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

Introduction

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 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 investigate whether initial diagnostic breadth is predictive of patient treatment too.

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 a 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?

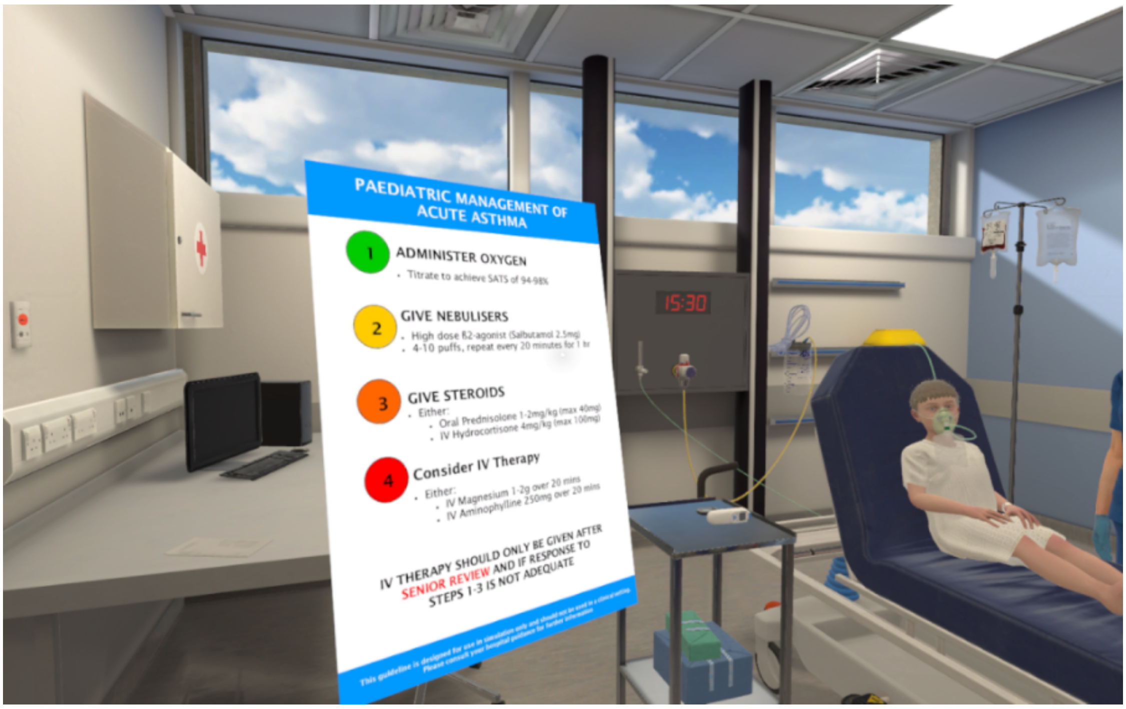
Methods

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 ward room 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.


PAEDIATRIC MANAGEMENT OF ACUTE ASTHMA
1 ADMINISTER OXYGEN
• Titrate to achieve SATS of 94-98%
2 GIVE NEBULISERS
• High dose ß2-agonist (Salbutamol 2.5mg)
• 4-10 puffs, repeat every 20 minutes for 1 hr
3 GIVE STEROIDS
• Either:
• Oral Prednisolone 1-2mg/kg (max 40mg)
• IV Hydrocortisone 4mg/kg (max 100mg)
4 Consider IV Therapy
• Either:
• IV Magnesium 1-2g over 20 mins
• IV Aminophylline 250mg over 20 mins
IV THERAPY SHOULD ONLY BE GIVEN AFTER SENIOR REVIEW AND IF RESPONSE TO STEPS 1-3 IS NOT ADEQUATE
This guideline is designed for use in simulation only and should not be used in a clinical setting. Please consult your hospital guidance for further information
15:30

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. The scenarios presented in each sessions 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 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 student subsequently switching roles for the other scenario. The VR participant was asked in the questionnaire to answer the follow (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.”)

The questionnaire was kept relatively short to minimise disruption to the scenario. This was due to the extra time that could be expended by asking participants to take off and put on the headset again to readjust to VR. Participants were given 20 minutes to complete the scenario, but could end the scenario early if they feel that they have completed the necessary care and tests for the patient. After completing the scenario, participants completed a second questionnaire on a separate sheet (this is considered time point 2). The second questionnaire featured the same three questions as the first questionnaire (see above), as well as the following questions:

- “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?”

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. This 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.

- 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 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. This overall measure then measures what proportion of the participants' likelihoods are dedicated to probable differentials. However, we also penalised participants for providing few differentials, such that high scoring sets of differentials are larger sets of likely or probably differentials.

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 the same action multiple times) classified under History, Examination or Testing across the scenario.
- Information Value: to calculate the value of each information sought across these categories, we calculate the difference in OMS performance score for participants with or without that information. We then sum all sought information values for each 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.

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.

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.

Scenario	n	Performance Score	Information Seeking	Initial Confidence	Confidence Cha
Asthma	37	68.14	19.73	5.70	1
DKA	30	57.30	19.47	7.10	1
Pneumonia	28	63.31	24.46	6.32	1
Seizure	22	63.81	24.09	5.55	2

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), Change in Confidence (difference in reported confidence between the pause point and end of the scenario) and Initial Diagnoses (the number of differentials reported at the pause point).

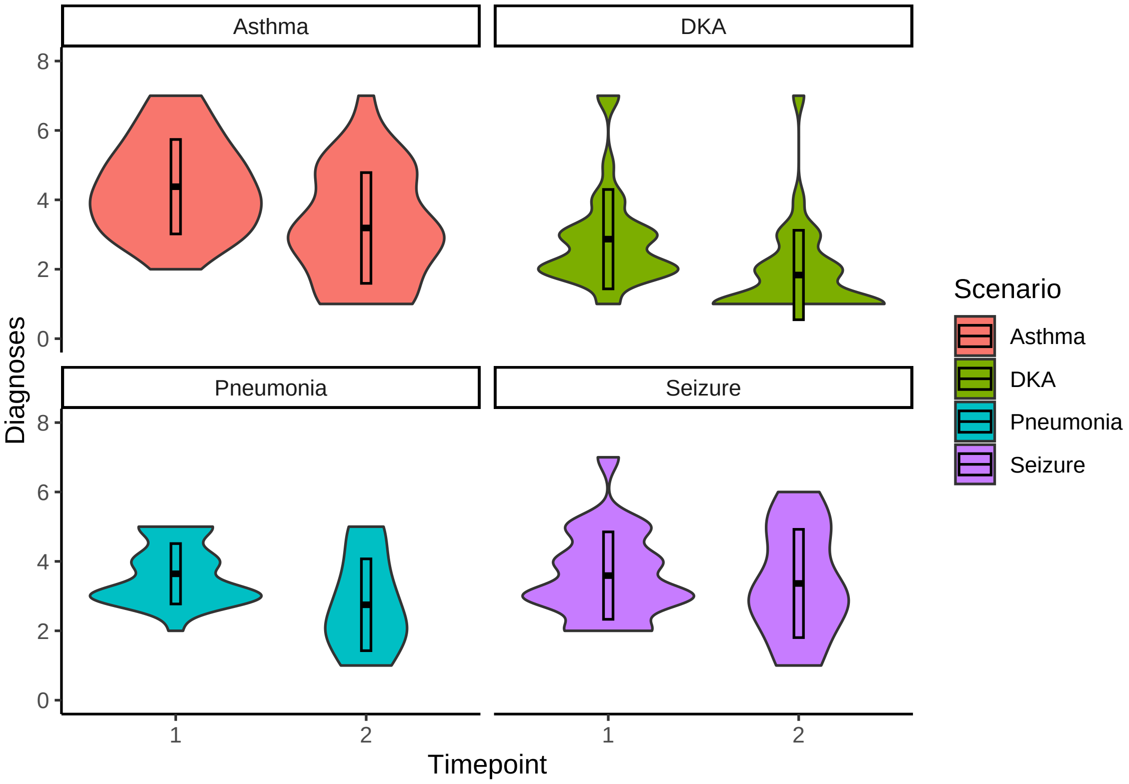


Figure 1: Violin plots showing the number of reported diagnoses at timepoint 1 (the pause point at 5 minutes into the scenario) and timepoint 2 (at the end of the scenario) by condition (Asthma = red, DKA = green, Pnuemonia = blue, Seizure = purple). The dark region of the box plot shows the mean value,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

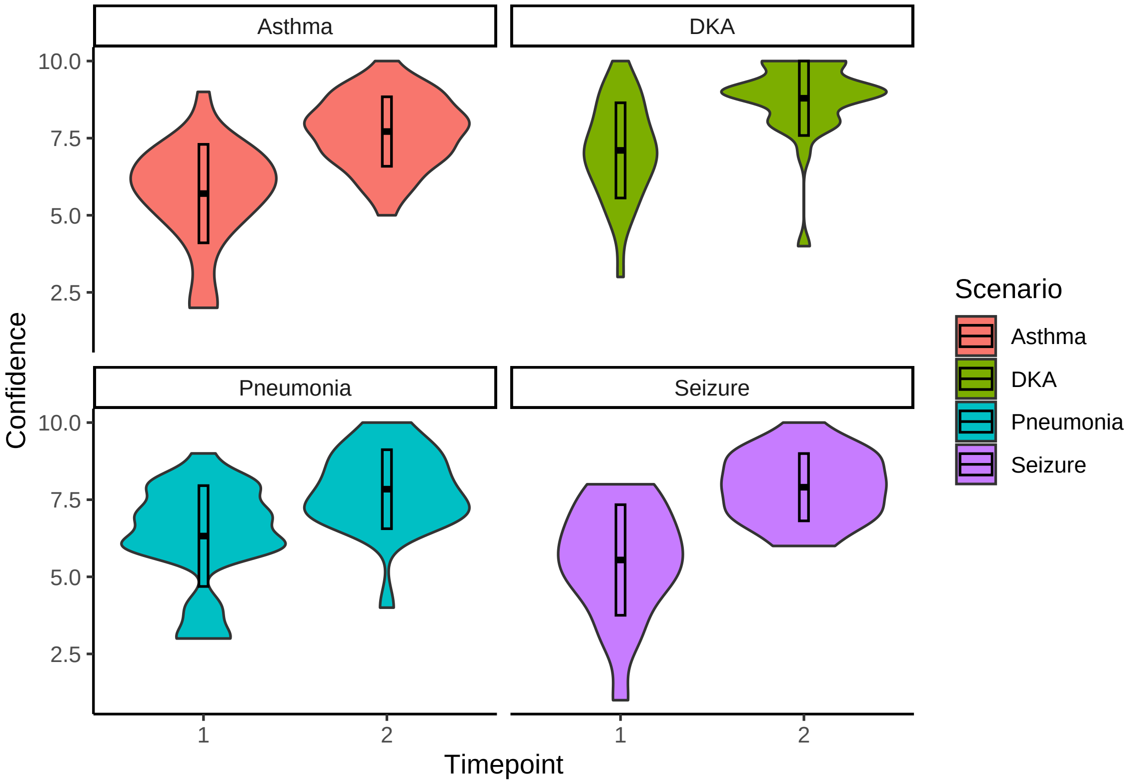


Figure 2: Violin plots showing confidence at timepoint 1 (the pause point at 5 minutes into the scenario) and timepoint 2 (at the end of the scenario) by condition (Asthma = red, DKA = green, Pnuemonia = blue, Seizure = purple). The dark region of the box plot shows the mean value, with the lines of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

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) was 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 linear 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 evidence that the initial diagnostic breadth is predictive of the amount of information seeking (β = 0.41, SE = 0.43 t = 0.96, p = 0.34) or changes in confidence (β = 0.14, SE = 0.12 t = 1.19, p = 0.24).

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 linear 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). We do not find evidence that initial confidence is predicted by prior information seeking related to Patient History (β = 0.07, SE = 0.07 t = 1.02, p = 0.31), Physical Examinations (β = 0.02, SE = 0.11 t = 0.15, p = 0.88) or Testing (β = 0.15, SE = 0.17 t = 0.9, p = 0.37).

We next look at if 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 evidence that initial confidence predicts subsequent history taking in a negative direction (i.e. that higher confidence is associated with lower subsequent history taking) (β = 0.4, SE = 0.19 t = 2.13, p = 0.04). We do not find evidence that initial confidence is associated with subsequent Physical Examinations (β = 0.05, SE = 0.18 t = 0.27, p = 0.79). We do find however that initial confidence was associated with higher amounts of Testing (β = 0.28, SE = 0.13 t = 2.13, p = 0.04).

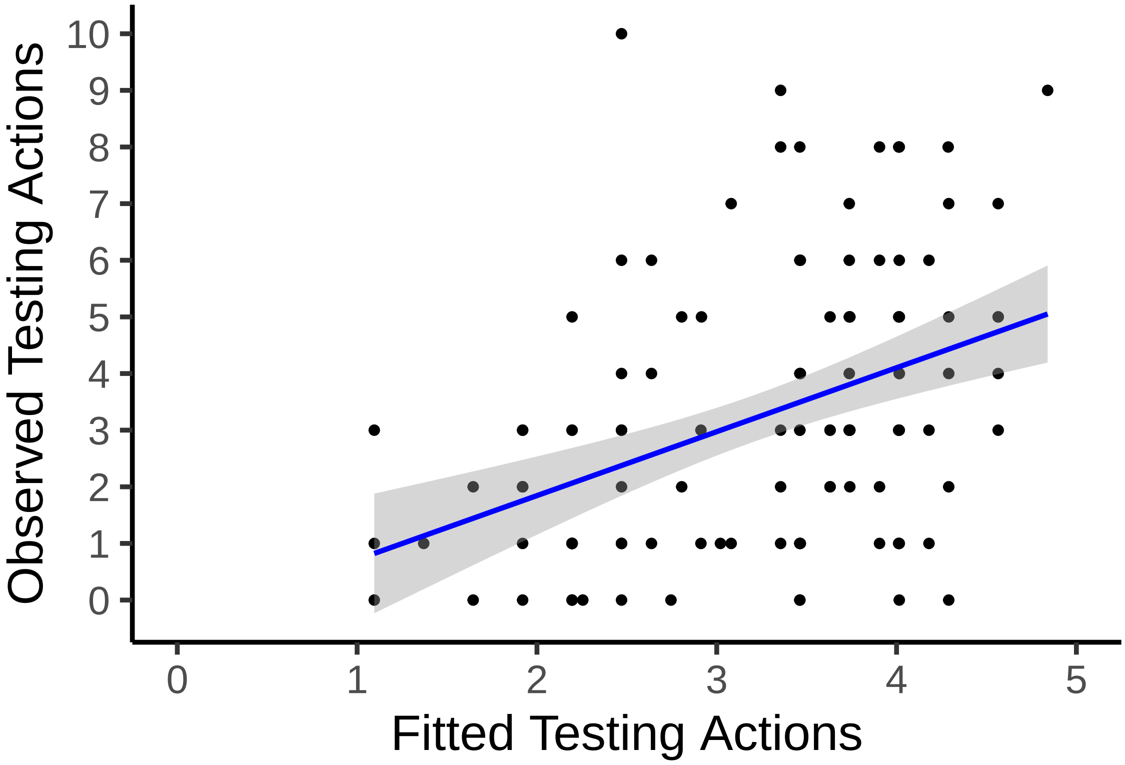


Figure 3: Plot of linear mixed effects model predicting the amount of testing actions (after the pause point) by the initial confidence reported (during the pause point). We show the fitted values for the number of testing actions (x-axis) against the actual observed number of testing actions (y-axis) 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 were predicted by the amount of treatment actions administered during the scenario (β = 0.38, SE = 0.13 t = 2.99, p = 0).

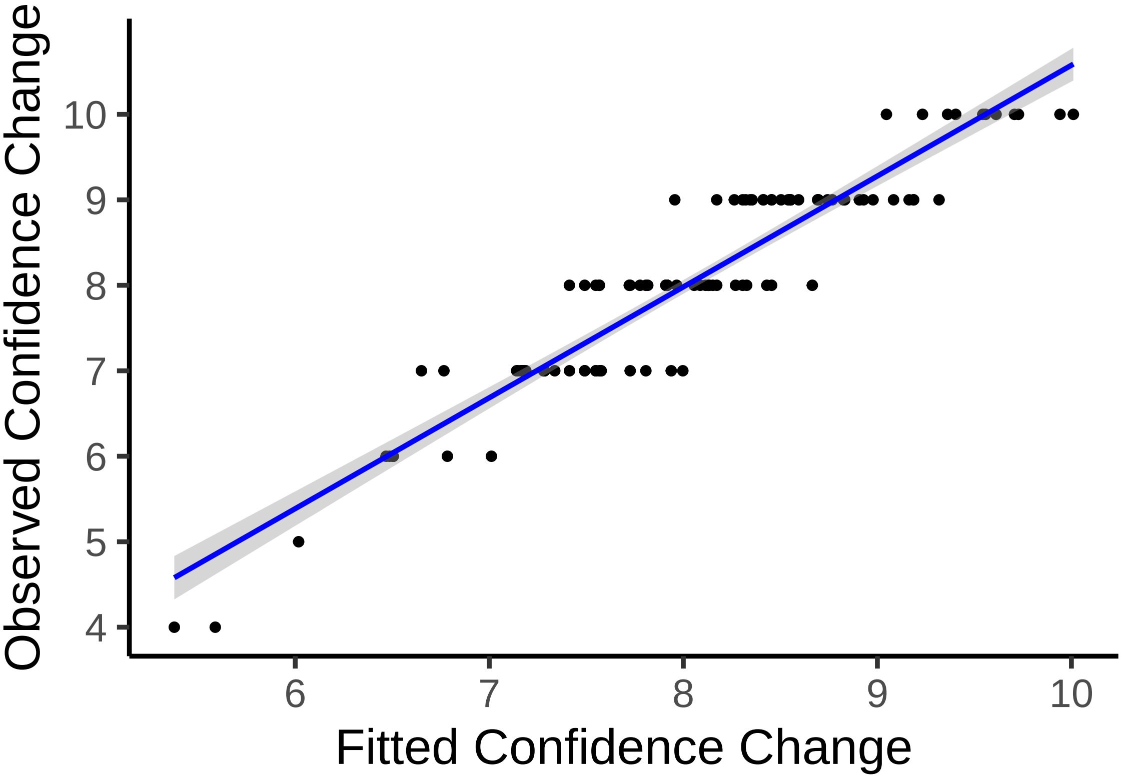
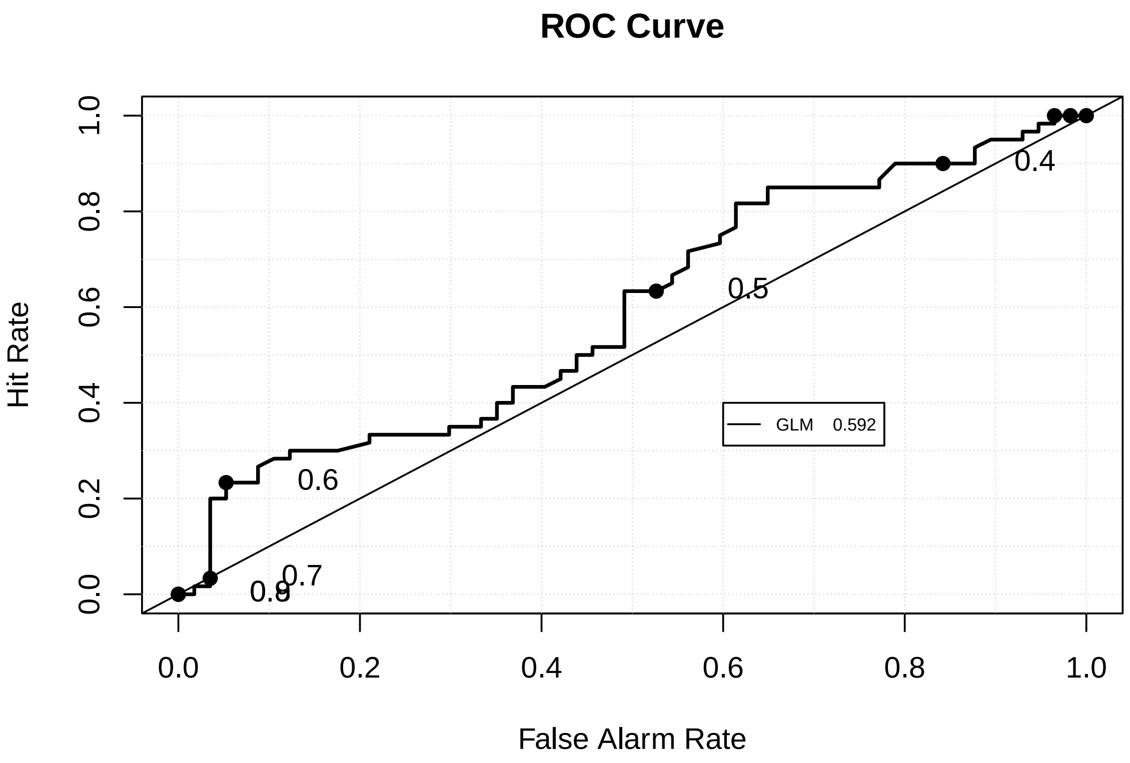


Figure 4: Plot of linear mixed effects model predicting final confidence (at the end of the scenario) by the amount of treatment actions. We show the fitted values for final confidence (x-axis) against the actual observed values for final confidence (y-axis) on each case (each data point representing a single case). We fit a linear model line of best fit with a 95% confidence interval denoted by the shaded region

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.



```
##       Model       Area    p.value binorm.area
## 1 Model  1 0.5918129 0.04367647          NA
```