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

Introduction

In the previous study, we presented results from an online vignette study where we investigated confidence and information seeking within evolving diagnostic decisions. Overall, we found that students provided confidence judgements that were well-calibrated with their objective accuracy. We found that higher information seeking was associated with larger increases in confidence over the course of diagnostic decisions. However, seeking more information was not associated with accuracy. Rather, accuracy was associated with selectivity in information seeking, such that certain information was beneficial to students regardless of the patient's condition. We also found that medical students tended to broaden the range of differentials they were considering with more information, as they rarely removed differentials from their consideration. However, students who considered more differentials early on tended to seek more information and increase their confidence to a greater extent. Given that we find that certain information seeking patterns were associated with accuracy, and that students were overall calibrated, we now aim to better understand how students approach the task in terms of their decision making strategies. As the previous study was conducted online, we are not able to ascertain why students sought certain information or why they considered certain differentials (as likely or unlikely). To address this, we present in this chapter a study using a modified version of the previous study's vignette methodology where students verbalise their thought process as they were performing their diagnoses.

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

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

Another focus of this study is to understand how differences in information seeking patterns arise. One possibility is that these differences stem from reasoning strategies adopted when making diagnoses. Clinical reasoning is a key skill that is taught, either explicitly or implicitly, within medical education. Medical decisions are frequently made in uncertainty, with clinical reasoning taught as a skill to navigate this uncertainty (with perhaps an intention to reduced the uncertainty perceived by clinicians). However, clinical reasoning has a broad remit and covers multiple different approaches to making medical decisions. Doctors may have different considerations when diagnosing a patient, and may draw on different approaches to making medical decisions accordingly. As we noted in the design of our previous study, doctors may have consider what conditions are likely for a patient and what conditions are too severe to not miss. A doctor's reasoning strategy may then reflect this dichotomy, such that the focus is either on determining what is most likely or being thorough such as to consider all possible diagnoses. Diagnostic decisions have traditionally been thought of as 'ideal' when using the hypothetico-deductive process (Kuipers & Kassirer, 1984), whereby hypotheses are initially formulated based on patient symptoms and established criteria for diagnoses. Further information is then gathered to test these hypotheses (Higgs et al., 2008) or eliminate others. One can think of this approach as akin to a 'process of elimination': by starting broad, clinicians then seek information to reduce the potential set of diagnoses to a more manageable set. This theory, that hypothetico-deductive processes are the gold standard for diagnostic decisions, was challenged by the results of Coderre et al. (2003). The authors found that reasoning strategies differed between novice and expert clinicians and that, crucially, a pattern recognition approach (rather than hypothethico-deductive) was associated with higher diagnostic accuracy. A pattern recognition approach would involve considering fewer diagnostic hypotheses and instead matching the symptoms to prototypical cases of a patient condition. Hypothetico-deductive reasoning represents

a more structured, systematic approach to make diagnoses, in which clinicians maintain a more open mind to different diagnostic possibilities, whilst pattern recognition is more directed and driven by intuition. The difference between these approaches can be related to past work that has connected medical decision making to the dual-system model of thought (Kahneman, 2011). System 1 thinking represents an automatic, intuitive mode for making decisions (akin to pattern recognition), whilst System 2 represents a more deliberative, rational mode for decision making (with hypothethico-deductive reasoning as an example).

In their paper, Coderre et al. (2003) asked novice and expert clinicians to provide diagnoses based on patient vignettes and asked them to verbalise their thought process during the diagnostic process. Using these verbalisations, the researchers categorised the clinician's reasoning strategy on each case. The researchers defined three reasoning strategies as follows (paraphrased from their paper):

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

These reasoning strategies may have been explicitly taught during a clinician's medical education or implicitly developed with experience. With these differing strategies, we can consider whether there are normatively 'better' strategies to use in certain clinical situations. As highlighted with the previous set of conflicting literature, there is currently not a consensus within medicine as to which strategy

is ideal for diagnostic accuracy. In their study, Coderre et al. found that novice clinicians tended to adopt a HD strategy more often, whilst experts tended to use a PR strategy more. In addition, using a PR strategy was associated with higher diagnostic accuracy, which in turn was used to explain why experts were more accurate than novices. In addition to explaining differences in accuracy, these strategies point to different approaches in how diagnostic differentials are generated and considered. A PR strategy, by definition, considers a clinician as seeking to identify the correct condition whilst SI and HD strategies reflect a more thorough, systematic consideration of all possible differentials. The work of Coderre et al. did not reveal whether these strategies result in differences in information seeking and confidence. We would expect that with a more thorough SI or HD reasoning strategy, participants would seek more information in order to consider a larger number of diagnostic possibilities. By contrast, we would expect that a PR strategy would result in selective (but less) information seeking in order to gather evidence in support of the single diagnosis being considered. By better understanding how different reasoning strategies manifest in differences to diagnostic decisions, we aim to inform future work within medical education and cognitive intervention design to consider when each strategy is appropriate and prompt clinicians accordingly to follow that strategy's decision process. This also can shift our understand of clinical reasoning as having a 'one size fits all' approach where one method for making decisions is 'ideal' for all possible medical decisions.

In order to pick apart these different reasoning strategies within our vignette-based paradigm, we adopted a think-aloud methodology similar to Coderre et al. (2003), whereby participants verbalise their thought process as they are doing the vignette-based diagnosis task. Think-aloud methodologies are useful for directly accessing ongoing thought processes during decisions (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, Bilalić, 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 at discrete time points, allowing us to investigate how diagnostic thinking develops with more information. We also are able to connect the verbalisations in our task to the exact information participants received to prompt that thought. As mentioned, a think-aloud methodology has previously been used to study the differences between novice and expert clinicians during diagnostic reasoning (Coderre et al., 2003). We build on the work of Coderre et al. here to further investigate how reasoning strategies contribute to differences in both accuracy and confidence, as well as understanding why certain cases result in differing strategies.

In order to bolster our findings from this study, we aim to use the reasoning strategies determined from this study to reanalyse the cases in Study 2. Whilst we record information seeking, confidence and differential/hypothesis generation behaviour during the previous study, we do not have an understanding of how participants are approaching each case from a reasoning strategy perspective. One way in which we can to some extent infer this is via the information seeking patterns that participants adopt. If our hypothesis is correct that reasoning strategies result in different patterns of information seeking, we should then we able to predict what strategy a participant is using solely from their information seeking. If we can also establish that these predictions are reliable, we can then study the properties of these reasoning strategies with the larger sample size afforded to us in Study 2. This work would then improve our understanding of how these reasoning strategies

not only affect diagnostic accuracy, but how they contribute to information seeking, confidence and differential generation.

Research Questions

In this study, we investigate the following research questions:

- What is the thought process of students as they performing our diagnosis task? Do students report ruling out differentials as they seek information on patient during diagnoses?
- Can we define different reasoning strategies based on the think-aloud utterances of medical students?
- If so, what reasoning strategies are medical students using when making diagnoses and weighing up differentials?
- How do differences in reasoning strategy manifest in terms of information seeking, both in terms of the quality and quantity of information sought?
- Are differences in reasoning strategy related to the individual or are they dependent on the case at hand? Do better performing medical students utilise specific reasoning strategies?
- What considerations do medical students report having whilst they are making diagnoses? And how these differ from how medical students reflect on their thought process after performing the task?

Methods

Participants

16 participants were recruited for this study. Participants were 5th or 6th year medical students at the University of Oxford (including 2nd year Graduate Entry Medical students) recruited via physical posters at Oxford's John Radcliffe Hospital and via a mailing list for students managed by the Medical Sciences Division at the University of Oxford. The study was conducted onsite at the John Radcliffe

hospital. Participants were recruited between July 5th 2023 and December 1st 2023. This study was reviewed and granted ethical approval as an amendment to our existing protocol to allow for audio recordings by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE004.

Materials

The same set of vignettes and a similar computer interface to Study 2 was used for this study, with the exception that participants no longer explicitly recorded their differentials at the end of each information gathering stage. Instead, participants' differentials were recorded in a more naturalistic way. Participants verbalised out loud their thought process as they worked through each diagnostic case. The study was conducted onsite using a laptop, with actions on screen recorded on video and the audio of participants' thinking aloud recorded via a microphone. Informed consent was obtained anonymously using an online electronic information sheet and consent form. Information, including experimental data and audio recordings, collected during the study were stored under anonymised IDs with no linkages to participants. Data was kept on a password-protected computer and hard drive.

Procedure

The general procedure was very similar to that of 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. Given that participants were encouraged to verbalise their considered differentials as they were performing the task, we did not show participants the screen where they explicitly listed the differentials they were considered. 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.

Aside from these differences, participants performed the same six patient vignettes (in a randomised order) from the first study using the same interface that allows them to seek information that they think is useful for that particular case.

Data Analysis

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

the process by which we code for reasoning strategies in the previous study's (Study 2) dataset.

Descriptive Quantitative Analysis

The variables defined for this study are similar to Study 2, as we utilised the same interface and vignettes. Specifically, the variables for confidence, subjective difficulty and information seeking are the same as in Study 2. We note some key differences however. Firstly, given that participants did not explicitly report the differentials they were considered at each information stage, we are not able to record the number of differentials in the same way at each stage. Secondly, we also define accuracy differently due to the lack of this differential reporting screen:

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

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

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

  Based on this, in lieu of the Number of Differentials variable from our

previous study, we define a new variable to look at the number of instances in which participants evaluate or reevaluate the differentials they are considering:

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

Coding of Reasoning Strategies

We aim to detect which reasoning strategies are used by students on each case. To code for reasoning strategies, we adopt a similar approach to Coderre et al. (2003). We define coding criteria 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 (HD)** - prior to selecting the most likely diagnosis, the participant analysed any alternative differentials one by one through something akin to a process of elimination.
- **Scheme Inductive Reasoning (SI)** - participant structures their diagnosis by pathophysiological systems or categories of conditions (e.g., infective vs cardiovascular causes) to determine root causes of patient symptoms rather than focusing on specific conditions.
- **Pattern Recognition (PR)** - participant considers only a single diagnosis with only perfunctory attention to the alternatives, or makes reference to pattern matching when using a prototypical condition to match its symptoms against the current observed symptoms for the patient (e.g., "these symptoms sound like X" or "this fits with a picture of Y").
- **None** - cases are defined as not having a clear reasoning strategy if there are insufficient utterances to make an inference that a participant is using a

particular reasoning strategy (as agreed by both coders).

We first code specific statements within each case that suggested one of these strategies, and then determined which strategy was most prevalent or influential for cases as a whole such that each case was categorised under one of these strategies. Coding of utterances and case-wise reasoning strategies were conducted with a second independent coder. For reasoning strategies, initial interrater reliability was low, with both coders agreeing on 58.3% of cases. When resolving these initial conflicts, changes were made to the coding criteria to prioritise strategies used early on in a case, as some participants were noted to utilise multiple strategies within a single case. For example, a participant may use a SI approach to focus on different pathophysiological systems and then adopt a PR approach to identify an appropriate diagnosis within a specific pathophysiological system. The coding criteria was also changed to allow cases to be coded as not having a clear strategy due to a lack of utterances. The coding criteria provided above are after these changes were made. Cases were then independently coded for a second time with these updated criteria. Both coders agreed on 78% of cases when coding for correctness, with conflicts resolved in consultation with a member of expert panel used to develop the vignettes (as mentioned in Study 2).

We hypothesise not only that this think-aloud methodology can be used to detect different reasoning strategies but also that usage of these reasoning strategies will vary. There are two possible ways in which they vary: they may vary as depdending on the individual medical student and/or they may vary as a function of the specific case/patient condition being treated. For the former, reasoning strategy would be more related to the individual medical student, in that each student will have their own strategy that they tend to use regardless of the patient. For the latter, there would be properties of specific patient conditions that prompt usage of certain reasoning strategies. We investigate both of these competing

theories within this study. To look at individual-level strategy, we ask participants about their diagnostic reasoning process during the debrief interview. This includes questions such as *"What's your general approach to making diagnoses?"* and *"Do you tend to keep a broad set of differentials in mind?"* (full list of questions available in the Appendices). Based on their responses, participants are each categorised as belonging to one of the three reasoning strategies. This is considered their 'subjective strategy'. To look at condition-level strategies, after classifying each case using independent coders, we find the most commonly used strategy for each condition. This is considered the condition's 'dominant strategy'. Once both of these are defined, we compare the strategies coded for each case against both the subjective and dominant strategies. By comparing the cases' reasoning strategies with both the subjective and dominant strategies (using Binomial Exact Tests), we can investigate whether it is the individual medical student or the patient's medical condition that is responsible for the choice of reasoning strategy on a given case.

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

Qualitative Thematic Analysis

The aim of this thematic analysis is to identify the reasoning strategies that medical students in this study reflectively report using in their medical practice, as well as understanding considerations made by students when making diagnoses. Similar to the think-aloud utterances by participants during the vignette task, we record and transcribe the responses given by participants during the debrief interviews (administered after the vignette task). With these interviews, we aimed to understand how participants report making diagnostic decisions, including how they seek information and weigh up differentials against each other. Based on these transcribed responses, 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 interview with predefined research questions of interest, rather than a dataset with a looser structure.

Recoding of Reasoning Strategies in Online Vignette Study

As laid out in the Introduction, we ask with this study is whether we are able to detect each participant's reasoning strategy on a case-by-case basis in Study 2. As Study 2 was an online study, we do not have access to participants' thought process via thinking aloud. However, we can ask whether the information seeking patterns in Study 3, in which we do have coded reasoning strategies based on think-aloud utterances, can be used to predict reasoning strategies in Study 2. In order for this to be possible, we assume that the reasoning strategies we have defined are differentiated from each other by their information seeking. To test this assumption, we establish whether the predictions of reasoning strategy are able to reliably correspond with the 'ground truth' labels of reasoning strategy via our independent interrater coding process. If these predictions of reasoning strategy are reliable within the think-aloud study, we can then apply the same method for predicting based on information seeking in the online study. If they are indeed reliable, we can determine, given medical students' reasoning strategies,

whether these strategies are associated with differences in information seeking and confidence (with the increased robustness of the online study's higher sample size). In order to apply reasoning strategies to the data from Study 2, we train a classifier using penalised multinomial logistic regression to classify cases as HD, PR or SI using the cases from the think aloud study (with Leave One Out Cross Validation). As mentioned, some cases are not coded with a reasoning strategy if there are insufficient utterances to infer a clear strategy, and these case are excluded from this training. The input parameters for the classifier are the 29 pieces of information as binary predictors (similar to the approach depicted in Figure 8 of the previous study) and the cases' condition. In other words, the cases from the think-aloud study make up the training data for the classifier whilst the cases from the larger online study make up the test dataset. The classifier was implemented using R's caret package with the train() and glmnet() functions. The testing data is then labelled with predicted strategies using R's predict function. We report accuracy values for the classifier when applied to think-aloud study in terms of how well the classifier is able to reproduce the coded reasoning strategies.

As well as predicting reasoning strategy on a case-by-case basis to determine differences in information seeking, confidence and differential generation, we also look at differences across cases with a given reasoning strategy. As we describe earlier, we use the think-aloud to determine the dominant strategy for a given case by observing which reasoning strategy was used by a majority of participants. We then look across cases for a given dominant strategy, how our key dependent variables are affected by the type of case. We group cases in Study 2 by the dominant strategy used in Study 3 for each condition. For example, if a majority of participants use a HD strategy in the think-aloud study for the UC case, we then consider UC to be HD-dominant.

Results

Descriptive Quantiative Results

Overall Performance and Calibration

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

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

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 some participants repeating information they see during the task, but participants also varied in terms of how much they externalised their thought process.

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 by condition, as can be seen above in Table 1. Similar to the previous study, we ask whether participants provided confidence judgements that were, on the whole, calibration to their objective accuracy. In order to investigate this for this study, we compare the final confidence reported by participants on cases when they were objective correct and when they were objectively incorrect. Participants would be considered calibrated if we found evidence of higher confidence when correct. Participants did indeed report higher confidence when correct (M = 50.24, SD =

23.89) compared to when they were incorrect (M = 42.32, SD = 26.71). Given that the samples from each of these groups are not independent, we test for a difference between these groups using a mixed effects model that predicts final confidence using the correctness of the case as a fixed effect and the individual participant as a random effect. We find evidence that the case correctness was predictive of final confidence (β = 0, SE = 0, t = -0.96, p = 0.34), indicating that participants provided confidence judgements that were well calibrated to their objective accuracy.

Differential Evaluations

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

Table 2: Descriptive Statistics for subcodes within the Differential Evaluation main code as detailed above in the Data Analysis section. Shown above are mean values for the number of instances/utterances for each of the following subcodes (from left to right): a new differential being considered, a differential being removed from consideration, a differential being seen as more likely given a piece of information, a differential being seen as less likely given a piece of information, the average total of these subcodes.

For utterances coded as Differential Evaluations, participants on average made 5.48 such utterances per case. The mean number of Differential Evaluations was relatively constant by condition except for the AD case, for which we observed a higher amount of Differential Evaluations (see table 2 above). As previously mentioned, Differential Evaluations can be further categorised into one of four subcodes: Differential Added, Differential Removed, Likelihood Increased and Likelihood Decreased. As found in the previous study, there is a general reticence to disregard differentials completely. Participants expressed significantly more statements adding differentials (M = 3.14, SD = 1.66) than removing differentials (M = 0.27, SD = 0.53) (t(15) = 14.14,

MDiff = 2.86, $p < .001$). Out of the 16 participants, 6 participants never recorded an utterance where they removed a differential from consideration. Participants expressed more statements of decreasing likelihoods (M = 0.99, SD = 1.09) rather than increasing likelihoods (M = 0.93, SD = 1.09) but we did not find evidence of a significant difference ($t(15) = 0.34$, MDiff = 0.06, $p = 0.73$).

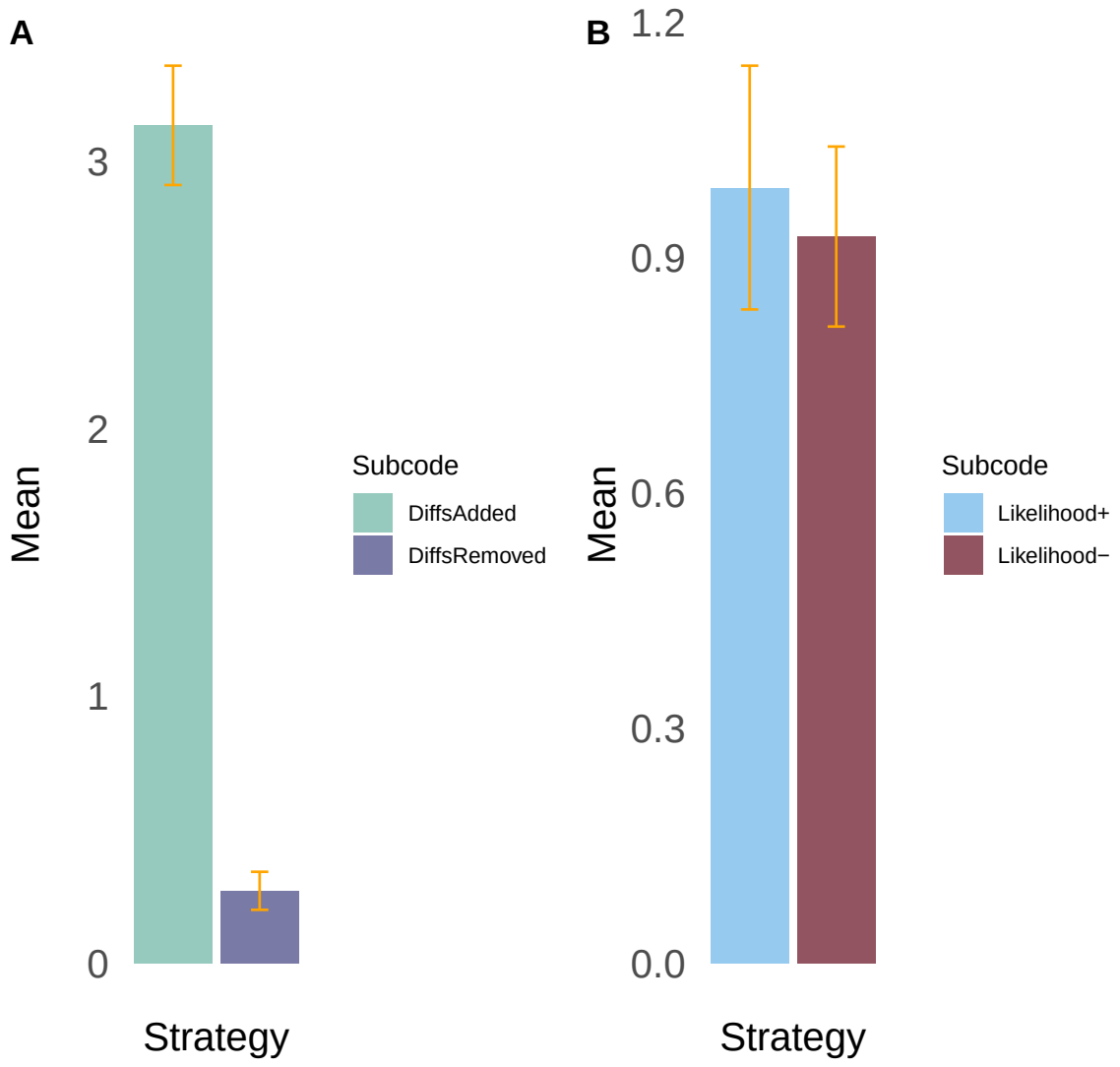

Figure 1: Bar graphs comparing incidences of each of the four subcodes within Differential Evaluations. We compare instances of differentials being added (green) and removed (purple) from consideration (Figure 1A) and compare instances of differentials decreasing (blue) and increasing (red) in likelihood (Figure 1B).

Reasoning Strategies

Incidence of Strategies

In Table 3 below, we show all 96 cases from the think-aloud strategy and the strategy coded to each after resolving all interrater conflicts. Of these cases, 6 were coded as not having a clear reasoning strategy due to both an insufficient amount of think-aloud utterances and no diagnostic differentials being mentioned. 43 cases

were coded as having a HD strategy, 29 cases were assigned a PR strategy and 18 cases were coded as SI. In Figure 2 below, we plot the proportion of cases for each patient condition that were categorised under each of the reasoning strategies. We note that the types of reasoning strategy used varies by condition (see Figure 2 above), with the MTB and TTP cases in particular exhibiting higher usage of PR than others, whilst HD was used by the majority of participants for the UC and AD cases in particular. In Table 4, we show examples of key quotes that resulted in coding of reasoning strategies.

ID	**AD**	**GBS**	**MTB**	**TA**	**TTP**	**UC**
3lkzjq	HD	SI	PR	SI	HD	HD
4khzxs	HD	HD	HD	SI	PR	HD
593ybw	PR	PR	PR	HD	HD	HD
5lvg8j	HD	HD	HD	SI	HD	HD
clhtyq	HD	HD	HD	HD	HD	HD
d9b1qf	HD	HD	HD	PR	PR	HD
dcjymb	SI	SI	PR	SI	SI	HD
gdq7tc	SI	HD	PR	PR	NONE	PR
gs6zbl	SI	HD	SI	PR	PR	PR
jdqsnf	HD	SI	PR	PR	PR	PR
k5376h	HD	SI	NONE	HD	PR	HD
l3jd8r	HD	HD	PR	SI	PR	HD
ly9kzg	HD	SI	PR	PR	PR	PR
rslkq8	NONE	NONE	NONE	PR	PR	NONE
y86m2n	HD	HD	HD	HD	HD	HD
ytpshg	PR	SI	PR	SI	SI	HD

Table 3: Table that shows the strategy coded for each case by participant (rows) and by patient condition (column) after resolving conflicts between both independent coders. Anonymised participant IDs are used.

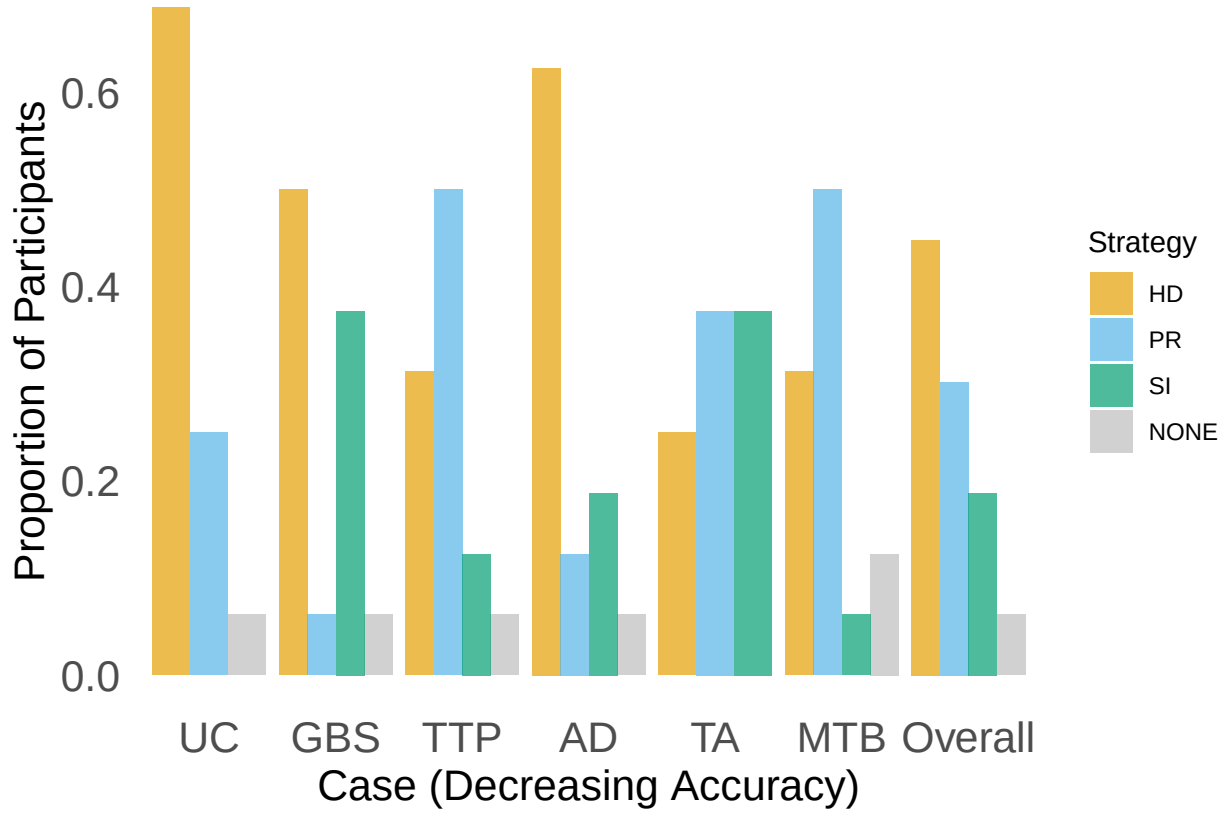

Figure 2: Proportion of participants who use each type of reasoning strategy for each condition/case, with the overall proportions across all cases shown by the rightmost bars. The strategies shown are: Hypothetico-Deductive (where multiple differentials are considered simultaneously, orange), Pattern Recognition (where a single differential is considered in turn, blue), Scheme-Inductive (where participants evaluate pathophysiological systems as causes of patients rather than specific conditions, green) and None (for cases where a clear differential is not mentioned or if there are not enough utterances to infer a clear strategy, grey).

Participant (Case)	Quote	Coded Strategy	Interpretation
3lkzjq (AD)	"A long time ago she had a total hysterectomy, doesn't really tell me much. Although... why did she have that? Was that for cancer? Has that come back? She must have been quite young when that happened. So yeah, it would be worth investigating further...so I would need more investigations before knowing how to treat her in this case. So I'm not very confident at all. If this is an infection, if this is cardiac related, is it a PE? Yeah, it could be many things or is this malignancy? Yeah. So hopefully the exam will tell us more."	HD	Considering multiple diagnostic possibilities before starting physical examinations, naming specific conditions (e.g. PE) rather than simply focusing on pathophysiological systems
593ybw (GBS)	"Reduced tone, reduced power, reduced coordination, reduced reflexes. Upper limbs normal. So it's looking like lower motor neuron given this reduced tone. And so could this lower motor neurone be something like Guillain Barre? What would cause that, would be sudden onset. Okay, but yeah, given that it's not upper motor neuron stroke."	PR	Use 'looks like' to denote pattern matching, points to a specific condition, mentions ruling out a previous differential before moving on to this current one.
ytpshg (TA)	"Wonder if this is an underlying rheumatology thing and this fever is something separate so let's just auscultate his lungs. Lungs are clear, get the heart. Okay neurological exam, normal, eyes, want to do abdo as well. Ok, fine so I think it's more of an infective picture but I don't know where the source of infection is...So testing, do CRP and ESR. CRP is raised, ESR is raised. So again could be rheumatological or infective."	SI	Weighing up two different groups of conditions that could be causing the patient's symptoms, but not mentioning specific conditions

(continued)

Participant (Case)	Quote	Coded Strategy	Interpretation
clhtyq (UC)	"So I think diverticulitis is probably at the top of my differential, IBD would be up there. Infective gastroenteritis. I would also just be worried generally about his hydration status because it sounds like he's losing a lot of water."	HD	Explicit mention of a 'list' of differentials, with one at the top of consideration
ly9kzg (MTB)	"Crackles on the right and left lower zones. So bilateral crackles could be congestive heart failure and then sounds like there's maybe a pneumonia there with the crackles in the right middle zone."	PR	Uses 'sounds like' to denote pattern matching, provides two concurrent conditions as a single diagnosis rather than separate possibilities
dcjymb (TTP)	"So the slurred speech, headache and weakness in the arms sounds neurological in origin. And his high blood pressure is concerning for such a young man...I don't have a definitive diagnosis. Now could be something infective or it could be something cardiac in origin."	SI	Mentions multiple pathophysiological systems that could be implicated, explicitly states that they do not have a definitive diagnosis

Table 4: Table that shows examples of cases that were coded for a particular reasoning strategy and key quotes that suggest the coded reasoning strategy. We provide the quotes from each case, as well as how we interpret the quotes in line with the particular reasoning strategy.

Reasoning Strategies' Effect on Dependent Variables

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

Table 5: Mean values for dependent variables broken down by the reasoning strategy coded after resolving conflicts between the two independent coders. From left to right, Accuracy refers to the proportion of cases where a correct differential was mentioned. Differential Evaluations refers to the number of coded utterances under one of the subcodes (Differential Added, Differential Removed, Increased Likelihood, Decreased Likelihood). Information Seeking refers to the proportion of available information sought across cases. Confidence Change refers to difference between initial confidence and final confidence.

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

Dominant Reasoning Strategies

We aim now to establish if participants are more accurate when using each condition's dominant strategy. This is based on the assumption that each medical condition has an 'optimal' reasoning strategy that should be used to maximise accuracy. We first categorise each of the 6 cases as having a 'dominant' reasoning strategy based on which was utilised the most across participants. Through this process, we categorise three conditions as HD (AD, GBS, UC), three conditions as PR (MTB, TTP, TA, we note that there was an equal number of PR and SI cases for TA condition, but we use PR as its dominant strategy to easily compare

HD and PR directly). HD was assigned to 62.5% of AD cases, 50% of GBS cases, and 68.75% of UC cases. PR was assigned to 50% of MTB cases, 50% of TTP cases, and 37.5% of TA cases. Accuracy was found to be higher for cases when participants matched the condition's dominant strategy (0.67) compared to when they did not (0.54). However, this difference was not found to be significant via a mixed effects logistic regression (on accuracy as a binary outcome measure with both condition and participant as random effects) ($\beta = 0.6$, SE = 0.5 t = 1.19, p = 0.23). Overall, participants matched the dominant strategy on 51 cases (56.7% of cases, excluding those cases without a clear reasoning strategy). When looking at whether participants use the case's dominant strategy more than chance (50%, given the two strategies of HD and PR), we compute a Binomial Test to test if the observed probability is significantly higher than expected by chance given the sample size. We do not find evidence that participants choose the case's dominant strategy significantly above chance (number of case-dominant strategies = 51, probability = 0.57, 95% CI = [0.47 , 1], p = 0.12).

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

Table 6: Table showing average accuracy values by cases where the participants used or did not use the dominant reasoning strategy for that case. Dominant strategies are decided based on which of the reasoning strategies was utilised by the majority of participants in the think-aloud study. Cases without a coded reasoning strategy are excluded from this table. The first column refers to the dominant strategy for that condition, whilst the second column refers to whether the cases' coded strategy matches the condition's dominant strategy.

Subjective Reasoning Strategies

In addition to reasoning strategies being coded based on the participants' think-aloud utterances, we also asked participants about their diagnostic process during the debrief interviews that can be used to infer the reasoning strategies participants think

they use in their regular medical practice. We use this to determine if participants are more accurate when using their subjectively preferred reasoning strategy. In Table 7 below, we categorise participants based on their subjective reflection of their diagnostic process. Through this process, we categorise 7 participants under a HD reasoning strategy, 6 participants as PR and 3 participants as SI. Given these categorisations of reasoning strategy based on subjective reflection by participants, we compare these participant-level strategies to the case-level strategies assigned by our independent coders. Accuracy was found to be higher for cases when participants matched their subjective strategy (0.68) compared to when they did not (0.58). However, this difference was not found to be significant via a mixed effects logistic regression (with accuracy as a binary outcome measure and both condition and participant as random effects) (β = 0.84, SE = 0.68 t = 1.25, p = 0.21). We find that there are 31 cases (34.44%, excluding cases without a coded reasoning strategy) where participants match the reasoning strategy during the case to their subjectively defined strategy that they tend to use for diagnostic decisions. As this is less than 50%, we overall find that participants do not match their subjectively preferred strategy. When computing a Binomial Test to determine if this probability is less than expected by chance, we find evidence that participants go against their subjective strategy significantly less than chance (number of cases with strategies matching subjective preference = 31, probability = 0.34, 95% CI = [0 , 0.44], p = 0.002).

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

Table 7: Dependent variables by cases where the reasoning strategy used (as categorised by the independent coders) matches the subjective strategy coded for that participant (as per responses to the debrief interview, see table 8 below).

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
3lkzjq	"I try to go based on kind of...the stock, that kind of differentials I have for the different presentations and then work down, kind of trying to narrow it down with further information and kind of rule out certain things in my head."	HD	I generate differentials based on the patient presentation and then narrow those down	Process of elimination
4khzxs	"Yeah, I think I would like to say that I would think about different systems and stuff, but I think I'm a bit more frantic...I'm not sure if I have a specific approach, I think I just wait for something into pop into my head, really, which is quite bad."	PR	I would like to be more structured but I tend to wait for a diagnosis to come to mind based on the information I have	Matching information with most likely diagnosis
593ybw	"But yeah, I definitely, like, often have an idea from the start. I think I probably do think about that the whole way through, which probably can be beneficial, but can also sometimes hold me back from looking at other options."	PR	Early differential used to guide rest of the diagnostic process, not much consideration of alternative diagnoses	Seeking information in favour of a single diagnosis

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
5lvg8j	"So going through like a system of starting with the history and sort of gathering as much information as I can there and thinking already what I think might be happening, and then examining them and seeing if that sort of changed anything, but then sort of getting investigations. And yeah, basically, I like to sort of, yeah, piece everything together, sort of, as I get information, and then sort of try to reach a diagnosis that I think might be going on sort of at the at the end of that...I think probably, especially as a medical students, we get taught to rule out like red flag stuff. So like a lot of my thinking is like, what like really worrying thing could this be that we need to rule out? And what tests do I need to rule it out? But then also like when I'm thinking or what what could this be, I'm also thinking about what investigations would help me to conclusively reach a diagnosis that this is what it is once I think I know what's going on."	HD	Seek as much information as possible to find what differentials are likely but also what differentials need to be ruled out. Based on what is ruled out, I can decide what is most likely.	Focus on ruling out differentials to find what is likely

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
clhtyq	"But I think most of the time I try and keep quite a broad set of ideas. And then narrow it down and try and think, like, system wise...like the one that was aortic dissection, that I was sort of like, this sounds like they're coming with abdo pain, but it could be something cardiac, it could be something like...trying to keep ruling out systems."	SI	Keeping broad set of ideas based on the possible systems that could be implicated in the presenting complaint and then ruling systems out	Structuring diagnosis by what patho-physiological systems can be ruled out
d9b1qf	"I think we need to rule out differentials, but I felt like a lot of points, I felt like this is the most likely but I still feel like I want rule everything else out first...But also for the last one I especially, I have very strong feeling it was GCA. Therefore I wasn't as keen to broaden my differentials."	PR	Ruling out differentials is important, but there tends to be a primary diagnosis that comes to mind	Focal diagnosis tends to come to mind, though some ruling out of differentials is also necessary
dcjymb	"So after taking a history, as long as the patient's not like acutely unwell, taking history and accurate family history, and their past medical history, creating a sort of list of most likely differentials in my head, and then seeing what would be useful in rolling some of those in and out."	HD	Generate a list of differentials first based on patient history and then decide what information is needed to either support or rule them out	Work from an initial set of differentials to guide information seeking

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
gdq7tc	"Trying to go from their main symptom and then trying to bring up general, maybe not specific conditions, but kind of what can go wrong that would lead to this picture...I think my brain can sometimes get stuck on an idea. And it's difficult to pull away from that."	PR	Think about what comes to mind based on the initial presentation, can be influenced by an early idea of what the patient has, less focused on specific conditions to rule out	Generate a general diagnosis based on what comes to mind from early patient presentation
gs6zbl	"Either spot diagnosis if it's a really typical case, like otherwise, basically, if I'm not sure I'll work through a surgical sieve."	PR	Come up with an idea for a diagnosis on the spot or go through a structured process if not sure	Default is to match the current patient to what comes to mind as a prototypical case
jdqsnf	"I guess, trying to go from their main symptom and then trying to bring up general, maybe not specific conditions, but kind of what can go wrong that would lead to this picture. And then trying to rule in or out things that fit with that...Like, if someone's coming in with a presentation that could be quite urgent and serious, then obviously you want to rule out like a stroke, you want to rule that out quite quickly."	HD	From the symptom, generate a set of differentials that could create this picture in the patient, then rule in and out from this set	Initial set of differentials to rule in or out

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
k5376h	"I guess you've always got to think about two things. What's common? And what do we have to rule out?... even just to rule out a few more things that perhaps then you don't even have to test for and consider later, if you've ruled them out in history."	HD	Generate differentials based on what's common and what's severe if missed, rule some differentials out during history taking so you don't have to test for them later	From an initial set of differentials, work on ruling some out as early as possible
l3jd8r	"Take a history, like detailed history, formulate my top differentials. And then basically, look at investigations and examinations to confirm or rule out these differentials or any that are sufficiently different? Consider different differentials."	HD	Generate a set of differentials based on a detailed history, seek information to rule them in or out	Work from an initial set of differentials to guide information seeking
ly9kzg	"I try and think about the, like the different systems that might be involved before thinking about...so for example, if someone presents with chest pain, there's lots of things in the chest that could lead to chest pain. So try and think about the systems more generally and then focus on the specific symptoms that align with this specific system if that makes sense."	SI	Based on the presenting symptoms, I think about what patho-physiological systems might be causing them and then go from there	Primary focus on determining the system involved before thinking about specific conditions

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
rslkq8	"Just sort of work through from the most like, basic things like history and examination within like least invasive tests, and then try to work up from there. trying to rule out the most serious things...Like some things, you just have to get the differentials broad because there's so many things that could cause it."	HD	Rule out serious differentials, important to remain as broad as possible	Focus on remaining open minded and broad to rule differentials out
y86m2n	"But I think generally the big things are is there a system under which it generally falls? Primarily, like, does it sound haemotological, cardiac, whatever."	SI	Trying to match symptoms to a particular pathophysio-logical system	Structures diagnosis by pathophysio-logical systems
ytpshg	"So probably getting history, I look at initial observations first...and then I look at ECG as well, then I want to get initial bloods being guided by what I think could be going on and then think about potential images. If it's quite obvious early on what's going on. And for example, it's some kind of infection, I'd want to start them on antibiotics early, but otherwise, I might get advice or wait until I get all my information, before starting treatments."	PR	Based on intial history and tests, see if it is obvious what is happening with the patient and if not, seek additional advice and information	Seek information until a clear diagnosis comes to mind matching the information

Table 8: Categorisation of participants under one of three possible reasoning strategies based on their responses during the debrief interview. We capture here the subjective reasoning strategy for each participant based on how they reflect on

how tend to make diagnostic decisions. In the second column are key highlighted quotes related to each of the participants' diagnostic decision making process. In the fourth column, we provide our summary of the quote and then in the fifth column, our interpretation of the quote that explains the choice of reasoning strategy for that participant.

To combine the two previous sections, we consider whether participants are more accurate when in fact their reasoning strategy matches both the case's dominant strategy and the participant's subjectively preferred strategy. For briefness, we refer to this as the "fully matched strategy" going forward. Participants used a fully matched strategy on 21 cases (23.33%). On these cases, participants were more accurate (0.71) than for cases when reasoning strategy matched neither the case's dominant strategy nor the individual's subjectively preferred strategy (0.52). We do not find evidence of an interaction via a mixed effects logistic regression (with accuracy as a binary outcome measure and both condition and participant as random effects) ($\beta = 0.26$, SE = 1.13 t = 0.23, p = 0.82).

Thematic Analysis from Debrief Questionnaire

In this section, we present key themes from the thematic analysis of participant responses to the debrief questionnaire. The questionnaire was designed to ask participants how they think they tend to make diagnostic decisions and what their main considerations are during the decisional process. We provide quotes from participants belonging to each of these themes. Participants are referred to by their anonymised identifiers.

Avoidance of Anchoring

A key consideration, as mentioned by six participants, was the concept of anchor bias. This was explained by one of these participants as follows:

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

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

"I'm quite rubbish, I often get fixated like 'I think this is this"'. . . but I'm not that good at thinking, 'oh, what else could it be' into like, 'I've got something that's proved to me it's not.' " (4khzxs)

"I try to, but I think my brain can sometimes get stuck on an idea. And it's difficult to pull away from that." (gdq7tc)

"I think I probably do think about that the whole way through, which probably can be beneficial, but can also sometimes hold me back from looking at other options" (593ybw)

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

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

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

"My knowledge isn't broad enough... to remember all the differentials for everything." (rslkq8)

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

Standard Tendencies

Participants reflected a few general tendencies (or rules of thumb) when making diagnoses. Firstly, seven participants mentioned that they prioritised any serious/emergency differentials early on when making diagnoses, which would affect the urgency with which they would approach ruling these differentials out. This suggests that some focus would be taken away from determining likely diagnoses and instead ruling out more serious diagnoses that would require more immediate medical attention. This also indicates that the manner in which medical students

approach diagnoses is dependent on the nature of the patient being treated and whether serious diagnoses are being considered. The need to consider serious diagnoses offers a potential reason for why participants do not simply utilise their preferred reasoning strategy on all cases, as some patients may display symptoms that prompt consideration of diagnoses that are harmful if missed.

"I do have an approach, the first (thing) I always want to think is if I miss something, is this patient gonna be bad? So, like, thinking about emergency stuff." (4khzxs)

"I think probably, especially as a medical student, we get taught to rule out red flag stuff... a lot of my thinking is like, what really worrying thing could this be that we need to rule out? And what tests do I need to rule it out?" (5lvg8j)

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

"Like, if someone's coming in with a presentation that could be quite urgent and serious, then obviously you want to rule out like a stroke, you want to rule that out quite quickly." (gdq7tc)

"If I think something's remotely possible, that's really like say, so like for GBS, I'm even thinking about Cauda Equina syndrome, like, regardless of how high my index of suspicion for it is, even if it's pretty low. If it's an urgent diagnosis, I'll just do it anyway." (gs6zbl)

"I'm trying to rule out the most serious things." (rslkq8)

"And essentially if there's any serious conditions, I make sure to rule those out... and then go from there. I probably should go through each one and weigh each one individually. Because that would avoid being as biased. But it's not something I do as much as I should... The other big things, are there any of the red flag symptoms that are really important that should influence what I'm thinking? Like fevers especially, that sort of thing... So generally, acute situations are where I narrow a little bit." (y86m2n)

Another tendency was for participants to report a form of progressive investigation that stems from the patient's history. In this sense, participants report a decision process that quite closely matches our experimental procedure of gradually seeking information based on patient's medical history to build up a picture of them. This illustrates the importance of a comprehensive medical history for the patient being available and how much it guides medical students' decisional process. We can also ascertain from this theme that the initial diagnostic differentials generated from the patient's history has a large influence on the subsequent diagnostic process:

"Definitely start like history... I think to go from there and like, kind of think about that in the context of the patient. Yeah, I feel I've definitely been taught in terms like that methodical, like do it in that order." (593ybw)

"I guess going through like a system of starting with the history and sort of gathering as much information as I can there and thinking already what I think might be happening. And then examining them and seeing if that sort of changed anything, but then sort of getting investigations." (5lvg8j)

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

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

"Take a history, like detailed history, formulate my top differentials. And then basically, look at investigations and examinations to confirm or rule out these differentials." (l3jd8r)

"I think, start, think systematically. So start with a thorough history, asking kind of about what's happening currently, and then going through the kind of past medical history and focusing on that, asking what the patient thinks might be going on. And then focus on a thorough examination which sometimes for the interest of time is focused on the, the kind of symptom at hand, but you should do a kind of formal full checkthrough as well... see if there's anything that points you towards a diagnosis. I think it (the experiment) was set up in the way that I go about things in the way that you do the history first, you do the examination, you do the investigations." (ly9kzg)

"Just sort of work through from the most like, basic things like history and examination, least invasive tests, and then try to work up from there." (rslkq8)

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

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

"I would always want to do like full blood count, VBG... Probably, as I said, I think like most people in the emergency department get a chest X ray" (4khzxs)

"With the examination, I think, normally, I would like... auscultating the heart and feeling the heart, abdomen, etc. These are things I think I would do in any patient, irrespective." (5lvg8j)

"And then for investigations... I'll take all the bloods, do an ECG, chest X Ray, just in case. Yeah, yeah. So I am a bit more like, on the side of caution." (d9b1qf)

"I think if you went to your senior and you said I'm really concerned about this patient, but I've not done an FBC, a U&E, an ECG, VBG... They'd be like,

what are you on about? So there's a few that you would do anyway, that are largely non invasive, in terms of... higher degree investigations, very much depends on anatomically, what you're seeing, what your differential is." (k5376h)

"I wasn't sure (during the experiment) whether I should try and be very focused to the presentation at hand... because for example, when I was going through the examinations, in reality, I would do a full exam on someone, even if they presented with something that was very specific, like a very specific symptom just so that you can have a full kind of clerking assessment." (ly9kzg)

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

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

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

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

"Basically, if I'm not sure I'll work through a surgical sieve." (gs6zbl)

"I think the other way I can do it is the sort of surgical sieve idea to make sure you've ruled out" (y86m2n)

When taken as a whole, there are differences between students in terms of how standardised they perceive the process of diagnosis to be. The extent to which it is based around intuition or a more logical, structured process is what may contribute to differences in how diagnostic decisions are made in terms of what information is sought and how clinicians/students generate diagnostic hypotheses.

Challenges of the Diagnostic Process

Participants cited a number of challenges related to our diagnostic task. These were related to ways in which our study did not emulate real-world aspects of the diagnostic process. Firstly, three participants noted that it was difficult to retain all of the information they needed during the task:

"I kept thinking that I couldn't like hold onto all of the information." (4khzxs)

"I'm quite a visual person. So... reading on a screen is quite different to I don't know, actually having seen a patient, seeing the exam findings, or even looking at the scans myself. I feel like I find that easier to retain the information. Whereas when it's, I find it, it's kind of hard to take it in when it's like just written down" (593ybw)

"I think just thinking on the spot and coming up with the diagnosis quite quickly, it's quite hard remembering the management afterwards as well." (rslkq8)

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

"I'd find as much information as possible and to ask for help... I think I'm someone who looks up stuff a lot. Like I rely a lot on looking up things. And that gives me a lot of comfort. I feel like when I don't have those tools, yeah, I feel a bit shaky... And I often kind of just take my phone out and look at that, and even just glancing at them kind of helps me structure my thoughts. So, yeah, I would wish it

would be kind of more, more of an organic process. But at the moment, I think I rely quite a lot on prompts and things like that, or guidelines, even if I don't read them thoroughly, I'd need kind of reminders, especially when I feel like there's so much that it could be and I lose myself a little bit in the possibilities." (3lkzjq)

"I guess in real life, I also have, like, Google. So at points where I forgot the disease, or like, what is the first line, I would have like checked before, before typing up my management plan... when I'm confused, I'm definitely going to approach the senior. So I wouldn't be the one making the diagnostic decision. So I think it's a bit harder to do it alone." (d9b1qf)

"When I'm not sure, I'll definitely running my my train of thought past my consultants." (l3jd8r)

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

Reasoning Strategies in Study 2

We next turn to the coding of reasoning strategies in Study 2. We train a multinomial logistic classifier to identify reasoning strategy in this think-aloud study, with each of the information requests as binary predictors (i.e. whether they were sought or not on each case). We then compare the predicted strategies from the classifier to the objective 'ground truth' strategies in order to assess the model's accuracy. Before we can interpret the reasoning strategies predicted for the Study 2 dataset, we first need to determine that the classifier reliably reproduces the reasoning strategies coded in the Study 3 dataset based on think-aloud utterances. In order to train the classifier, we account for imbalance in the training data (due to the larger number of HD cases and the lower number of SI cases) using downsampling and limited regularisation. In summary, we are not able to train a classifier that performs better than chance at predicting reasoning strategy. The accuracy of our classifier is 0.39, which is significantly lower than the No Information Rate of 0.48 (i.e. the accuracy of the classifier if HD was predicted on all cases). We compare our classifier accuracy to a bootstrapped null distribution where the ground truth labels are repeatedly

shuffled (i.e. such that the labels of reasoning strategies are meaningless). Based on 10 shuffles, we find an average accuracy of 0.37. We also compare Balanced Accuracy values (i.e. the accuracy within each reasoning strategy, in order to account for the data imbalance) to those of the null classifier. We find that the Balanced Accuracy of our trained classifier (HD = 0.5, PR = 0.65, SI = 0.49) does not exceed that of the null classifier (HD = 0.5, PR = 0.48, SI = 0.51). To summarise, we are not able to predict reasoning strategy reliably based on information seeking alone using on our think-aloud dataset. Hence, we do not perform further analysis based on the predicted reasoning strategies in Study 2's dataset.

We are not able to reliably predict reasoning strategy on a case-by-case basis. We instead look at behaviour as a function of the dominant reasoning strategy for the patient conditions based on the coded reasoning strategies in the think-aloud study. Our aim to is to look at whether cases result in different diagnostic behaviour assuming that each patient condition is associated with a particular reasoning strategy. We find that three of the six conditions (UC, GBS, AD) were approached using a HD strategy by a majority of participants. The remaining cases tended to be performed using a PR strategy (except for the TA case, where there were equal numbers of PR and SI coded participants). We consider these cases (TTP, TA, MTB) as PR-dominant cases for ease of comparison with HD-dominant cases. As our observations are not independent, we calculate the mean values across conditions within each group of cases (HD-dominant or PR-dominant) in the online dataset from Study 2 such that there is a single observation per participant. This allows us to investigate broadly if reasoning strategy on a case-wise level affects diagnostic behaviour.

First, we look at whether our key variables vary as a function of the dominant strategy for cases in order to determine if strategy is associated with a difference in behaviour. As we have two groups of cases, we use paired Wilcoxon Signed

Rank to compare median values between these groups (averaged across conditions), as these variables are not normally distributed. We observe higher accuracy for HD-dominant cases (Mdn = 0.47) when compared to PR-dominant cases (MDn = 0.4) (V = 2508.5, pseudomedian difference = 0.11, 95% CI = [0.05, 0.17], p < .001). We also observe that more initial diagnoses being considered during HD-dominant cases (MDn = 3.33) when compared to PR-dominant cases (MDn = 3) (V = 2127.5, pseudomedian difference = 0.51, 95% CI = [0.33, 0.83], p < .001). When looking at differences in information seeking, we observed that more information was sought on PR-dominant cases (MDn = 0.63) when compared to HD-dominant cases (MDn = 0.6) (V = 723, pseudomedian difference = 0.04, 95% CI = [0.06, 0.03], p < .001). We also observed that informational value was higher for PR-dominant cases (MDn = 2.98) when compared to HD-dominant cases (MDn = 1.67) (V = 2.5, pseudomedian difference = 1.18, 95% CI = [1.27, 1.1], p < .001). We did not observe a significant difference between groups of cases in terms of changes in confidence.

Finally, we look at predictors of diagnostic accuracy, hypothesising that the dominant reasoning strategy for a case interacts with the number of initial diagnoses when predicting accuracy. Given that HD has been previously associated with more diagnoses being considered early on, we expect that diagnostic accuracy is determined by using the optimal reasoning strategy given the initial diagnostic breadth. In order to investigate this hypothesis, we fit a linear mixed effects model that predicts diagnostic accuracy with an interaction between the number of initial diagnoses and case-dominant strategy, with participant as a random effect. As in the previous analysis, variables are averaged per participant across all conditions within each case group (i.e. for HD-dominant cases, accuracy and initial diagnoses are averaged across the UC, GBS and AD cases for each participant). We find evidence for an interaction between the number of initial diagnoses and case-dominant reasoning strategy (F(1,108.95) = 10.7, p = .001). As depicted in Figure 3 below, we observe lower accuracy for PR-dominant cases compared to HD-dominant cases with lower initial diagnostic breadth, but accuracy is highest for PR-dominant cases

with higher initial diagnostic breadth.

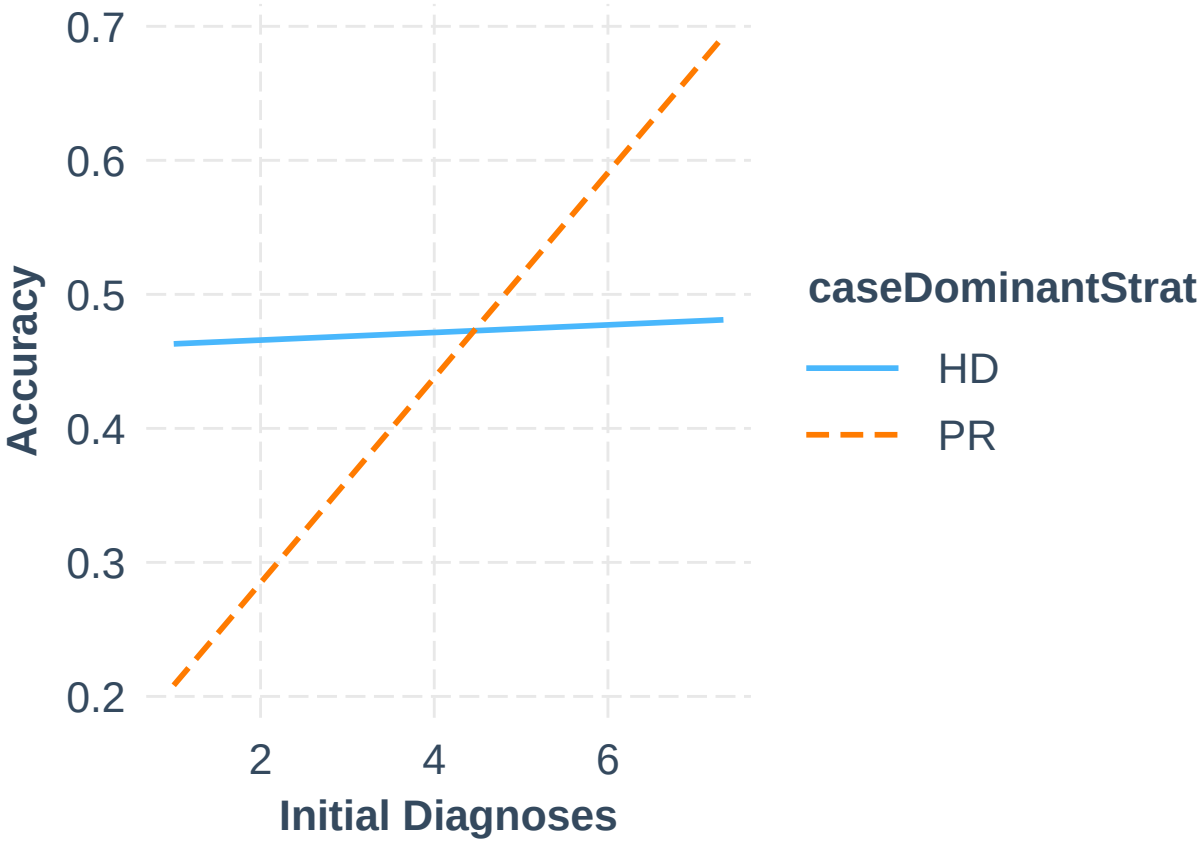

Figure 3: Fitted regression line for a linear mixed effects model that predicts accuracy (y-axis) with an interaction between the number of initial diagnoses/differentials (x-axis) and dominant reasoning strategy (HD in blue averaged across the UC, GBS and AD cases, PR in orange averaged across the MTB, TA and TTP cases). Data shown here is from the online dataset of Study 2.

Discussion

This study of 16 medical students explored the usage of a think-aloud methodology to understand thought processes during medical diagnoses. Using our online interface and recorded verbalisations by students, we aimed to detect clinical reasoning strategies based on criteria adapted from Coderre et al. (2003). The strength of this paradigm is in qualitatively recording medical students' thought process as it involves with information as per our flexible, evolving vignette-based interface

design. By recording how participants consider different diagnoses in real time, we are able to understand the reasoning approach students are applying for each case and how this affects their information seeking and confidence behaviour. We are also able to investigate if these reasoning strategies affect diagnostic accuracy, both in the context of this current study and in the previous online study.

In terms of performance and calibration, we found that participants' confidence judgements were well calibrated to their objective accuracy, similar to the previous study. The measure of accuracy was different in this study by necessity, in that a case was considered 'correct' with the mention of a correct differential at some point during the case. This measure is most similar to the lenient measure of accuracy from our previous study, in that participants were marked as correct if they considered a correct differential without taking into account its relative likelihood compared to other differentials. However, given that accuracy and our test of calibration being different in this study, it is difficult to compare these findings directly. Our finding of calibrated confidence in both studies however is an indication that medical students express uncertainty appropriately. Similar to the previous study, we also find that medical students are reticent to remove differentials from consideration. In this study, participants report low occurrences of disregarding differentials. This would correspond with our assumption that medical students attempt to remain open minded in their diagnoses. By rarely removing differentials from consideration, students are then observed to broaden their differentials with more information (as the number of differentials being considered at once only increases).

In terms of differential evaluations, we observed via think-aloud utterances a similar tendency for medical students to very rarely remove differentials from their consideration. This would support the finding from the previous study that students tend to broaden the differentials they are considering as they receive more information. We find in this study that a HD reasoning strategy is the most common among students, which is similar to Coderre et al. finding that novice clinicians

tended to use a HD reasoning strategy over PR). The overall pattern of broadening differentials in Study 2 can explained by students tending to use a HD strategy but rarely removing differentials from consideration (assuming that there was a majority of students using a HD strategy in Study 2). Prior to conducting both of these studies, we may have expected a 'process of elimination' to be used by students but this does not appear to be the case across both of these studies. Removing differentials seems to be a clear tendency for medical students, rather than a quirk of our interface from the previous study. This has implications for medical education in terms of whether students should be taught situations where it makes sense to disregard differentials or if remaining open-minded at all times is useful all patients.

On reasoning strategies, we were able in this study to use think-aloud utterances to detect reasoning strategies on the part of the medical students. We considered three different strategies: Hypothetico-Deductive (HD), Pattern Recognition (PR) and Scheme-Inductive (SI). These strategies represent different approaches to diagnosis, either seeking to be comprehensive in both the information sought and differentials considered or focusing in on a single diagnosis. We found that these choices in reasoning strategies were not determined by either an individual's general decision making approach or by specific patient conditions. On the former, we found evidence that participants do not choose their subjectively preferred strategy on average. This could be because of the practicalities of patient cases meant that students were forced to make decisions in ways that they were not used to. This begs the question of what the properties are of a patient case that determine the choice of reasoning strategy on a given case. One account is that reasoning strategy is determined by how much experience/familiarity the student/clinician has with that type of patient presentation. If they had seen a similar patient before (during their education or practice), they may be more likely to use pattern recognition to identify the patient's condition. Future work can elucidate this by asking participants to report how much familiarity they have with the patient presentation and symptoms before they proceed with determining a diagnosis.

Whilst these reasoning strategies carry some differences qualitatively, our study was used to investigate how these strategies actually manifest in differences to information seeking and confidence. We found that HD reasoning was associated with reevaluating the diagnoses considered more often when compared to the other approaches, as well as higher diagnostic accuracy. This is different to the results to Coderre et al. (2003), who found PR was associated with accuracy. We would interpret our findings as HD being a 'better' approach for medical students and PR would be more suitable for experienced clinicians who have more cases to draw from. Our findings suggest that the reasoning strategy used by a medical decision maker should not only be appropriate for the case at hand, but also appropriate for the decision maker's experience/expertise (either for cases of that nature or in terms of general medical experience). This implies pattern recognition should be employed when possessing enough experience to use it accurately. On information seeking, whilst we do not find evidence for differences in information seeking when looking at individual cases (in this think-aloud study), we do find evidence for reasoning strategy affecting information seeking when considering whether a case/condition tends to be approached with a HD or PR approach by a majority of students. We observed higher information seeking and informational value for PR-dominant cases when compared to HD-dominant cases. Given that we averaged across individual cases/conditions however, we exercise caution in interpreting this findings, as difference in information seeking could also be a result of these being different cases as opposed to involving reasoning strategies. Disentangling reasoning strategy from the medical conditions being treated is where looking at individual reasoning strategies on each case is useful, though we would require a much larger sample size in our think-aloud study to investigate this. We were unable to reliably predict reasoning strategies on an individual-case level for the online study where we did not have access to think-aloud utterances, which was likely a result of our think-aloud study dataset being both underpowered and imbalanced in terms of the incidences of reasoning strategies. Looking at reasoning strategies on a case-by-case level with

larger samples (and different levels of expertise) would be incredibly useful and we recommend future work to adopt such a methodology to do so, as the think-aloud methodology provides useful insight into the diagnostic decision process.

Based on our qualitative findings, we provide support for findings from our previous study. Firstly, several participants reported progressively investigating patient symptoms based on the patient's history and their initial set of diagnostic differentials. This corresponds with our finding in the previous study that the number of initial differentials considered based on the patient history was predictive of information seeking and changes in confidence. This supports evidence for the large weighting on early information received by clinicians, especially to do with history taking, because early information is responsible for the initial set of diagnoses that then guide subsequent information seeking. Secondly, we find a qualitative theme that participants report certain information being standard to seek regardless of the patient case. This corresponds with the finding from our previous study that lower information seeking variability was associated with higher accuracy, with certain information requests in our task being associated with higher accuracy when sought across cases. This further corroborates our evidence for a degree of standardisation in information seeking being useful for maximising diagnostic accuracy. In particular, we had found in Study 2 that lower variability in information seeking was associated with higher accuracy, with specific pieces of information (e.g. full blood count, assessing extremities) being useful to seek regardless of the patient's condition. We can then find an agreement between these studies that diagnosis is a decision process where there is considered to be 'optimal' information to seek for any given patient. Future work from medical professionals should focus on designing cognitive aids for prompting information seeking, such as providing a 'checklist' of what information should be needed for particular conditions. Whilst this is not feasible for all possible patient conditions, some standardisation in information seeking would be beneficial for diagnostic accuracy. Finally, we note that a number of participants mention that they aware of 'anchoring bias' and that they should avoid it. The conscious

attempts to avoid narrowing on a diagnosis too early could explain our findings of differentials rarely being removed consideration. There has been past work studying how to reduce such instances of 'premature closure' (Voytovich, Rippey & Suffredini, 1985, Eva & Cunnington, 2006, Krupat et al., 2017), with the former paper finding that experienced clinicians were more susceptible to settling on a diagnosis too early than those who were less experienced. As this work was conducted using controlled experimental paradigms, they do not consider environmental factors that may exacerbate tendencies toward anchoring/premature closure (e.g. work stress, busyness in managing multiple patients at once). Future work should hence study how such anchoring arises and interventions that can mitigate it.

Finally, we identified the dominant reasoning strategy for each of the patient conditions used in our vignettes, and then investigated their effect on accuracy within our previous study's (larger) dataset. We sorted cases into two groups: HD-dominant (where HD was used by a majority of students in the think-aloud study) and PR-dominant (where PR was used by a majority of students in the think-aloud study). Whilst we observed higher accuracy and higher initial diagnostic breadth for HD-dominant cases, we observed higher information seeking and informational value for PR-dominant cases. It seems likely that as a result of our task design, naming more diagnoses increases the chance that a correct diagnosis is mentioned. A HD reasoning strategy being associated with greater diagnostic breadth corresponds with the nature of HD being that of considering a broad set of differentials to either add to or subtract from. Higher information seeking for PR-dominant cases is surprising however, as we may have expected a PR strategy to result in selective information seeking in order to confirm a focal diagnosis. An alternative explanation could be participants seek more information because they are less able to generate a broad set of differentials. However, we find through modelling that accuracy was highest when participants had initial diagnostic breadth on a PR-dominant cases (whilst increasing diagnostic breadth had less of impact on accuracy). Whilst reasoning strategy does seem to vary as a function of the patient condition, an optimal strategy

for diagnostic accuracy seems to be generating a larger set of initial differentials and then selectively choosing from this set the diagnosis that closely resembles the patient's symptoms and observations. In other words, medical students performed best by starting broad and then narrowing their differentials. This is predicated on students and clinicians being able to identify plausible diagnoses early on, whilst remaining open minded to other possibilities. An assumption we make from this result however is that the dominant strategy for each patient conditions is the same in the think-aloud study as it was in the online-study. As our multinomial classifier was unable to accurately predict strategies for individual cases in the online study, we are unable to verify this assumption. This is likely because of the relatively small and unbalanced think-aloud dataset used to train the classifier. This could be rectified by future work that uses think-aloud methodologies for larger scale data collection for quantitative analysis such as this.

These two studies together provide a nuanced and in-depth look of the diagnostic process as demonstrated with our vignette-based task. We should consider however the generalisability and ecological validity of these studies. By using a vignette-based paradigm, participants do not actually interact with, observe and treat a patient. We are also limited in terms of the information that is available for clinicians to seek. In addition, participants completed the studies in relatively controlled environments, outside of their usual medical context. Our next study hence aims to study the link between information seeking and confidence, but with a more naturalistic paradigm. To alleviate these concerns of generalisability, we require a paradigm that allows for more open-ended information seeking, observation of a patient that can be treated during cases and the use of a clinical environment akin to the one in which clinicians operate. As previously explored in our systematic review, the use of in-situ research lacks objective markers of accuracy that we utilise. To this end, we use virtual reality (VR) in our next study to simulate a realistic medical environment, as well as the patients themselves. This allows for a realistic, interactive paradigm where participants observe a (virtual) patient in real-time and

can administer treatment (and observe reactions to this treatment in the patient). There is also more openness in terms of the information that can be sought and clinical actions taken, making its use more analogous to real medical contexts.