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

Introduction

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 their information seeking patterns and changes in confidence over the course of diagnostic decisions. We utilise a very similar experimental procedure to our previous online study (using the same patient vignettes), but instead prompted students to think out loud as they were performing the task. Everything that was said by participants were audio-recorded, transcribed and then coded for both quantitative and qualitative analysis. We also investigate why reasoning strategies may vary on a case-by-case basis.

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 number of differentials reported at each information stage increasing over each case on average. 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 that participants to use a 'process of elimination', which would manifest in decreasing numbers of differentials 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 the narrowing the 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 encouraged to disregard differentials within our paradigm.

Another focus of this study is to understand how differences in information seeking patterns arise. One possibility is that these differences stem from different reasoning strategies when making diagnoses. Clinical reasoning is a key skill that is taught, either explicitly or implicitly, within medical education. Medicine is based in uncertainty and reasoning as a skill is taught to navigate this uncertainty. However, clinical reasoning has a broad remit and covers multiple different methods by which a decision can be reached. Doctors may have different considerations when diagnosing a patient. As we noted in the design of the 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 diagnostic possibilities. Diagnostic decisions have traditionally been thought of as 'ideal' when using the hypothetico-deductive process (Kuipers & Kassirer, 1984), whereby hypotheses are formulated based on patient symptoms and established criteria for a diagnosis, with further information gathering to test these hypotheses (Higgs et al., 2008) or eliminate others. This theory was challenged by the results of Coderre et al. (2003), who 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. A hypothetico-deductive represents a more structured, systematic method for diagnosis where clinicians maintain a more open mind to different diagnostic possibilities, whilst pattern recognition is more directed and intuitive as a process.

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. Using these verbalisations, the researchers categorised the clinician's reasoning strategies. 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 the alternatives.

In their study, the researchers found that novice clinicians tended to adopt a HD strategy more often, whilst experts tended to use a PR strategy more. In addition, a addition a PR was associated with higher diagnostic accuracy, which in turn explains 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 possible differentials. The work of Coderre did not reveal whether these strategies result in differences in information seeking and confidence. We would expect that with a more thorough SI 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 more selective information seeking in order

to provide evidence in support of the single diagnosis being considered. Finally for a HD strategy, participants would be seeking information to rule out differentials, but it is unclear as to how selective participants would be with the information they seek.

In order to pick apart these different reasoning strategies within our vignette-based paradigm, we adopted a think-aloud methodology similarly to Coderre et al. (2003), whereby participants verbalise their thought process as they are doing the 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 with time, allowing us to investigate how diagnostic thinking develops with more information. A think-aloud methodology has previously been used to study the differences between novice and expert clinicians during diagnostic reasoning (Coderre et al., 2003). This study found a general trend that experts tend to use a 'pattern recognition' approach to diagnosis more than novices, who tended to use a 'hypothetico-deductive' process (which is aforementioned to be the 'textbook' description of the diagnostic process), but this was highly dependent on the case presented. We build on the work of Coderre et al. (2003) here to further investigate how reasoning strategies contribute to accuracy and why certain cases result in differing strategies.

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 Oxford University (including 2nd year Oxford University Graduate Entry Medical students) recruited via posters in the John Radcliffe Hospital in Oxford and via a mailing list for students managed by the Medical Sciences Division at the University of Oxford. The study was conducted onsite at John Radcliffe hospital. Participants were recruited between July 5th 2023 and December 1st 2023. Data was reviewed on an ongoing basis to cease recruitment when emerging themes were exhausted. This study was reviewed and granted ethical

approval as an amendment to our existing protocol to allow for audio recordings by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE004.

Materials

The same set of cases and a similar computer interface from Study 1 were used for this study, with the exception that participants no longer recorded their differentials in a specific screen at the end of each information gathering stage. Instead, participants' differentials were recorded as a more naturalistic part of their diagnostic process as they reported aloud their thoughts as they worked through each diagnostic case. The study was conducted onsite using a laptop, with actions on screen recorded on video and the audio of participants' thinking aloud recorded via a microphone. Informed consent was obtained anonymously using an online electronic information sheet and consent form. Information, including experimental data and audio recordings, collected during the study were stored under anonymised IDs with no linkages to participants. Data was kept on a password-protected computer and hard drive.

Procedure

The general procedure was very similar to that of Study 1, except that participants were given the following instructions at the start of the study:

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

The experimenter occasionally prompted participants with content-neutral probes: *"can you tell me what you are thinking?"* in cases of periods of long silence, and *"can you tell me more?"* when the participant said something vague that may warrant further detail. We emphasise that these are non-leading questions. The audio of the participants' verbalisations was recorded and then transcribed. An initial transcript was generated using Microsoft Office's transcription feature, but the transcript was checked and modified for accuracy by listening through the audio recordings again. The screen of the experimental interface was also recorded, such that the audio could be linked to specific actions within the task. The focus of this study is on verbal utterances rather than any non-verbal or inferential aspects of the participants' qualitative data. 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 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.

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

- **Differential Evaluation**: any time that the participant (each of the following is considered a separate subcode):
- – *Differential Added:* - Mentions a new condition that they are considering
- – *Differential Removed:* - Rules out or eliminates a condition from consideration

- – *Likelihood Increased:* - Mention of increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition

- – *Likelihood Decreased:* - Mention of decreased likelihood of a previously mentioned condition, or that information seems to contradict with a condition

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

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

- **Hypothetico-Deductive Reasoning (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 of 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. Conflict resolution led to changes made to the coding criteria by prioritising strategies used early in a case, as some participants were noted to utilise multiple strategies within a single case, as well as allowing some cases to be coded as not having a clear strategy due to a lack of utterances. The coding criteria provided above is after these changes were made. Conflicts were then resolved with these updated criteria. Both coders agreed on 78% of cases when coding for correctness, with conflicts resolved in consultation with a member of expert panel used to develop the vignettes (as mentioned in Study 2).

We hypothesise not only that this think-aloud methodology can be used to accurate detect different reasoning strategies but also usage of these reasoning strategies will vary. There are two possible ways in which they vary: they may vary as per 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 cases 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. Based on this, 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 based on how similar the cases' reasoning strategies are to the subjective and dominant strategies. We then compare our dependent variables (Accuracy, Confidence, Information Seeking, Differential Evaluations) as a function of reasoning strategy.

Qualitative Thematic Analysis

The aim of this thematic analysis was to identify the reasoning strategies that medical students in this study reflectively report using in their medical practice, as well as understanding considerations made by students when making diagnoses. Similar to the think-aloud utterances by participants during the vignette task, we record and transcribe the responses given by participants to our 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 response, 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 (e.g. interview recordings).

Recoding of Reasoning Strategies in Online Vignette Study

We aim to determine, given medical students' reasoning strategies, whether these strategies are associated with differences in information seeking and confidence. To improve the robustness of our findings, we not only look at quantitative differences by reasoning strategy within this study, but also using the previous study's dataset (where we have a larger sample size, as well as access to recorded differentials). In order to apply reasoning strategies to the data from Study 2, we train a classifier using penalised multinomial regression to classify cases as HD, PR or SI using the cases from the think aloud study (with Leave One Out Cross Validation). The input parameters for the classifier are the 29 pieces of information as binary predictors (similar to the approach depicted in Figure 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 nnet package (version 7.3-19). The testing data is then labelled with predicted strategies using R's predict function.

Based on these predicted reasoning strategies, we first test for differences in information seeking, differentials and confidence between the HD and PR cases in particular, as they represent relatively opposite approaches for making diagnoses (i.e broad vs narrow differentials respectively). We test for these differences using linear mixed effects models that control for participants as random effects, given that the samples from each group are not independent (as participants contribute multiple data points). We also fit linear mixed effects model to cases belonging to each strategy to predict changes in confidence (i.e. the difference in reported confidence at the Patient History and Testing stages). We hypothesise that based on the reasoning strategy employed, reported confidence would be a readout of different aspects of the diagnostic process, be it the differentials considered or the information sought. These models are defined as follows (variables are defined in the Data Analysis section of the previous study):

Confidence Change = Number of Initial Differentials + Proportion of Information Seeking + Change in Differentials

Random Effects: Participant, Patient Condition

We also aim to determine how reasoning strategies affect diagnostic accuracy (defined as the likelihood assigned to the correct differential if included. We hypothesise that the efficacy of a reasoning strategy is dependent on how many differentials are generated based on the patient's history, given that reasoning strategies have differing focuses on broadening or narrowing the differentials considered. To this end, we fit a linear model that predicts Accuracy by an interaction between the Number of Intial Differentials and Reasoning Strategy. This is because the initial number of diagnoses should, in theory, dictate the type of reasoning strategy that clinicians should use. With a broad set of differentials generated from the patient history, clinicians would be expected to focus on narrowing their set in order to determine a more definitive diagnosis (via a PR approach), whilst they would be expected to focus on broadening their differentials given fewer initial differentials in order to cover more possibilities (via a HD approach).

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 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 0.5733333 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 = 6.57, SE = 4.63 t = 1.42, p = 0.159), 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 (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, $p < .001$). 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, $p = 0.73$).

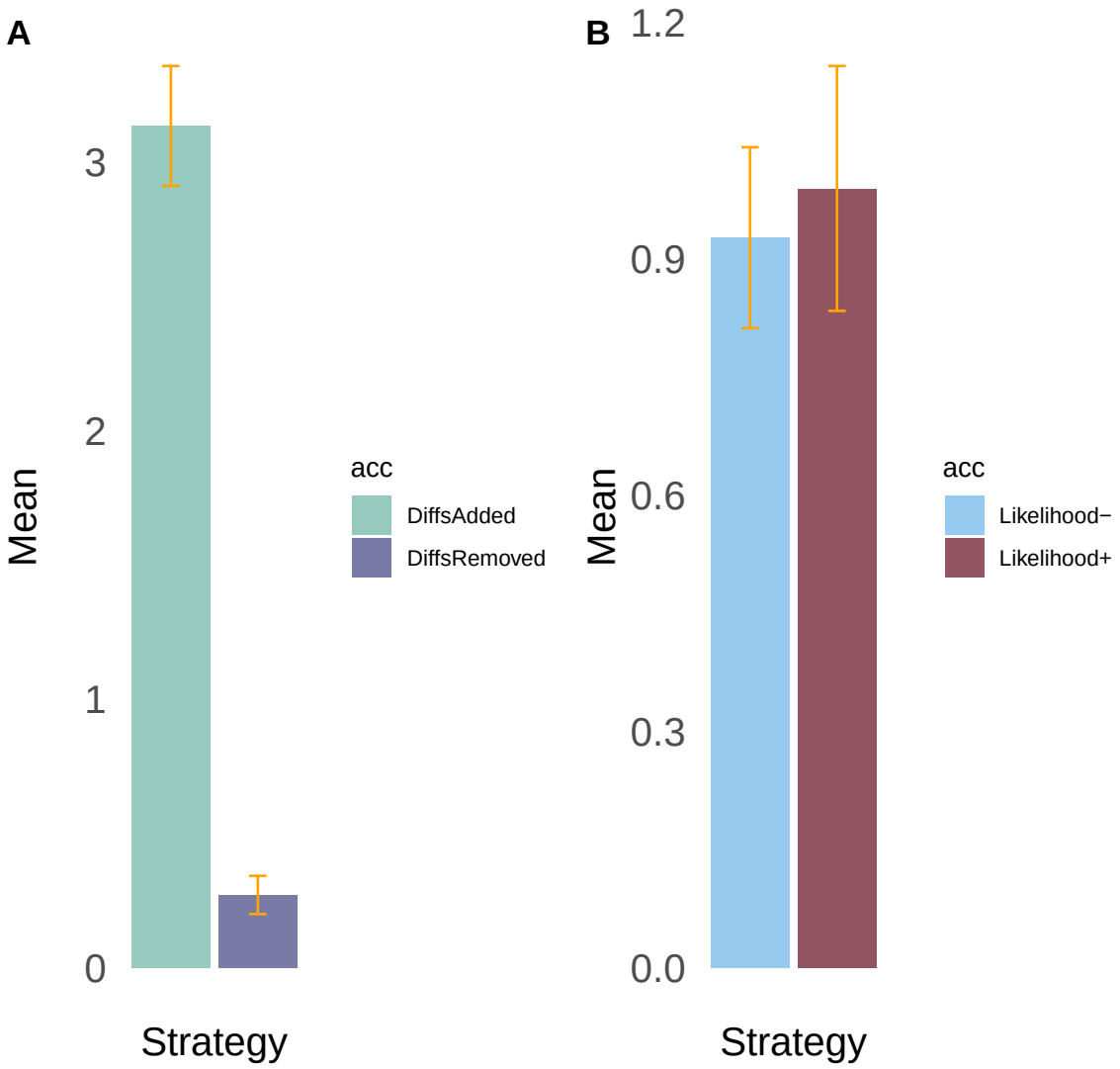

Figure 1: Bar graphs comparing incidences of each of the four subcodes within Differential Evaluations. We compare instances of differentials being added (green) and removed (purple) from consideration (Figure 1A) and compare instances of differentials decreasing (blue) and increasing (red) in likelihood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

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.

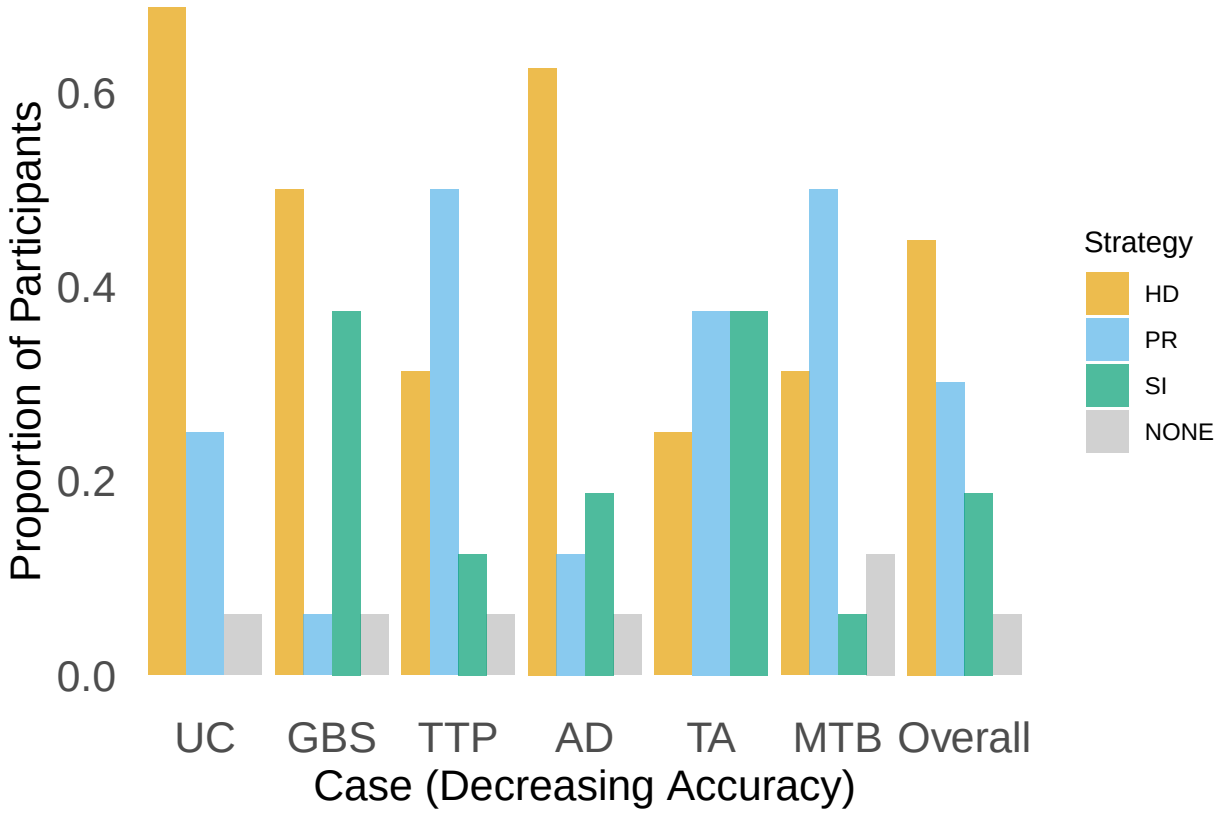

Figure 2: Proportion of participants who use each type of reasoning strategy for each condition/case, with the overall proportions across all cases shown by the rightmost bars. The strategies shown are: Hypothetico-Deductive (where multiple differentials are considered simultaneously, orange), Pattern 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).

Reasoning Strategies' Effect on Dependent Variables

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

Table 4: 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 4 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). 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. Information seeking was fairly consistent across reasoning strategies.

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. Overall, participants matched the dominant strategy on 51 cases (56.67% of cases, excluding those cases without a clear reasoning strategy). 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 participant as a random effect) ($\beta = 0.54$, SE $= 0.44$ t $= 1.23$, p $= 0.22$).

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 5: 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 6 below, we categorise participants based on their subjective reflection of their diagnostic process. Through this process, we categorise 0 participants under a HD reasoning strategy, 0 participants as PR and 0 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. 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. Accuracy was found to be higher for cases when participants matched their subjective strategy (0.68) compared to when they did not (0.58) (see Table 5 below). However, this difference was not found to be significant via a mixed effects logistic regression (on accuracy as a binary outcome measure with participant as a random effect) ($\beta = 0.59$, SE = 0.57 t = 1.03, p = 0.3).

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 6: 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 7 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

(continued)

Participant	Full Quote	Coded Strategy	Condensate	Interpretation
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 7: 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.

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 consideration alternative diagnoses early on in their decisional process. 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 more experienced clinicians and may focus on diagnoses that they have more familiarity with:

"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)*

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:

"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.

"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, six participants (including the quote above from participant ytpshg) noted that there are pieces of information or test that would seek for all patients (regardless of their condition) as part of a routine diagnostic approach:

"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)

Challenges of the Diagnostic Process

Participants cited a number of challenges related to our diagnostic task. These were related to ways in which study that 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

Coded Strategies

Strategy	Number of Cases	Accuracy	Information Amount	Information Value	Change in Confidence	Initial Diagnoses	Change in Diagnoses
HD	187	0.50	0.58	2.09	0.16	3.48	0.63
PR	100	0.40	0.54	2.20	0.15	2.64	1.10
SI	223	0.35	0.67	2.49	0.18	3.22	1.09

Table 8: Dependent variables in the Study 2 dataset broken down by the reasoning strategy assigned to each case via a multinomial classifier.

We show a breakdown of cases by their coded reasoning strategy in Table 8. We trained our multinomial classifier using all cases with a coded reasoning strategy in the think-aloud study. When comparing predicted strategies from our model with the actual strategies assigned by the independent coders, our model had an accuracy of 0.75, which is higher than the chance level of 33% accuracy (given that there are three possible reasoning strategies that could be assigned). This indicates that our model is accurate at identifying reasoning strategies in the training data.

We now look to compare our key dependent variables by strategy, in particular comparing PR and HD cases. We observed higher accuracy for HD cases (M = 0.5, SD = 0.36) on average when compared to PR cases (M = 0.4, SD = 0.34) (β = 1.01, SE = 0.43 t = 2.38, p = 0.02). In line with our expectations based on our definitions of the reasoning strategies, we observed more initial differentials were considered during HD cases (M = 3.48, SD = 1.75) relative to PR cases (M = 2.64, SD = 1.31). Change in confidence was not significantly different between PR and HD cases (p = 0.78) and neither was the amount of information seeking (p = 0.6).

Predictors of Confidence

As discussed in the Data Analysis, we next look at predictors of changes in confidence within each strategy's cases. We fit a linear mixed effects model for cases categorised under each of the three reasoning strategies that predict Confidence Change using fixed effects of the Number of Initial Differentials, Change in Differentials and Amount of Information Seeking. We use participant and patient

condition as random effects. We find that for HD cases, there was a marginal effect of Change in Differentials on Confidence Change (β = 2.76, SE = 1.49, t = 1.86, p = 0.06). For PR cases, we find that the Number of Initial Differentials was predictive of Confidence Change (β = 0.26, SE = 1.14, t = 0.23, p = 0.82). Finally, for SI cases, we find that the Amount of Information Seeking was predictive of Confidence Change (β = 5, SE = 10.33, t = 0.48, p = 0.63). All other coefficients for these models were not significant (ps > .1).

Predictors of Accuracy

We next look at reasoning strategy as a predictor of accuracy. In particular, we hypothesise that reasoning strategy has an interaction effect with the Number of Initial Differentials which then affects accuracy. This is based on the assumption that it makes more sense for clinicians to narrow or broaden their set of diagnostic hypotheses based on if the patient presentation and history brings to mind a larger or smaller number of initial hypotheses. Below in Figure 3, we plot fitted regression lines that show the relationship between Initial Differentials and Accuracy. We fit a mixed effects model to predict Accuracy in cases from Study 2 via an interaction effect between the Number of Initial Differentials and Reasoning Strategy. We test for an interaction effect using stepwise model comparison. Starting from a null model with no predictors and fixed intercept, we compute chi-square and change in AIC for each increase in model complexity. Table 9 below shows the model comparison metrics for each added model parameter. We find marginal evidence for an interaction effect between Initial Differentials and Reasoning Strategy.

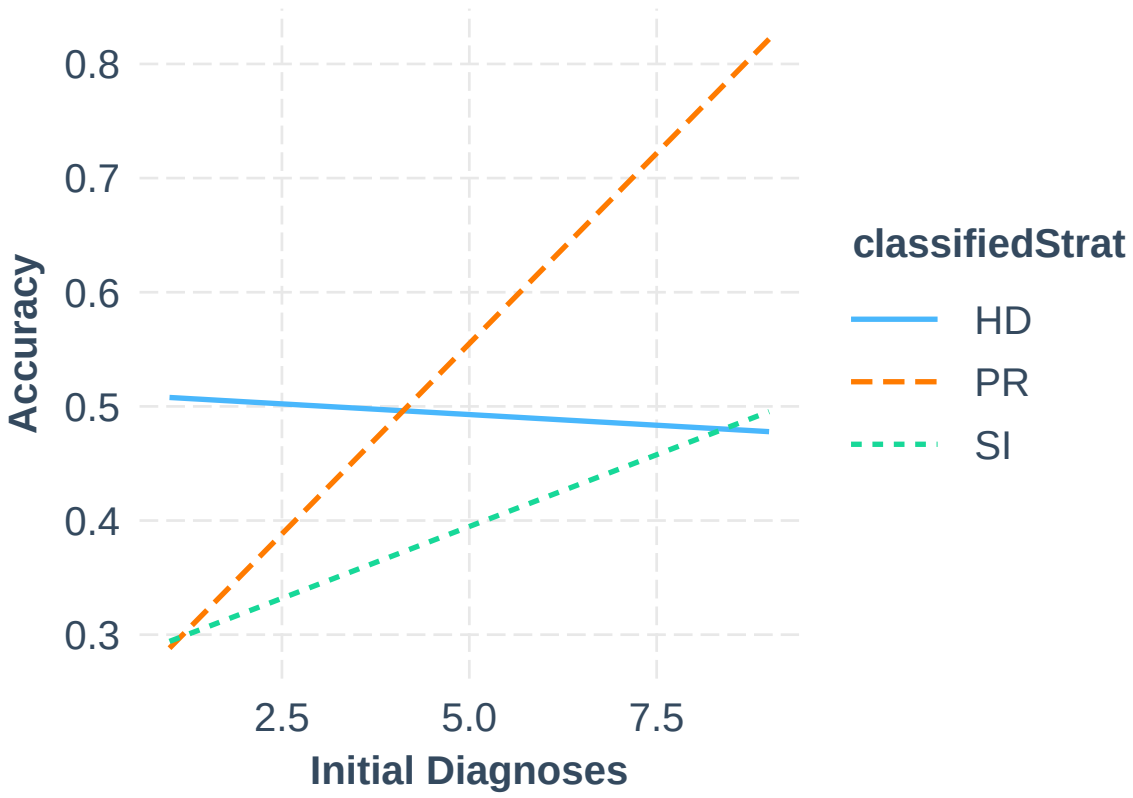

Figure 3: Fitted regression lines for model to explore the interaction between the number of initial diagnoses/differentials (x-axis) and reasoning strategy (as per our multinomial classifier, HD = blue, PR = orange, SI = green) for the online dataset from Study 2 and this interaction's effect on diagnostic accuracy (y-axis).

	Num. Parameters	AIC	BIC	Log Likelihood	Chi Square	p value
Null Model	3	379.20	391.90	-186.60	NA	NA
Strategy Main Effect	4	376.30	393.24	-184.15	4.90	0.03
Diagnoses Main Effect	5	364.16	385.33	-177.08	14.14	0.00
Strategy + Diagnoses Main Effects	6	362.76	388.17	-175.38	3.40	0.07
Strategy + Diagnoses Interaction	8	360.82	394.69	-172.41	5.94	0.05

Table 9: Incremental model comparisons of linear mixed effects model to predict Accuracy using the Number of Initial Diagnoses and Reasoning Strategy as fixed effects and participants as a random effect. From top-down, the model forms (as expressed in model syntax) are as follows: 'Null Model' <- Accuracy ~ 1 + (1|Participant), 'Strategy Main Effect' <- Accuracy ~ Strategy + (1|Participant), 'Diagnoses Main Effect' <- Initial Differentials + (1|Participant), 'Strategy + Diagnoses Main

*Effect' <- Strategy + Initial Differentials + (1/Participant), 'Strategy + Diagnoses Interaction' <- Strategy * Initial Differentials + (1/Participant). From left to right, we show values for each model: Number of model parameters (complexity), Aikake Information Criterion, Bayesian Information Criterion, Log Likelihood, Chi Square, p value of Chi Square test.*

Discussion