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. We utilise a very similar experimental procedure to our previous vignette-based study, but instead prompted students to think out loud as they were performing the task. Everything that was said by participants were audio-recorded, trascribed and then coded for both quantitative and qualitative analysis.

Opening up differentials and not ruling out What are the different reasoning strategies How do we expect strategies to affect information seeking

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.

0.0.1 Research Questions

In this study, we investigate the following research questions:

- Do students report ruling out differentials as they seek information on patient during diagnoses?
- 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?
- What considerations do medical students report having whilst they are making diagnoses?

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 (as depicted in Figure X in the previous section). At the end of the experiment, the researcher administered a semi-structured interview to better understand what the participants feel their diagnostic reasoning approach tends to be. These questions are provided in the Appendices.

Data Analysis

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

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

- **Differential Evaluation**: any time that the participant (each of the following is considered a separate subcode):
- – *Differential Added:* - Mentions a new condition that they are considering
- – *Differential Removed:* - Rules out or eliminates a condition from consideration
- – *Likelihood Increased:* - Mention of increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition
- – *Likelihood Decreased:* - Mention of decreased likelihood of a previously mentioned condition, or that information seems to contradict with a condition
- **Information Seeking Strategies**: any time the participant expresses why they may or may not request a particular piece of information in relation to ruling out or confirming a condition.

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

- **Hypothetico-Deductive Reasoning** - prior to selecting the most likely diagnosis, the participant analysed any alternative differentials one by one through something akin to a process of elimination.
- **Scheme Inductive Reasoning** - participant structures their diagnosis by pathophysiological systems or categories of conditions (e.g., infective vs cardiovascular causes) to determine root causes of patient symptoms rather than focusing on specific conditions.

- **Pattern Recognition** - participant considers only a single diagnosis with only perfunctory attention to the alternatives, or makes reference to pattern matching when using a prototypical condition to match its symptoms against the current observed symptoms for the patient (e.g., "these symptoms sound like X" or "this fits with a picture of Y").

We first code specific statements within each case that suggested one of these strategies, and then determined which strategy was most prevalent or influential for cases as a whole such that each case was categorised under one of these strategies. In addition to coding each case under one of these strategies, we also code participants on an overall level based on their subjective perception of how they make diagnostic decisions. This is based on responses provided during the debrief interview (as described in the Procedure section). Hence, reasoning strategy codes are at the case level and also at the participant level.

Coding of utterances and case-wise reasoning strategies were conducted with a second independent coder. For reasoning strategies, initial interrater reliability was low, with both coders agreeing on 58.3% of cases. Conflict resolution led to changes made to the coding criteria by prioritising strategies used early in a case, as some participants were noted to utilise multiple strategies within a single case, as well as allowing some cases to be coded as not having a clear strategy due to a lack of utterances. Conflicts were then resolved with these updated criteria. Both coders agreed on 78% of cases when coding for correctness, with conflicts resolved in consultation with a member of expert panel used to develop the vignettes (as mentioned in Study 1).

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

Results

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

As previously mentioned, Differential Evaluations can be further categorised into one of four subcodes: Differential Added, Differential Removed, Likelihood Increased and Likelihood Decreased. As found in the previous study, there is a general reticence to disregard differentials completely. Participants expressed significantly more statements adding differentials (M = 3.14, SD = 0.89) than removing differentials (M = 0.27, SD = 0.28) ($t(15) = 14.14$, MDiff = 2.86, $p < .001$). Participants expressed more statements of decreasing likelihoods (M = 0.99, SD = 0.62) rather than increasing likelihoods (M = 0.93, SD = 0.46) but we did not find evidence of a significant difference ($t(15) = 0.34$, MDiff = 0.06, $p = .73$).

Reasoning Strategies

Next we look at our coding of reasoning strategies at a case level. As mentioned, our criteria for each code was applied to each individual case based on the transcribed utterances. When looking at reasoning strategies by case, 43 cases were coded as Hypothetico-Deductive, 29 were coded as Pattern Recognition and 18 were

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

We note, rather unsurprisingly, that we observe a higher number of average Differential Evaluations when cases are correct (M = 5.85, SD = 0.38) compared to when they are incorrect (M = 4.34, SD = 0.39). Given our methodology for defining accuracy, participants are more likely to mention a correct differential if they mention more differentials. The procedure used in the previous study for collecting data on which differentials participants were considering at each information stage was not present here and hence we are not able to operationalise accuracy in the same manner as before. While we look at which differentials are mentioned, we cannot observe how participants weigh up differentials against each other in the same way as in the first study.

To connect the results of this study to those of Study 1, we break down the same dependent variables (as operationalised in that study) by reasoning strategy. We do not apply statistics to this study due to the lower sample size. We first categorise each of the 6 cases as having a 'dominant' reasoning strategy based on which was utilised the most across participants. Through this process, we categorise three conditions as HD (AD, UC, GBS), three conditions as PR (MTB, TTP, TA). The proportions of participants who use each reasoning strategy for each condition can be

viewed in Figure 10. We then compare the individual case classifications of strategy to this reasoning strategy that is most commonly used for that medical condition. Table 2 shows how dependent variables are affected by reasoning strategy. We find that the amount of information seeking was fairly consistent across reasoning strategy, but that PR cases were associated with higher value in information seeking. In order to derive informational value, we used the same values of each piece of information for each case that were derived in Study 1. This higher informational value does not translate into higher accuracy for PR cases, though we should note that the manner in which accuracy was defined for this study limits the analysis only to statements made out loud of specific conditions rather than formally recorded differentials as we did in Study 1. In order to formally replicate this finding with the larger dataset, we use the cases from this study and the coding of strategies to apply the same coding to our online dataset from Study 1.

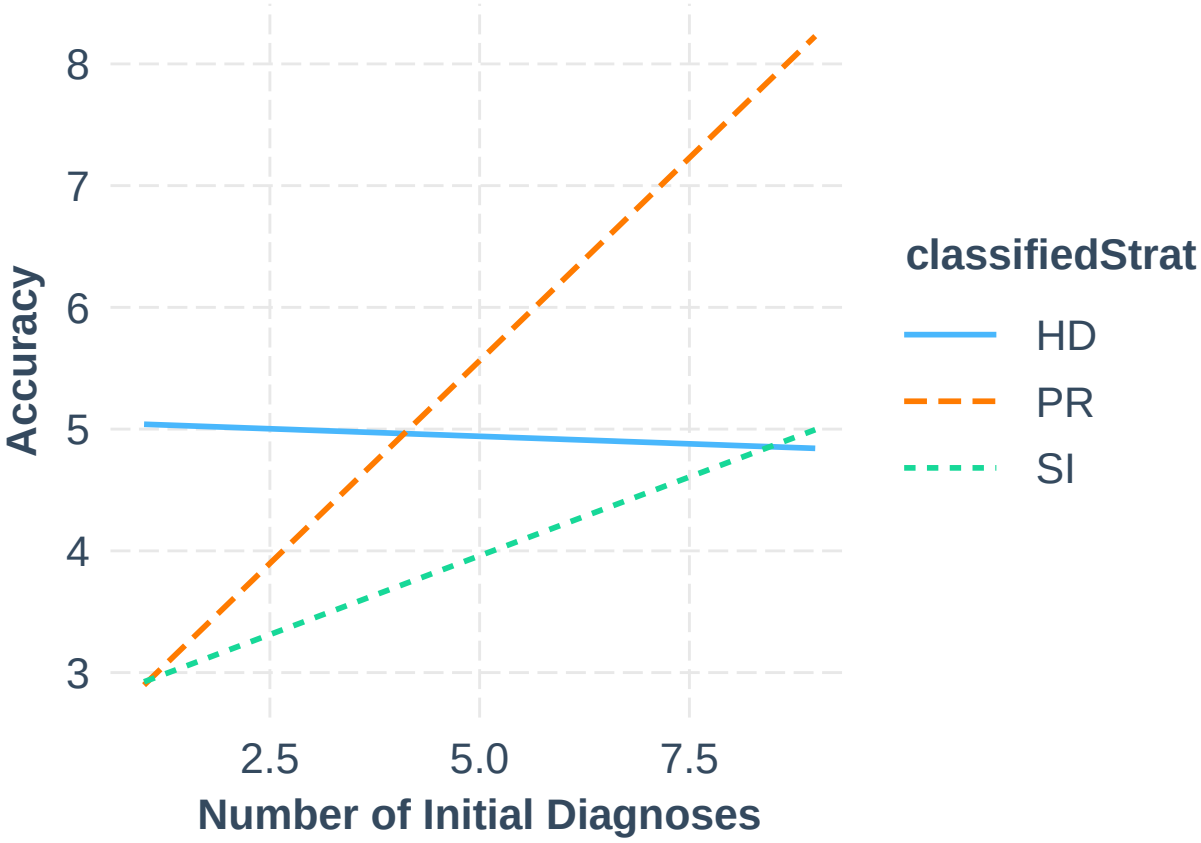
Reasoning Strategies in Study 1 Dataset

In order to apply reasoning strategies to the data from Study 1, we train a classifier using penalised multinomial regression to classify cases as HD, PR or SI using the cases from the think aloud study (with Leave One Out Cross Validation). The input parameters for the classifier are the 29 pieces of information as binary predictors (similar to the approach depicted in Figure 7) and the cases' condition. In other words, the cases from the think-aloud study make up the training data for the classifier whilst the cases from the larger online study is the test dataset. The classifier was implemented using R's nnet package (version 7.3-19). The testing data is then labelled with predicted testing strategies using R's predict function. We note that the training data was initially labelled with reasoning strategies using the think-aloud utterances and thus is separated from the information sought during the case.

We show a breakdown of cases by their coded reasoning strategy in Table 4. We now look to compare our key dependent variables by strategy, in particular comparing PR and HD cases. In line with our expectations based on the definitions of HD and PR reasoning approaches, we find that HD cases are associated with

more differentials being considered (M = 3.37, SD = 1.64) average when compared to PR cases (M = 2.84, SD = 1.58) and find evidence of a difference between the two via a Welch Two Sample t-test (t = 2.89, MDiff = 0.53, p = .004). We find that PR cases are associated with higher informational value (M = 2.35, SD = 1.07) when compared to HD cases (M = 2.15, SD = 1.32) (t = 1.48, MDiff = 0.20, p = .14). However we do find evidence of higher amounts of information seeking for HD cases (M = 0.63, SD = 0.21) when compared to PR cases (M = 0.50, SD = 0.21), (t = 5.28, MDiff = 0.13, p < .001). Overall, this indicates that PR reasoning were associated with lower but more selective information seeking when compared to HD reasoning.

We hypothesised that an interaction with reasoning strategy is associated with accuracy on the task. This is because a single reasoning strategy is considered unlikely to be more accurate for all cases. Different patient conditions seem to result in varying reasoning strategies being utilised, which begs the question of what properties of a condition contribute to changes in strategy and in accuracy. One possibility is that reasoning strategy interacts with the diagnostic uncertainty of a case (i.e. the breadth of conditions that a patient could have given their current symptoms and history, with some conditions presenting in a more apparent way than others), as captured by the number of initial differentials reported by participants. To test this hypothesis, we fit a linear model to predict accuracy with an interaction between the number of initial diagnoses and reasoning strategy.


8
7
6
5
4
3
Accuracy
2.5
5.0
7.5
Number of Initial Diagnoses
classifiedStrat
HD
PR
SI