Study 2 - Information Seeking and Confidence in Diagnosis

Methods

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

Participants

We recruited final year medical students within the UK. 85 medical students completed the study, including 32 males, 52 females and 1 participant who identified as non-binary. Their ages ranged between 22-34 years (M = 24.2). Participants were recruited between July 11th 2022 and April 6th 2023 via email sent to UK medical students via a UK Medical Schools Council mailing list. Participants were emailed with a study information sheet and a link to access the experiment, where they first provided consent via an anonymous online form. After doing so, the participant provided demographic information (age, gender and years of medical experience). The study was conducted online, with participants able to run the experiment in a browser on a desktop computer or laptop (and not a phone or tablet) in a location of their choice. The experiment was coded using the JSPsych Javascript plugin. The code is publicly available on Github: https://github.com/raj925/Diag

nosisParadigm. Ethical approval was granted by the Oxford Medical Sciences Interdivisional Research Ethics Committee under reference R81158/RE001.

Materials

This study involved patient vignettes that we adapted from anonymised past cases used by Friedman (2004). Six cases were chosen, each designed to indicate a specific underlying condition the patient had: Aortic Dissection (AD), Guillain-Barre Syndrome (GBS), Miliary TB (MTB), Temporal Arteritis (TA), Thrombotic Thrombocytopenic Purpura (TTP) and Ulcerative Colitis (UC). The order in which the cases were presented was randomised for each participant. We also included a practice case (Colon Cancer) to familiarise the participants with the experimental procedure and the interface. Cases were chosen to reflect a variety of affected pathophysiological systems and to test medical students on medical conditions that they were expected to know given their level of education/training.

A panel of 3 subject matter experts (practising doctors and researchers within the NHS and the OxSTaR centre www.oxstar.ox.ac.uk) were recruited to design the vignettes used in this study. These medical professionals were at differing experience levels, with their medical roles at the time of this study as follows: Speciality trainee (ST7) in Anaesthetics, Foundation (F1) Doctor and Gastroenterology Consultant. The panel assisted with translating terms (e.g., medication names, tests etc.) from US to UK doctors' vernacular, updated patient details to be more current and provided input on the choice and complexity of the cases chosen.

Procedure

The goal of the task was to determine a diagnosis, or diagnoses, for each presented patient (Figure 1). Information on the patient was split into a series of discrete stages to control what information the participants had access to at any given point in the experiment. each point of new information was termed an "information stage". Participants were able to seek information freely until they were ready to move on.

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

When a participant clicks on any of these requests, the information for that request is shown on screen after a 3 second delay. It was emphasised during the task instructions that participants should only request information that they believe will help them with diagnosing the patient for that specific case. Participants are free to request the same piece of information multiple times, including information from a previous stage. At any point, they can choose to stop gathering information for that stage. They are then taken to a new screen where they report a list of all differential diagnoses that they are considering for that patient at that stage. For each differential, participants report a "level of concern" for that differential, which is how concerned they would be for that patient if this differential really was the patient's underlying condition. This is reported on a 4 point scale, with labels of "Low", "Medium", "High" and "Emergency". Participants also reported a likelihood rating for each differential, ranging from 1 (very unlikely) to 10 (certain). In subsequent stages, the list from the previous stages is available for participants

to update concern/likelihood ratings, or to add/remove differentials from the list. Even at the last information stage, participants can report multiple differentials.

After recording their differentials, participants are then asked to report their confidence that they are "ready to start treating the patient" on a 100 point scale, ranging from fully unconfident to fully confident. Participants also indicate using a checkbox whether they are ready to start treating the patient, at which point a text box appears for them to report what further tests they would perform, any escalations they would make to other medical staff and treatments they would start administering for the patient. Once all three stages are complete, participants report how difficult they found it to determine a diagnosis for that case, on a scale from 1 (trivial) to 10 (impossible). At the end of all six patient cases, participants are told the 'true' conditions for all the patients. The session took approximately 40-60 minutes to complete.

Data Analysis

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

correlations between our dependent variables were tested using Pearson's product moment correlation tests (an alpha value of less than 0.05 was regarded as statistically significant). Our sample of 85 participants is calculated have 80.4% power to detect a medium effect size of $r = 0.3$ (using an approximate arctangh transformation correlation power calculation). Our key dependent variables are as follows:

Case-Wise Measures

- *Accuracy*: Our main measure of diagnostic accuracy is computed as the likelihood value assigned to the correct differential for the case (and scored as 0 if this differential is not listed). For a case to be considered 'correct', the participant should have reported the correct condition for that case within

their list of differentials regardless of the number of differentials provided. Likelihoods range from 1-10 when a correct differential is included and has a value of 0 when a correct differentials is not included. The value is then rescaled to range from 0 and 1, where 1 corresponds to a correct differential assigned maximum likelihood. If multiple differentials that are considered correct were provided, then the likelihood value of closest differential to the true condition was used.

- *Confidence*: Participants reported their confidence that they are ready to start treatment at each information stage. Initial Confidence refers to the reported confidence after the first stage of information seeking (Patient History), whilst Final Confidence refers to the reported confidence after the third and last stage of information seeking (Testing). As with accuracy, confidence is rescaled to fall between 0 and 1 to allow for direct comparison between the two variables. We can then use these two variables to calculate Confidence Change, by subtracting the participants' Initial Confidence from their Final Confidence. Hence, a positive value for Confidence Change means that the participant has gained confidence over the course of the patient case.

- *Number of Differentials*: This measure captures the breadth of diagnoses considered by participants. The number of items in the list of differentials was recorded at each stage. Initial Differentials refer to the number of differentials after the first stage of information seeking (Patient History), whilst Final Differentials refer to the number of differentials after the third and last stage of information seeking (Testing).

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

Derived Information Seeking Measures Across Cases

- *Proportion of Information Seeking*: This measure captures the amount of information that participants seek on cases relative to how much they could have sought if seeking all avilable information. We take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage, tests that had been requested before that stage and excluding repeat tests) and divide this by the number of possible tests available.

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

Results

Overall Performance

Across cases, accuracy increased with each stage of information gathering as per our Accuracy measure ($F(1, 254) = 8.52$, $\Eta^2G = 0.03$, $p = 0.004$). Participants had lower accuracy during the Patient History stage ($M = 0.31$, $SD = 0.14$) than during the Physical Examination ($M = 0.4$, $SD = 0.16$) and Testing stages ($M =$

0.41, SD = 0.15). Table 2 shows overall accuracy (at the Testing stage) by case, indicating that there was variability in performance due to cases varying in difficulty.

Calibration of Confidence to Accuracy

Confidence also increased as participants received more information (F(1, 254) = 7.93, $\Eta$^{2}G = 0.03, p = 0.005). Participants reported lower confidence during the Patient History stage (M = 0.3, SD = 0.15) than during the Physical Examination (M = 0.41, SD = 0.17) and Testing stages (M = 0.47, SD = 0.47). We note here that confidence was on average below 50% even at the end of each case, which indicates that participants were not highly confident to start treatment.

```
## # A tibble: 6 x 5
##   caseCode Proportion of Participants who Incl~1 Accuracy `Perceived Difficulty`
##   <chr>                                    <dbl>    <dbl>                  <dbl>
## 1 AD                                        0.6      0.28                    5.9
## 2 GBS                                       0.75     0.41                    6.9
## 3 MTB                                       0.42     0.24                    6.7
## 4 TA                                        0.74     0.5                     6.2
## 5 TTP                                       0.61     0.34                    6.8
## 6 UC                                        0.99     0.72                    5.3
## # i abbreviated name:
## #   1: `Proportion of Participants who Included a Correct Differential`
## # i 1 more variable: `Mean Final Confidence` <dbl>
```

Table 1: Showing statistics across participants for each case (leftmost column, AD = Aortic Dissection, GBS = Guillain Barré Syndrme, MTB = Miliary Tuberculosis, TA = Temporal Arteritis, TTP = Thrombotic Thrombocytopenia Purpura, UC = Ulcerative Colitis). Accuracy refers to the average likelihood (on a 1-10 scale) assigned to a correct differential if included. Both of these measures, as well as Final Confidence, are calculated at the final information stage of each case (i.e. the Testing stage).

When comparing Accuracy (taking into account the likelihood assigned to correct differentials) to Confidence, we find, across stages, participants' Confidence was fairly well aligned to their Accuracy (see Figure 4). To determine whether confident participants tended to be more accurate, we compared a paired t-test between Average Confidence and Average Accuracy (across cases) at each stage. There was no evidence of a difference between the two at the Patient History ($t(84) = 0.32$, MDiff = 0.006, $p = 0.75$) and Physical Examination stages ($t(84) = 0.75$, MDiff = 0.01, $p = 0.45$), but there was a statistically significant difference between the two at the Testing stage ($t(84) = 2.4$, MDiff = 0.06, $p = 0.02$). This indicated well-calibrated confidence after Patient History and Physical Examination, but a slight overconfidence across participants after Testing.

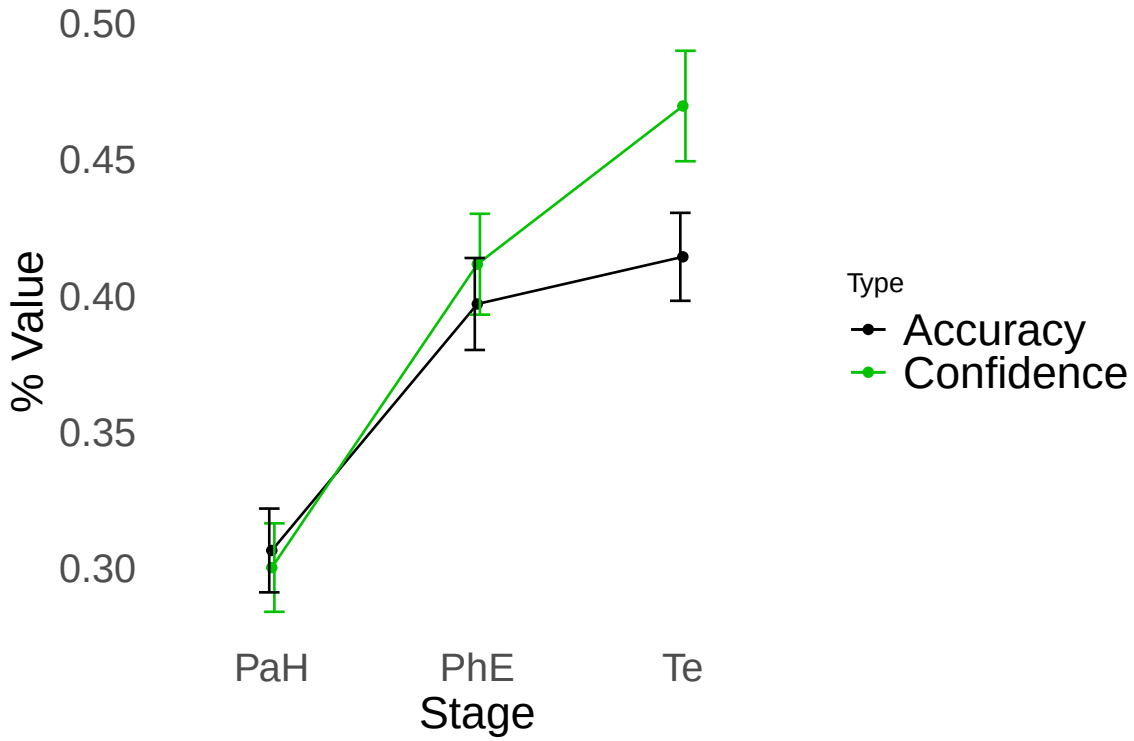

Figure 1: Graph showing Accuracy (black) and Confidence (green) at each of the three information stages (PaH = Patient History, PhE = Physical Examinations, Te = Testing).

Differentials

Analysis of the number of differentials considered by participants at each stage provides little evidence for an overall strategy of deductive narrowing towards a single differential. Instead, participants overall increased the number of the differentials they reported as they received more information ($F(1, 254) = 25.29$, $\$\backslash Eta\$^2G = 0.09$, $p < .001$). Participants reported fewer differentials during the Patient History stage ($M = 3.2$, $SD = 1.11$) than during the Physical Examination ($M = 3.88$, $SD = 1.33$) and Testing stages ($M = 4.12$, $SD = 1.43$). The majority (74/85) did not decrease the number of differentials between Patient History and Testing on any case, indicating a tendency to widen rather than narrow the set of considered diagnoses through the evolving decision process (even while, on average, growing increasingly certain of the correct diagnosis).

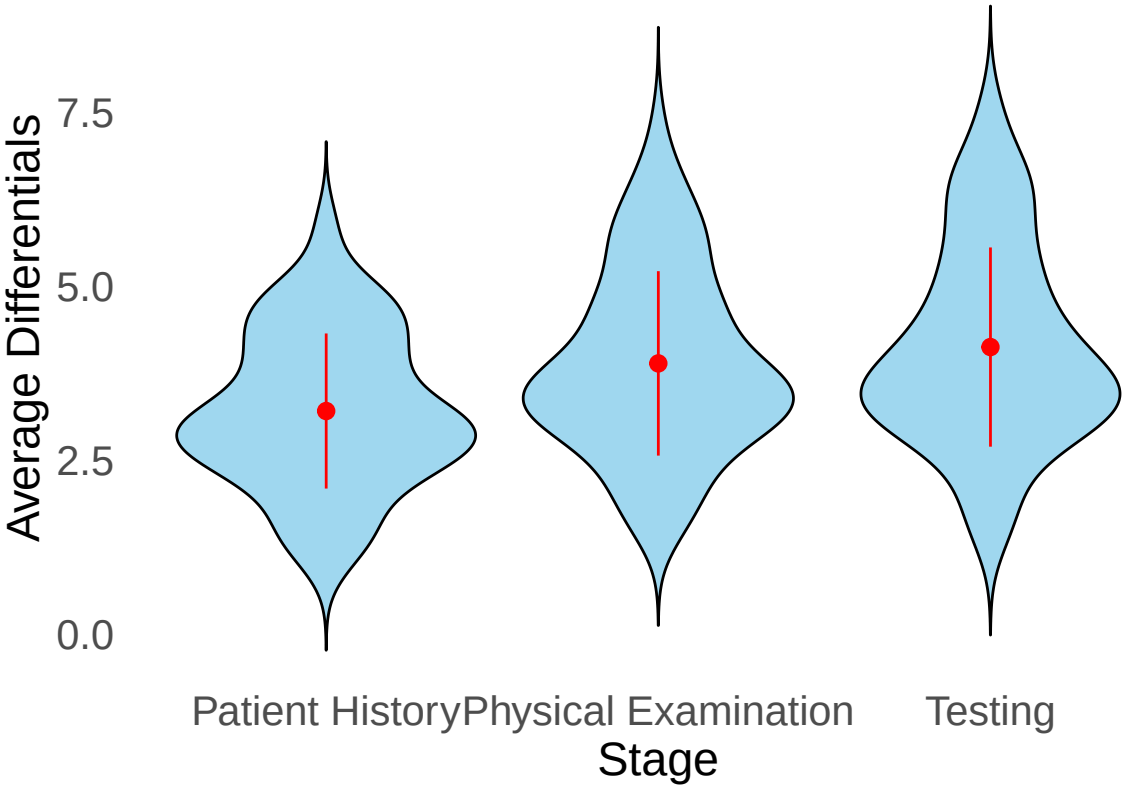

Figure 2: The average number of differentials after each stage of information seeking.

We then ask if participants who generate more differentials early in the diagnostic process go on to seek more information by conducting a Pearson's Correlation test on individual differences. We find an association (see Figure 3) between the average number of differentials generated from the Patient History and the average amount of information sought during cases ($r(83) = 0.3$, 95% CI = [0.1, 0.49], $p = 0.005$). As previously discussed, participants rarely seem to remove differentials from consideration. Therefore, one can surmise here that higher information seeking is associated with the consideration of more diagnostic differentials. We also find evidence for a positive association between the number of initial differentials and the change in confidence (i.e. the difference in confidence reported during the Patient History stage and the Testing stage) ($r(83) = 0.24$, 95% CI = [0.03, 0.43], $p = 0.03$)

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

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

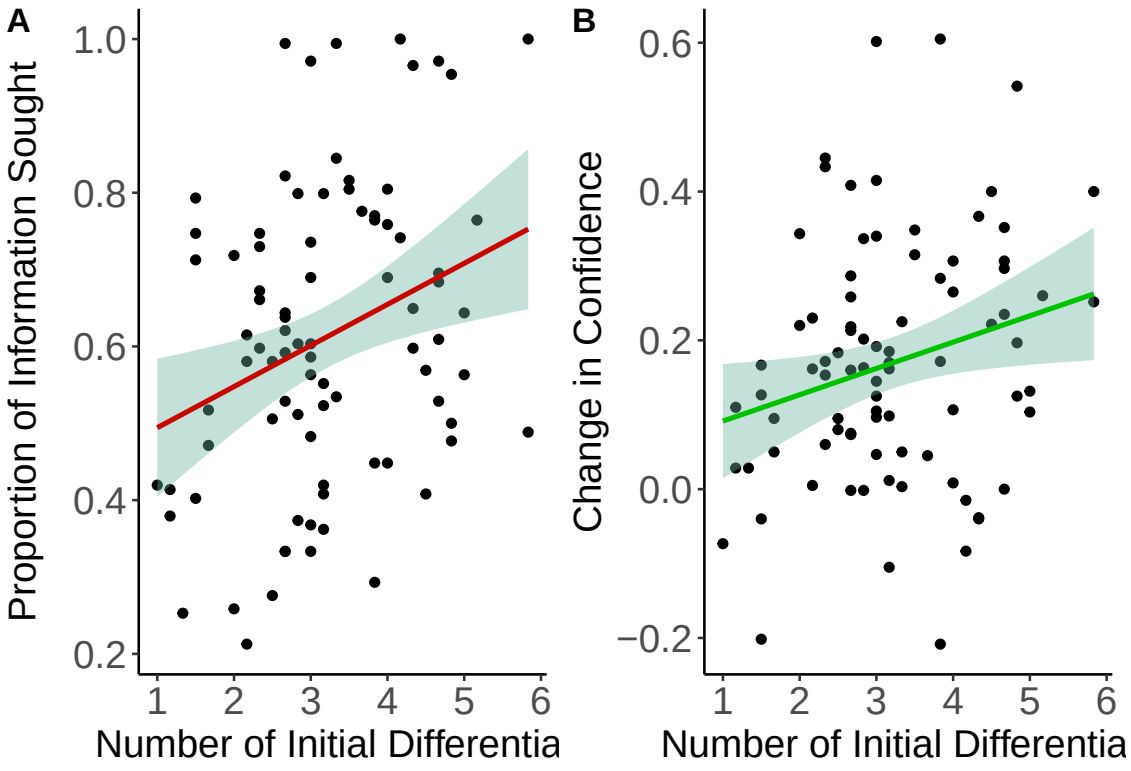

Figure 3: Scatter plot showing the relationship between the number of initial differentials reported at the Patient History stage (x-axis) and both the proportion

of available information sought (y-axis, figure 3A) and change in confidence (y-axis, figure 3B). Each point represents a single participant with all three variables averaged across the six cases that each participant performs. The x-axis refers to the average number of differentials that participants report in their list at the Patient History stage. The y-axis in 3A refers to the average proportion of available information sought, with each case containing 29 pieces of information across the Patient History, Physical Examination and Testing stages. The y-axis in 3B refers to the difference in confidence reported at the Patient History and Testing stages, such that a positive represents that the participant on average increased in their confidence over the course of the cases. The line of best fit is plotted using the geom_smooth function in R with a linear model. The shaded region shows the 95% confidence interval of the correlation.

Information Seeking

When investigating whether participants became more selective in their information seeking over the course of cases, we find that the Proportion of Information Seeking decreased with each information stage ($F(1, 253) = 100.12$, $\$\backslash Eta\$^2G = 0.28$, $p < .001$). Participants sought more of the available information during the Patient History stage ($M = 0.85$, $SD = 0.2$ than during both during the Physical Examination ($M = 0.59$, $SD = 0.24$) and Testing stages ($M = 0.5$, $SD = 0.22$).

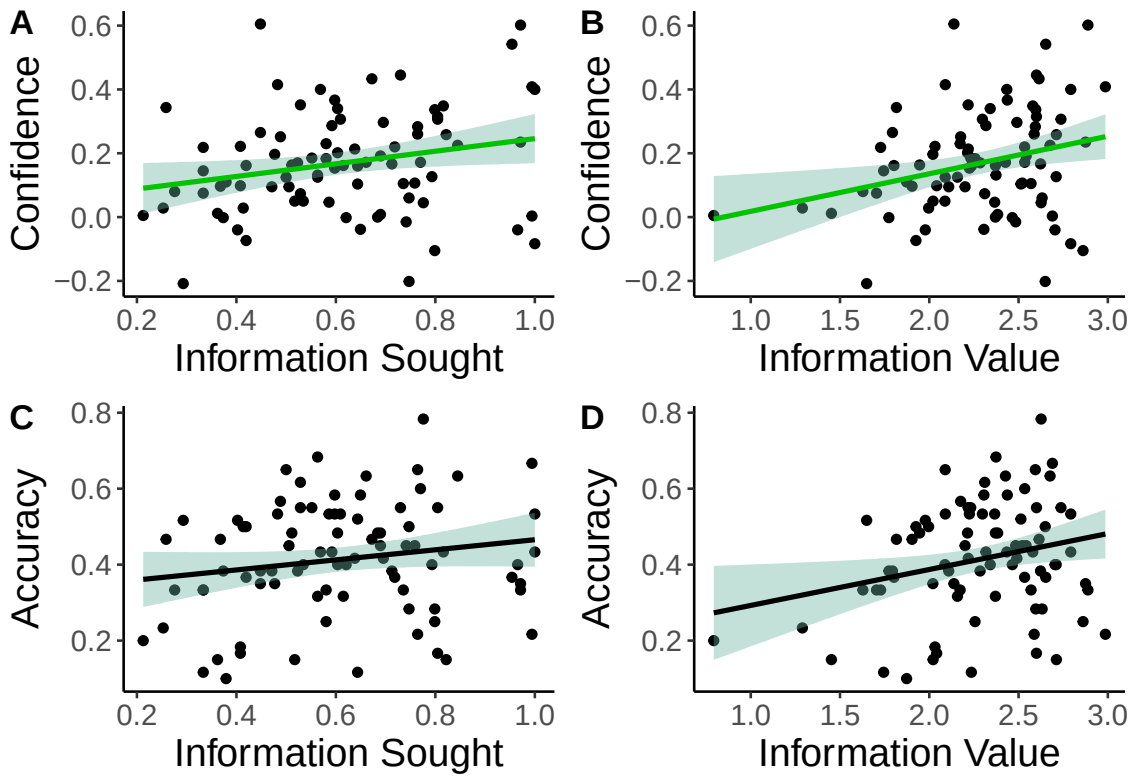

Figure 4: Scatter plots showing our information seeking variables (amount in figures 4A & 4C and value in 4B & 4D) against our key dependent variables of change in confidence (difference between final confidence and initial confidence, figures 4A & 4B) and accuracy (the likelihood assigned to a correct differential if provided, figures 4C & 4D). Information Sought refers to the proportion of available information sought across cases. Information Value refers to the sum of all mean information values across all 6 cases for a given participant. All data points are for a single participant where variables are averaged across all 6 cases they completed.

We do not find that participants who sought more information across cases were also more accurate in their diagnoses ($r(83) = 0.17$, 95% CI = [-0.04, 0.37], $p = 0.11$). However, participants who sought more information did tend to increase their confidence more over the course of a case on average ($r(83) = 0.24$, 95% CI = [0.02, 0.43], $p = 0.03$). While seeking more information may imbue students with a greater level of confidence, it does not necessarily translate into more accurate diagnoses. This links to the results presented in Figure 1, in which confidence and accuracy were related to one another but imperfectly (especially during the Testing stage).

In order to examine more specifically what differences in information seeking are driving differences in both accuracy and confidence, we look at their relationship with informational value. We assess the degree to which each participant's accuracy is predicted by the quality of the information they sought and find evidence for a positive relationship between accuracy and information value ($r(83) = 0.25$, 95% CI = [0.04, 0.44], $p = 0.02$), as well as between confidence and information value ($r(83) = 0.28$, 95% CI = [0.07, 0.47], $p = 0.01$).

The amount of information sought does not seem to be predictive of accuracy. However, it may be that there are identifiable 'fingerprints' reflected in information seeking patterns that differentiate between high and low accuracy diagnosticians. If this is the case, participants who are high and low accuracy participants could be predicted based on their information seeking patterns.

In order to test this, we investigate whether information seeking is predictive of participants who are higher or lower in their diagnostic accuracy using binary classification and receiver operating characteristic (ROC) analysis. ROC is a form of analysis that assess how well a model performs at predicting a binary outcome (in this case, whether a case was performed by a high or low performing participant). We trained a binary classification algorithm using a generalised logistic regression (GLM) model to identify if participants exhibited high or low accuracy based on the information they sought. We first split all cases into two groups by whether they were performed by a high and low Accuracy participant. This was done using a median split by participants' average Accuracy across the six cases. By doing this, we can look at whether participants who perform better at diagnoses seek information in a markedly different way to participants who performed worse.

We train the classifier using a Generalised Linear Model (GLM) by treating the 29 binary variables for each information as predictors (with a 1 signifying that the information was sought for that case and 0 when the information was not sought) to predict the binary outcome of whether the participant is a low or high accuracy participant. We used Leave One Out Cross Validation, such that each case is predicted by training the algorithm on all other cases. When plotting an

ROC curve, the area under the curve (AUC) is indicative of how well a model performs at correctly categorising cases. An AUC of 0.5 would signify that our model is performing at chance and is not able to predict participant accuracy in any meaningful way. By plotting an ROC curve for our model, we find an AUC value of 0.72 (plotted in Figure 5). When conducting a DeLong test, to test the null hypothesis that the AUC is equal is 0.5 (i.e. that the classifier is completely unable to predict high and low accuracy participants), we find $p < .001$, indicating that the AUC differs significantly from 0.5 and that the classifier is able to reliably predict high and low accuracy participants.

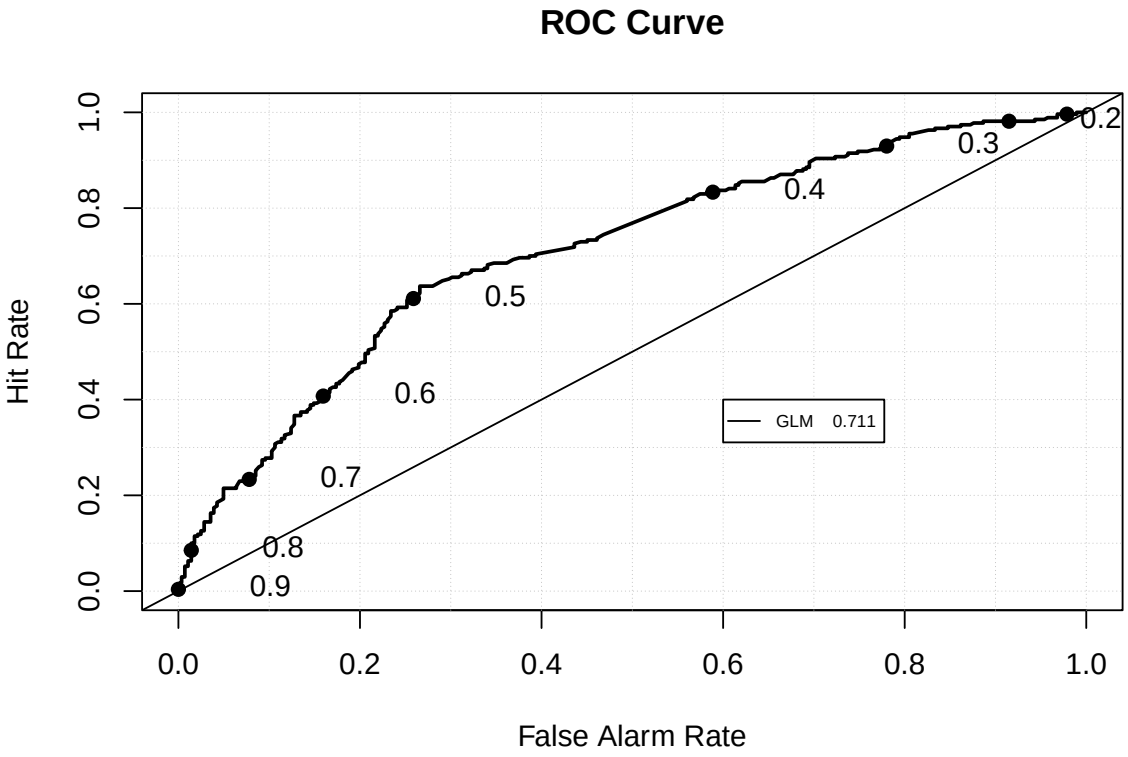

Figure 5: Receiver-Operator Characteristic (ROC) curve using a Generalised Linear Model to classify individual cases as being performed by either high or low accuracy participants. The models are trained on the raw binary predictor variables for each of the 29 available pieces of information, with 0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

This result indicates overall that differences in information seeking are indeed predictive of a difference in participant ability at above chance at a broad level. Essentially, information seeking patterns separate high and low accuracy participants, but this analysis does not tell us what aspects of information seeking in particular are predictive of accuracy. We next seek to characterise the specific differences in information seeking that contribute to higher diagnostic performance.

Discussion

This study of 85 medical students explored the interplay between confidence and information seeking in a novel medical diagnosis task. Using a novel online interface, we explored how medical students work through diagnostic scenarios via information seeking to develop and test sets of possible differentials.

We found that participants become more accurate as they received more information, though cases varied in difficulty as reflected in participant accuracy. In particular, the AD and MTB cases were more difficult based on lower observed accuracy across participants. Using our measure of accuracy, which is obtained by using the likelihood values assigned to correct differentials (if included), we find that accuracy tracks confidence quite closely at each information stage. Participants exhibited a general pattern of broadening the differentials they were considering as they received more information. The initial breadth of diagnoses considered from the patients' history was seen to be predictive of subsequent information seeking and changes in confidence. Relatedly, information seeking and confidence was associated, such that participants who sought more information tended to increase their confidence more over the course of the diagnoses. However, the amount of information sought was not predictive of diagnostic accuracy, which was instead associated with seeking more valuable/appropriate information for a given patient condition.

Previous work (e.g. Meyer et al., 2011) have noted a gap between subjective confidence and objective accuracy. In particular, there has been demonstrated to

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

The main strength of this study's paradigm is in allowing us to investigate the diagnostic process as an evolving process over time and with more information, rather than as a single decision at a single point in time. By tracking how both confidence and the diagnoses considered by participants changes over time, we gain a better understanding of how the manner in which information sought is key to the diagnostic process.

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

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