Study 4 - Diagnostic Uncertainty and Information Seeking in Virtual Reality Paediatric Scenarios

Introduction

In the previous two studies, we used experimental paradigms that made use of textual vignettes, in which patient information was described to participants as per standardised case descriptions adapted from the work of previous researchers (Friedman et al., 2004). The first study was conducted online with medical students from across the UK, whilst the second study used an in-person think-aloud version of a similar paradigm with Oxford medical students. When taking these studies together, we find that the students' diagnostic confidence was associated with the amount of information they sought about the patient (i.e. about the patient's medical history, physical examinations and testing). Both of these variables were associated with the students' initial diagnostic breadth, which is the number of diagnostic differentials considered early on (based on Patient History information). This suggests the importance of early information during the diagnostic process and that the initial set of differentials is influential on the subsequent decisional process with regards to information seeking.

Through our think-aloud study, we used verbalisations during diagnostic decisions to classify each of the students' case under one of three reasoning strategies (adapted from the work of Coderre et al., 2003): Hypothetico-Deductive (HD), Pattern Recognition (PR) and Scheme-Inductive (SI). We found that using a HD strategy was associated with participants reevaluating their diagnosis more often and, perhaps

relatedly, higher accuracy when compared to the other strategies. We also found that different cases had 'dominant' strategies based on a majority of participants choosing a particular strategy. HD-dominant cases were associated with higher accuracy and higher initial diagnostic breadth, whilst PR-dominant cases were associated with higher information seeking. This hints not only at differences in these strategies in terms of how students approach the diagnostic process, but also that there are properties of patient cases that contribute to uses of particular reasoning strategies.

A common finding across both studies was that medical students showed a reticence to remove differentials from consideration, and instead broadened the differentials they were considering as they received more information. This brings up a question of whether such a tendency would be exhibited in real medical practice. One can imagine that when a patient requires treatment, the situation requires clinicians to commit to a working diagnosis at some point and remove others from consideration. There is also a limitation with our work thus far, in that the use of textual vignettes has limits in terms of naturalism. As we found during our systematic scoping review, the majority of past work on confidence during diagnoses made use of textual vignettes in their experimental paradigms. Past papers seemed to make limited use of high-fidelity simulations or other types of naturalistic paradigms (although some papers used in-situ questionnaires whilst clinicians are treating actual patients). Use of such simulation-based paradigms would have increased naturalism when compared to textual vignettes, as well as allowing researchers to look at how confidence impacts the actual treatment of patients in a controlled experimental environment. In the case of our work, we can use simulation of patient scenarios to better understand how confidence impacts diagnoses for patients who actually require treatment (rather than simply being described as per our vignettes) and develop over time in terms of condition.

In this study, we aim to extend our previous findings using a virtual reality (VR) experimental paradigm that is more naturalistic to real medical practice.

Participants were unable to see the patient in the vignette task, which is important given that the visual state (or distress) of a patient can be informative for a doctor in diagnosing the patient. By simulating this, we aim to investigate the link between information seeking and confidence in a more open-ended clinical situation that has a wider range of possible options for history taking, physical examination, testing and treatment options when compared to our vignette paradigm (which constrained the amount of information available on each case for usability). Given this increased flexibility, we can look at more fine-grained aspects of information seeking, as well as the effect of ongoing treatment of patients on confidence. Our vignette task was static in time, in that the patient does not change over the course of a case (i.e. improving or deteriorating over time). This VR paradigm then allows for doctors to start managing the patient's symptoms and even using reactions to their treatment plan in order to change their understanding of the patient.

In our previous two studies, we have found evidence for a general tendency for medical students to broaden the range of differentials they are considering as they receive more information. These studies made use of patient vignettes where there was no requirement to treat the patient and no subsequent observation of improvement or deterioration in the patients' state. This begs the question of whether medical students still show a tendency to broaden the differentials they are considering when beginning a treatment plan for a patient. This is a situation that intuitively requires a degree of narrowing of diagnoses. We predict then that with the use of VR scenarios where patients are deteriorating and require treatment, medical students will be more likely to narrow the differentials they are considering due to the medical situation demanding it. Similar to our previous studies, we measure the range of diagnostic differentials that students are considering at multiple points during the scenarios. Our online study found the initial diagnostic breadth of students was predictive of their subsequent information seeking and changes in confidence. We not only look to replicate this finding in a more naturalistic medical context but also to investigate whether initial diagnostic breadth is predictive

of patient treatment too. VR has seen limited use in previous work on clinical decision making but has potential for studying and improving clinical reasoning and decision making (Jans et al., 2023).

One of the aforementioned benefits of using a VR paradigm is that we are able to simulate a real medical environment. This includes the wide range of possible actions available to a clinician. Using our paradigm, we are able to record every action or information request made by participants. These actions can then be categorised into a number of areas: Patient History, Physical Examinations, Testing and Treatment. Our findings around initial diagnostic breadth and the qualitative theme from the previous study on the importance of an in-depth history to base diagnosis on necessitate a deeper look at history taking during diagnoses. Our vignette paradigm used fairly limited patient histories, with a perceived lack of detail potentially explaining why some participants expressed diagnostic uncertainty. In the VR paradigm, there is much more detail available on the patient's medical history, including follow-up questions to patients to access more detail on their condition. For example, if a patient is feeling pain, the interactive nature of VR allows participants to ask about the nature of the pain (e.g. whether it is a dull or sharp pain, whether anything makes the pain better/worse etc.). With the wider range of actions available to participants, we not only look at information seeking as a whole, but also information seeking with each of these categories. In particular, we are interested in how the comprehensiveness of participants' history taking affects their subsequent confidence and considered diagnoses. Given that VR also simulate active medical situations, participants can be graded based on the information they seek, the tests they run and treatment they administer.

Due to our VR methodology being substantially different to our vignette methodology, the manner in which we think of accuracy has to change to reflect this. In the previous studies, we operationalised accuracy given that there was a specific condition/diagnosis that participants were tasked with identifying. In this

task however, determining a diagnosis is not the primary focus of the task (although we do ask participants to report the set of diagnostic differentials that they are considering). Instead, participants are required to begin treatment for the patient in the scenario and handover the case to a senior. Given this, there are two ways in which performance can be measured for participants: performance in terms of the clinical actions (e.g. testing, treatment etc.) they take or in terms of the diagnoses they report. For the former, we hence make use of predetermined criteria for which clinical actions are considered optimal for each patient scenario. For the latter, given that the scenarios are more naturalistic, there is not a correct (ground truth) condition that the patients have. For example, one of the scenarios sees the patient having a febrile convulsion/seizure. Identifying this as such, as a focal diagnosis, is expected of most medical students due to there being a lack of diagnostic uncertainty. Identifying the causes of this seizure however, is associated with more diagnostic uncertainty as there are several possible causes of such a medical episode. When it comes to identifying these causes, there is not a set correct answer, as the scenario does not comprise of later stages of the patient's care pathway. Because of this, we instead consider a measure of Diagnostic Appropriateness, where we measure how suitable the recorded set of diagnostic differentials is as a whole in terms of whether participants record differentials that would be considered plausible or likely given the patient's condition.

Research Questions

With this study, we investigate the following research questions:

- Do medical students narrow or broaden their diagnostic differentials in a naturalistic medical scenario where patient treatment is required?
- Is information seeking, in terms of quantity and quality, linked to more appropriate sets of diagnoses?
- How do specific types of information seeking (i.e. around Patient History, Physical Examinations and Testing) relate to confidence, both in terms of information seeking preceding confidence, and as a result of confidence?

Methods

Participants

We recruited medical students based at the University of Oxford in their second year of clinical training (which equates to three or four years of educational experience). 76 students completed this study.

Materials

We used VR scenarios implemented by Oxford Medical Simulation (OMS, https://oxfordmedicalsimulation.com/), a company that implements bespoke VR software for medical education and simulation. Participants in this study were medical students based in Oxford who were at the time taking part in VR-based teaching sessions as part of their medical degrees. Students performed the scenarios using Oculus Quest 2 VR headsets. Scenarios were based in paediatrics, meaning that the patients in the scenario were children who were attending the hospital with their legal guardian. Each scenario features a visual 3D implementation of a basic wardroom in a hospital. Participants are shown a (child) patient, their guardian and a nurse who can help with certain treatment and testing. All of the 'avatars' in the scenario can be questioned by the participant using a predefined set of requests/actions (e.g. asking the nurse to check blood pressure, asking the patient/child about if they are in pain). The scenarios have full sound (e.g. being able to hear the patient's lung auscultation) and the avatars are voiced.

Figure 1: Screenshot from the VR software, implemented by Oxford Medical Simulation. Depicted here is the patient/child, their parent/guardian and a nurse (who can be asked to seek tests or administer treatment)

Figure 2: Screenshot from the VR software. Depicted here is the participant consulting available guidelines on management of asthma patients.

Each participant completed two scenarios over two separate VR sessions. The sessions were held around one month apart. During each session, the participants each performed one scenario in VR and observed their partner during their scenario. Participants also engaged in peer-to-peer feedback discussions as part of their education. We chose medical scenarios that were considered fairly common to arise for paediatric patients. The scenarios presented in each session are described below (students are split into two groups, shown below as groups A and B, each performing a different pair of scenarios in a fixed order):

- Session One:
- – *Group A:* patient/child is a 6-year-old-girl presenting with a 1-day history of central abdominal pain and thirst. She was generally unwell for 2 days prior, with a reduced appetite and a sore throat. Collateral history reveals Type 1 Diabetes and erratic blood sugars. (**Underlying Condition: Diabetic Ketoacidosis**)
- – *Group B:* patient/child is a 5-year-old boy presenting with worsening shortness of breath, wheeze, and signs of respiratory distress, on the background of 2 days of likely viral illness. He has a medical history of asthma and has had similar exacerbations in the past. (**Underlying Condition: Acute Severe Exacerbation of Asthma**)
- Session Two:
- – *Group A:* patient/child is a 5-year-old boy presenting with shortness of breath and drowsiness (**Underlying Condition: Chest Sepsis/Pneumonia**)
- – *Group B:* patient/child is a 5-year-old girl with a 1-day history of sore throat and fever. She starts having a generalised tonic-clonic seizure during the scenario. (**Underlying Condition: Febrile seizure on background of tonsillitis**)

Procedure

The aim for students in the scenarios was to diagnose the patient, begin treatment and hand over the case to a senior with appropriate understanding of the patient (handovers were conducted using a standardised framework known as SBAR, meaning that clinicians have to brief the senior on the Situation, Background, Assessment and Recommendation for the patient). They were expected to take a clinical history, complete a physical examination, start emergency treatment to stabilise the patient and escalate to a senior clinician for further input. Whilst in the scenario, participants can learn about the patient's medical history, check key parameters (such as temperature, pulse, blood pressure, respiratory rate etc), perform physical exams/tests and begin certain treatment actions (such as administering oxygen or prescribing medication). Participants were also expected by the end of the scenario to be able to give an explanation of the situation to the patient's parent/guardian. All participants have the same starting point in each scenario and the patient in the scenario deteriorates in an identical way if the participant takes no action. If participants undertake certain actions, the patient improves both in terms of vital signs (e.g. blood pressure, heart rate, oxygen saturation etc.) and in their response to questions (e.g responding "Yes, I feel a bit better" to a question of how they are feeling). If participants select irrelevant actions, the patient does not improve, whilst some actions will result in the patient's state deteriorating.

After 5 minutes in the scenario (by which point it is expected that participants would have a history of the patient and have started some early assessment of the patient), participants are asked to pause the scenario (taking off their VR headset) and fill in a brief questionnaire on paper. Multiple VR participants were performing the scenario simultaneously and were paired with another student who would watch their performance. This other student would aid with administering the questionnaire, with the students subsequently switching roles for the other

scenario. The VR participant was asked in the questionnaire to answer the following (this is considered time point 1):

- "Please say all the conditions that you are currently considering or are concerned about for this patient. Include any/all common, rare or contributing conditions you are considering. For each, please rate how likely you think they are on a scale of 1 (low) to 5 (high)."
- "On a scale of 1-10, how confident are you that you understand the patient's condition?"
- "How severe do you think the patient's condition is on a scale of 1 to 10?" (Each point of the scale represented a different clinical action/course, with 1 representing "Discharge in <4 hours, no follow up" and 10 representing "Requires arrest/peri arrest team.")
- "To what extent would you be prepared to leave the patient prior to a senior review" (this question was answered using a visual analogue scale)
- "Did you complete all the history, examinations and investigations necessary? If not, what else would you do if given more time?"

During the scenario, participants had access to a phone on the wall that could be used to call external staff members for help. In terms of general options for external second opinions, participants could call the on-call doctor or the day team on the ward. For more serious situations, participants could call the rapid response team, crash team or push the emergency button (in very severe situations). Participants were expected to at least handover to another staff member using the SBAR protocol but could request the help of a senior before this point. Crucially, the structure of the scenario (including the amount of time spent in the scenario) is dictated by the participant, who can handover the case when they feel they have done enough to stabilise the patient and understand enough to handover to a senior. Once the participant is ready to finish, they 'leave the room' in the VR ward. However, the scenario automatically ends after 20 minutes if the participant has not finished it themselves.

Data Analysis

The dependent variables that we derive are as follows:

- **Performance Score**: The OMS software implements a series of objectives for each scenario, which are tasks or actions that the participant is expected to have completed within the allotted time. These objectives are comprised of a range of actions, including examinations that they were expected to conduct, history questions they should have asked and treatment that should have been administered. These objectives can include administering oxygen, prescribing a particular medication or calculating the Patient Early Warning Score (PEWS). The proportion of completed objectives is used as a score of the participant's performance during the scenario. See Figure 3 below for an example of objectives used for the Pneumonia scenario.

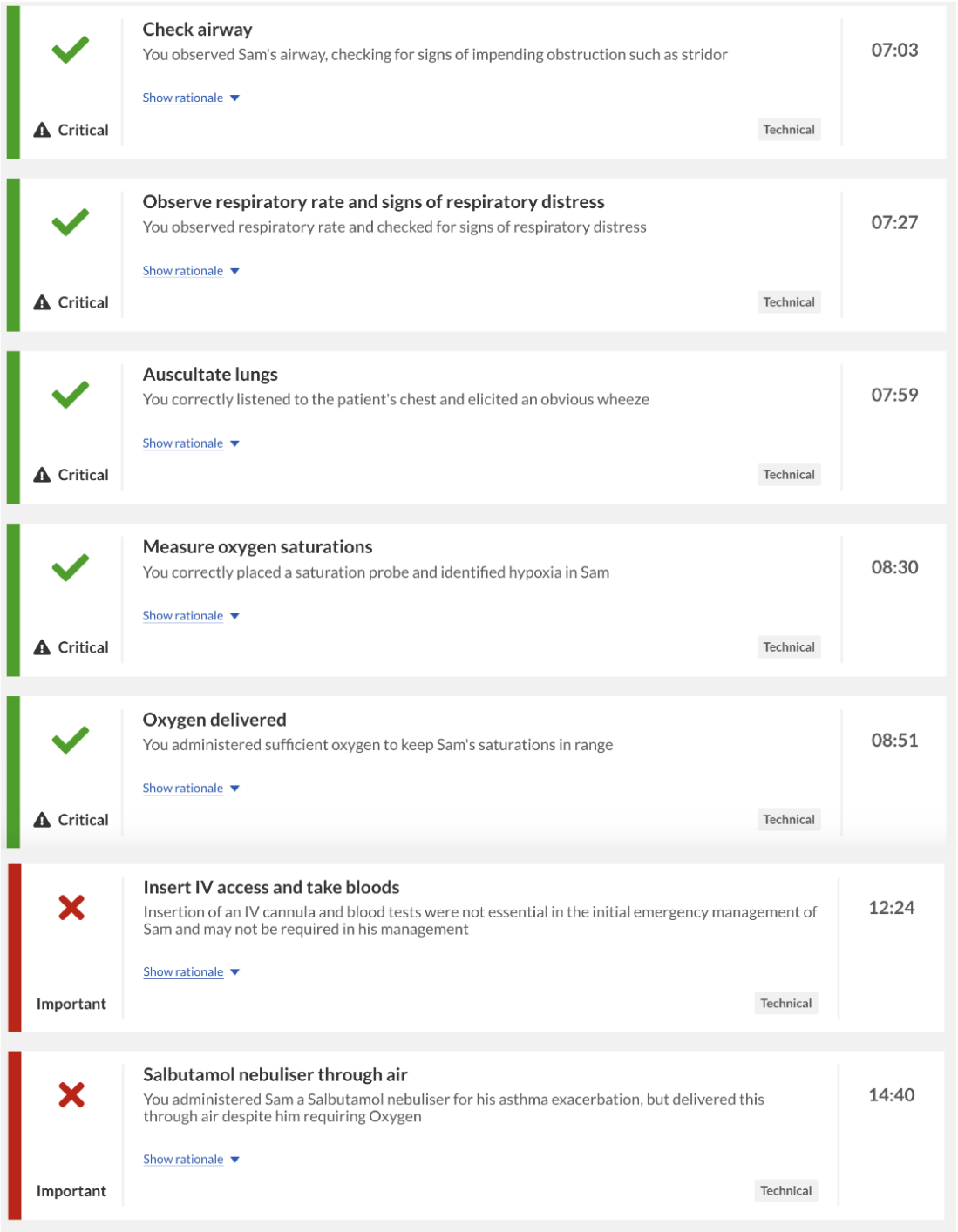

Figure 3: An example of scoring criteria used when calculating the Performance Score. The criteria are pre-defined specifically to each scenario. This example shows objectives that have been met (as denoted by a green tick) and those that have not been met (as denoted by a red cross). This example is taken from the OMS software, which calculates Performance Score internally for each case.

- **Confidence Change**: the participants' confidence in their understanding of the patient's condition is recorded at two time points, with the first being after 5 minutes (out of the 20-minute time limit) and the second being after the participant has finished the scenario. Confidence at each stage is recorded on a 10-point scale (1-10). The difference between the second and the first confidence rating is taken, such that a positive value indicates that the participant has increased their confidence over the course of the scenario.

- **Number of Differentials**: participants are asked to record all the diagnostic differentials that they are considering at the two aforementioned time points. Hence, the total number of differentials is recorded at each stage. The Initial Number of differentials is the number of diagnoses provided at the pause point.

- **Diagnostic Appropriateness**: each participant's set of differentials are assessed for how appropriate they are for the scenario. Each scenario has a set of differentials that are considered most likely, probable and improbable (with any others considered incorrect). To calculate a score for how appropriate the diagnoses are, we consider what proportion of likelihood values are assigned to likely differentials. We sum the likelihood values provided for all differentials that were marked as most likely or probable. We then add these to the sum of likelihood values for improbable differentials divided by two. This sum is divided by the total sum of all differentials. However, we also penalised participants for providing fewer differentials, such that high scoring sets of differentials are larger sets of likely or probable differentials. This is because we expect participants to provide a wider set of differentials that could be contributing to the patients' conditions. We only calculate this score for the initial set of differentials recorded during the pause point, as the number of differentials changes between the two time points.

To define this metric formally, we express **Diagnostic Appropriateness (A)** as follows:

Let $L = \{l_1, l_2, \ldots, l_m\}$ be the set of all likelihoods provided across participants, where each array of likelihoods provided for a given scenario l_j has a length $|l_j|$ (i.e. the number of differentials recorded).

We define:

- $n_{\max} = \max(|l_1|, |l_2|, \ldots, |l_m|)$ (length of the longest array in L)
- The penalty for a lower number of provided differentials:

$$\lambda_j = n_{\max} - |l_j|$$

Diagnostic Appropriateness for a given set of differentials and likelihoods A_j for each array l_j can be expressed as:

$$A_j = \frac{S_{p,j} + \frac{1}{2} S_{u,j}}{S_{L,j} + \lambda_j}$$

Where:

- $S_{p,j} = \sum_{l_k \in l_j, l_k = p} l_k$ (sum of likelihoods for probable/possible differentials for that scenario/condition)
- $S_{u,j} = \sum_{l_k \in l_j, l_k = u} l_k$ (sum of likelihoods for unlikely differentials for that scenario/condition)
- $S_{L,j} = \sum_{l_k \in l_j} l_k$ (total sum of likelihoods across all differentials in l_j)

We also derived measures of information seeking similar to previous studies. The VR scenarios are far richer in terms of the available set of information for participants when compared to the vignette paradigm. For our analysis, we record all actions (or 'clicks) made by participants whilst in the scenario. Actions are categorised into a number of groups. The main categories are labelled as History, Examination or Testing, similar to in the vignette study. This set of information is mostly similar across scenarios though there are minor differences especially in the History category. Across scenarios, there are 35 possible History actions, 29 Examination actions and 18 Testing actions. This especially means that in comparison to the vignette paradigm, participants can take more detailed patient histories and can receive very different pieces of information depending on what they request from patient documentation and from asking the patient/guardian in the scenario. Outside of these categories, there are other actions available to participants, such as administering medication for the patient, calling for help or providing reassurance to the patient/guardian, but these are not used for our analysis. After categorising the participants' actions, we define a number information seeking measures:

- **History Taking**: this is the number of History actions for a given scenario that take place before the pause point.
- **Total Information Seeking**: this is the number of unique actions (i.e. does not include requesting the taking of the same action multiple times) classified under History, Examination or Testing across the scenario.
- **Information Value**: to calculate the value of each piece of information sought across these categories, we calculate the difference in OMS performance score for participants with or without that information on a given scenario. We then sum all values across all information sought by the participant within each of the information categories (History, Examination, Testing).
- **Amount of Treatment**: this is the number of actions classified as treatment of the patient across the scenario.

Given that, unlike our vignette paradigm, participants are able to administer treatment and ask for help from a senior member (using the telephone in the VR ward), we measure how long it takes for students to do both of these:

- **Time to Treatment**: this is the amount of time (in seconds) between the start of the scenario and the point at which the first treatment action is performed.
- **Time to Call for Help**: this is the amount of time (in seconds) between the start of the scenario and the point at which the participant uses the phone to call a senior for help.

We expect that confidence will predict how long it takes for participants to both start treatment and call for help, with higher confidence being associated with quicker times to start treatment and slower times to call for help (given that they do call for help at some point during the scenario). We exclude participants from these analyses if they do not ask for help or do not administer any treatment respectively.

As all actions are recorded with timestamps in the output dataset, we categorise whether actions occurred before or after the pause point (5 minutes in). Hence, we can investigate information seeking before and after the pause point where participants record their initial diagnoses and confidence. For this study, we are particularly interested in the relationship between confidence and information seeking as it follows the timecourse of a diagnostic decision. To this end, we look at whether information seeking up until the pause point predicts initial confidence (as reported during the pause point). We then look at whether this initial confidence predicts subsequent information seeking (after the pause point). This allows us to look at the relationship between information seeking and confidence in both directions: respectively, how information informs subsequent confidence and how confidence informs subsequent information seeking. This is important given that

the directionality of this relationship was unclear from the previous studies we conducted. Finally, we look at whether the amount of patient treatment during the scenario is predicted by confidence, both before and after participants administer treatment. We investigate this by looking at the number of treatment actions performed and its relationship with both initial and final confidence. These analyses are performed on a case-wise level, and thus we used generalised mixed effects modelling due to the lack of independence between observations. As such, we control for individual participants and the patient condition/case as random effects.

Given that Diagnostic Appropriateness is our variable for diagnostic accuracy in this study, we look at information seeking as a predictor of accuracy similar to Study 2. We do not use the OMS performance score as a measure of accuracy to relate to information seeking, as this score is in itself calculated using the information sought and actions taken by students in the scenario. Given the much larger set of possible information that could be sought in the VR scenarios, we use Principal Component Analysis (PCA) to reduce the dimensionality of the information seeking data.

Results

Overall Performance

We report data from 76 participants. As shown in Table 1, some participants only completed a single scenario rather than two. 41 participants completed two scenarios (as part of either Scenario A or B as explained in the Procedure section). Overall, 37 participants completed the Asthma scenario, 30 participants completed the DKA scenario, 28 participants completed the Pneumonia scenario and 22 participants completed the Seizure scenario.

In terms of overall performance, the average OMS score across all scenarios and all participants was 63.35 (ranging from 38 to 79), which indicates the percentage of predefined 'objectives' successfully completed by participants during each of the

scenarios (each scenario has its own set of objectives, with some overlap). The mean OMS score for each scenario was as follows: Asthma = 68.14, DKA = 57.3, Pneumonia = 63.31, Seizure = 63.81. The average Diagnostic Appropriateness score across all scenarios and participants was 0.55 (ranging from 0.07 to 0.85). The mean Diagnostic Appropriateness score for each scenario was as follows: Asthma = 0.61, DKA = 0.53, Pneumonia = 0.48, Seizure = 0.59. These performance measures are weakly correlated as per a Spearman's Rank Correlation test ($rs(111) = 0.19$, $p = 0.04$).

We next look at confidence reported by participants and its calibration to objective performance. Across all scenarios, participants increase their confidence between the two timepoints. Average initial confidence (recorded on a 10-point Likert scale) across scenarios was 6.17 and average final confidence was 8.06. By scenario, initial confidence and final confidence respectively were: Asthma = [5.7, 7.71], DKA = [7.1, 8.79], Pneumonia = [6.32, 7.84], Seizure = [5.55, 7.9].

Scenario	n	OMS Score	Information Seeking	Initial Confidence	Final Confidence	Initial Diagnoses	Diagnostic Score
Asthma	37	68.14	19.73	5.70	7.71	4.38	0.61
DKA	30	57.30	19.47	7.10	8.79	2.87	0.53
Pneumonia	28	63.31	24.46	6.32	7.84	3.64	0.48
Seizure	22	63.81	24.09	5.55	7.90	3.59	0.59

Table 1: Average values for dependent variables by scenario. The n column denotes the number of participants (or 'observations') per scenario. We show mean values for the Performance Score, Amount of Information Seeking, Initial Confidence (as reported at the pause point in the scenario), Final Confidence (reported at the end of the scenario), Initial Diagnoses (the number of differentials reported at the pause point) and Diagnostic Appropriateness Score.

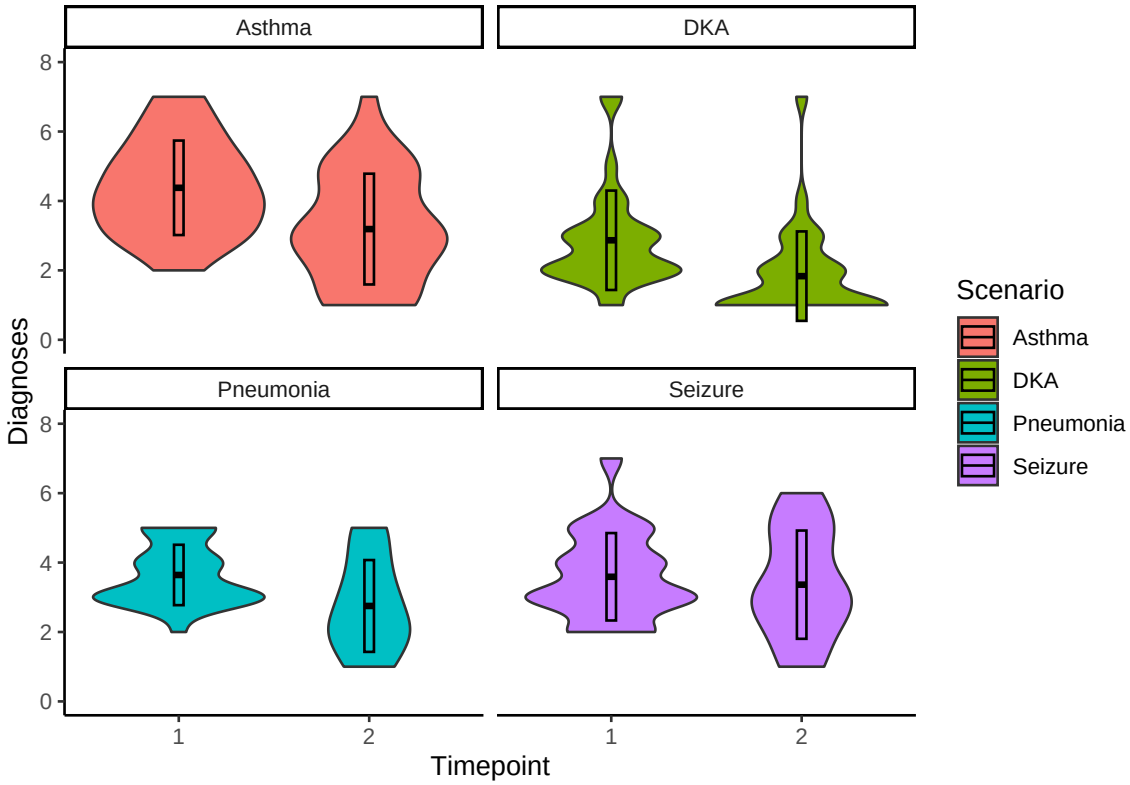

Figure 4: Violin plots showing the number of reported diagnoses at timepoint 1 (the pause point at 5 minutes into the scenario) and timepoint 2 (at the end of the scenario) by condition (Asthma = red, DKA = green, Pnuemonia = blue, Seizure = purple). The dark region of the box plot shows the mean value, with the lines of the box plots showing standard deviation.

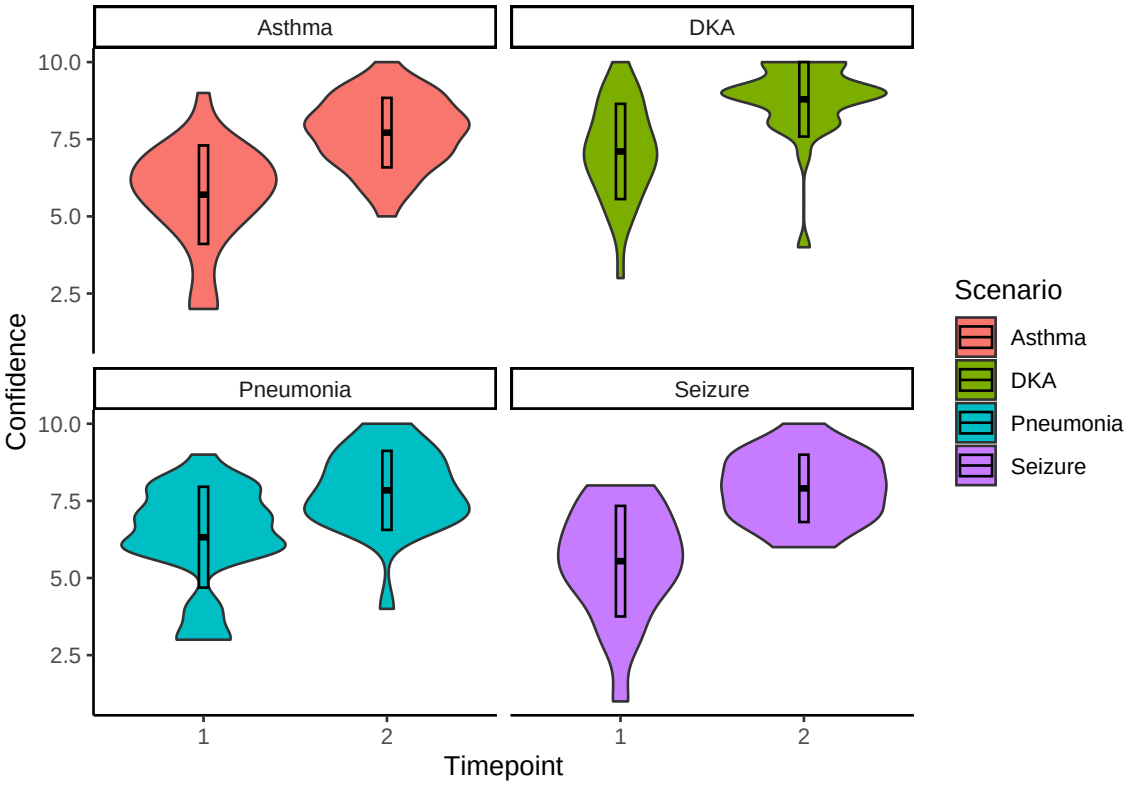

Figure 5: Violin plots showing confidence at timepoint 1 (the pause point at 5 minutes into the scenario) and timepoint 2 (at the end of the scenario) by condition (Asthma = red, DKA = green, Pnuemonia = blue, Seizure = purple). The dark region of the box plot shows the mean value, with the lines of the box plots showing standard deviation.

As with previous studies, we now look at whether participants provide confidence judgements that are calibrated to their objective performance. We separately determine how calibrated initial and final confidence judgements are. Initial Confidence is compared against the Diagnostic Appropriateness (which is calculated based on the differentials provided as the pause point), whilst Final Confidence is compared against Performance Score (which is calculated at the end of the scenario based on the actions performed and information sought across the scenario). We median split cases into two groups of low and high Diagnostic Appropriateness and also into groups of low and high Performance Score. Participants would be considered calibrated if we found evidence of higher confidence when performance

is higher. Contrary to expectations, participants indeed reported lower Initial Confidence when Diagnostic Appropriateness was higher (M = 6, SD = 1.74) compared to when it was low (M = 6.36, SD = 1.7). Given that the samples from each of these groups are not independent, we test for a difference between these groups using a binomial mixed effects model that predicts performance group using initial confidence as a fixed effect and both the individual participant and condition as random effects. We do not find evidence of confidence being predictive of Diagnostic Appropriateness (β = -0.09, SE = 0.13, z = -0.68, p = 0.5). For final confidence, participants reported lower Final Confidence when Performance Score was higher (M = 7.95, SD = 1.28) compared to when it was low (M = 8.19, SD = 1.21). We test for a difference between these groups using a binomial mixed effects model that predicts performance group using final confidence as a fixed effect and both the individual participant and condition as random effects. We do not find evidence of confidence being predictive of Performance Score (β = -0.06, SE = 0.2, z = -0.31, p = 0.75). Overall, we do not find evidence that participants provide calibrated confidence judgements at either timepoint with either of our measures of accuracy/performance.

Initial Diagnostic Breadth

We now look at whether the initial diagnostic breadth (i.e. the number of diagnostic differentials being considered early in the scenario) is predictive of information seeking and change in confidence over the course of the scenario (as we found evidence for such an association in Study 2). We fit mixed effects models to predict each of these with the number of initial diagnoses as a fixed effect and both the scenario and participant as random effects. We do not see evidence that the initial diagnostic breadth is predictive of the amount of information seeking (β = 0.02, SE = 0.02 z = 0.99, p = 0.32) or changes in confidence (β = 0.14, SE = 0.12 t = 1.19, p = 0.24). As a result, we are not able to replicate findings from Study 2 on a case-level, in which initial diagnostic breadth was predictive of information seeking and changes in confidence when averaged across each of the participants' cases.

Information Seeking and Confidence

We now ask whether confidence is related to the amount of information sought on a given case. To investigate this, we look at information seeking before and after the pause point and look at both initial and final confidence. This allows us to look at this association in both directions: whether the amount of information seeking predicts subsequent confidence and whether confidence predicts subsequent information seeking. We fit generalised mixed effect models using the amount of information seeking in each of the three categories (Patient History, Physical Examinations and Testing).

We first look at whether initial confidence is predicted by information seeking prior to the pause point (i.e. prior to when this initial confidence was reported). For this, we use a generalised mixed effect model with a Poisson distribution. We do not find evidence that initial confidence is predicted by prior information seeking related to Patient History ($\beta = 0.01$, SE = 0.02, z = 0.67, p = 0.5), Physical Examinations ($\beta = 0.01$, SE = 0.03, z = 0.23, p = 0.82) or Testing ($\beta = 0.03$, SE = 0.04, z = 0.85, p = 0.4).

We next look at whether initial confidence predicts subsequent information seeking. To investigate this, we fit separate linear mixed effect models for each type of information seeking as the outcome variable. We find marginal evidence that initial confidence predicts subsequent history taking in a negative direction (i.e. that higher confidence is associated with lower subsequent history taking) ($\beta = 0.08$, SE = 0.04, z = 1.91, p = 0.06). We do not find evidence that initial confidence is associated with subsequent Physical Examinations ($\beta = 0.01$, SE = 0.03, z = 0.42, p = 0.67). We do find however that initial confidence was associated with higher amounts of Testing ($\beta = 0.09$, SE = 0.04, z = 2.34, p = 0.02). We plot this model's fitted values against the actual observed values below in Figure 6.

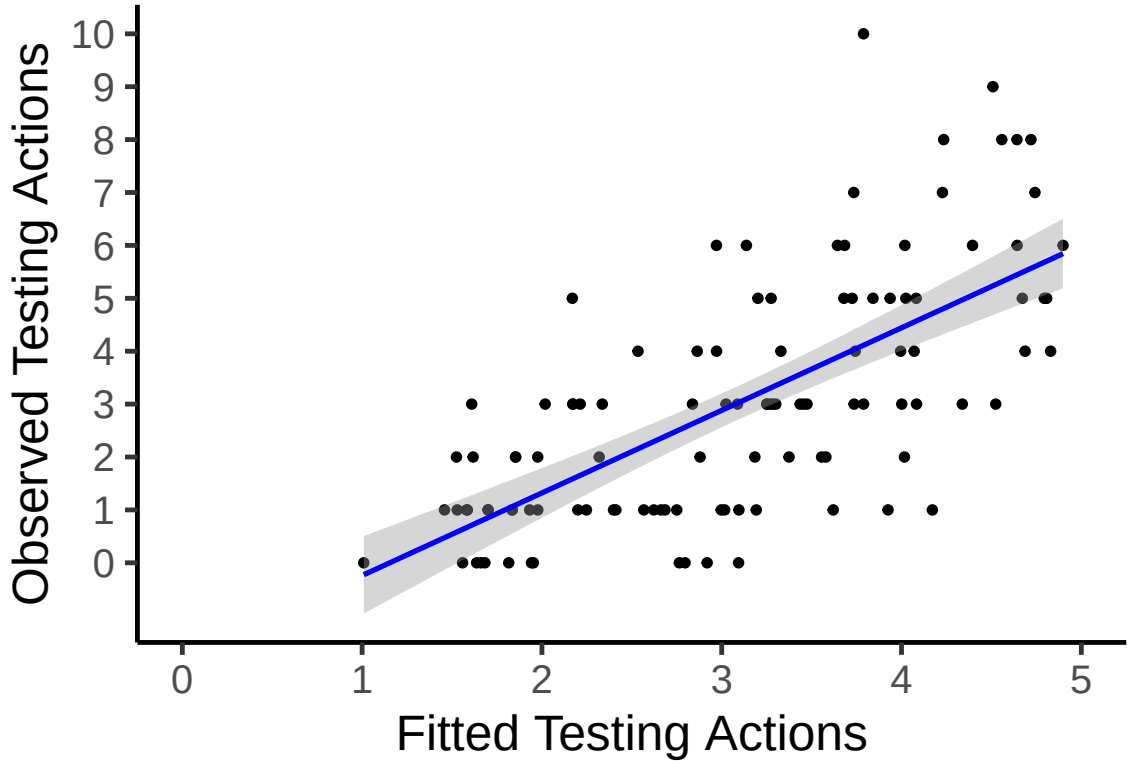

Figure 6: Plot of linear mixed effects model predicting the amount of testing actions (after the pause point) by the initial confidence reported (during the pause point). We show the fitted values for the number of testing actions (x-axis) against the actual observed number of testing actions (y-axis) on each case (each data point representing a single case). We fit a linear model line of best fit with a 95% confidence interval denoted by the shaded region

We finally look at whether final confidence (as reported at the end of the scenario) is predicted by the number of treatment actions performed by participants during the scenario. We fit a linear mixed effects model with the number of treatment actions as a fixed effect and both scenario and participant as random effects. We find evidence that final confidence was predicted by the amount of treatment actions administered during the scenario ($\beta = 0.38$, SE $= 0.13$ t $= 2.99$, p $= 0.003$). We plot the model's fitted values against the actual observed values below in Figure 7.

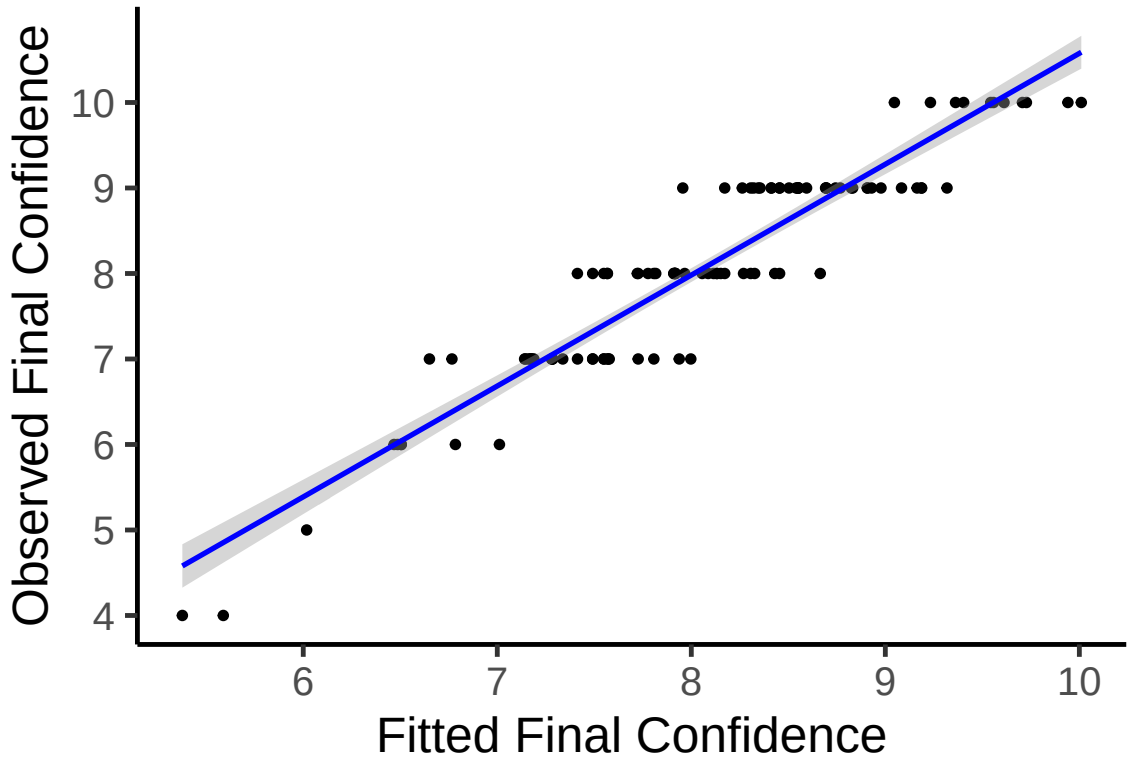

Figure 7: Plot of linear mixed effects model predicting final confidence (at the end of the scenario) by the amount of treatment actions. We show the fitted values for final confidence (x-axis) against the actual observed values for final confidence (y-axis) on each case (each data point representing a single case). We fit a linear model line of best fit with a 95% confidence interval denoted by the shaded region

Time to Treat and Call for Help

We next turn to the time taken for participants to start treatment of patients and to call for help from a senior in the scenarios. We use linear mixed effect modelling to determine if the time taken to start treatment and to call for help is predicted by both confidence judgements and ratings of patient severity. We control for individual participants and scenarios as random effects. We do not find evidence that Initial Confidence predicts the time taken to start treatment or ask for help (ps $>$ .1). 16 cases were excluded from the analysis involving time taken to treat, as no treatment actions were recorded for these cases. 9 cases were excluded from the analysis involving time taken to call for help, as no help actions were recorded

for these cases. We do not find evidence that initial ratings of severity predict the time taken to start treatment (β = -11, SE = 8.27, t = -1.33, p = 0.19), but we do find that severity is predictive of the time taken to call for help (β = -26.13, SE = 10.79, t = -2.42, p = 0.02). Each unit increase in severity rating was associated with participants seeking help 26.13 seconds quicker.

Diagnostic Appropriateness

We next ask whether the diagnoses provided by participants is a result of 'better' information seeking. If this were the case, we would be able to differentiate between low and high quality diagnoses (as per our Diagnostic Appropriateness measure) solely from the information sought by participants. This is similar to the analysis presented in Study 2 (in Figure 8). We split participants into low and high diagnostic accuracy via a median split. Given the large amount of information available (82 unique information requests across History, Physical Examinations and Testing), we reduce the dimensionality of the information seeking data using Principal Component Analysis. When computing a scree plot and observing eigenvalues for each cumulative component that exceeded 1, we find that reducing the 82 information factors to 26 components is recommended.

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. When plotting an ROC curve, the area under the curve (AUC) is indicative of how well a model performs at correctly categorising cases as having high or low diagnostic appropriateness. 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 0.77 (plotted below). When conducting a DeLong test, to test the null hypothesis that the AUC is equal to 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. This indicates that diagnostic appropriateness is predicted by information seeking during the scenarios.

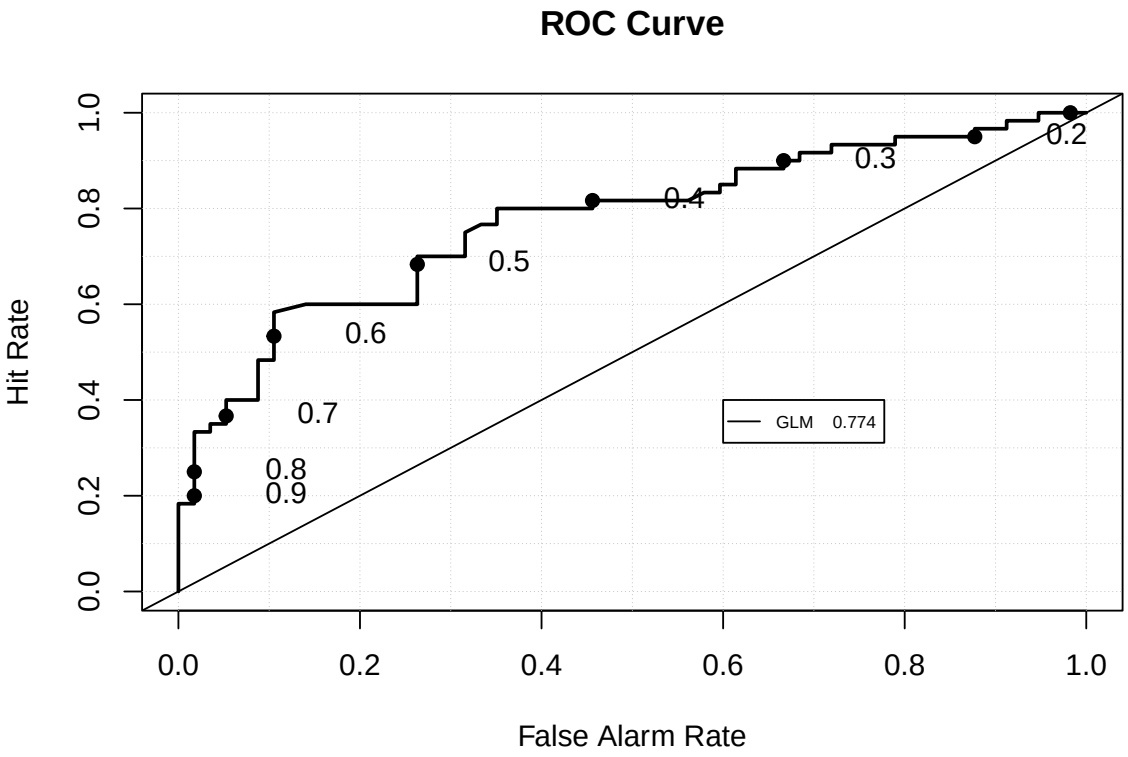

Figure 8: Receiver-Operator Characteristic (ROC) curve using a Generalised Linear Model to classify individual cases as having either high or low diagnostic appropriateness reported based on the initial set of differentials. The model is trained PCA components of the total available information requests, with the 82 information requests reduced to 26 components. Cases were sorted as high or low diagnostic appropriateness based on a median split.

Given that information seeking is broadly associated with diagnostic appropriateness, we seek to better understand the specific aspects of information seeking that relate to appropriate diagnoses. As in previous studies, we look at two aspects of information seeking: the amount of information seeking and informational value. To this end, we look at whether the information seeking amount and value for each of the three categories (History, Physical Examinations, Testing) predicts diagnostic appropriateness score on each case. We fit generalised mixed effect models with

inverse Gaussian distributions by controlling for scenario and participant as random effects. We find that diagnostic appropriateness is not predicted by the number of actions in any of the information seeking categories (Fs < 1, ps $> .1$). We do however find evidence that diagnostic appropriateness is associated with the value of Patient History taking ($F = 4.85$, $p = 0.03$), but not with Physical Examination value ($F = 2.79$, $p = 0.09$) or Testing value ($F = 0.54$, $p = 0.46$). The relationship between Patient History Value and Diagnostic Appropriateness is plotted below in Figure 9.

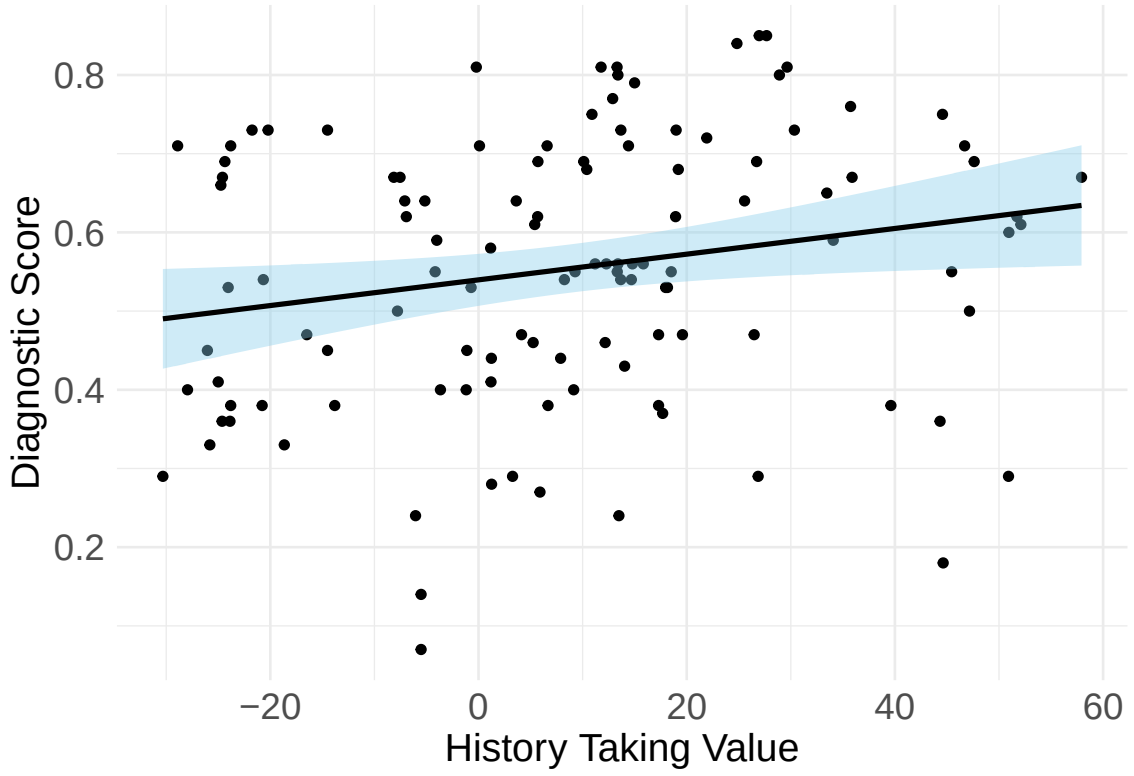

Figure 9: Scatter plot showing the relationship between information value for information requests in the Patient History category (x-axis) and Diagnostic Appropriateness score (y-axis) calculated based on the initial set of differentials provided at the pause point in the scenario. Each data point in this plot represents a single case/scenario.

Discussion

In this study, we used paediatric Virtual Reality scenarios to study how medical students seek information and consider diagnostic differentials in a naturalistic

manner. During the study, we used 4 scenarios that represent common paediatric cases within real medical practice. In each scenario, medical students were tasked with diagnosing the patient, beginning treatment and handing over the case to a senior with an appropriate understanding of the patient. Participants were paused after 5 minutes in the scenario to report the diagnoses they were considering for the patient and how confident they were that they understood the patient's condition. These questions were then answered again at the end of the scenario to examine how their diagnostic thinking had changed over the course of the scenario. We recorded all information sought by participants, especially on Patient History, Physical Examinations and testing, as well as Treatment actions taken. We were especially interested in understanding how both history taking and patient treatment affects both confidence and diagnostic accuracy.

On overall performance, we used two different measures of performance that took into account each of the four scenarios. We defined Diagnostic Appropriateness, which assesses the differentials provided at the first timepoint and how likely they are for the patients' condition. We also defined Performance Score, which is based on the proportion of pre-defined objectives met by participants during the scenario (such as requesting specific tests, starting certain treatment etc.). We did not find evidence for confidence judgements being calibrated to either performance measure, which marks a point of difference when compared to our previous vignette studies. A potential reason for this could be that medical students were less introspective when formulating their confidence and instead relying on markers from the patient, such as if the patient's condition improves over the course of the scenario. In the previous studies, as there was no visible patient and no treatment administered, confidence judgements could be solely based on the participants' own understanding and knowledge of the patient's condition. By relying on patient observations in this study however, this could lead clinicians away from their own self-reflection of what knowledge they have about the patient and how certain they are. This does hint at differences in behaviour between the controlled, 'medical theory' based vignette

studies and the naturalistic practically-focused VR paradigm. This account could explain how miscalibration could arise in medical practice, but this requires further study to understand how patients themselves contribute to diagnostic confidence.

On diagnostic breadth, we did not replicate our finding from Study 2 that the initial diagnostic breadth of medical students (i.e. the number of differentials recorded during the pause point at 5 minutes in) did not predict information seeking or changes in confidence. Given we observed this relationship in Study 2, we consider the difference in analysis. In Study 2, we looked at this relationship with a correlation of individual differences averaged across each participant's cases. In this study however, we looked at this relationship on a case-by-case level. This could then suggest that initial diagnostic breadth has different properties when studied as an individual-level factor or a case-level factor. Future work could elucidate this further by studying how individual decision makers differ in their tendencies to consider a broad or narrow set of differentials (similar to our work on individual reasoning strategies).

On confidence and information seeking, we were able to look at information seeking in a more fine-grained manner in comparison to our previous studies due to the paradigm's open-ended nature and greater availability of information requests, testing and treatment in the VR scenarios. As a result, we were able to study information seeking not just as a whole, but related to specific types of information that clinicians use to formulate diagnoses. We did not find initial confidence was predicted by information seeking prior to that point. We do however find that initial confidence then predicted the amount of subsequent testing that medical students requested. If we intuitively consider the different stages of the diagnostic process, as we observed in Studies 2 and 3 with our vignette methodology, clinicians tend to request tests when they are honing in on a particular diagnosis and want to either confirm their beliefs or rule out an alternative diagnosis. Given we observed little evidence of the latter in our previous studies, it is then more likely that clinicians

perform tests to focus in on a particular diagnosis. This would explain why, with higher confidence, medical students in this study subsequently request more tests as they seek to confirm their diagnostic hypotheses. This is especially supported by the higher fidelity of information available in VR (e.g. ECG traces, blood gas) when compared to textual vignettes. We also found that confidence was predicted by the amount of treatment administered by students. The consideration of treatment in this study is certainly an important aspect of real medical practice to emulate. When formulating a diagnosis, clinicians then use this diagnosis to guide their future treatment and care pathway. By administering treatment, both in real medical practice and in our VR scenarios, clinicians can observe the patient changing in terms of their condition. If a clinician decides to administer oxygen to the patient, they may then observe the patient's oxygen levels increase if successful.

We also use the aspect of treatment and patient improvement to explain our overall finding in this study that differentials narrowed between the two timepoints, rather than broadening as in the previous studies. The act of administering treatment and observing the patient's reaction to this treatment is a key part of the diagnostic decisional process, as it provides clinicians with a form of feedback on their decisions. When participants could not administer treatment in the vignette studies and observe the patient's change in condition (either improving or deteriorating), they then do not receive feedback that can be used to support or rule out diagnoses. Taken together, this provides an important consideration for future work looking at diagnostic uncertainty, in that methodologies without treatable patients (e.g textual vignettes) may result in different behaviour to how clinicians would approach such diagnoses in everyday practice.

On the time taken to treat the patient and call for help, we do not find evidence that confidence was predictive of how long it took for participants to provide treatment or ask for help. We found evidence that perceptions of patient severity were related to asking for help, such that higher severity ratings were associated with quicker

requests for help. The lattermost finding makes intuitive sense, as doctors are likely required to escalate more serious cases to other staff with more urgency if patients require inputs from other specialists or other departments.

On diagnostic appropriateness, we used a score for diagnostic accuracy that took into account the range of differentials that medical students considered. We adopted this measure to assess diagnostic thinking as a whole, rather than simply identifying a focal diagnosis/condition correctly. This was important to do given that the diagnostic uncertainty came not from the focal condition, but from identifying its source and causes. We found that information seeking was predictive of differences in diagnostic appropriateness. More specifically, we found more informative/valuable history taking was associated with higher diagnostic appropriateness. There is a heuristic taught within medicine that history taking alone determines between 70% and 90% of diagnoses (Keifenheim, 2015). This would then explain why we observe the positive effect of optimal history taking on diagnostic performance. We show that with a more appropriate patient history, participants are better able to understand the patient's condition and its possible causes. This suggest that future work and interventions could be especially effective when focused on history taking and early information seeking by clinicians. We note that with this measure, we operationalise diagnostic accuracy quite differently to previous work. Such a measure is analogous to real practice as clinicians may not always be able to identify a focal diagnosis, or such a task is not the central priority for their practice. Rather, their priority is on starting an appropriate treatment plan and being thorough in considering possible causes of the patient's condition. As we noted in previous chapters however, diagnostic accuracy can be defined in many different ways. We revisit this line of discussion in the Overall Discussion section.

Across our systematic scoping review and three experimental studies, we have investigated the cognitive mechanisms of diagnostic decision making using patient vignettes, simulation-based experiments (via Virtual Reality) and in-situ. Before

synthesising our findings together, we present a reflective chapter based on in-situ observations in two medical settings: Intensive Care and Emergency Department. We then use these observations as a grounding to explore our findings from the previous studies and suggest implications and recommendations for future cognitive psychology research, medical education and clinical practice.