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

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

Methods

Participants

We recruited medical students within the UK who were in the year of study before their foundation programme. 85 medical students completed the study, including 32 males, 52 females and 1 participant who reported as non-binary. Their ages ranged between 22-34 years (M = 24.2). Participants were recruited between July 11th 2022 and April 6th 2023. The UK Medical Schools Council distributed the study to UK medical students using a mailing list. Participants were emailed with a study information sheet and a link to access the experiment, where they first provided consent via an anonymous online form. After doing so, the participant provided demographic information (age, gender and years of medical experience). The study was conducted online, with participants able to run the experiment in a browser on a desktop computer or laptop. Participants were not able to run the experiment on their phone or tablet. The experiment was coded using the

JSPsych Javascript plugin. The code is publicly available on Github: https://github.com/raj925/DiagnosisParadigm. This study was reviewed and granted ethical approval by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE001. Informed consent was obtained anonymously using an online electronic information sheet and consent form.

Materials

This study involved patient vignettes that have been adapted from actual past cases. We adapted scenarios from a bank of patient cases from Friedman (2004). Our study involved 6 patient cases, each designed to indicate a specific underlying condition that the real patient had. These conditions were: Aortic Dissection (AD), Guillain-Barre Syndrome (GBS), Miliary TB (MTB), Temporal Arteritis (TA), Thrombotic Thrombocytopenic Purpura (TTP) and Ulcerative Colitis (UC). The order in which the cases were presented was randomised for each participant. We also included a practice case (Colon Cancer) to familiarise the participants with the experimental procedure and the interface.

A panel of 3 subject matter experts (practising medical staff and researchers within the NHS) who work in collaboration with the OxSTaR Centre (Oxford Simulation, teaching and Research Centre, based at Oxford's John Radcliffe Hospital) were recruited to design the vignettes used in this study. These medical professionals were at differing experience levels, with their medical roles at the time of this study as follows: Speciality trainee (ST6) in Anaesthetics, Foundation (F1) Doctor and Gastroenterology Consultant. The panel assisted with translating terms (e.g., medication names, tests etc.) from US to UK doctors' vernacular and updated patient details to be more current. The panel also crucially gave input on which medical conditions to choose from the provided bank of vignettes that medical students would be expected to know, whilst having some variance in both difficulty and the affected pathophysiological system.

Procedure

The goal of the task was to determine a diagnosis, or set of diagnoses, for each presented patient (Figure 1). Information on the patient is split into a series of discrete stages so that the researchers are able to control what information the clinicians have access to at any given point in the experiment. We can call each point of new information an "information stage". Participants are able to seek information freely until they are ready to move on, similar to the paradigm adopted by Kämmer et al. (2019).

The procedure of a single case is as follows. The participant is asked to imagine that they are working in a busy district hospital and they encounter patients in a similar way to how they would in their real medical practice. At the start of each case, the participant is shown a description of a patient, which includes the patient's gender, age and their presenting complaint. An example of this is: "patient is a 68 year old male presenting with fever and arthralgia". Each case is split into three information stages: Patient History, Physical Examination and Testing (in this order). The set of information requests for each stage is the same for all cases. The Patient History stage includes information on "Allergies", "History of the Presenting Complaint", "Past Medical History" and "Family History". The Physical Examination stage includes 'actions' that a doctor may take when examining a patient, such as "auscultate the lungs", "abdomen examination", "take pulse" and "measure temperature". Finally, the Testing stage involves information on any bedside tests or tests they may request from another department. This includes "Chest X-Ray", "Venous Blood Gas", "Urine Dipstick" and "Clotting Test". In total, 29 possible tests that can be requested across the three information stages.

When a participant clicks on any of these tests, the information for that test is shown on screen after a 3 second delay. It was emphasised during the task instructions that participants should only request information that they believe will help them with diagnosing the patient for that specific case. Participants are free to request the same piece of information multiple times, including information

from a previous stage. At any point, they can choose to stop gathering information for that stage. They are then taken to a new screen where they report a list of all differential diagnoses that they are considering for that patient at that stage. For each differential, participants report a "level of concern" for that differential, which is how concerned they 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). In subsequent stages, the list from the previous stages is available for participants to update concern/likelihood ratings, or to add/remove differentials from the list. Even at the last information stage, participants can report multiple differentials.

After recording their differentials, participants are then asked to report their confidence that they are "ready to start treating the patient" on a 100 point scale, ranging from fully unconfident to fully confident. Participants also indicate using a checkbox whether they are ready to start treating the patient, at which point 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' conditions for all the patients.

Data Analysis

Responses were coded for correctness manually with help from a medical consultant, who looked at all the information available for each case and determined which diagnoses could be valid answers. All lists of differentials were 'marked' for correctness manually using the criteria found in Table S1 of the Supplemental Materials.

We test for correlations between our dependent variables using Pearson's product moment correlation tests, and interpret an alpha value of less than 0.05 as indicative of a statistically significant effect. Our sample of 85 participants is

calculated have 80.4% power to detect a medium effect size of r = 0.3 (using an approximate arctangh transformation correlation power calculation). Our key dependent variables are as follows:

Case-Wise Measures

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

- *Confidence*: Participants 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). As for accuracy, confidence is rescaled to fall between 0 and 1 to allow for direct comparison between the two variables. We can then use these two variables to calculate Confidence Change, by subtracting the participants' Initial Confidence from their Final Confidence. Hence, a positive value for Confidence Change means that the participant has gained confidence over the course of the patient case.

- *Number of Differentials*: We record the number of items in the list of differentials at each stage. Initial Differentials refer to the number of differentials

after the first stage of information seeking (Patient History), whilst Final Differentials refer to the number of differentials after the third and last stage of information seeking (Testing).

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

Derived Information Seeking Measures Across Cases

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

- *Information Seeking Value*: We compute the average value of sought information across cases. To do this, we take each of the 29 pieces of information in turn by case and split trials into two groups: trials of that case where that information was sought and trials of that case where that information was not sought. For each group, we compute the proportion of trials where the students included a correct differential, and then take the difference between these two values. A positive value would indicate that students were more likely to identify the correct condition with that information rather than without that information. This difference can be considered that information's 'value'. For each of the participants' cases, we compute this difference for each piece of information that the participant sought (for information they did not seek, the informational 'value' would be 0) and then calculate the sum of all information values for each case. We then take the mean information value across all cases for each participant. This gives an overall measure of, on average, how useful the information was that participants sought on each case.

Results

0.0.1 Overall Performance

Across cases, the proportion of trials where participants include a correct differential within their set of differentials increased with each stage of information gathering. ($F(2, 128) = 59.52$, $^2G = .08$, $p < .001$). Participants included a correct differential on fewer trials during the Patient History stage ($M = 0.54$, $SD = 0.23$) than during the Physical Examination ($M = 0.66$, $SD = 0.22$) and Testing stages ($M = 0.69$, $SD = 0.21$). Table 2 shows overall performance by case, indicating that there was variability in performance due to cases varying in difficulty.

0.0.2 Calibration of Confidence to Accuracy

Confidence also increased as participants received more information ($F(1, 123) = 75.45$, $^2G = .15$, $p < .001$). Participants reported lower confidence during the Patient History stage ($M = 0.30$, $SD = 0.15$) than during the Physical Examination ($M = 0.41$, $SD = 0.17$) and Testing stages ($M = 0.47$, $SD = 0.19$). We note here that confidence was on average below 50% even at the end of each case, which may indicate that participants were not highly confident to start treatment.

```
caseBreakdown <- studentCaseDf %>%
  group_by(caseCode) %>%
  dplyr::summarise(`Proportion of Participants who Included a
↪  Correct Differential` = mean(correct),
                   Accuracy = mean(likelihoodOfCorrectDiagnosis)/10,
                   `Perceived Difficulty` =
                    ↪  mean(subjectiveDifficulty,na.rm=T),
                   `Mean Final Confidence` =
                    ↪  mean(finalConfidence)/100)

caseBreakdown
```

```
## # A tibble: 6 x 5
##   caseCode Proportion of Participants who Incl~1 Accuracy `Perceived Difficulty`
##   <chr>                                   <dbl>    <dbl>                   <dbl>
```

```
## 1 AD                                    0.605    0.278                5.93
## 2 GBS                                   0.756    0.407                6.87
## 3 MTB                                   0.419    0.237                6.64
## 4 TA                                    0.744    0.492                6.14
## 5 TTP                                   0.616    0.341                6.81
## 6 UC                                    0.988    0.724                5.28
## # i abbreviated name:
## #   1: `Proportion of Participants who Included a Correct Differential`
## # i 1 more variable: `Mean Final Confidence` <dbl>
```

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

When comparing Accuracy (taking into the likelihood assigned to correct differentials) to Confidence, we find, across stages, participants' Confidence was fairly well aligned to their Accuracy (see Figure 1). To determine whether confident participants tended to be more accurate, we compared a paired t-test between Average Confidence and Average Accuracy (across cases) at each stage. We did not evidence of a difference between the two at the Patient History ($t(84) = 0.32$, MDiff = 0.01, $p = .75$) and Physical Examination stages ($t(84) = 0.75$, MDiff = 0.01, $p = .45$), but we do see evidence of a difference between the two at the Testing stage ($t(84) = 2.40$, MDiff = 0.06, $p = .02$). This indicated well-calibrated confidence after Patient History and Physical Examination, but a slight overconfidence across participants after Testing.

```
nPpts <- nrow(studentAggData)
rootN <- sqrt(nPpts)

xb <- c("Patient History","Physical Exams", "Testing")
yb <- c(mean(studentAggData$meanInitialConfidence)/100,
 ↪  mean(studentAggData$meanMiddleConfidence)/100,
 ↪  mean(studentAggData$meanFinalConfidence)/100)
zb <- c(mean(studentAggData$meanInitialAccuracy),
 ↪  mean(studentAggData$meanMiddleAccuracy),
 ↪  mean(studentAggData$meanFinalAccuracy))
```

```
val <- c(yb,zb)
typ <- c(rep("Confidence",3),rep("Accuracy",3))

secon <- c(sd(studentAggData$meanInitialConfidence/100)/rootN,
↪  sd(studentAggData$meanMiddleConfidence/100)/rootN,
↪  sd(studentAggData$meanFinalConfidence/100)/rootN)
selik <- c(sd(studentAggData$meanInitialAccuracy)/rootN,
↪  sd(studentAggData$meanMiddleAccuracy)/rootN,
↪  sd(studentAggData$meanFinalAccuracy)/rootN)

ses <- c(secon,selik)

dataV <- data.frame("Stage" = xb, "Value"= val, "Type"= typ, "se"
↪  = ses)

p <- ggplot(dataV, aes(x = Stage, y = Value, group = Type, color =
↪  Type )) +
  geom_line() +
  geom_point() +
  geom_errorbar(aes(ymin=Value-se, ymax=Value+se), width=.2,
   ↪  position=position_dodge(0.05)) +
  labs(title="    ",x="Stage",y="% Value") +
  theme_classic() +
  scale_color_manual(values = c(accuracyColour,confidenceColour)) +
  theme(axis.text=element_text(size=16),
                              axis.title=element_text(size=18),
                              plot.title=element_text(size=20,face=
                               ↪  "bold"),
                              legend.text = element_text(size = 18),
                              line = element_blank())

print(p)
```

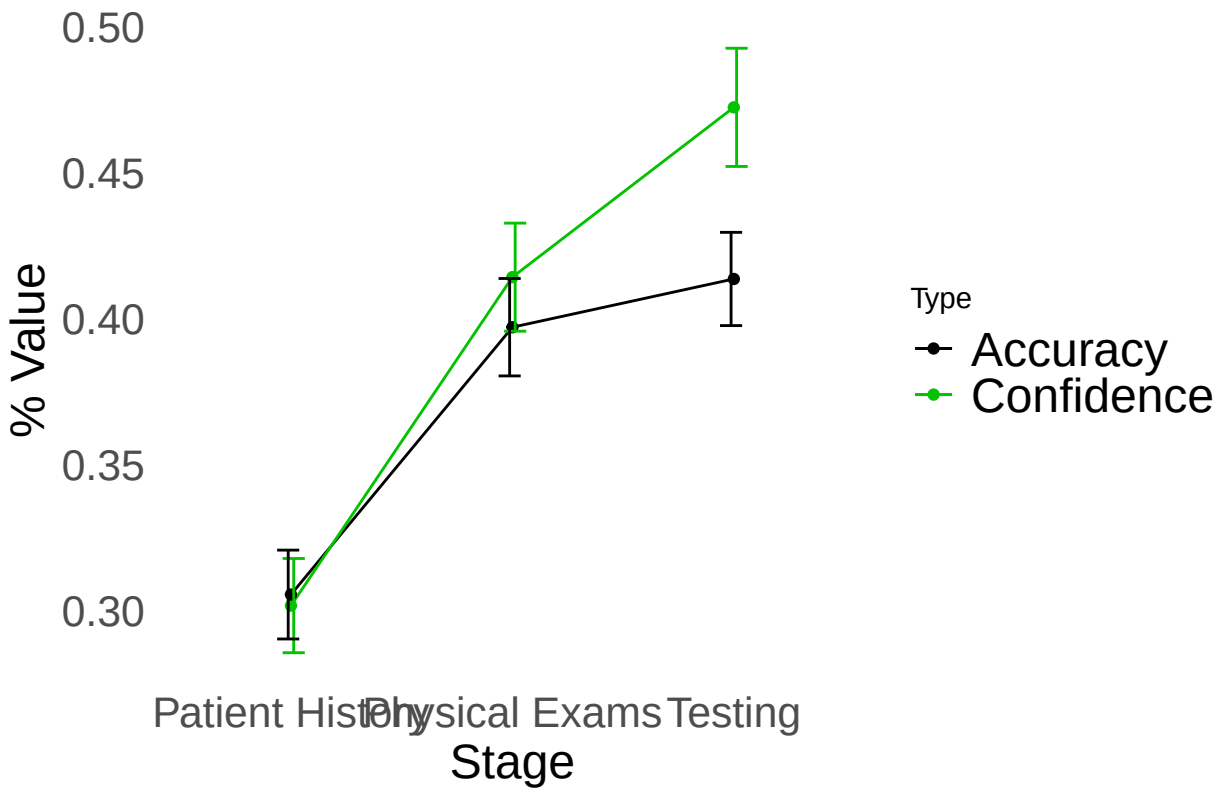


Figure 1: Graph showing Accuracy (black) and Confidence (green) at each of the three information stages.

0.0.3 Differentials

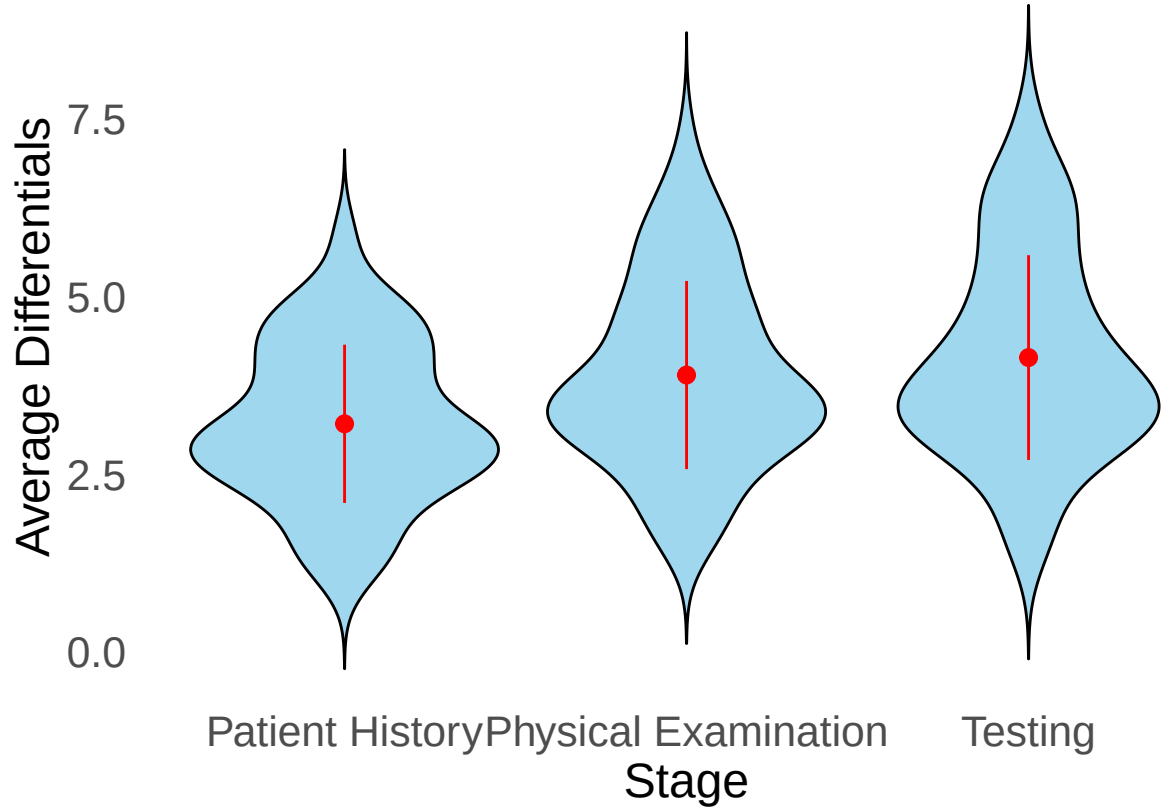
We first look at the number of differentials that participants report at each stage. Participants overall increased the number of the differentials they reported as they received more information ($F(1, 107) = 94.02$, $^2G = .08$, $p < .001$). Participants reported fewer differentials during the Patient History stage ($M = 3.20$, $SD = 1.11$) than during the Physical Examination ($M = 3.88$, $SD = 1.33$) and Testing stages ($M = 4.12$, $SD = 1.43$). The majority (74/85) did not decrease the number of differentials between Patient History and Testing on any case.

```
nPpts <- nrow(studentAggData)

xb <- c(rep("Patient History",nPpts),rep("Physical
 ↪  Examination",nPpts), rep("Testing",nPpts))
yb <- c(studentAggData$meanInitialDiffs,
 ↪  studentAggData$meanMiddleDiffs, studentAggData$meanFinalDiffs)
```

```
dataV <- data.frame("Stage" = xb, "Mean"= yb)
dataV$Stage <- as.factor(dataV$Stage)
diffs <- ggplot(dataV, aes(x=Stage, y=Mean)) +
  geom_violin(colour="black", fill=differentialColour, alpha=0.8,
  ↪  trim=FALSE) +
  # geom_dotplot(binaxis='y', stackdir='center',
  ↪  dotsize=0.5,colour="white") +
  stat_summary(fun.data=data_summary, colour="red")

print(diffs +
        labs(x = "Stage", y = "Average Differentials") +
        theme_classic() +
        theme(axis.text=element_text(size=16),
              axis.title=element_text(size=18),
              plot.title=element_text(size=18,face="bold"),
             line = element_blank()
        )
      )
```




```
###dataV$Stage <- as.numeric(dataV$Stage)
###model <- lm(Mean ~ Stage, data=dataV)
###print(summary(model))
```

Online Study

Figure 2: The average number of differentials after each stage of information seeking.

To look at whether the number of initial differentials generated the amount of information sought, we conducted a Pearson's Correlation test on individual differences. We find an association (see Figure 3) between the average number of differentials generated from the Patient History and the average amount of information sought during cases ($r(83) = 0.30$, 95% CI = [.10, .49], $p = .005$). As previously discussed, participants rarely seem to remove differentials from consideration. Therefore, one can surmise here that higher information seeking is associated with the consideration of more diagnostic differentials.

```
### Correlation between initial differentials and overall
↪  confidence change

cor <- cor.test(studentAggData$meanInitialDiffs,studentAggData$
↪  proportionOfInfo,method="pearson")

diffCon <- ggplot(data = studentAggData, aes(x=meanInitialDiffs,
↪  y=proportionOfInfo)) +
  geom_point() +
  geom_smooth(method=lm , color=confidenceColour, fill="#69b3a2",
  ↪  se=TRUE) +
  theme_classic()

print(diffCon +
      labs(y="Proportion of Possible Information Requested", x =
      ↪  "Number of Initial Differentials") +
      theme(axis.text=element_text(size=16),
            axis.title=element_text(size=16),
            plot.title=element_text(size=14,face="bold")
      ))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

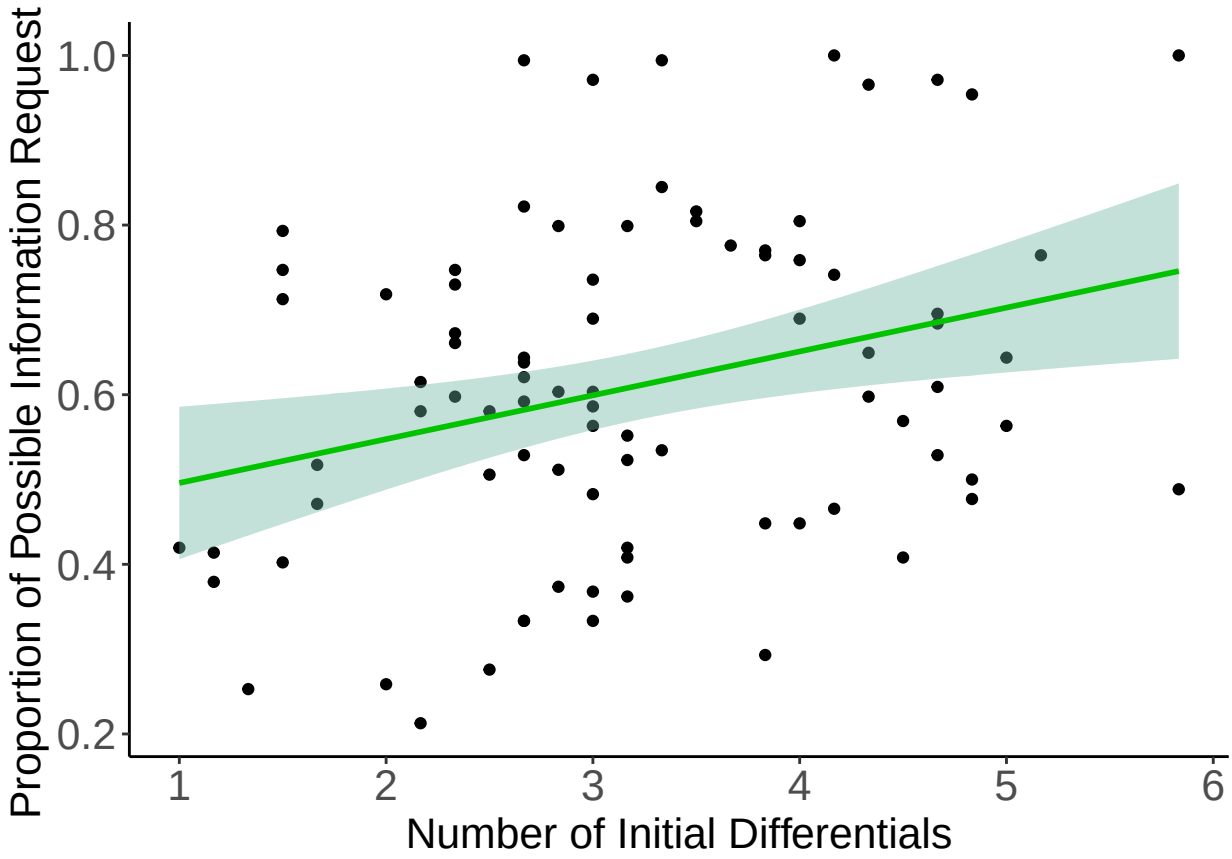


Figure 3: Scatter plot showing the relationship between the number of initial differentials (x-axis) and the proportion of available information sought (y-axis). Each point represents a single student with both variables average across the six cases that each student performs. The x-axis refers to the average number of differentials that participants report in their list at the Patient History stage. The y-axis refers to the average proportion of available information sought, with each case containing 29 pieces of information across the Patient History, Physical Examination and Testing stages. The line of best fit is plotted using the geom_smooth function in R with a linear model. The shaded region shows the 95% confidence interval of the correlation.

0.0.4 Information Seeking

The Proportion of Information Seeking decreased with each information stage ($F(2, 151) = 122.0$, $^2G = .30$, $p < .001$). Participants sought more of the available information during the Patient History stage ($M = 0.85$, $SD = 0.20$) than during

both during the Physical Examination (M = 0.59, SD = 0.24) and Testing stages (M = 0.50, SD = 0.22).

When conducting a Pearson's Product Correlation test, we do not find that participants who sought more information across cases were also more accurate in their diagnoses (r(83) = 0.17, 95% CI = [-.04, .37], p = .11). However, participants who sought more information did tend to increase their confidence more (r(83) = 0.24, 95% CI = [.02, .43], p = .03). This is distinct from their final confidence, for which we did not find evidence of an association with the amount of information sought (r(83) = 0.11, 95% CI = [-.11, .31], p = .33). While seeking more information may imbue students with a greater level of confidence, it does not necessarily translate into more accurate diagnoses. This links to the results presented in Figure 1, in which confidence and accuracy were related to one another but imperfectly (especially during the Testing stage).

```
### Correlation between initial differentials and overall
↪  confidence change

cor <- cor.test(studentAggData$meanInitialDiffs,studentAggData$
↪  meanConfidenceOverallChange,method="pearson")

diffCon <- ggplot(data = studentAggData, aes(x=meanInitialDiffs,
↪  y=meanConfidenceOverallChange)) +
  geom_point() +
  geom_smooth(method=lm , color=confidenceColour, fill="#69b3a2",
  ↪  se=TRUE) +
  theme_classic()

print(diffCon +
      labs(y="Final Confidence - Initial Confidence", x = "Number of
      ↪  Initial Differentials") +
      theme(axis.text=element_text(size=16),
            axis.title=element_text(size=16),
            plot.title=element_text(size=14,face="bold")
      ))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

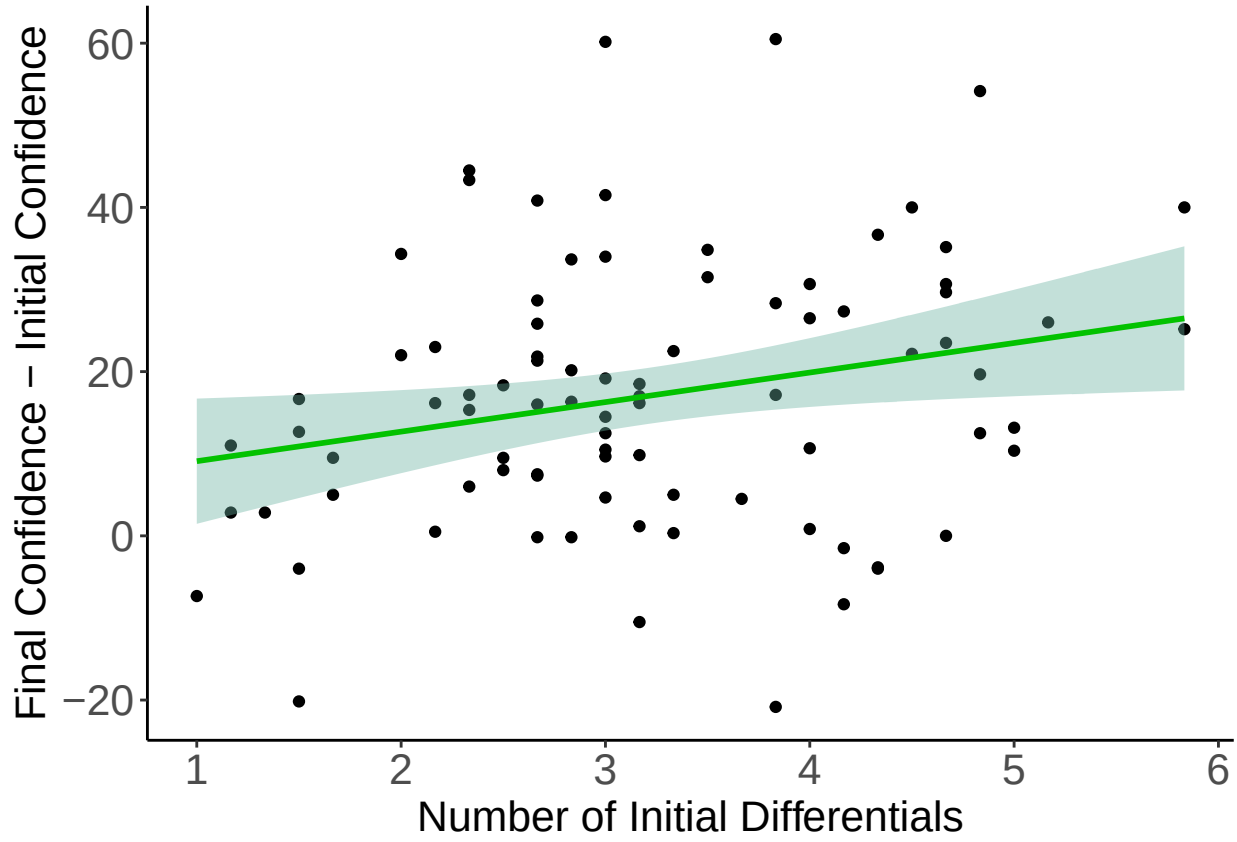


Figure 4: Scatter plot showing the relationship between the number of initial differentials (x-axis) and the proportion of available information sought (y-axis). Each point represents a single student with both variables average across the six cases that each student performs. The x-axis refers to the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

The amount of information sought does not seem to be predictive of accuracy. However, it may be the case that patterns of information sought are instead predictive of differences in accuracy on this task. In order to test this, we investigate whether information seeking is predictive of participants who are higher or lower in their diagnostic accuracy using binary classification and receiver operating characteristic (ROC) analysis. We trained a binary classification algorithm using

a generalised logistic regression model to identify if participants of high or low accuracy based on the information they sought. We split all cases by whether they performed by a high and low Accuracy participant using a median split of participants by their average Accuracy across the six cases. We train the classifier using a Generalised Linear Model (GLM) by treating the 29 binary variables for each information as predictors (with a 1 signifying that the information was sought for that case and 0 when the information was not sought) to predict the binary outcome of whether the participant is a low or high accuracy participant. We used Leave One Out Cross Validation, such that each case is predicted by training the algorithm on all other cases. By plotting an ROC curve of our classifier, we find an area under the curve (AUC) value of 0.72 (plotted in Figure 5). When conducting a DeLong test, to test the null hypothesis that the AUC is equal is 0.5 (i.e. that the classifier is completely unable to predict high and low accuracy participants), we find $p < .001$, indicating that the AUC differs significantly from 0.5 and that the classifier is able to reliably predict high and low accuracy participants.

This indicates overall that differences in information seeking are indeed predictive of a difference in participant ability at above chance at a broad level. Essentially, information seeking patterns separate high and low accuracy participants, but this analysis does not tell us what aspects of information seeking in particular are predictive of accuracy. We next seek to identify and better characterise these specific differences in information seeking that contribute to this relationship with diagnostic ability by correlating behavioural information seeking variables with accuracy.

```
set.seed(1000)

classifierData <- infoSeekingFullMatrix[,c(2:29,38)]
classifierData$AccuracyGroup <-
↪  as.integer(as.logical(classifierData$AccuracyGroup>2))
classifierData$AccuracyGroup <-
↪  as.factor(classifierData$AccuracyGroup)
colnames(classifierData)[1:29] <- c("T2",  "T3",  "T4",  "T5",
↪  "T6",  "T7",
                                    "T8",  "T9", "T10", "T11",
                                    ↪  "T12", "T13", "T14",
```

```
                                        "T15", "T16", "T17", "T18",
                                        ↪  "T19", "T20", "T21",
                                        ↪  "T22",
                                        "T23", "T24", "T25", "T26",
                                        ↪  "T27", "T28",
                                        ↪  "T29","Group")


thresh<-seq(0,1,0.001)
#specify the cross-validation method
ctrl <- trainControl(method = "LOOCV", savePredictions = TRUE)

# Shuffle rows in case there are order biases
classifierData <- classifierData[sample(1:nrow(classifierData)),]
modelglm<-train(Group ~ T2 + T3 + T4 + T5 + T6 + T7 + T8 + T9 + T10
↪  +
                T11 + T12 + T13 + T14 + T15 + T16 + T17 +  T18 + T19
↪  + T20 +
                T21 + T22 + T23 + T24 + T25 + T26 + T27 + T28 + T29,
↪  method = "glm", family = binomial(link=probit), data =
↪  classifierData, trControl = ctrl)
prediglm<-predict(modelglm,type = "prob")[2]

# Plot all test results on one ROC curve
rocPlot <- roc.plot(x=classifierData$Group=="1",pred=cbind
↪  (prediglm),legend =
↪  T,
                    leg.text = c("GLM"),thresholds = thresh)$roc.vol
```

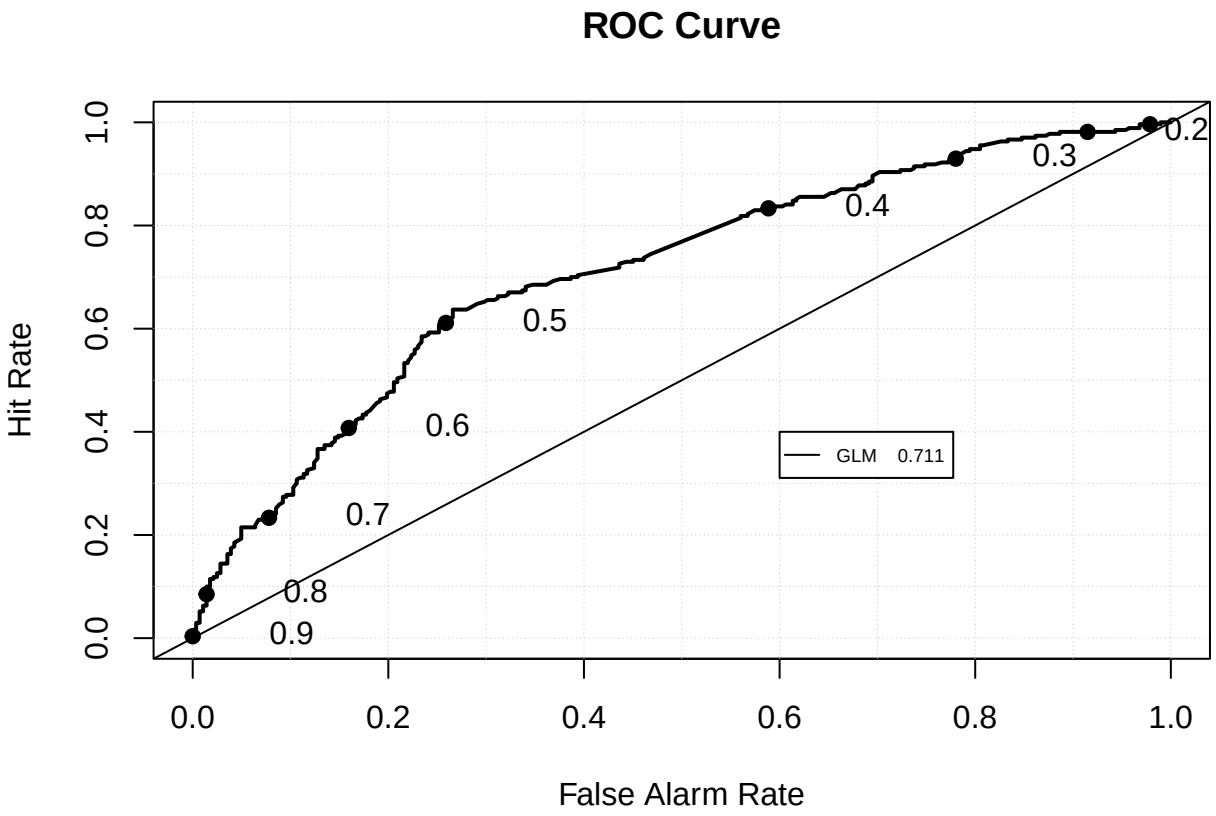


Figure 5: Receiver-Operator Characteristic (ROC) curve using a Generalised Linear Model to classify individual cases as being performed by either high or low accuracy participants. The models are trained on the raw binary predictor variables for each of the 29 available pieces of information, with 0 indicating that the information was not sought for the case and 1 indicating that the information was sought. Participants were sorted as high and low accuracy based on a median split on their average Accuracy value across the six cases.

In order to examine more specifically what differences in information seeking are driving differences in both accuracy and confidence, we look at the relationship between accuracy and informational value. We assess the degree to which each participant's accuracy is predicted by the quality of the information they sought and find evidence for a positive relationship between accuracy and information value ($r(83) = 0.25$, 95% CI = [.04, .44], $p = .02$), as well as between confidence and information value ($r(83) = 0.28$, 95% CI = [.07, .47], $p = .01$).

Discussion

This study of 85 medical students found that accuracy and confidence are well calibrated. Previous work (e.g. Meyer et al., 2011) have noted a gap between subjective confidence and objective accuracy. In particular, there has been demonstrated to be a general tendency for less experienced medical trainees to be underconfident and for more experienced medical professionals to be overconfident (Yang and Thompson, 2010). Part of this discrepancy between our findings and past findings could stem from the diagnostic uncertainty expressed by students in this study, which they do in two ways. Firstly, students broaden, rather than narrow, their considered diagnostic differentials with more information and still report a broad range of differentials after receiving all available information for a given case. There is a general adage in healthcare that medical students come across which says that "history is 80% of the diagnosis". It is therefore worth considering whether there is a specific facet of diagnostic decisions whereby students are taught not to disregard diagnostic possibilities easily. Secondly, students reported fairly low confidence overall to treat patients, with an average confidence of below 50% even after receiving all available information. This may indicate that part of ensuring appropriate confidence, or expressions of uncertainty could be related to properly evaluating all possible diagnostic differentials rather than forcing decisions to focus on a single diagnosis, which has been cited previously as a problematic tendency (Redelmeier & Shafir, 2023).

We find the amount of information sought informed confidence, whilst accuracy was associated with seeking more useful information on each case. This hints at the richness of this dataset in picking on information seeking and differential generation behaviour. We note however that whilst predictors of diagnostic by information seeking behaviour were found, they do not tell us how overarching differences in such behaviour arise. One possibility is that these differences stem from reasoning strategies that we cannot infer from this current dataset. In order to ascertain these strategies, we conduct a follow-up study using a similar diagnostic paradigm

conducted in-person where students think out loud as they make diagnoses. We use criteria taken from Coderre et al. (2003) to code case by the reasoning strategy employed. We hypothesise that different reasoning strategies for generating differentials are useful for some cases more than others and that information seeking varies as a function of strategy. This coding of reasoning strategies is then subsequently used to classify the same reasoning strategies in the online dataset from study 1 (where we do not have access to the participants' thought process) by using the information they sought.

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 Appendices.

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

- **Information Seeking Strategies**: any time the participant expresses why they may or may not request a particular piece of information in relation to ruling out or confirming a condition.

We also define a group of codes that indicate 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. 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, found in the Appendices).

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.

Think Aloud

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.

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.

Methods

Participants

Materials

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.

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)

Procedure

The aim for students in the scenarios 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?"

Data Analysis

The dependent variables that we derive are as follows:

- Performance: The OMS software 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 or 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).
- Amount of Treatment: this is the number of actions classified as treatment of the patient across the scenario.

As all actions are recorded with timestamps in the output dataset, we categorise whether actions occurred before or after the pause point (5 minutes in). Hence, we can investigate information seeking before and after the pause point where participants record their initial diagnoses and confidence.

Results

Overall Performance

Initial Diagnostic Breadth

Predictors of Confidence

Information Seeking

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

Vignette Marking Scheme (Studies 1 and 2)

B

Think Aloud Study - Debrief Questionnaire

C

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