstly find that in our think-aloud study, PR cases were associated with the highest informational value (despite fairly constant amounts of information seeking across reasoning strategies). SI cases were associated with the highest increase in confidence across the information stages. Whilst the latter finding was similar in the online study, PR cases saw higher informational value than HD cases in this dataset. In accordance with our definitions of reasoning strategies, we find that PR cases were associated the fewest differentials of the three strategies. These findings indicate that different reasoning strategies result in behavioural differences when performing diagnoses, both in terms of information seeking and weighing up differentials.

We also find evidence of an interaction effect between reasoning strategy and the number of initial diagnoses on accuracy. Intuitively, it makes sense to broaden or narrow differentials based on the number of differentials being considered. Given that reasoning strategies differ by a function of the case/condition, it might be that the case-level factor affected reasoning strategy is how 'apparent' the underlying condition of the patient is based on the initial patient presentation/history. We operationalise this as the number of initial differentials, which captures how clear the patient's condition for a given participant based on the patient history. In the case of the interaction depicted in Figure 14, we find that reasoning strategy and the number of initial differentials interact. With a lower number of initial differentials, participants exhibited increased diagnostic accuracy by broadening

their consideration of differentials via a hypothetico-deductive process (i.e. "what else could be causing these symptoms?"). With a higher number of initial differentials, higher diagnostic accuracy was found by narrowing differentials via a pattern recognition process (i.e. "which of these differentials does this patient most resemble?"). Whilst past work has tended on focusing on designing cognitive interventions that aim to fit all diagnostic scenarios, this result indicates that a flexibility in reasoning strategy based on the patient case is key for increased accuracy. Future research should hence focus on prompting the right reasoning strategy based on the initial patient presentation.

We can consider both studies together to provide a nuanced discussion of the diagnostic process among medical students. We find that information seeking patterns and evaluation of differentials during the diagnostic process contribute to diagnostic accuracy. When students generated a greater number of differentials from a patient history, they sought a greater amount of information. We then observe an association between information seeking and confidence, but not with accuracy. Instead, accuracy was characterised by more selective information seeking during the diagnostic process. This is important to note as it demonstrates that being selective in information seeking is a better marker of performance and giving a lower ability participant all available information does not necessarily translate into accurate diagnoses even though it increases diagnostic confidence (Gruppen, Wolf & Billi, 1991).

This has interesting implications for medical practice, as the ordering of unneeded tests or patient examinations may not contribute to better decisions and is not cost effective. Given the constraints within most hospitals and healthcare to obtain certain tests, being selective with information seeking is already a necessity and results from this study seem to show evidence that it is also a good marker of diagnostic performance. There has been increased research on overtesting, such as requesting costly imaging scans when they may not be medically necessary (Carpenter, Raja & Brown, 2015). 'Overtreatment' has been estimated to cost the US healthcare system between 158 and 226 billion dollars

in 2011 (Berwick & Hackbarth, 2012). Seeking more information during the task made students more confident but not more accurate, which is important to note as it corresponds with previous findings from the cognitive psychology literature (Ko, Feuerriegel, et al., 2022).

The finding of evidence for an interaction between strategy and the number of diagnoses with regards to accuracy is an interesting one for future medical education. Past work that aim to teach diagnostic reasoning or administer cognitive interventions/aids has tended to assume that a single aid can be optimal for all types of patient cases. However, this results hints at the fact that reasoning strategies' effects on accuracy depend on the initial diagnostic uncertainty associated with the case. In particular, PR seems to result in lower accuracy for fewer initial diagnoses and higher accuracy for more initial diagnoses. This makes intuitive sense when considering how reasoning strategies relate to reducing or expanding the space of possible diagnoses. For instance, if a clinician has a large set of possible differentials in mind from the initial patient presentation, they should narrow their range of possibilities using a pattern recognition approach ("which of these conditions does this most match?"). Conversely, if a clinician is struggling to bring multiple differentials to mind, they should broaden their thinking to consider more conditions using a hypothetico-deductive approach ("what other conditions should I be concerned about?"). This account of the results is bolstered by our operationalisation of accuracy, whereby participants are more accurate by not only considering the correct condition but also considering it as highly likely amongst the considered differentials.

Coderre et al. (2003) found the pattern recognition was utilised more as clinicians increased in experience. On the one hand, this makes sense given that having more experience of disease presentations would improve a diagnostician's ability to match symptoms to a condition. However, as alluded to by students in this study, knowledge and experience brings with it the ability to generate more differentials than a less experienced clinician. One cannot adopt a hypothetico-deductive reasoning process, whereby multiple differentials are considered and then

eliminated, if the clinician lacks sufficient knowledge to generate a set of differentials based on the observed patient. This may be where the complexity/difficulty of the case has a bearing on reasoning process too, whereby harder cases are harder because one cannot easily generate differentials for them. However, the inverse could also be true, whereby a set of conflicting symptoms may cast a wider net of potential differentials that are more challenging to narrow down. As we noted in the online study, the number of initial differentials has an impact on information seeking behaviour, but as we explain here, differentials are themselves a result of a particular reasoning strategy. Ascertaining the exact interaction between reasoning strategy, case difficulty and differential evaluation is hence important for us to focus on in the following study, as it informs how diagnosis is characterised as a cognitive process and how cognitive interventions are designed to aid the process.

We can also observe that reasoning strategies may in turn bring with them differences in information seeking patterns. The choice of information or tests within the diagnostic has been understudied to date given its role in real-world clinical settings. Our results in Study 1 indicate that information seeking patterns are associated with accuracy, specifically around greater selectivity and less variability in information seeking. This relates well to real-world clinical decisions where information and tests can require sizeable time and even other staff or technology to request, as mentioned earlier. Hence, in a setting where all possible information is not always readily available to clinicians, being selective is advantageous. In addition, being more standardised in information seeking can also make comparisons between patients easier given that the information being compared is more similar. We can assume that a big part of gaining medical experience is by using past patient cases personally experienced by the clinician and then drawing upon that experience for a new incoming patient. This may explain the findings of Coderre et al. (2003) that pattern recognition was utilised more with experience: because experienced clinicians have more past patients to draw patterns from.

Study Strengths

We note a number of strengths of both of these studies. To our knowledge, this is the first research of this kind to use both a mixed-methods approach to understand the cognitive underpinnings of diagnosis as a decisional process. Our paradigm also emulates diagnosis in a manner that is not simply a single decision. When creating a task that emulates diagnosis, we in a sense conceptualise what diagnosis looks like in a fairly static manner, when really diagnosis is a more fluid and nebulous process in medicine. In our study, diagnosis is modelled as a process that develops over time with more information and is constantly shifting doctors' thought processes. In particular, we note linking diagnostic accuracy, confidence, information seeking and differential generation has not been attempted in prior work and should be considered for future work to consider a more complete study of diagnostic decisions. Future work could build upon on this work to investigate more flexible and open information seeking by emulating naturalistic decision making processes. This can be used to investigate how contextual limitations on information seeking impact confidence and accuracy (such as time pressures or testing being unavailable). The use of a think-aloud paradigm also brings a number of strengths by allowing for an ongoing recording of medical students' thought processes, again showing the dynamic, evolving nature of the diagnoses. While participants varied in terms of their ability to verbalise their thoughts, it provides a clear access to how they approaching their decisions that would otherwise be difficult to determine in real-time. We therefore encourage future researchers to consider think-aloud methodologies in their work.

Limitations

We note that the use of a think-aloud methodology brings with it a couple of limitations. Firstly, participants may behave differently to how they otherwise would, given that they are being observed and recorded by a researcher. Hence, there may be a tendency toward medical students behaving in a manner that

they believe to be judged as better by others, such as being thorough in their information seeking and differential evaluation. Relatedly, we found that medical students naturally differed from one another in terms of the amount of verbalising they did during the task, which could be related to differences in verbalisation skills (van Somersen, Barnard & Sandberg, 1994). By not explicitly asking students for their diagnostic differentials as we did in Study 1 (and minimising the amount of input that the researcher had during the task), we are constrained to analysing only what students say out loud. Given that some students do not verbalise their thoughts as naturally, we may not aware of the aspects of their thought process that they did not verbalise. For example, participants may not explicitly say out loud that a differential is no longer under consideration when in actual fact it has been dropped from their thought process, leading to a lower number of removed differentials as we observed in the data. Similarly, participants may have multiple differentials in mind but some may be subconsciously considered too unlikely to be even worth mentioning. This would then contribute to fewer overt instances of a hypotethico-deductive reasoning process as differentials are underreported. Future work should utilise more structured methods for eliciting clinicians' thought process during diagnoses in order to ensure accurate reporting of differentials in a more naturalistic, evolving manner. We also could have recruited a larger sample in order to gain a better range of participants and reasoning strategies, increasing the power of our analyses, as well as participants from differing experience levels.

Implications for Medical Practice and Education

There is real value in teaching metacognition and uncertainty within medical education (Royce, Hayes, & Schwartzstein, 2019), such as with the use of cognitive aids (Chew, Durning & Van Merriënboer, 2016, Ely, Graber & Croskerry, 2011), especially given that doctors can be reticent to express their uncertainty (Katz, 1984). A more structured aid is needed, as simply looking at a case for a second time may not be sufficient to improve diagnostic accuracy (Monteiro et al., 2015) and current cognitive forcing strategies have not been found to be effective enough

(Sherbino et al., 2014). The reason for this might that past distinction between System 1 (automatic, quick) and System 2 (deliberate, slow) thinking for prompting diagnostic reasoning may be overly simplistic, in that one solution may not fit all possible cases and all clinicians. Future work should focus on understanding when certain reasoning styles and which cognitive aids may be more useful for a given clinical situation. In particular, our findings indicate that cognitive aids should prompt reasoning strategies based on how clear a patient presentation is. This can be thought of as whether the patient's initial symptoms suggest a narrow or wide range of differentials, which is likely a combination of the clinician's knowledge and case's complexity.

Past work on cognitive interventions have not tended to focus on prompting appropriate information seeking, and we show here that different facets of information seeking contribute uniquely to both confidence and accuracy. While the most relevant information that should be afforded to clinicians will differ depending on the medical discipline, interventions can focus on standardising which is the most valuable information to be presented to clinicians in the first place. This could not only improve diagnostic accuracy but ensure more appropriate expressions of confidence and uncertainty by reducing a tendency toward overtesting. We emphasise that such recommendations are highly dependent and variable depending on the specific medical context, but this acts an important facet for medical education to consider around how seeking information relates to reasoning styles and how important these non-technical skills are to integrate into the educational context of medicine.

Study 4 - Diagnostic Uncertainty and Information Seeking in Virtual Reality Paediatric Scenarios

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

We used VR scenarios implemented by Oxford Medical Simulation (OMS), a company that uses VR for medical education and simulation, in their bespoke software. Participants in this study were medical students based in Oxford who were at the time taking part in VR-based teaching sessions as part of their medical degrees. Students performed the scenarios using Oculus Quest 2 VR headsets. Scenarios were based in paediatrics, meaning that the patients in the scenario were children who were attending the hospital with their legal guardian. Each scenario features a visual 3D implementation of a basic ward room in a hospital. Participants are shown a (child) patient, their guardian and a nurse who can help with certain treatment and testing. All of the 'avatars' in the scenario can be questioned by the participant using a predefined set of requests/actions (e.g. asking

the nurse to check blood pressure, asking the patient/child about if they are in pain). The scenarios have full sound (e.g. being able to hear the patient's lung auscultation) and the avatars are voiced.

The aim for students was to diagnose the patient, begin treatment and hand over the case to a senior with appropriate understanding of the patient (handovers were conducted using a standardised framework known as SBAR, meaning that clinicians have to brief the senior on the Situation, Background, Assessment and Recommendation for the patient). Whilst in the scenario, participants can learn about the patient's medical history, check key parameters (such as temperature, pulse, blood pressure, respiratory rate etc), perform physical exams/tests and begin certain treatment actions (such as administering oxygen or prescribing medication). Compared to the previous studies, participants have a much wider array in terms of the information and tests they can request, as well as being able to begin a treatment plan.

After 5 minutes in the scenario (by which point it is expected that participants would have a history of the patient and have started some early assessment of the patient), participants are asked to pause the scenario (taking off their VR headset) and fill in a brief questionnaire on paper. Multiple VR participants were performing the scenario simultaneously and were paired with another student who would watch their performance. This other student would aid with administering the questionnaire, with the student subsequently switching roles for the other scenario. The VR participant was asked in the questionnaire to answer the follow (this is considered time point 1):

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

- "How severe do you think the patient's condition is on a scale of 1 to 10?" (Each point of the scale represented a different clinical action/course, with 1 representing "Discharge in <4 hours, no follow up" and 10 representing "Requires arrest/peri arrest team.")

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

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

Each participant completed two scenarios over two separate VR sessions. The sessions were held around one month apart. During each session, the participants each performed one scenario in VR and observed their partner during their scenario. Participants also engaged in peer-to-peer feedback discussions as part of their education. The scenarios presented in each sessions are described below (students are split into two groups, shown below as groups A and B, each performing a different pair of scenarios in a fixed order):

Session One: Group A: patient/child is a 6-year-old-girl presenting with a 1 day history of central abdominal pain and thirst. She was generally unwell for 2 days prior, with reduced appetite and a sore throat. Collateral history reveals Type 1 Diabetes and erratic blood sugars. (True Condition: Diabetic Ketoacidosis) Group

B: patient/child is a 5-year-old boy presenting with worsening shortness of breath, wheeze, and signs of respiratory distress, on the background of 2 days of likely viral illness. He has a medical history of asthma and has had similar exacerbations in the past. (True Condition: Acute Severe Exacerbation of Asthma)

Session Two: Group A: patient/child is a 5-year-old boy presenting with shortness of breath and drowsiness (True Condition: Chest Sepsis/Pneumonia) Group B: patient/child is a 5-year-old girl with a 1 day history of sore throat and fever. She starts having a generalised tonic clonic seizure during the scenario. (True Condition: Febrile seizure on background of tonsillitis)

The dependent variables that we derive are as follows:

Performance: OMS implements a series of objectives for each scenario, which are tasks or actions that the participant is expected to have completed within the allotted time. This can include administering oxygen, prescribing a particular medication or calculating the Patient Early Warning Score (PEWS). The proportion of completed objectives is used as a score of the participant's performance during the scenario. Confidence Change: the participants' confidence in their understanding of the patient's condition is recorded at two time points, with the first being after 5 minutes (out of the 20 minute time limit) and the second being after the participant has finished the scenario. Confidence at each stage is recorded on a 10 point scale (1-10). The difference between the second and the first confidence rating is taken, such that a positive value indicates that the participant has increased their confidence over the course of the scenario. Number of Differentials: participants are asked to record all the diagnostic differentials that they are considering at the two aforementioned time points. Hence, the total number of differentials is recorded at each stage. Initial Diagnostic Breadth: this is the number of diagnostic differentials reported by the participant at the pause point. Diagnostic Appropriateness: each participant's set of differentials are assessed for how appropriate they are for the scenario. Each scenario has a set of differentials that are considered most likely, probable and improbable (with any others considered incorrect). To calculate a score for how appropriate the diagnoses are, we sum the likelihood values provided

for all differentials that were marked as most likely or probable. We then add these to the sum of likelihood values for improbable differentials divided by two. This sum is divided by the total sum of all differentials. This overall measure then measures what proportion of the participants' likelihoods are dedicated to probable differentials. However, we also penalised participants for providing few differentials, such that high scoring sets of differentials are larger sets of likely or probably differentials.

We also derived measures of information seeking similar to previous studies. The VR scenarios are far richer in terms of the available set of information for participants when compared to the vignette paradigm. For our analysis, we record all actions (or 'clicks) made by participants whilst in the scenario. Actions are categorised into a number of groups. The main categories are labelled as History, Examination or Testing, similar to in the vignette study. This set of information is mostly similar across scenarios though there are minor differences especially in the History category. Across scenarios, there are 35 possible History actions, 29 Examination actions and 18 Testing actions. This especially means that in comparison to the vignette paradigm, participants can take more detailed patient histories and can receive very different pieces of information depending on what they request from patient documentation and from asking the patient/guardian in the scenario. Outside of these categories, there are other actions available to participants, such as administering medication for the patient, calling for help or providing reassurance to the patient/guardian, but these are not used for our analysis. After categorising the participants' actions, we define a number information seeking measures:

History Taking: this is the number of History actions for a given scenario that take place before the pause point. Total Information Seeking: this is the number of actions classified under History, Examination and Testing across the scenario. Information Value: to calculate the value of each information sought across these categories, we calculate the difference in OMS performance score for participants with or without that information. We then sum all sought information

values for each participant within each of the information categories (History, Examination, Testing).

Reflective Based on Observations in Intensive Care

ICU Reflective

Presented here is a reflective chapter that contextualises the findings from this thesis within a real-world medical setting, namely that of Intensive Care. The account presented here is based on observations during multiple handovers and ward rounds at an intensive care unit, as well as discussions with staff working at the unit.

Firstly, some context is required for an Intensive Care Unit (ICU) as a medical setting. ICU is first and foremost a support unit that is relatively agnostic with regards to medical subdisciplines. The primary aim of the unit is to provide ongoing care for acutely unwell patients in a supportive capacity rather than a remedial one. Hence, clinicians and nurses in ICU are limited in what they can do for patients in their care. ICU can be hugely beneficial for patients by providing urgent care for patients in hopes of aiding their road to recovery. Patients then tend to move elsewhere in the hospital, such as the main ward or to theatre for surgical intervention. As mentioned earlier, ICU sits outside of other medical subdisciplines. It is hence very frequent that individuals working in ICU are required to bring in external advice from other departments in the hospital, such as Rheumatology, Neurology, Surgery, Vascular or Trauma. ICU can hence act as a central coordinator of several decision makers who are involved with a particular patient's care whilst clinicians within ICU itself will not be able to do too much without the involvement of these other departments whilst still having primary responsibility for that patient whilst they are in the unit. As one clinician put it, "someone who has trauma is longer Trauma's responsibility." In brief, ICU is usually a point of transition for patients within their medical pathway through the

hospital, with other departments feeding into and being fed from ICU. But ICU can also be the last point in their patient journey (either positively or negatively).

ICU is then a department that involves many individuals, both from within and outside its remit. A key tenet is then quickly and temporarily formed teams that have to collaborate on a patient and align their mental models. It is very common for teams of individuals to work together despite having little to no prior experience with each other.

Perhaps the most focal decisions that consultants within ICU have to do with monitoring ICU capacity in the present and in the future. Every ICU unit has a limited capacity in terms of the number of beds available and hence the number of patients who can be cared for at any given time (this was 22 beds for the unit observed). A patient is able to leave ICU and hence free up a bed if they either improve enough to transition to another in the hospital or if they unfortunately die in ICU. However, because ICU is merely a support unit, patients can also find themselves in a longer period of little change where they neither improve or deteriorate significantly. As a result, patients can sometimes stay in ICU for weeks or even months on end. Clinicians and nurses in ICU have to balance what they can realistically do for a patient within their remit whilst being cognizant of the longer term outcome of the patient. This is best summed up by one clinician who said during observations: "there is balance of what we can do and what is kind (to the patient)."

Making decisions about the current and future capacity of ICU is hence extremely complex, as it involves an understanding of each patient's condition not only in the current moment but in the future. Essentially, how likely is the patient to improve or deteriorate? There is a projection of future state that occurs. This occurs at the individual patient level, where clinicians imagine how well/unwell a patient will be in the short or long term future. This involves looking at the trend of treatment and what the upcoming milestone/endpoint for that patient might be. This can include simply getting the patient to eat solid food again or get up from their bed, or it could be tied to specific patient parameters (e.g. raise blood sugar to

above 4). This projection also occurs at the unit level though, as the combination of each patient's situation produces an overall picture of the unit's available capacity to admit new patients. Finally, the projection can also take place over the entire trust/region. During observations, the start of a morning shift began with the announcement that there was 'no capacity across the trust', meaning any incoming requests from other departments to admit patients to ICU would have to be refused.

These issues to do with capacity are of course related to actions of those in ICU but are also inexorably linked to wider environmental factors. This includes funding for increased ICU capacity and staffing as well as structural or technological issues within the hospital and region/trust as a whole. During one of our observation sessions, the unit was understaffed relative to the required number of staff needed to manage the unit. While these observations took place in the UK within the wider context of the UK's National Health Service (NHS), environmental factors will look very different in other countries, especially those less economically developed. There are even aspects of Human Factors at play. In our observations, the ICU unit was split over two floors that each had their own consultant to manage them, which would likely be different to if all beds were on a single floor. These other environmental factors are outside of the scope of this thesis and will be briefly revisited at the end of this chapter.

Part of ICU's coordination with other departments are incoming requests for the admitting of new patients. This could include a patient who has experienced a complication during surgery or a patient who has been admitted from an outside hospital in need of urgent care. Capacity is constantly at a premium and it becomes the forefront of an ICU consultant's thinking. Ideally, the unit should be able to operate with a spare buffer capacity of one or two beds in case of an emergency. This spare capacity can be fairly rare to obtain however, as it can be due to factors outside of the control of ICU clinicians.

Decision making here is hence extremely difficult and high-pressure. Decisions have to be made of when to admit patients and when patients are likely to be discharged. What underscores these decisions is the likelihood of a patient realistically

improving within ICU. There is only so much that ICU can do to help a patient who may be past the point of recovery. This demonstrates the aforementioned balance of what can be done and "what is kind." These kinds of decisions are difficult to make for everyone, be it the clinicians in ICU or the patient's next of kin. Being realistic about a patient's prospects is incredibly hard but is required in order to adequately manage the ICU's capacity in the future.

There can be diagnostic uncertainty for patients around what pathophysiologically may be driving their current set of symptoms. However, the real uncertainty stems not from the patient's condition now, but the patient's condition in the future, such as in 24 hours or 48 hours time. An ICU consultant may consider the following questions:

- How bad 'could' this patient's condition be relative to how unwell the patient is now?
- What realistic milestones/goals can we set for this patient's recovery plan?
- Is the patient 'wardable'? (i.e. is the patient well enough to be discharged from ICU and sent to the main hospital ward for continued care that is not as acute)

The state of a patient can change fairly quickly as a sudden development in their situation can occur over a single shift. This is why, at least in the unit that we observed, there is a regular communication cadence between individuals working in ICU. This comprises a morning handover, where the consultant during the night shift hands over to the morning shift consultant and reports patient developments that occurred during the night. This also comprises morning, afternoon and evening ward rounds, during which consultants visit each patient bed to receive updates on the patient by the caring nurses and (when possible) talk to the patient. During these ward rounds, the consultant will collaborate with the registrar, nurses and any individuals from other relevant departments to formally record an assessment of the patient and recommend a short term action plan to be taken for that patient to be coordinated with the attending nurses. This includes a formal assessment of

whether the patient is "clinically fit for Critical Care Discharge'' (wording taken from the computerised system used to record ward round documentation during observations). During observations, these ward rounds took several hours due to the amount of detail and attention afforded to each patient but this can vary depending on the consultant and the unit.

We shall now look at how the research questions within this thesis relate to the setting of ICU. On confidence, On information seeking, On differential evaluation, What is not covered,

Appendices

A

R Environment and Packages

```
# print("R version:")
# version$version.string
#
# print("Rstudio version:")
# rstudioversion <- rstudioapi::versionInfo()
# rstudioversion$version
#
# print("Citations for packages used:")
# get_pkgs_info(pkgs = required_packages, out.dir = getwd())
# pkgs <- scan_packages()
# get_citations(pkgs$pkg, out.dir = getwd(), include.RStudio =
↪  TRUE)
# cite_packages(pkgs = required_packages, output = "table",
↪  out.format = "Rmd", out.dir = getwd())
#
# required_packages %>%
#   map(citation) %>%
#   print(style = "text")
```

References