Chapter 3 - Information Seeking and Confidence During Medical Diagnoses

Introduction

In the previous chapter, we presented a systematic scoping review of the extant literature on certainty and confidence during medical diagnoses. One output from this review was that we identified one gap in the literature in that past work has not studied the association between the ongoing receipt of information and confidence. Past work has tended to frame information seeking as a single action/choice taken after diagnosis, rather than an ongoing activity that causes regular reconsideration of a diagnosis and course of treatment. In this chapter, we aim to fill this gap by presenting results from an empirical study that investigates the interaction between confidence and information seeking during medical diagnoses.

In this study, we aim to retain the control and simplicity of vignette-based approaches while incorporating some of the complexities that characterise real diagnostic decision making. By doing this, we aim to study diagnostic confidence and accuracy as it develops over the course of a diagnostic decision. In this chapter, we first introduce our paradigm and its flexibility in allowing free information seeking and updating of diagnostic differentials over time. We then introduce the research questions that such a paradigm allows us to investigate.

Our systematic scoping review on confidence during diagnoses revealed two main findings. Firstly, past work that measured confidence and accuracy during diagnostic decisions found that confidence was rarely calibrated to objective accuracy, leading

to overconfidence (Friedman et al., 2005, Fernández-Aguilar et al., 2022, Garbayo et al., 2023) or underconfidence (Mann, 1993, Yang, Thompson & Bland, 2012, Brezis et al., 2019). Secondly, confidence was associated with many aspects of the patient care process, such as prescriptions (Levin et al., 2012, Garbayo et al., 2023), referrals (Calman, Nyman & Licht, 1992) and requesting investigations (Tabak, Bar-Tal & Cohen-Mansfield, 1996, Gupta et al., 2023). The lattermost of these factors is of particular relevance to our research interests. As we identify in our scoping review, the link between information seeking and confidence is a current gap in the extant literature. This link currently made between confidence and information seeking in the literature is that confidence is a subjective judgement that then guides a clinician's subsequent testing and requests for information (Tabak et al., 1996, Gupta et al., 2023).

Crucially, this past work studies information seeking by asking participants/clinicians a single question of whether they would (hypothetically) seek further information or not. However, the link between confidence and information seeking can be expanded upon in three ways. Firstly, we can study how information seeking prior to the point at which confidence is reported affects this confidence. Secondly, we can look at specific aspects of information seeking that are linked to confidence aside from merely the intention to seek or not seek. This includes the amount of information sought, how relevant the information is to the patient case, and the degree to which clinicians vary their information seeking on a case-by-case basis. Finally, confidence and information seeking are likely to influence each other over time. Past work has tended to study diagnosis by asking clinicians to provide a single diagnosis/condition after delivering all available information. Whilst this is a useful simplification for the sake of empirical study, it leaves open the key aspects of how the diagnostic process unfolds in real clinical work. In everyday practice, clinicians engage with diagnosis as an active, ongoing decision process that develops with more time and as more information about the patient becomes available. With a more open-ended paradigm, we investigate how diagnoses evolve over time. For

instance, does a clinician reach an initial diagnosis and then change their mind when they received unexpected information (e.g. a test result)? And does a clinician tend to have a single diagnosis in mind or do they tend to keep an open mind by having considering several diagnostic possibilities at once? Then, with these questions in mind, how does a clinician seek information to further validate their diagnosis?

With these points in mind, we aimed to design a paradigm that better reflects the evolving nature of diagnosis and allows to us to study aspects of the information seeking process. Our paradigm is then a step towards more realistic diagnostic decisions, as it retains the simplicity and control of vignette-based diagnosis whilst allowing more flexibility in information seeking and committing to a diagnosis (or set of diagnoses). This allows us to investigate more fine-grained aspects of information seeking and how they impact diagnoses. Specifically, is clinician confidence informed by the quantity and quality of information sought during the diagnostic process?

For this study, we designed and implemented a novel vignette-based experimental paradigm where participants are asked to provide a list of all diagnostic differentials they are considering based on the information they have received. We ask clinicians to update this list and their confidence at each of a series of stages related to the information sought about the patient: Patient History, Physical Examinations and Testing. We then ask participants to update this list in light of new information by adding or removing differentials. This allows us to more comprehensively capture their thought process in terms of how differentials are being weighed up against each other. Participants report how severe and likely each of their differentials are to draw a more nuanced distinction between differentials. Whereas past work has tended to provide a preset amount of information to clinicians, we instead prompt participants to actively seek out information that they feel is useful for diagnosing the patient they are presented with. This is more analogous to real medical practice where all the required information is not immediately available to clinicians when

presented with a patient. We can then look at information seeking patterns within participants to study how they impact confidence.

Past work from cognitive psychology has shown a link to the quantity of information received and confidence, even if the information is disconfirmatory of one's beliefs (Ko et al, 2022). We can hence investigate in this experiment if this holds during medical diagnoses; if so, we would observe that higher amounts of information seeking would be associated with higher confidence. Information seeking could also be a marker of accuracy in addition to confidence though, as we can study whether clinicians who make more accurate diagnoses seek more appropriate information for the patient. This is important to study as some tests/information are less relevant than others for helping to reach a diagnosis for a patient, resulting in instances of overtesting. With all this in mind, allowing clinicians to freely seek information was then an important tenet for designing this experimental paradigm.

Another aspect of past work we aimed to expand on was on generating differentials (a term used in medicine to refer to hypotheses for diagnoses that a patient could have). Past work has tended to frame diagnosis as a single decision where a clinician responds either to a single diagnosis (Redelmeier & Shafir, 2023) or a limited number of conditions that a patient could have (Meyer et al., 2013). In the latter case, clinicians may report multiple differentials when prompted to consider alternative differentials via a cognitive intervention that encourages clinicians not to miss other diagnoses (Feyzi-Behnagh et al., 2014). These experimental approaches do not necessarily represent the manner in which clinicians make diagnoses in their everyday medical practice. While clinicians may focus on a single differential at a time, they may also generate multiple diagnostic possibilities that past experimental paradigms do not capture. For instance, a clinician usually has to weigh up differentials (Schiff et al., 2009), based on their likelihood (taking into account the base rate of medical conditions within a given patient population) and severity (which may be less likely for a given patient, but would be more dangerous if not considered by the clinician

as a possibility). In this sense, a clinician may have, at least, a primary diagnosis that is most likely for the patient and a more serious diagnosis that is less likely but can be dangerous if missed. Our paradigm should then allow clinicians to report multiple differentials at a time without constraints, in order to capture both the primary differentials being considered and the differentials that clinicians keep 'in the back of their mind'. We can then use the breadth of differentials considered by clinicians as another marker of uncertainty that may guide their subsequent information seeking. By allowing participants to record a list of all differentials they are considering at each stage, we can capture their thought process as it pertains to the information they have received prior to that point.

As our paradigm is designed to capture the diagnosis process as evolving over time, we can also study confidence differently to past work. Rather than seeing confidence as a static quantity, confidence may shift to reflect the current relative strength of evidence in favour of a decision alternative (Vickers & Packer, 1982). Our paradigm then records confidence alongside the participants' list of differentials as it is being updated. We can not only use this facet to link confidence to the breadth of diagnoses considered but also to examine how confidence changes over the course of a case. For instance, a clinician may receive a surprising or inconclusive test result for a patient, causing them to reduce their confidence and seek more information as a result to increase their confidence. Our measure of confidence is also distinct from measures used in past work as we aim to capture the diagnostic process as it pertains to subsequent treatment of patients. An ideal diagnostic process would involve a clinician seeking information to formulate a diagnosis of a patient and, in the process, create a treatment plan to address this diagnosis. We then capture confidence in this study specifically to measure how ready the clinician is to treat the patient, as opposed to past studies that have tended to ask clinicians how confident they are that their diagnosis is the correct one.

There are multiple ways we can define how calibrated participants' confidence is.

To recap, measuring calibration requires a subjective judgement of confidence and an objective measure of accuracy to compare this confidence judgement against. For past work where a single differential is provided by clinicians when they are asked to make a diagnosis, accuracy is relatively easy to measure, as it simply requires marking the provided differential as either correct or incorrect. In our paradigm however, participants not only provide all possible differentials that they are considering but also provide assessments of how likely each differential is. We must then consider how to assess each set of differentials as being accurate or not. A lenient definition of accuracy is to simply mark a set of differentials as accurate if it includes a correct differential. Henceforth, we refer to this measure as Differential Accuracy. However, this measure does not take into account the likelihoods assigned to differentials, so it does not consider how clinicians weigh up differentials against each other. Participants are also more likely to be correct by simply including more differentials in their list. A stricter definition of accuracy would be to look at whether the most likely differential (as rated by the participant) is correct and use the likelihood value assigned to this. Henceforth, we refer to this measure as Highest Likelihood Accuracy. However, this penalises participants who consider the correct differential as likely but not as their primary diagnosis. We therefore use the following measure of accuracy as our primary measure: we look at the likelihood rating assigned to the correct differential if it is present in the participant's list. This provides a more nuanced measure of accuracy that takes into account how differentials are weighed up against each other, which marks a difference from accuracy as it is defined in past work. We should note however that assessing the calibration of participants' confidence judgements is potentially contingent on the accuracy measure used. We therefore measure calibration using our primary measure of accuracy (the likelihood of the correct diagnosis), but also provide results using the other two measures mentioned here (Differential Accuracy and Highest Likelihood Accuracy).

For our studies, we chose to focus on medical students who were relatively advanced

in terms of their medical education but were still early in their clinical experience. Medical students are yet to settle on a particular medical subdiscipline to specialise in, which allows our vignettes to cover a variety of medical conditions and pathologies. We also focus on students as findings from our work could have implications for future medical education in terms of how clinical reasoning and cognitive psychology is taught. Finally, recruiting students allows us to collect a relatively large sample to facilitate detailed analysis of information seeking patterns.

Research Questions

With this study, we investigated the following research questions:

- **Is confidence calibrated to accuracy within medical students?** - Whilst past work has found disassociations between diagnostic confidence and accuracy, these were found in the context of simple tasks with limited flexibility in terms of information seeking and recording multiple diagnostic differentials. We therefore investigate if similar miscalibrations of confidence occur within a more flexible experimental paradigm.
- **How do medical students weigh up competing differentials during the diagnostic process** - Past work has considered that clinicians may have multiple differentials in mind when diagnosing a patient, but such research has not studied how the differentials being considered changes with the receipt of new information. Specifically, do medical students tend to narrow the differentials over time (i.e. akin to a process of elimination) or do they tend to broaden their thinking as new information on the patient is received?
- **How do confidence and information seeking interact in the diagnostic process?** - We expect that confidence would predict information seeking, such that confidence in diagnoses is predicted by both the quality and quantity of information sought.

- **Do differences in confidence and information seeking predict differences in diagnostic accuracy?** - We expect accuracy to associated with the quality/suitability of information seeking but not the quantity of information sought or by confidence (as per the aforementioned miscalibrations of confidence).

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 characterise their diagnostic process and potentially to inform future 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 emails sent to all UK medical students within 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/DiagnosisParadigm. 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 developed by Friedman et al. (2001). 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 (see procedure in Figure 3.1 below). 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.

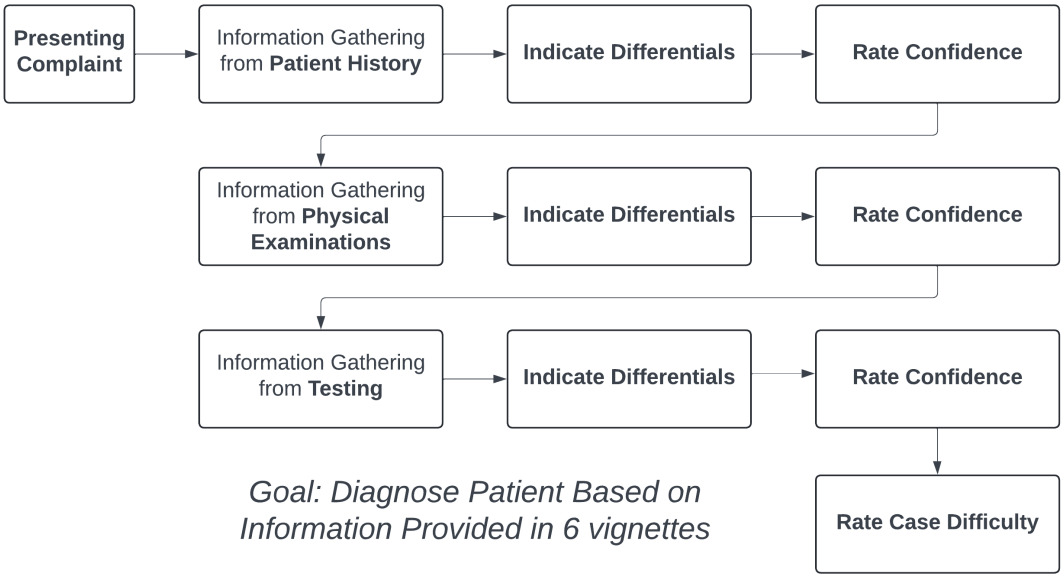

Figure 3.1: Paradigm of the online vignette study, showing the procedure for a single patient case.

Figure 3.2: Screenshot of the interface. Shown here is the screen in which the participants seek information during the Testing stage.

The procedure of a single case is as follows. The participant was 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 was shown a description of a patient, which includes the patient's gender, age and their presenting complaint. An example of this was: "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). This structure has been utilised in past work as being prototypical for a diagnosis (Hampton et al., 1975, Peterson et al., 1992). The Patient History stage included information on "Allergies", "History of the Presenting Complaint", "Past Medical History" and "Family History". The Physical Examination stage included 'actions' that a doctor may take when examining a patient, such as "auscultate

the lungs", "abdominal examination", "take pulse" and "measure temperature". Finally, the Testing stage involved 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 were 29 possible information requests across the three stages, with the available set of information being the same for all patient cases.

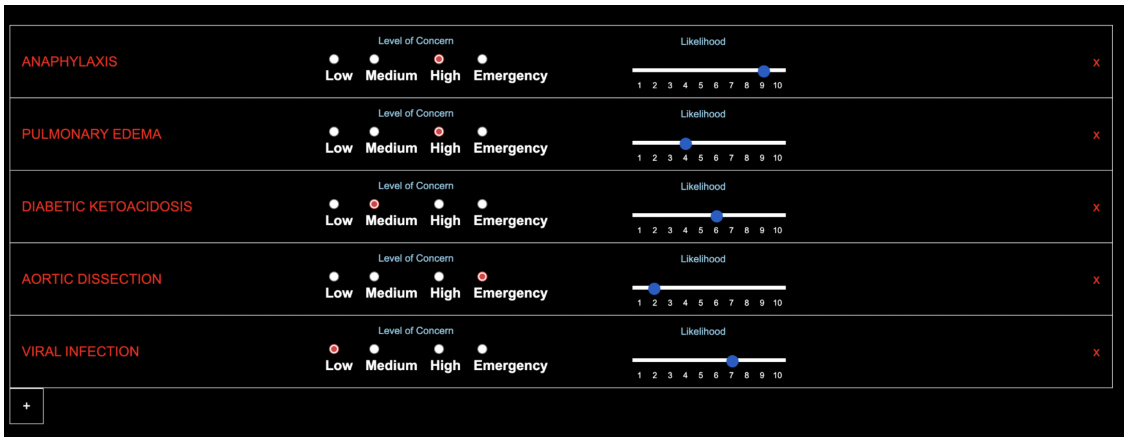

Figure 3.3: Screenshot of the interface. This is the screen in which participants report their current list of differentials, including the name of each condition as well as the severity and likelihood ratings for each condition. Participants remove conditions by clicking the red cross on the right hand side of each differential. Participants add a new differential by clicking the plus icon below the list.

When a participant clicked on any of these requests, the information for that request was shown on screen after a 3 second delay. This delay was added after pilot testing (with 10 participants) revealed that participants tended to select most, if not all, of the information available to them. We mitigated this tendency by adding this delay and by emphasising to participants during the task instructions that they should only request information that they believe will help them with diagnosing the patient for that specific case. Participants were free to request the same piece of information multiple times, including information from a previous

stage. At any point, they could choose to stop gathering information for that stage. They were then taken to a new screen where they reported a list of all differential diagnoses that they were considering for that patient at that stage. For each differential, participants reported a likelihood rating, ranging from 1 (very unlikely) to 10 (certain), and a "level of concern" (which was how concerned they would be for that patient if this differential really was the patient's underlying condition) on a 4 point scale (labels of "Low", "Medium", "High" and "Emergency"). In subsequent stages, the list from the previous stages was available for participants to update concern/likelihood ratings, or to add/remove differentials from the list. Even at the last information stage, participants could report multiple differentials.

After recording their differentials, participants were then asked to report their confidence that they were "ready to start treating the patient" on a 100 point scale, ranging from not at all confident to fully confident. Participants also indicated using a checkbox whether they are ready to start treating the patient, at which point a text box appeared 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. This allowed participants to express what actions they would take that were not covered by our set of available information requests. Once all three stages were completed, participants reported 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 were told the 'true' conditions for all the patients. The session took approximately 40-60 minutes to complete.

Data Analysis

During analysis, no sought information was recorded for three cases across participants (i.e. at all three stages during a case, the participant did not appear to seek any information). These cases were excluded from analysis. We now describe the key dependent variables for this study. The first set of the measures (Case-Wise Measures) are calculated at each of the three information stages (except for Perceived

Difficulty). When averaging these variables within a participant, we use the values obtained at the final stage (i.e. Testing). The second set of measures (Derived Information Seeking Measures) are based on information seeking by participants on each case across all three information stages.

Case-Wise Measures

- *Correct Differential Included*: This measure captures whether participants consider a correct diagnostic differential. 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. Each case is marked as correct if the list of differentials provided includes the correct condition or a differential considered correct as per our marking scheme in Table A1 of the Appendices. Otherwise, the case is considered incorrect if a 'correct' differential is not included.

- *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 differential is not included. The value is then rescaled to range from 0 to 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 the closest differential (as per our marking criteria with help from a medical consultant) to the true condition was used.

- *Highest Likelihood Accuracy*: This stricter measure accuracy is computed as the likelihood value assigned to the differential with the highest likelihood (in comparison to other differentials provided in the participant's list) if this

differential is considered correct. If not, a value of 0 is assigned. Again, likelihoods range from 1-10 for correct differentials, so this is rescaled to range from 0 and 1.

- *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).

- *Change in Differentials*: This measure captures how much participants change the differentials they consider over the course of the case. In other words, we can look at how much participants have narrowed or broadened their list of differentials as they receive more information. This is calculated by taking the absolute value of the difference between the number of Initial Differentials and the number of Final Differentials.

- *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

- *Amount 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 available information. We take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage and excluding repeat tests) and divide this by the number of possible tests available.

- *Information Value*: We calculate a measure of information value to capture how appropriate the information sought for a case is for the patient's 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 at any stage 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.

- *Information Seeking Variability*: We calculate a measure of how much, for a given set of cases, information seeking varies across cases. This is operationalised as the average dissimilarity between cases' information seeking (by taking the average of all pairwise comparisons) using each piece of information as a binary variable (i.e. whether it was sought or not). This measure is calculated both within participants, to tell us how much each participant varied the information they sought across their cases, and between participants, to tell us how dissimilar participants are to each other in terms

of the information sought for a given condition. We calculate this value using the Dice coefficient (Dice, 1945), due to it being well suited specifically for binary data, as well as its increased weighting on discordant pairs (ie a piece of information being sought in one case but not the other). A higher value between two cases indicates that the information sought on those trials are more dissimilar to each other.

We used statistical analyses to consider differences in confidence, accuracy and information seeking. When looking at how our variables change over the three information stages, we used Analysis of Variance models with Bonfferoni-corrected pairwise T-tests on all pairwise comparisons. We test if there is a relationship between confidence and information seeking (Amount, Value Variability) and between accuracy and information seeking using Pearson's product moment correlation tests (an alpha value of less than 0.05 was regarded as statistically significant). These help us answer how confidence and information seeking interact during the diagnostic process and whether differences in diagnostic are predicted by information seeking and confidence. Our sample of 85 participants is calculated as having 80.4% power to detect a medium effect size of $r = 0.3$ (using an approximate arctangh transformation correlation power calculation). In order to test if information seeking patterns are predictive of differences in accuracy, we used generalised logistic regression to classify cases as being performed by high or low accuracy participants (via a median split). To test if information seeking patterns are predictive of the case (i.e., whether participants tailor their information seeking to each patient case), we use penalised multinomial regression to classify cases by their patient condition. Both models were trained on the information requests as binary variables (with a 1 signifying that the information was sought for that case and 0 when the information was not sought). We used Leave One Out Cross Validation for both models, such that each case is predicted by training the algorithm on all other cases.

Results

Overall Performance and Calibration

We first look at our research question as to whether confidence is calibrated within medical students. When comparing Accuracy (taking into account the likelihood assigned to correct differentials) to Confidence, we find, across stages, participants' Confidence was aligned to their Accuracy (see Figure 3.4 below). To determine whether there is any systematic discrepancy between subjective confidence and objective accuracy across stages, we compute 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 ($t(84) = 0.29$, $MDiff = 0.01$, $p = 0.77$) and Physical Examination stages ($t(84) = 0.74$, $MDiff = 0.01$, $p = 0.46$), but there was a statistically significant difference between the two at the Testing stage ($t(84) = 2.35$, $MDiff = 0.05$, $p = 0.02$). This indicated well-calibrated confidence after Patient History and Physical Examination, but a slight overconfidence across participants after Testing.

To investigate the dynamics of confidence and accuracy further, we look at how both variables change over the course of the information seeking stages. Across cases, accuracy increased with each stage of information gathering as per our Accuracy measure, ($F(2, 252) = 21.6$, $\eta^2G = 0.15$, $p < .001$). Participants had lower accuracy at the Patient History stage ($M = 0.31$, $SD = 0.14$) than during the Physical Examination ($M = 0.04$, $SD = 0.15$) and Testing stages ($M = 0.41$, $SD = 0.15$). Pairwise comparisons between the History stage and each of the other two stages are significant ($ps < .001$). Table 3.2 shows overall accuracy (at the Testing stage) by case, indicating that there was variability in performance between cases.

Confidence also increased as participants received more information ($F(2, 252) = 21.6$, $\eta^2G = 0.15$, $p < .001$). 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). Pairwise comparisons between History and each of the other two stages are significant (ps $< .001$). 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. This is reflected in participants expressing their readiness to treat the patient in the vignette, which allows them to enter a treatment plan for the patient. In 38% of cases, participants reported they were ready to treat the patient and entered a treatment plan.

Case	Differential Accuracy	Accuracy	Highest Likelihood Accuracy	Final Confidence	Difficulty	Information Seeking
AD	0.60	0.28	0.12	0.49	5.9	0.59
GBS	0.75	0.41	0.30	0.37	6.9	0.63
MTB	0.43	0.24	0.10	0.45	6.7	0.64
TA	0.74	0.50	0.45	0.49	6.2	0.62
TTP	0.61	0.34	0.20	0.41	6.8	0.66
UC	1.00	0.73	0.69	0.62	5.2	0.55

Table 3.1: Average statistics across participants for each case (leftmost column, AD = Aortic Dissection, GBS = Guillain Barré Syndrome, MTB = Miliary Tuberculosis, TA = Temporal Arteritis, TTP = Thrombotic Thrombocytopenia Purpura, UC = Ulcerative Colitis). Differential Accuracy (0-1) refers to the proportion of participants who correctly included the correct condition or a condition considered correct for that case based on our marking criteria. Highest Likelihood Accuracy refers to the likelihood assigned to the differential with the highest likelihood if it is correct (1-10), otherwise the value for a given case is 0 if this differential is incorrect. This value is then rescaled to range between 0-1. Accuracy refers to the average likelihood (on a 1-10 scale, rescaled to range between 0-1) assigned to a correct differential if included. Confidence refers to the confidence provided by participants on their readiness to treat the patient at the Testing stage (on a scale of 0-100, rescaled

to fall between 0-1). All these measures are calculated based on values observed at the final information stage of each case (i.e. the Testing stage). Difficulty refers to the subjective rating provided at the end of each case of how difficult participants found the case to be in terms of determining a diagnosis (on a scale of 1-10).

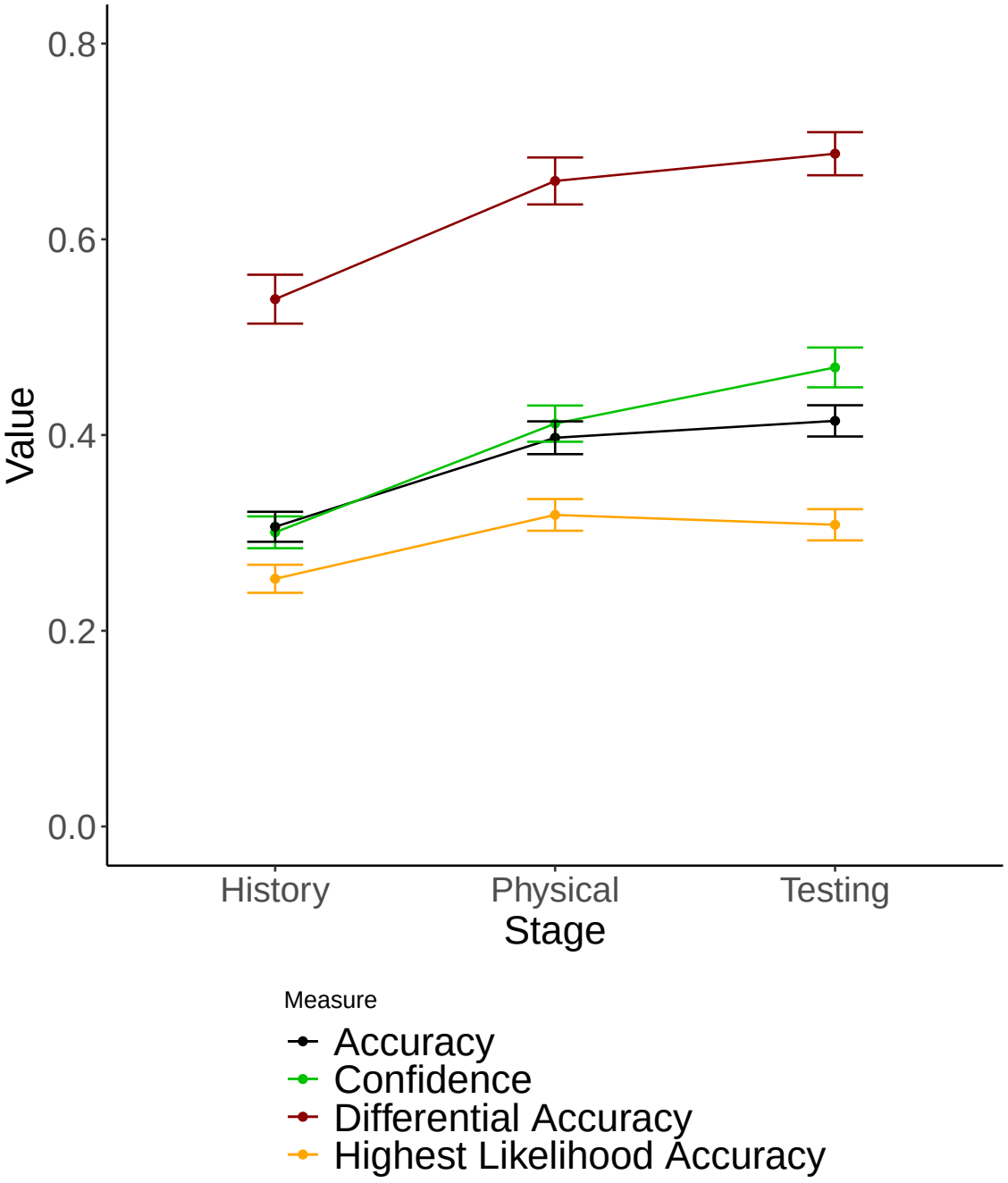

Figure 3.4: Graph showing Confidence (green) at each of the three information stages (History = Patient History, Physical = Physical Examinations, Testing

= Testing) in comparison to our main accuracy measure (black, likelihood value assigned to the correct diagnosis), the more lenient measure of the proportion of trials where a correct differential was included (dark red) and the stricter measure of the value assigned to the highest likelihood differential if it is correct (orange). Values shown are averaged across participants and cases, with the error bars representing standard error.

In order to examine the observed overconfidence in more granularity, we compare confidence and our primary accuracy measure by case (the mean values of which can be found in Table 3.1). We conducted paired t-tests for each condition's cases by comparing accuracy and confidence values (at the final Testing stage) to observe if they significantly differ from each other. A p value of less than .05 is interpreted as evidence for overconfidence or underconfidence (depending on the direction of the effect). We observed overconfidence for the AD case ($t(84) = 4.71$, MDiff = 0.21, $p = < .001$) and for the MTB case ($t(83) = 4.31$, MDiff = 0.21, $p = < .001$). We observe underconfidence for the UC case ($t(82) = -3.51$, MDiff = -0.12, $p = < .001$). The remaining cases did not yield a significant effect, indicating calibrated confidence judgements across participants. The overall overconfidence after Testing that we observe in Figure 3.4 is then driven by the AD and MTB cases, for which accuracy was lowest compared to other cases and confidence was not sufficiently adjusted to reflect this.

Differentials

We analysed the number of differentials to provide insights into the diagnostic decision process across stages, specifically the degree to which it follows a process of deductive narrowing (decreasing differentials) or open-minded broadening (increasing 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 ($F(2, 252) = 11.66$,

η^2G = 0.08, p < .001). Participants reported fewer differentials during the Patient History stage (M = 3.2, SD = 1.12) than during the Physical Examination (M = 3.89, SD = 1.32) and Testing stages (M = 4.13, SD = 1.43). Pairwise comparisons between the History stage and each of the other two stages are significant (ps < .05). The majority of participants (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). As can be observed in Figure 3.5 below, there is general consistency in terms of participants broadening their differentials with more information despite some inter-participant variability, with a small minority of participants narrowing their differentials on average.

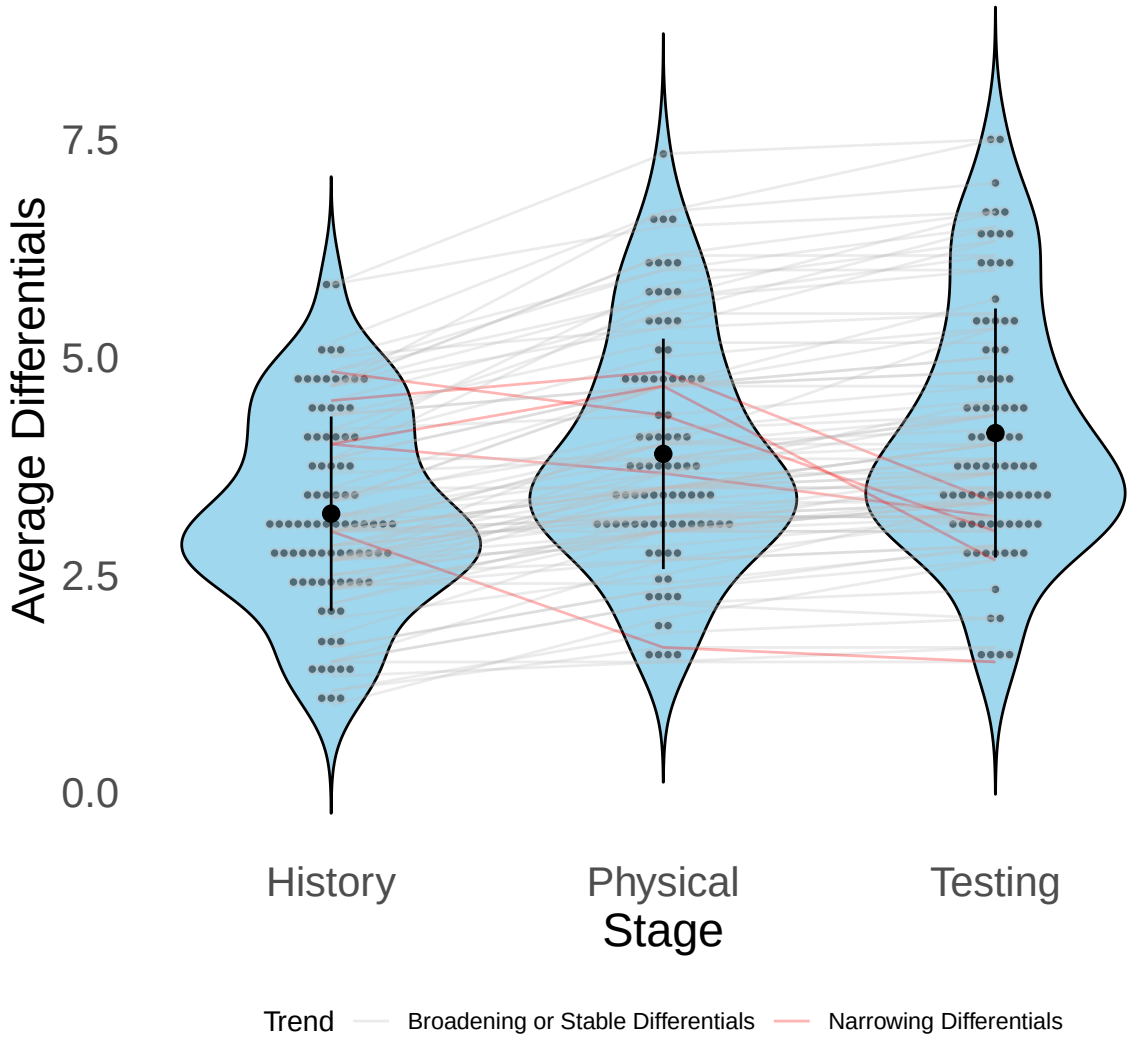

Figure 3.5: The average number of differentials after each stage of information seeking (x-axis, History = Patient History, Physical = Physical Examinations, Testing = Testing). The width of the blue area corresponds to the amount of data points that fall within that part of the y-axis, with a wider area meaning a higher concentration of data points. The larger black dots indicate the mean values, whilst the larger black vertical lines indicate standard deviations. The grey dots show individual values at each stage, with lines connecting the dots at each stage to represent individual participants' trend across the information seeking stages. The participants who show a narrowing of differentials (i.e. recording fewer differentials at the Testing stage compared to the Patient History stage) are marked with a red line, whilst the remainder of participants are marked with a grey line.

As a first probe of the dynamics of the diagnostic process, we analysed whether participants who generated 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 a positive correlation (see Figure 3.6) 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 = [0.09, 0.48], $p = 0.005$, Figure 3.6a). 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) ($r(83) = 0.23$, 95% CI = [0.02, 0.42], $p = 0.04$, Figure 3.6b).

Given that we observe an broadening (increasing number) of differentials across participants, we ask how this change in differentials related to information seeking and changes in confidence. As well the initial diagnostic breadth of participants, we are also interested in whether information seeking and changes in confidence relate to how much participants change the number of differentials they consider over the course of the case. This allows us to capture how much their diagnostic differentials have changed based on the information received. We find the amount of Differential Change was associated with both the amount of information sought ($r(83) = 0.3$, 95% CI = [0.09, 0.48], $p = 0.005$, Figure 3.6c) and change in confidence ($r(83) = 0.39$, 95% CI = [0.19, 0.56], $p = < .001$, Figure 3.6d). These results indicate that participants who tended to increase differentials also tended to seek more information and increase their confidence to a greater extent. If broadening of differentials was a reflection of diagnostic uncertainty, we may have expected a decrease in confidence, but this does not appear to be the case.

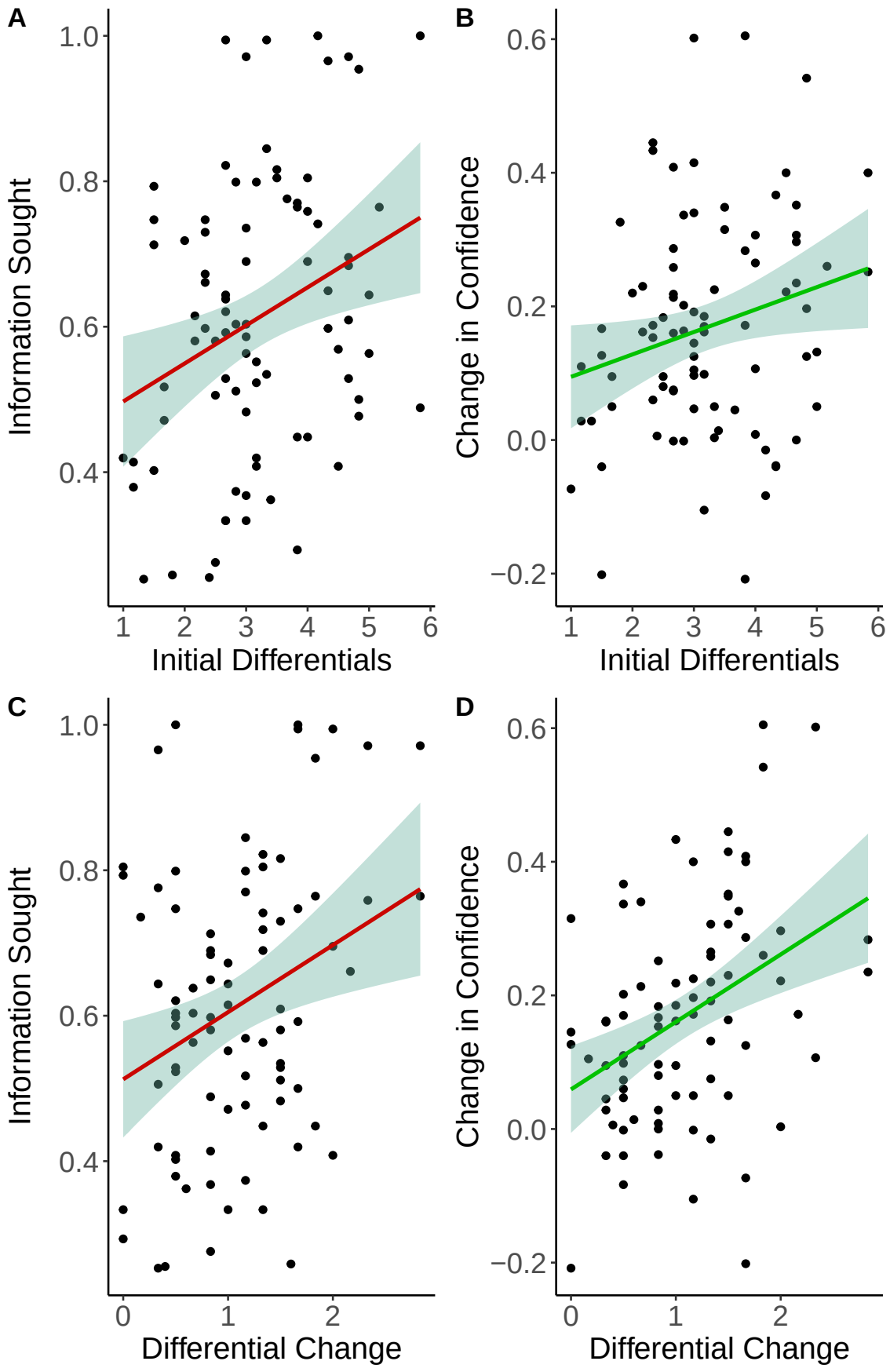
A
B
C
D
Information Sought
Change in Confidence
Initial Differentials
Differential Change
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1
2
3
4
5
6
0

Figure 3.6: Scatter plot showing the relationship between the number of initial differentials reported at the Patient History stage (x-axis, figures 3.6A and 3.6B) and the change in differentials (x-axis, figures 3.6C and 3.6D) against both the proportion of available information sought (y-axis, figures 2.6A & 3.6C) and change in confidence (y-axis, figures 3.6B & 3.6D). Each point represents a single participant with all three variables averaged across the six cases that each participant performs. Initial Differentials refers to the average number of differentials that participants report in their list at the Patient History stage. Differential Change refers to the absolute difference in the number of Initial Differentials (at the Patient History Stage) and the number of Final Differentials (at the Testing Stage). Information Sought 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. Change in Confidence refers to the difference in reported confidence 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

To investigate our research questions of how both confidence and accuracy interact with information seeking during the diagnostic process, we first look at broad characteristics of information seeking and then ask if they are predictive of differences in confidence and accuracy. 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 ($F(2, 252) = 57.26$, $\eta^2G = 0.31$, $p < .001$). Participants sought more of the available information during the Patient History stage ($M = 0.85$, $SD = 0.19$ than during both during the Physical Examination ($M = 0.59$, $SD = 0.24$) and Testing stages ($M = 0.5$, $SD = 0.22$). All pairwise comparisons are significant ($ps < .05$). This selectivity in information seeking does not seem to reflect participants being less certain about their diagnoses, which the general pattern of broadening differentials may have indicated.

Given the design of our task, we ask if seeking all available information is in fact a helpful strategy for increasing diagnostic accuracy by testing for a correlation between the two. We do not find that participants who sought more information across cases were also more accurate in their diagnoses ($r(83) = 0.16$, 95% CI = [-0.05, 0.36], $p = 0.13$, Figure 3.7A). However, participants who sought more information tended to have increased their confidence more during cases ($r(83) = 0.24$, 95% CI = [0.02, 0.43], $p = 0.03$, Figure 3.7C). While seeking more information may imbue students with a greater level of confidence, we do not find evidence that this translates consistently into more accurate diagnoses. This finding links to the results presented in Figure 3.4, in which confidence and accuracy were related to one another but imperfectly (especially during the Testing stage, during which subjective confidence was higher than objective accuracy across participants).

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 ($r(83) = 0.22$, 95% CI = [0, 0.41], $p = 0.05$, Figure 3.7B), as well as between confidence and information value ($r(83) = 0.29$, 95% CI = [0.08, 0.47], $p = 0.01$, Figure 3.7D). When comparing the correlations between both information amount and information value to accuracy via a Fisher's z-Test of dependent correlations, we find they are not significantly different from one another ($z = 1.08$, $p = 0.28$). This means that we cannot make a valid comparison between the correlations with information amount and information value with respect to accuracy.

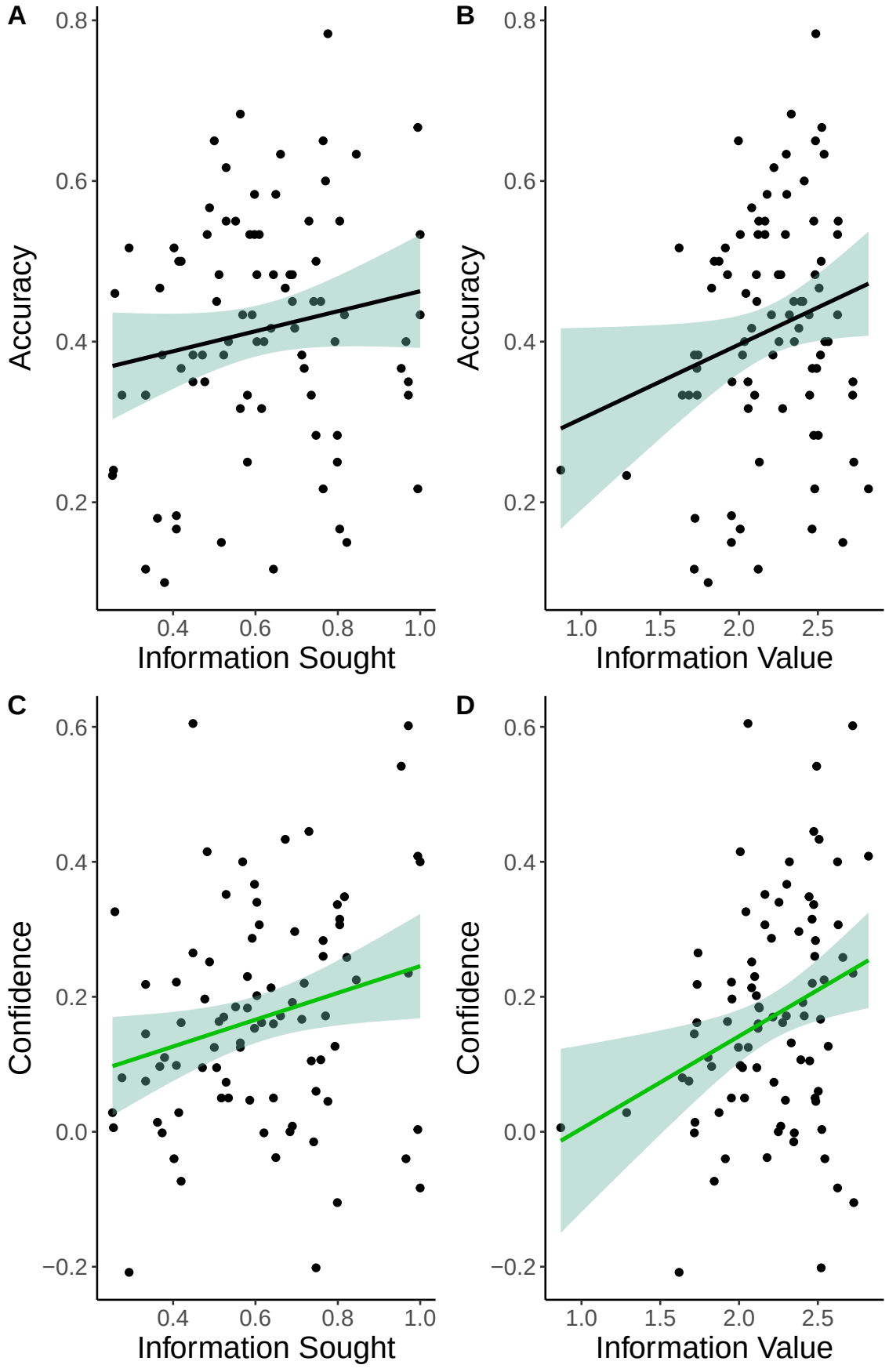
A
B
C
D
Accuracy
Confidence
Information Sought
Information Value
0.8
0.6
0.4
0.2
0.0
−0.2
1.0
1.5
2.0
2.5

Figure 3.7: Scatter plots showing our information seeking variables (amount in figures 2.7A & 2.7C and value in 3.7B & 3.7D) against our key dependent variables of accuracy (the likelihood assigned to a correct differential if provided, figures 3.7A & 3.7B) and change in confidence (difference between final confidence and initial confidence, figures 3.7C & 3.7D). 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.

Whilst we do not find evidence that the amount of information sought is predictive of accuracy, 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 assesses 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 with Leave One Out Cross-Validation (LOOCV) to identify if participants exhibited high or low accuracy based on the information they sought. LOOCV is where our classifier is trained on all data except one case to ask if, based on the learnt patterns from this data, the classifier is able to predict the participant's accuracy (high or low) on the remaining case. This process is then repeated with each case being left out of training and used as this 'test' case. 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.

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 3.8). When conducting a DeLong test, to test the null hypothesis that the AUC is equal is 0.5 (i.e. that the classifier is unable to differentiate between 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.

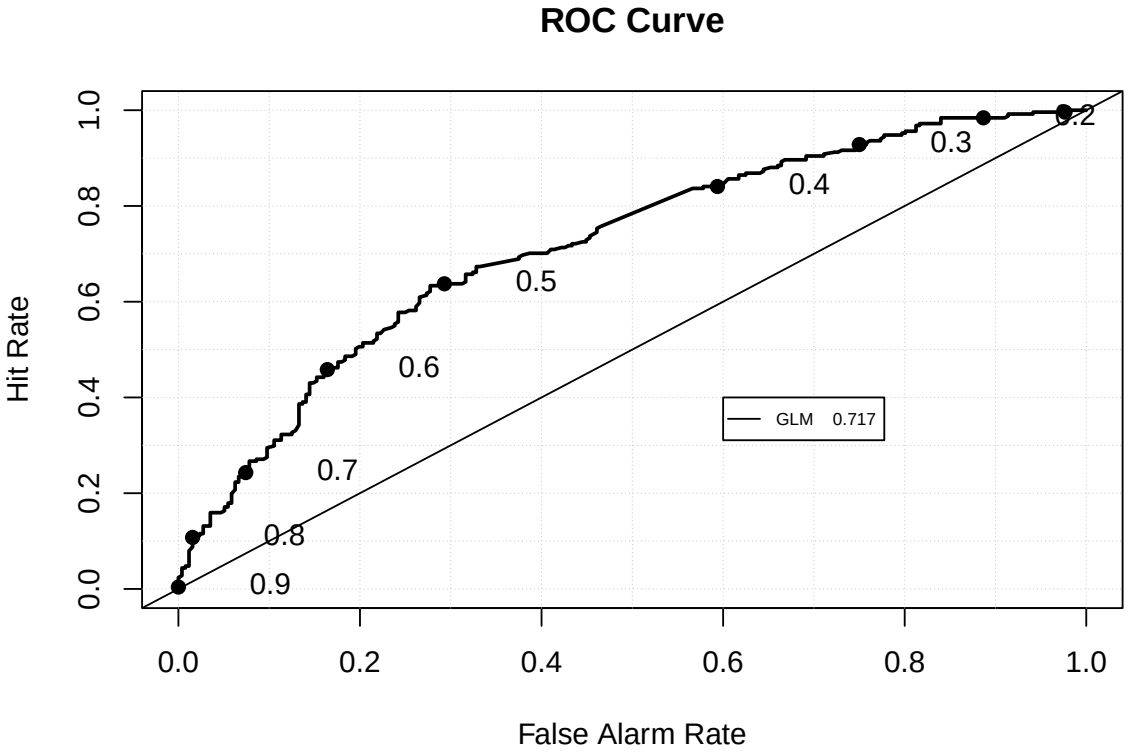

Figure 3.8: Receiver-Operator Characteristic (ROC) curve using a Generalised Linear Model to classify individual cases as being performed by either high or low accuracy participants. The models are trained on the raw binary predictor

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, in terms of high and low accuracy participants seeking different sets of information. Essentially, information seeking patterns are different between high and low accuracy participants. This analysis alone, however, does not tell us what aspects of information seeking in particular are predictive of accuracy. We know from figure 3.7D that seeking more valuable information is associated with higher accuracy. We next seek to characterise the specific differences in information seeking that contribute to higher diagnostic performance.

By looking at the extent to which participants vary the information they seek by case, we can ask the following: is diagnostic accuracy characterised more carefully tailoring information seeking to each individual case, or is it characterised by adopting a more consistent information seeking approach regardless of the patient case? With our measure of how much participants vary in their information seeking across cases, we can see if the variability in information seeking is associated with higher diagnostic accuracy. If higher variability is associated with higher accuracy, this would indicate the former approach being more beneficial (tailored information seeking). If lower variability is associated with higher accuracy, this would indicate the latter approach being more beneficial (consistent information seeking).

We find marginal evidence for a negative association between Information Seeking Variability and Accuracy ($r(83) = -0.22$, 95% CI = [-0.42, -0.01], $p = 0.04$). This data is plotted below in Figure 3.9. We can also look at variability between groups of participants for each case to ask: are higher performers (in terms of accuracy)

more alike in their information seeking than lower performers? To do this, we median split participants into high and low overall accuracy across cases (similar to the ROC analysis in Figure 3.8). We then look at variability in information seeking between participants for each case. If variability is higher, this would indicate that for a given case, participants adopt information seeking approaches that are more different from one another. A plot of variability by case is shown in Figure 3.10. When performing a t-test across conditions, we find that higher performers are more alike in their information seeking (i.e. exhibit lower variability) ($t(10) = 2.64$, MDiff = 0.36, 0.31, $p = 0.02$). As can be seen in Figure 3.10, better performing participants show less variability in their information seeking patterns for 5 out of 6 cases, indicating that higher diagnostic accuracy is associated with a consistent 'optimal' information seeking strategy.

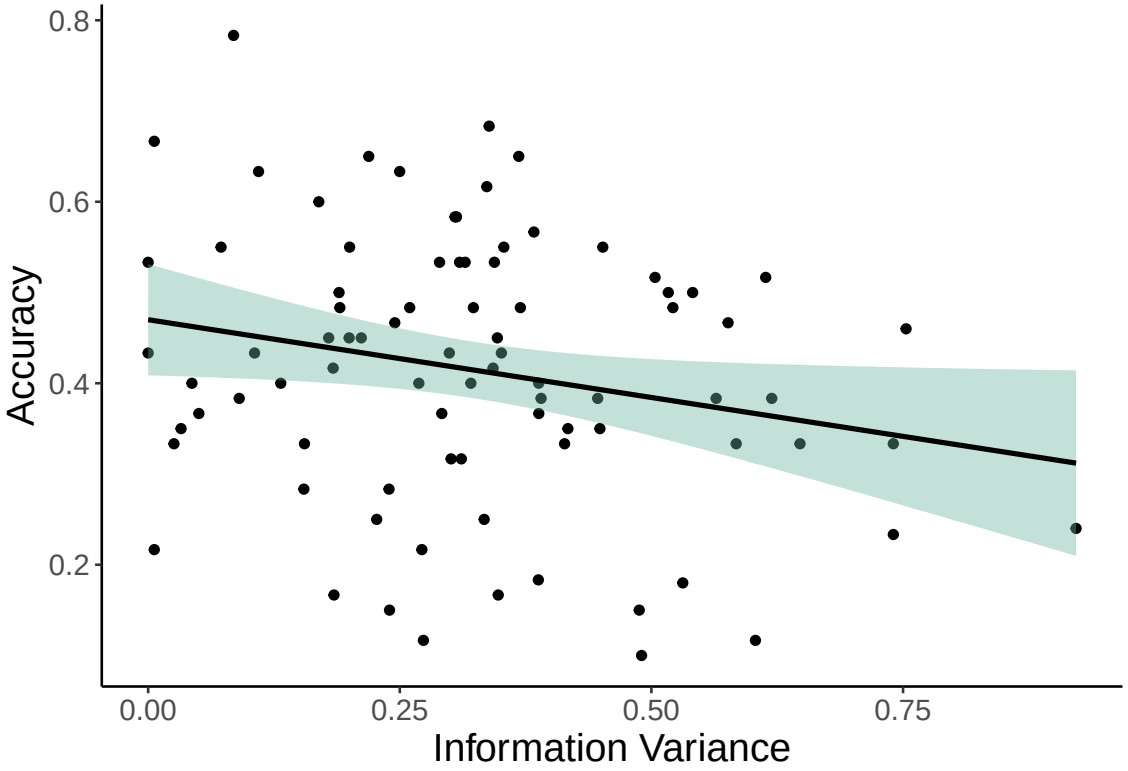

Figure 3.9: Scatter plot showing the relationship between Information Seeking Variability (x-axis, quantified as the average Dice Distance between all pairwise comparisons of cases for a given participant) and Accuracy (y-axis). Each data point represents a single participant.

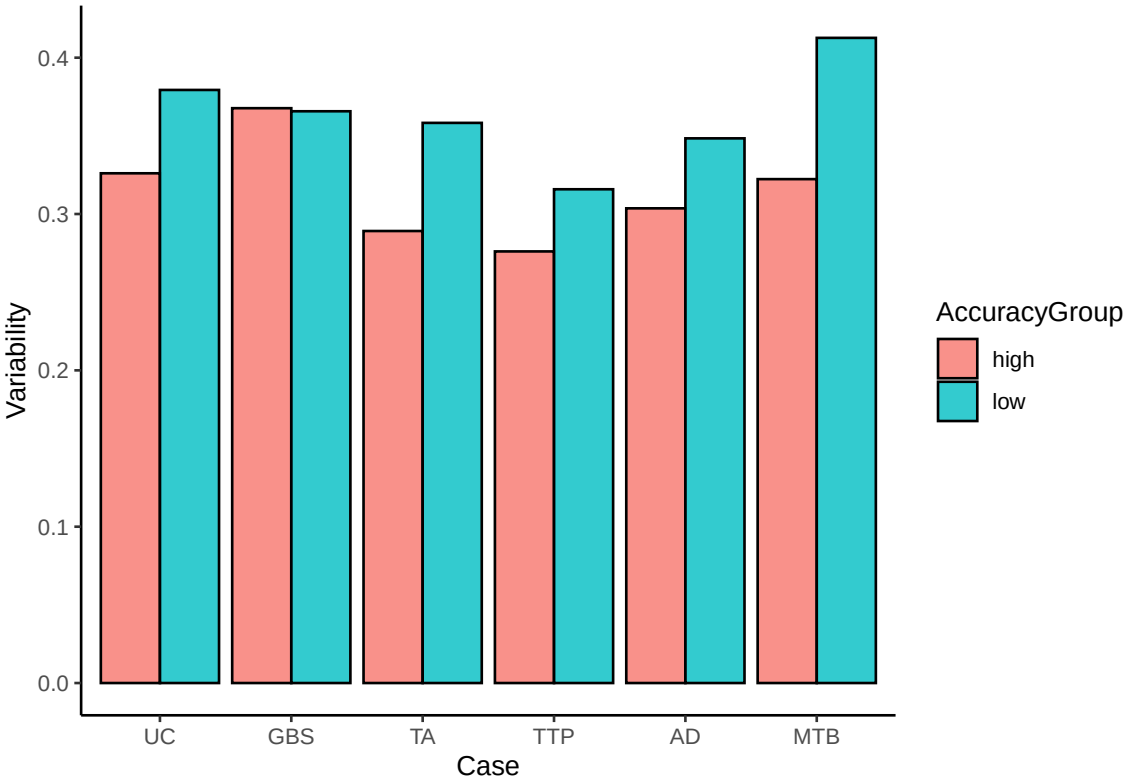

Figure 3.10: Information Seeking Variability (y-axis) for all cases of a given condition (x-axis), with cases median split by participant accuracy. Red bars indicate high performers and blue indicating lower performers. Cases are in descending order (UC = Ulcerative Colitis, GBS = Guillain Barré Syndrome, TA = Temporal Arteritis, TTP = Thrombotic Thrombocytopenic Purpura, AD = Aortic Dissection, MTB = Miliary Tuberculosis) by their average accuracy across participants. Higher variability values signify that participants were less alike one another in terms of the information they sought on a given case.

Given that information seeking variability has a weak negative association with accuracy on our task, we next ask if information seeking is also specific to patient conditions. If so, we would expect the information sought to be predictive of which case the participant is performing. To investigate this, we train a classifier using Penalized Multinomial Regression and Leave One Out Cross Validation (i.e. we train the classifier on all data except one case, and ask if, based on learnt patterns

in the remaining data, whether we are able to predict which case/condition it is based on information seeking patterns). Our input parameters are the available information requests as binary predictors (i.e. to denote whether they were sought on each case or not). The outcome variable of the classifier is the patient condition. We generate model predictions and then look at whether they correctly match the actual condition for that case. Across 510 cases (85 participants performing 6 cases each), the accuracy of the classifier was 57%, which is higher than the chance level of 16.6%. When breaking down accuracy of our classifier by condition, we find accuracy to be above chance across all conditions (see Table 3.2 below).

Condition	Prediction Accuracy
AD	0.65
GBS	0.46
MTB	0.48
TA	0.47
TTP	0.56
UC	0.78

Table 3.2: The accuracy of our multinomial classifier that predicts patient condition for each case based on the information sought/not sought as binary predictors. We then test the accuracy of the classifier by comparing the predicted condition from the model against the actual patient condition for each case. We then split cases by condition to look at accuracy on a case-by-case level. Given that participants perform 6 cases each, accuracy would be 1/6 (16.6%) when at chance.

Taking these findings together, keeping information seeking more constant (i.e. requesting similar high-value information) across cases was found to have an association with accuracy whilst there also being some information that is useful for clinicians to know for patients with specific conditions. To reconcile these, we derive which information requests were most weighted in our classifier models to find which were considered markers of accuracy (by being sought across cases) and which were

considered markers of identifying specific cases. We extract coefficients from the logistic classifier of accuracy (the ROC curve for which was shown in Figure 3.8) and the multinomial classifier (the accuracy of which was depicted above in Table 3.2). We identify the highest weighted information requests as input parameters for each model. The five highest weighted information requests for each model are shown below in Table 3.3. We also show how often each piece of information was sought for each of the cases in Figure 3.11 below. Viewing this figure shows individual tests that are useful for specific cases. For example, an ECG is sought by most participants for the AD (a heart condition) case.

	Accuracy			**Condition**		
Rank	Test Name	Coefficient	Odds Ratio	Test Name	Coefficient	Odds Ratio
1	FBC	0.45	1.57	Neurologic Exam Record	0.22	1.25
2	Venous Blood Gas	-0.44	0.65	Measure Blood Pressure	0.22	1.25
3	Urine Dipstick	-0.39	0.67	UREA and Electrolytes	0.2	1.22
4	Assess Extremities	0.38	1.47	Rectal Examination	0.17	1.19
5	Other Biochemistry Tests	-0.36	0.7	Urine Dipstick	0.16	1.17

Table 3.3: The five highest weighted parameters (by the absolute value of the coefficient values) for our logistic classifier of participant accuracy (under the "Accuracy" heading above) and our multinomial classifier of patient condition (under the "Condition" heading above). We also show coefficient values and odds ratio values for each parameter.

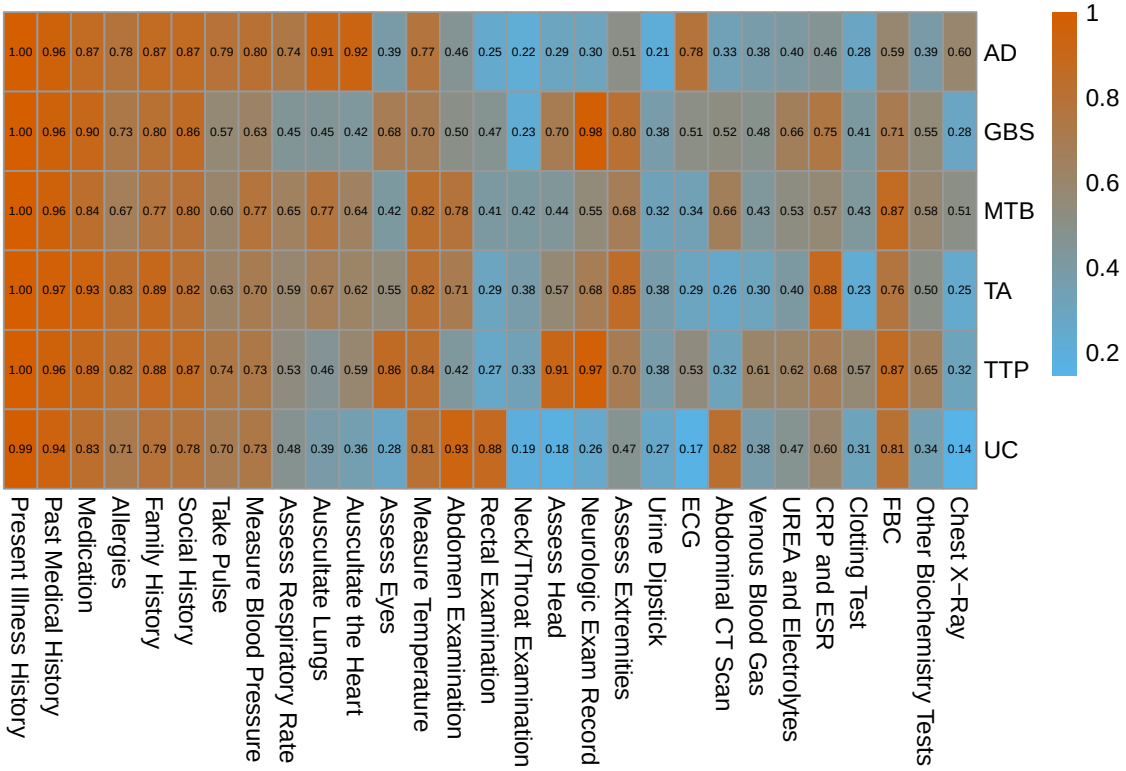

Figure 3.11: Visualisation of the proportion of participants who sought each available piece of information (columns, x-axis) broken down by case (rows, y-axis). Lighter blue colours indicate that fewer participants sought that information for a given case (i.e. towards 0%), whilst lighter orange colours indicate more participants sought that information for a given case (i.e. towards 100%)

Discussion

This study of medical students explored the interplay between confidence, accuracy and information seeking in a novel medical diagnosis task. Using an online interface, we explored how medical students work through diagnostic scenarios, freely seeking information to develop and test sets of possible differentials. Our aim was to look at how different aspects of information seeking impacts both diagnostic confidence and accuracy. The main strength of this study's paradigm is in allowing us to investigate the diagnostic process as it evolves 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 and to clinicians' subjective confidence.

On the question of whether medical students provided confidence judgements that were calibrated to their objective accuracy, we found that students become more accurate across successive stages of information seeking as well more confident. However, cases varied in difficulty as reflected in participant accuracy. In particular, the AD and MTB cases exhibited lower observed accuracy across participants. We observed overconfidence for these two cases, and underconfidence for the UC case (for which accuracy was highest). This indicates a classical hard-easy effect of confidence (Lichtenstein, Fischhoff & Phillips, 1977), whereby individuals have a greater tendency to be overconfident for more difficult decisions when compared to easier decision (Merkle, 2009). Confidence also increased as participants received more information. However, 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). This may also be a function of undertaking the diagnostic process in isolation (i.e. without being able to discuss with colleagues, as would be the case in naturalistic medical environments). Such a reduction in confidence when making a decision alone rather than in group would be justified from a calibration perspective, as combining medical students' diagnoses has been found to improve accuracy (Kämmer et al., 2017).

Previous work (e.g. Meyer et al., 2011) has revealed a gap between subjective confidence and objective accuracy. In particular, a general tendency has been demonstrated for less experienced medical trainees to be underconfident and for

more experienced medical professionals to be overconfident (Yang and Thompson, 2010). Part of the discrepancy between our findings and past findings could stem from the way that diagnostic uncertainty is expressed by students in this study. Using our primary 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. We note however that our finding of calibrated confidence is highly contingent on the measure of accuracy used. When using a more lenient measure, the proportion of cases where a correct differential was reported (as used in previous papers, Friedman et al., 2004, Meyer et al., 2013, Lambe et al., 2018, Küper et al., 2023), participants were found to be underconfident. When using a stricter measure, the likelihood value assigned to the most likely differential if it is correct, participants are found to be overconfident. Calibration also varied across cases, with participants sometimes showing overconfidence and sometimes showing underconfidence. While we therefore temper our finding of calibration, this has implications for further research that looks at calibration during diagnoses, given that accuracy can be defined in multiple ways when participants record multiple differentials. In addition, our confidence measure is related to the participants' subjective readiness to treat the patient, rather than confidence in the set of differentials. Such a measure of confidence is novel to our study and has not been used in previous studies of diagnostic confidence. This limits the extent to which we can compare accuracy and confidence directly. However, rather than confidence being a subjective judgement, we connect to it to clinical action that would be taken by participants if the patient presented were real. This is similar to one paper in which confidence was measured as the subjective likelihood of seeking assistance to reach a diagnosis (Friedman et al., 2005), with the authors finding that medical students had a lower tendency toward confidence than both medical residents and faculty. When considering this result alongside our own finding of low confidence across medical students, it is possible that tempering overconfidence may stem from tying judgements to specific clinical actions. Given that medical students lack the experience of more senior clinicians, they may generally be less

confident as a result: the lower reported confidence is partly a reflection of their general aptitude/experience with the clinical action being prompted during the confidence judgement. Future work could then measure how confidence relates to specific aspects of the patient care pathway and differences in calibration.

On the question of whether participants tend to broaden or narrow their differentials with new information, participants exhibited a general pattern of broadening the range of differentials they were considering across successive information seeking stages. In addition, we observed that participants did not tend to remove differentials from consideration despite having the option to do so. This marks a novel finding when situated within past research, which has not studied how the differentials being considered evolves over time. We can interpret this as students being careful not to miss differentials from consideration, indicating a focus on being comprehensive in their generation of differentials rather than a focus on narrowing in on a single diagnosis. It is therefore worth considering whether students are explicitly taught not to disregard diagnoses completely, instead focusing on remaining open-minded to new possibilities for differentials. Joseph & Patel (1990) found that clinicians with lower domain knowledge generated accurate hypotheses but were unable to differentiate eliminate hypotheses when receiving more information, unlike clinicians with higher domain knowledge who were able to confirm and eliminate hypotheses using the information received. This may help explain the broadening pattern of medical students, as their relative inexperience meant that they were not able to easily eliminate hypotheses.

We also found that the initial breadth of diagnoses considered from the patients' history was predictive of the amount of subsequent information seeking and changes in confidence. We also find that how much participants change the number of differentials they are considering is predictive of 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, with accuracy instead being associated with seeking more valuable/appropriate information for a given patient condition. When taken together, these findings give an interesting picture of the diagnostic process as we capture it within our task. Our account of how participants approach this task can be summarised as follows (note that this account requires follow-up study to elucidate further):

- Medical students generate an initial set of differentials from the patient history and use this to guide their information seeking.
- With more differentials to consider, students seek more information to 'test' each of these hypotheses.
- Seeking more information increases the likelihood that new differentials are brought to mind, resulting in more differentials being added as under consideration.
- When participants have more information and have considered a wider range of differentials, they are likely to increase their confidence due to being more comprehensive (i.e. considering more differentials) during their thought process.
- With more differentials being considered, participants are more likely to consider a 'correct' differential. However, considering a larger set of differentials makes it more difficult to focus on finding a differential that is most likely.

Given the flexibility afforded by our paradigm, we are able to monitor fine-grained aspects of how participants seek information. We find that the accuracy and confidence gained over the course of cases was related to the quality of the information sought. We also find that higher accuracy was associated with less variability in information seeking (i.e. seeking a similar set of information regardless of the patient case). Higher accuracy participants were found to be more alike in

their information seeking compared to lower accuracy participants. Putting these findings together, we can surmise that each patient condition has associated valuable pieces of information that are worth seeking, but that there is a consistent set of information that accurate participants tend to seek across cases. When combined, each case can be seen to have an 'optimal' set of information that participants should seek. In addition, while seeking more information may increase confidence, having more information may be problematic for weighing up differentials against each other. This is because it can be harder to synthesise more information into a cohesive account of the patient. While past work has called for greater standardisation within healthcare (Wears, 2015), what seems to constitute accurate diagnoses in our task is a degree of standardisation with certain selectivity of information given the patient condition. As depicted in Table 3.3, certain information is useful regardless of patient condition whilst others are useful for specific medical conditions. While we show certain tests/examinations as being most useful across patient cases or for specific cases, we recommend caution in interpreting these as representative of all diagnostic decisions outside of this task. These specific information requests were found to be useful for our task, but may not generalise to other patient conditions or diagnostic decisions.

Given these results, we know what information is sought by medical students, but only have a limited insight into why they sought certain information and how it directly affected the diagnosis they provided. For one, are all students using a similar decision making process when making diagnoses? As of now, we are inferring the participants' thought processes from data of their differentials and information seeking without context of how they are thinking about the task. One possibility is that there are differences in how medical students approach diagnoses that stem from differing reasoning strategies, which we cannot infer from this current dataset. In order to ascertain this, we would need to record the students' thought processes as they are doing the task. To this end, we conduct a follow-up

study using a similar diagnostic paradigm conducted in-person where students think out loud as they make diagnoses.

Coderre et al. (2003) used a think-aloud paradigm to characterise distinct diagnostic reasoning strategies: a "hypothetico-deductive" strategy that is closest to the idealised process of elimination that is the typical characterisation of diagnosis, a "pattern matching" strategy where clinicians draw similarities between the current patient and either a past patient or prototypical case of a particular condition, and a "scheme-inductive" strategy in which clinicians follow a structured framework for diagnoses (e.g. a surgical sieve, that considers each pathophysiological system in turn). Of interest to our work was whether we would observe similar variation in reasoning strategies in our medical trainees and, if so, how these strategies related to patterns of information seeking and confidence We hypothesise not only that we can detect reasoning strategies based on the verbalisations of participants' thought process, but that different reasoning strategies for generating differentials are useful for some cases more than others. We also hypothesise that information seeking and changes in confidence vary as a function of the reasoning strategy employed.

Given the recording of qualitative data during this task, we can understand both how medical students are thinking about diagnoses as they are making them but also how they reflect on their thought process outside of the task. This detection of reasoning strategies, if successful, can then subsequently be used to detect the same reasoning strategies in this online study dataset (where we do not have access to the participants' thought process) based on the information sought. Given the higher sample size afforded by the online study, we can more robustly look at differences between reasoning strategies and whether they can tell us about what makes more accurate and more confident diagnoses.