**Study 1:  Investigating Confidence and Information Seeking in Medical Diagnosis Decisions**

Introduction

Our first study aimed to extend the past work of Friedman et al (2004) and Meyer et al (2013) among others. Methodologically, these studies share a lot of commonalities with each other and with the current study presented here. Our paradigm involve the use of patient vignettes adapted from real patient cases. The goal of the task was to determine a diagnosis, or set of diagnoses, for each presented patient. Information on the patient is split into a series of discrete stages so that the researchers are able to control what information the clinicians have access to at any given point in the experiment. We can call each point of new information an “information stage”. Each information stage includes information on the patient’s medical history, observations from physical examinations and generalised/bedside tests on the patient. At each information stage, participants are asked to provide their diagnosis for the patient, as well as reporting their decision confidence. Kämmer et al (2021) built upon the aforementioned studies by allowing the clinicians to freely gather information in each stage. This added the amount of information sought as a key dependent variable to investigate alongside confidence and ability/accuracy. It was hypothesised that differences in ability would be reflected in information seeking patterns. For example, a clinician with lower ability might seek more information less discriminately, as they would have higher uncertainty. This has been investigated in the past by relating to clinical experience to risk aversion and further information seeking behaviour (Lawton et al, 2019).

There are aspects of this past work that we sought to extend. Participants are often asked for a single diagnosis in previous studies. The diagnosis provided is then compared to the true underlying condition that the patient had in order to establish a measure of objective accuracy. This then allows researchers to investigate miscalibrations of confidence, as these are usually derived via the difference between reported confidence (one’s subjective probability of being correct) and accuracy (one’s objective proportion of correct answers). While this is useful from an empirical perspective, this has limitations in terms of ecological validity. It is a fairly artificial constraint to limit clinicians to a single diagnosis as their final answer. In addition, a reported diagnosis does not consider the conditions that the clinician has ‘in the back of their mind’. If a clinician is asked for a single diagnosis, it is not possible to know whether other diagnoses have been eliminated from their train of thought. Being able to know what diagnoses are considered possibilities allows for a more nuanced investigation of premature closure and hypothesis elimination. We therefore allow clinicians to report multiple diagnoses, whilst also allowing for a differentiation between the more likely diagnoses and the more serious but rarer diagnoses (the ‘back of the mind’ diagnosis). We also wanted to further study differences in information seeking patterns as investigated by Kammer et al (2019).

Methods

**Participants**

The study was conducted online, with participants able to run the experiment in their browser. The experiment was coded using JSPsych, which is a Javascript plugin used specifically for psychology experiments. We recruited fifth or sixth (foundation) year medical students within the UK. The UK Medical Schools Council distributed the study to UK medical students using a 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 age ranged between 22-34 (M = 24.2). 85 medical students completed the study, including 32 males, 52 females and 1 participant who self-reported as non-binary. Participants were recruited between July 11th 2022 and April 6th 2023.

**Materials**

This study involved the usage of patient vignettes. As before, these are simulated patient cases that have been adapted from actual past cases. We adapted scenarios from a bank of patient case from Friedman (2004). These vignettes were developed by a team of researchers based in the US, meaning that certain medical terms (eg medication names, tests etc) had to be ‘translated’ into the vernacular used by doctors based in the UK. This was done via consultation with three different researchers working with the OxSTaR Centre who were also practising medical staff and students within the NHS. The aim was to consult medical professionals who had completed their clinical education and were at differing experience levels and medical subdisciplines. Their medical roles at the time of the development of this study were as follows: Speciality trainee (ST6) in Anaesthetics, Foundation (F1) Doctor and Gastroenterology Consultant.

As well as translating the language of the cases from US to UK-orientated (e.g. drug names, conditions etc), there were also changes made based on time, given that the original vignettes were developed for a paper published in 2004. Cases made occasional references to specific years in the patient’s history where they had previous medical conditions or hospitalisations. These years were updated to make sense for a contemporary patient (i.e by adding 18 years such that any referenced years were relative to 2022 rather than 2004). Whilst a sizable bank of vignettes were provided by Friedman, certain conditions were considered too rare (either for the current time or for the UK) to be used. Our goal was to test the clinicians’ ability to deal with diagnostic uncertainty, rather than testing their declarative knowledge of obscure medical conditions. In consultation with the aforementioned doctors, cases were therefore chosen based on their underlying medical conditions being those that medical students would be expected to know. We also aimed to choose cases that displayed a variety of both difficulty and of condition type (i.e. involving different anatomical systems).

Our study involved 6 patient cases, each with a true underlying condition. These conditions were: 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.

**Procedure**

The procedure of a single case (or ‘trial’) is as follows. The participant is asked to imagine that they are working in a busy district hospital and they encounter patients in a similar way to how they would in their real medical practice. At the start of each case, the participant is shown a presenting complaint for a patient, which includes the patient’s gender, age and their presenting complaint. An example of this is: “patient is a 68 year old male presenting with fever and arthralgia”. This presenting complaint remains on screen throughout the entire case. Each case is split into three information stages: Patient History, Physical Examination and Testing. This order of stages is fixed for all participants. At each stage, the participant sees pieces of information or tests that they can request. Participants can view information from a previous stage but cannot see information for a future stage (e.g. if a participant is at the Physical Examination stage, they will be able to see information pertaining to Patient History and Physical Examination, but not information pertaining to Testing). The set of information requests for each stage is the same for all cases. The Patient History stage includes information on “Allergies”, “History of the Presenting Complaint”, “Past Medical History” and “Family History”. The Physical Examination stage includes ‘actions’ that a doctor may take when examining a patient,  such as “auscultate the lungs”, “abdomen examination”, “take pulse” and “measure temperature”. Finally, the Testing stage involves information on any bedside tests or tests they may request from another department. This includes “Chest X-Ray”, “Venous Blood Gas”, “Urine Dipstick” and “Clotting Test”. There was a total of 29 possible tests that can be requested across the three information stages. When a participant clicks on any of these tests, the screen shows a loading icon for 3 seconds before showing the information for that test on screen. During this loading time, other tests cannot be requested. When any subsequent test is requested, the previous test result is removed from the screen such that participants can only view one piece of information at a time. The time delay for receiving information was added after piloting the study, where the lack of time delay meant that participants were likely to request most information without being selective. It was also emphasised during the task instructions that participants should only request information that they believe will help them with diagnosing the patient for that specific case. The information shown for each test is pre-defined as per the medical vignettes and is the same for all participants. Participants are free to request the same piece of information multiple times in order to remind themselves, including information from a previous information stage.

At any point, the participant could choose to stop gathering information for that stage. They do so by clicking the “Enter Differentials” button. At this point, they are taken to a new screen where they can report a list of all differential diagnoses that they are considering for that patient at that stage. Participants could report as many diagnoses in their list as they wanted to. For each differential, participants report a “level of concern” for that differential, which we describe as how concerned the participants would be for that patient if this differential really was the patient’s underlying condition. This is reported on a 4 point scale, with labels of “Low”, “Medium”, “High” and “Emergency”. Participants also reported a likelihood rating for each differential, ranging from 1 (very unlikely) to 10 (certain). When reporting differentials at the first information stage (Patient History), the list of differentials is blank and participants must add at least one differential to proceed. In subsequent stages, the list from the previous stages is available for participants to update concern/likelihood ratings, add differentials or remove differentials from the list. Participants are asked to carefully consider which differentials they have in mind in light of the new set of information they have received. Even at the last information stage, participants can report multiple differentials if they do not prune their list down to a single diagnosis. Participants are not penalised for reporting a wide set of differentials at any stage.

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

**Data Analysis**

Responses were coded for correctness manually with help from a medical consultant, who looked at all the information available for each case and determined which diagnoses could be accepted answers. This depended on the nature of the case, as a case may sometimes have a vague set of information such that determining the exact correct diagnosis was considered too challenging. For example, for the TTP case, making a diagnosis of TTP (even with all information requested by the participant) was seen as too difficult given that the information provided was not discriminant enough. Hence, other conditions like ITP and Meningitis were also accepted as correct answers. All lists of differentials were ‘marked’ for correctness manually using the following criteria (the correct condition is followed by the list of accepted diagnoses to be considered correct):

**Temporal Arteritis**: any inflammatory arteritis is accepted

**Ulcerative Colitis**: infectious colitis, ischemic colitis and diverticulitis are also accepted answers.

**Miliary TB**: any TB or lymphoma type is accepted

**Aortic Dissection**: pulmonary embolism and coarctation of the aorta are also accepted answers.

**Guillain-Barre Syndrome**: Cauda Equina Syndrome is also accepted

**TTP**: ITP or Meningitis are also accepted.

For all cases, acronyms or spelling mistakes on the accepted answers were marked as correct.

There are a number of key dependent variables that we are able to derive from our data:

* **Confidence**: the reported confidence 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). 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**: we record the number of items in the list of differentials at each stage. **Initial Differentials** refer to the number of differentials after the first stage of information seeking (Patient History), whilst **Final Differentials** refer to the number of differentials after the third and last stage of information seeking (Testing).
* **Perceived Difficulty**: the subjective rating by participants at the end of each case for how difficult they found it to determine a diagnosis for that patient case. This is reported on a scale from 1 (trivial) to 10 (impossible).
* **Accuracy**: 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. Given that differentials are provided via free text, cases are manually coded as correct or incorrect using the aforementioned criteria. Our main accuracy measure is computed by the taking the likelihood value assigned to the correct differential if it is included in the list of differentials. This means that accuracy ranges from 1-10 when a correct differential is included and has a value of 0 when a correct differentials is not included.
* **Proportion of Information Seeking**: we take the number of unique tests requested at a given information stage (i.e. not including any tests from a previous stage or including tests that had been requested before during that stage) and divide by the number of possible tests available.
* **Information Seeking Variance**: We compute a vector of length 29, which is made up of 0s and 1s where for each of the piece of information available for a case, a value of 1 is assigned if that information is requested and 0 is assigned if that information is not requested during the case. The vectors for all cases for a given participant are combined to produce a 29 x 6 matrix. We calculate the Euclidean distance between each row of the matrix (trial) using R’s dist function (in the proxy package). The computation of all pairwise distances produces a 6 x 6 matrix where each trial is given a Euclidean distance value relative to every other trial. A lower distance value between two trials indicates that the information sought on those trials are more similar to one another. In order to look at the similarity of information seeking across all six trials, we compute the variance (the standard deviation squared) of the participant’s Euclidean distances. A lower variance value indicates that participants seek similar information across the cases whilst a higher value indicates that information seeking is varied more by case.
* **Information Seeking Value**: We use the information seeking of the experienced participants as a reference to compare the student participants against, meaning that we use the information seeking of the experienced doctors to derive the value of each information and then compute the average value of sought information for the medical students using these values. To do this, we take each of the 29 pieces of information in turn by case and split trials into two groups: trials of that case where that information was sought and trials of that case where that information was not sought. For each group, we compute the proportion of trials where the experienced participants included a correct differential, and then take the difference between these two values. A positive value would indicate that participants 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’. For each of the student participants’ trials, we calculate the mean of information values for all information that the participant did seek. This gives an overall measure of how useful the information was that participants sought on a case-by-case basis.

Results

***Overall Performance***

We first look at the number of differentials that participants report at each stage. Participants overall increased the number of the differentials they reported as they received more information (F(1, 107) = 94.02, η2G = .08, p < .001). Participants reported fewer differentials during the Patient History stage (M = 3.20, SD = 1.11) than during the Physical Examination (M = 3.88, SD = 1.33) and Testing stages (M = 4.12, SD = 1.43). We especially note that out of 85 participants, 74 did not decrease their number of differentials between Patient History and Testing on any case. Confidence also increased as participants received more information (F(1, 123) = 75.45, η2G = .15, p < .001). Participants reported lower confidence during the Patient History stage (M = 0.30, SD = 0.15) than during the Physical Examination (M = 0.41, SD = 0.17) and Testing stages (M = 0.47, SD = 0.19). We note here that confidence was on average below 50% even at the end of each case, indicating that participants were not highly confident to treat the presented patients on the whole. The Proportion of Information Seeking decreased with each information stage (F(2, 151) = 122.0, η2G = .30, p < .001). Participants sought more of the available information during the Patient History stage (M = 0.85, SD = 0.20) than during both during the Physical Examination (M = 0.59, SD = 0.24) and Testing stages (M = 0.50, SD = 0.22).

We now report the proportion of trials where participants include a correct differential within their set of differentials (henceforth referred to as Differential Accuracy). Participants increased their Differential Accuracy overall with more information (F(2, 128) = 59.52, η2G = .08, p < .001). Participants had lower Differential Accuracy during the Patient History stage (M = 0.54, SD = 0.23) than during the Physical Examination (M = 0.66, SD = 0.22) and Testing stages (M = 0.69, SD = 0.21).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Case** | **Differential Accuracy** | **Accuracy** | **Perceived Difficulty** | **Final Confidence** |
| UC | 0.92 | 7.32 | 5.30 | 61.08 |
| GBS | 0.69 | 5.44 | 6.88 | 37.03 |
| TA | 0.66 | 6.68 | 6.16 | 48.96 |
| TTP | 0.55 | 5.50 | 6.80 | 40.86 |
| AD | 0.53 | 4.67 | 5.92 | 48.87 |
| MTB | 0.42 | 5.67 | 6.69 | 44.91 |

A chart of different levels of performance

Description automatically generated

Previous work (such as from Meyer et al, 2011) have noted a gap between subjective confidence and objective accuracy. In particular, there has been demonstrated to be a general tendency for less experienced medical trainees to be underconfident and for more experienced medical professionals to be overconfident (Yang and Thompson, 2010). We sought to investigate whether we observe a similar pattern by comparing participants’ final confidence on this task with their objective performance. We operationalise performance in two ways: Differential Accuracy being a similar performance measure to that used in previous studies (i.e. a trial is marked as correct if a correct differential is provided) and Accuracy taking into account the likelihood values assigned to correct differentials on this task. We find that when using Differential Accuracy, we do observe that participants (i.e. medical students) are underconfident by reporting lower confidence than their performance across all information stages. However, when using the Accuracy measure, we find that confidence is fairly well calibrated to objective performance except for a slight deviation between the two during the Testing stage (t(84) = 2.40, MDiff = 0.056, p = .02). This could imply that past findings on miscalibration of confidence are dependent on how objective performance is operationalised.

***Information Seeking***

Using information seeking data, we sought to identify if differences in information seeking patterns were predictive of differences in diagnostic accuracy on this task. To investigate the differences in information seeking, we first trained a binary classification algorithm using a generalised logistic regression model to predict accuracy. We aimed to identify if participants of different diagnostic ability on this task exhibited differences in information seeking. To do this, we first split all trials into high and low ability participant trials with a median split of participants by their average accuracy across the six cases. We train the classifier by treating the 29 binary variables for each information as predictors (with a 1 signifying that the information was sought for that case and 0 when the information was not sought) to predict the binary outcome of whether the participant is a low or high accuracy participant. We used Leave One Out Cross Validation, such that each trial is predicted by training the algorithm on all other trials. By plotting an ROC curve of our classifier, we find an area under the curve (AUC) value of 0.72 (with p < .001 when comparing the ROC curve to AUC = 0.5). We also use a Classification Tree Algorithm to classify trials by participant accuracy and find an ROC AUC of 0.62. This indicates overall that differences in information seeking are indeed predictive of a difference in participant ability. Next, we seek to identify and better characterise the specific differences in information seeking that contribute to this relationship with diagnostic ability.

When conducting a Pearson’s Correlation test, we found evidence for a positive correlation between the change in confidence (the difference in confidence in the first and last stages) and the proportion of information sought (r(83) = 0.24, p = .03), such that seeking more information was associated with higher gains in confidence. Interestingly, we also find an association between the number of differentials generated from the Patient History and the amount of information sought during a given case (r(83) = 0.30, p = .005). As previously discussed, participants rarely seem to remove differentials from consideration, leading to a close relationship between the number of the differentials at the Patient History and Testing stages (r(83) = 0.81, p < .001). Therefore, one can surmise here that higher information seeking is associated with the consideration of more diagnostic differentials, though the direction of causality is unclear as this juncture.

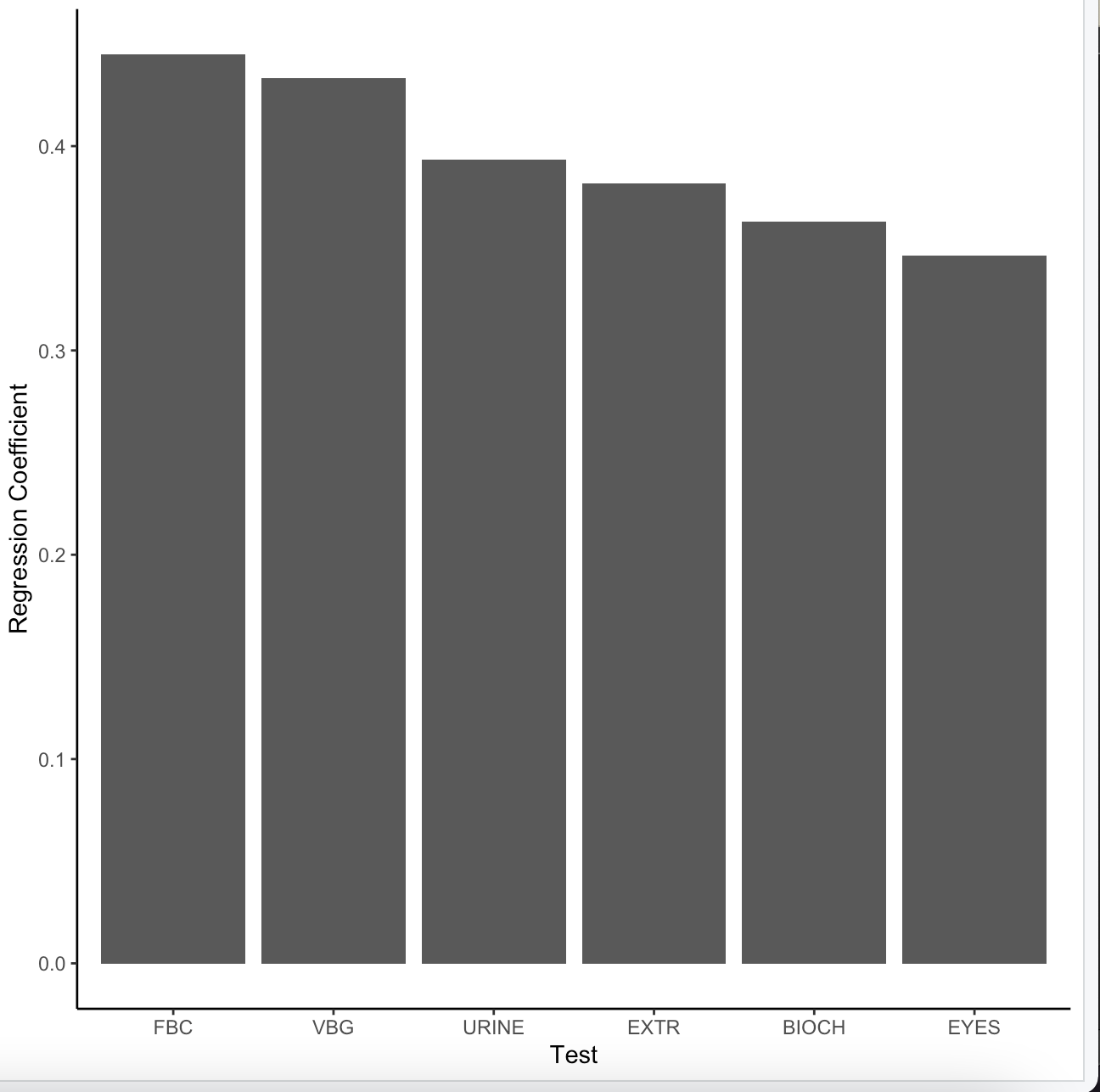
We then look at how ability on the task relates to information seeking behaviour. To do this, we calculated the average accuracy for each participant (across cases) and then sorted participants into four groups by quantiled accuracy. We then look at mean information seeking variance for each group of participants. We find that participants with higher overall accuracy have a lower variance in information seeking. In other words, students with a higher diagnostic ability are found to have varied the information they sought across cases less, seeking more similar information for each case when compared to students of a lower diagnostic ability. We can also test the same hypothesis by treating participant accuracy as a continuous measure, and we find evidence for a negative correlation between accuracy and information seeking variance (r(83) = -0.23, p = .04). We apply a similar analysis to look at how information value varies as a function of participant ability. We find evidence for a positive relationship between accuracy and information value (r(83) = 0.25, p = .02), but not between confidence and information value (r(83) = 0.15, p = .18). We also find that the proportion of available sought is not shown to correlate with accuracy on at the final stage (r(83) = 0.17, p = .11) but does correlate with the participants’ change in confidence, which is the difference in confidence between the first and final stages (r(83) = 0.24, p = .03). While seeking more information may imbue students with a greater level of confidence, it does not necessarily translate into more accurate diagnoses. This is important to note as it demonstrates that being selective in information seeking is a better marker of performance and giving a lower ability participant all available information does not necessarily translate into accurate diagnoses. In addition, students with a higher diagnostic ability seek better information but also approach each case in a more similar manner. This could indicate a base of information kept constant across cases alongside a more selective set of useful information related to that patient. Meanwhile, participants with a lower diagnostic ability are not selective with their information seeking and hence do not seem to have a set framework or plan for what information to seek.

This has interesting implications for medical practice, as the ordering of unneeded tests or patient examinations may not contribute to better decisions. Given the constraints within most hospitals and healthcare to obtain certain tests, being selective with information seeking is already a frequent necessity and results from this study seem to show evidence that it is also a good marker of diagnostic performance.

With these differences in information seeking emerging from the data, we next look at specific tests/information that participants are requesting and how this varies by participant accuracy. To do this, we plot the proportion of trials where each piece of information was requested across all trials with each quartile of participant accuracy. We exclude information from the Patient History stage of the task, as all participants tended to seek all information during this stage. Within each accuracy quartile, tests are then ordered in descending order of the proportion of trials in which each information was sought. The ranking of tests in this manner was quite similar across groups (all Kendall Tau Distances < 0.2 across all pairwise comparisons of accuracy groups). Visually however, one can notice that while there is mostly agreement on which tests are more or less worth seeking, higher accuracy groups tend to have more of a numerical separation between high and low ordered tests, indicating they are better differentiating between more valuable and less valuable tests.

A graph of different colored bars

Description automatically generated



Discussion

Story structure:

* Accuracy and confidence are well-calibrated here, so can look at drivers of accuracy to in turn inform confidence
* Diagnostic uncertainty in the task even at final stage
* Doing well at the task is characterised by selective and standardised information seeking
* Initial differentials might mean greater information seeking to reduce differentials
* Greater information seeking increases confidence, not accuracy
* If using mean of value, then greater information seeking results in less average in value
* Preview of think aloud study: why broaden differentials? Is standardisation/reduced variance related to reasoning strategy? What are the reasoning strategies that may affect initial differential generation?

There are a few limitations with our study. We did not use more naturalistic stimuli, such as images of scans/test results or audio cues (such as the sound of lung auscultation) and instead used solely textual results for all tests. While this may make the experiment more ecologically valid, it takes away the interpretation of complex stimuli which could affect information seeking. For example, if two participants requested a chest X-ray, they may interpret the X-ray image in different ways. While this difference in perception may be interesting, it adds a potential confound for the purposes of this study. That is why, for this study, if a participant requested a chest X-ray, they instead see a result that reads something like “no abnormalities found”, such that the interpretation of the image has already been done for the participant. These limitations will be addressed in Study 3, where we use virtual reality (VR) to investigate similar research questions around diagnosis but in a much more naturalistic paradigm where there is a visual representation of a patient that can be interacted with.

Our experiment also assumed that all tests were equal in terms of how long they take for results to be shown. If the tests were analogous to real medical practice, certain tests would take longer to produce results after being requested. Some tests (e.g. a chest X-ray) are not performed by the doctor themselves at the patient’s bedside and require staff and technology from another department. We should also note that our experiment was run via an internet browser, meaning that study participants were taken out of the setting within which they would usually make these decisions. This means that participants may act differently than they might do in their regular medical practice.  In addition, we attempted to make the patient cases as realistic as possible whilst having a moderate degree of difficulty. The original researchers removed certain findings from the cases that may give away the patient’s condition in a fairly obvious manner. In that sense, the patient cases may not replicate the set of information that might be available to clinicians in a similar scenario during medical practice. However, using a paradigm similar to past research does extend and build upon empirical experiments on diagnosis. As previously mentioned, information was chosen in order to be general to all cases and was not very discriminant.

Within this discussion, it is worth mentioning a few general observations in the data and how they might inform the design of future studies of this nature. Firstly, participants did not tend to use the ability to remove differentials from their list. In our study, participants could remove a differential in the interface by clicking the X button on a differential. One explanation is that the button is not very prominently placed on the screen. However, this feature was explicitly explained in the tutorial to the experiment. This tendency is reflected in the overall pattern of the average number of differentials increasing over the three stages of a case. What this may indicate then is an attachment to hypotheses and unwillingness to remove them from consideration. There is a general adage in healthcare that medical students come across which says that “history is 80% of diagnosis”. The fact that diagnostic differentials do not change that much between stages is supportive of this. Indeed, accuracy does not improve by a large amount between stages (from 52.2% after Patient History to 65.9% after Testing). It is indeed striking that in over half of all cases, students are able to include the correct condition in their differentials by the patient’s history alone. It is therefore worth considering whether there is a specific facet of diagnostic decisions whereby clinicians are taught not to disregard diagnostic possibilities easily. This also corresponds with participants tending to request most, if not all, information during the Patient History stage (86.1% across all participants) and then becoming more selective in information seeking during later stages. Hence, this indicates a general behaviour to gain the majority of diagnostic differentials from Patient History and to not easily disregard diagnoses.

Another aspect of note is the manner in which participants reported their differentials. Given that differentials were provided via free text, there is a lot of freedom in the diagnostic differentials that participants can report. What this can mean however is there are differences in the specificity of differentials provided. For example, one participant may report “lymphoma” as a differential whilst another may report “Hodgkin’s Lymphoma”, “Non-Hodgkin’s Lymphoma” and “Chronic Lymphocytic Leukaemia” within the list (all of which are different types of lymphoma). Both participants essentially capture the same ‘differential’ but do so in different manners. When looking at the number of differentials however, the former produces one differential whilst the latter produces three. This example illustrates that participants differ in how specific they are when reporting their differentials and how this affects our ability to analyse the number of differentials that participants report.

We should also note the manner in which accuracy was coded manually for each case. This depended on the nature of the case, as a case may sometimes have a vague set of information such that determining the exact correct diagnosis was considered too challenging. For example, for the TTP case, making a diagnosis of TTP (even with all information requested by the participant) was seen as too difficult given that the information provided was not discriminant enough. This ties into one of the main challenges of designing these vignettes and this study: the set of information available for participants to request were chosen such that they were reasonable to be requested in any of the cases. The participants may have wanted to request more specialised, discriminant tests (e.g. lumbar puncture, biopsy), but including these could clue participants into the nature of the patient’s condition. In addition, these types of highly specialised tests that target a specific type of diagnosis tend to take much longer to come back to doctors with results after they request them in a real healthcare setting. Hence, having results available at the touch of a button for these may seem unrealistic unless we alter the design to have patient cases unfold over a longer time period.

Previous research that allows participants to report multiple diagnostic differentials would consider a trial correct if the correct condition is within the list of differentials. However, accuracy would be conflated with the number of differentials provided. If a participant simply continues to add more differentials, they are more likely to be correct. Therefore, we calculated Accuracy using a more fine-grained measure referred to as the Correct Diagnosis Likelihood. To calculate this, we first identify the correct differential if provided in the list (as per the earlier marking scheme) and find the likelihood rating assigned to that differential. The highest possible value here would be 10 if the participant included the correct condition in their differentials and assigned it the maximum likelihood rating. If a correct differential is not provided, a value of 0 is assigned for that trial. The mean of these values across the six cases produces an overall measure of **Accuracy** for each participant. If multiple differentials that are considered correct were provided, then the likelihood value of closest differential to the true condition was used.

**Study 2: Thinking Aloud During Diagnostic Decisions**

The main results from Study 1 were better diagnosis on our task was characterised by more standardised information seeking and that participants were increasing the number of differentials they were considering with more information. Both of these results were surprising and hence it had to be considered that the results were due to the nature of our specific task. When creating a task that emulates diagnosis, we in a sense conceptualise what diagnosis looks like in a fairly static manner, when really diagnosis is a more fluid and nebulous structure in medicine. For example, a doctor’s approach to a patient is not always going to fit within the idealised structure of taking a patient history, conducting physical examinations and then requesting tests in this order. There are environmental or even patient factors that necessitate information being processed out of order, as well as different diagnostic approaches by doctors. This taken together brings up the question of whether the observed results on reduced variance in information seeking being associated with accuracy was a result of our strict task structure. In addition, it was striking that participants in our study rarely removed differentials from their list of suspected conditions despite having the ability to do so. This lack of removing differentials was what drove our observed effect of the number of differentials increasing with more information. We wanted to hence see if signs of these results would be evident in the thought process of medical students. Are doctors seeking information to confirm their existing set of differentials, to rule out differentials or to expand their set of considered possibilities? And are these different approaches interleaving or are they more dependent on individual diagnostic decision making styles?

In order to provide more context to the results from study 1, we conducted a follow-up study that utilised a very similar experimental procedure, but instead prompted participants to think out loud as their performing the task. Their utterances were then transcribed and coded to conduct both quantitative and qualitative analysis.

**Participants**

Due to the richness of qualitative data, we recruited a smaller sample of participants. In total, 16 participants were recruited for this study. Participants had to be 5th or 6th year medical students at Oxford university in order to take part. Participants were recruited using posters in John Radcliffe Hospital in Oxford and via a mailing list for students managed by the Medical Sciences department at the University of Oxford. The study was conducted onsite at John Radcliffe hospital. Participants were recruited between July 5th 2023 and December 1st 2023.

**Materials**

The same set of cases and interface from Study 1 was also used for this study. The study was conducted onsite using a laptop, with actions on screen recorded on video and the audio of participants’ thinking aloud recorded via a microphone.

**Procedure**

The general procedure was very similar to that of study 1, whereby participants were shown six patient scenarios and were tasked with diagnosing the patient. However, we removed the screen where participants record their list of differentials. Instead, the experiment was run in-person so that participants could think aloud as they were doing the task. Participants were given the following instructions at the start of the study:

*“Whilst you are doing the task, you will be asked to think aloud. This means that you verbalise what you are thinking about, especially how you interpret the information you receive and what conditions or diagnoses you are considering or are concerned about for each patient case. If you have nothing to say or nothing on your mind, there’s no need to say anything but do say whatever is on your mind once it pops up. If you are unsure about anything you see or do not know about what something means, you will not receive any help but verbalise when you are unsure about anything during the task. Please make sure that you speak clearly ‘to the room’.”*

The researcher in the room was to remain mostly silent, aside from asking the participant “can you tell me what you are thinking?” if there is a period of long silence and asking the participant “can you tell me more?” if the participant says something vague that may warrant further detail. This was so that any utterances by participants were not swayed or prompted to say certain things by the researcher. The audio of the participants’ verbalisations was recorded and then transcribed. An initial transcript was generated using Microsoft Office’s transcription feature, but the transcript was checked and modified for accuracy by listening through the audio recordings again. The screen of the experimental interface was also recorded, such that the audio could be linked to specific actions within the task. The focus of this study is on verbal utterances rather than any non-verbal or inferential aspects of the participants’ qualitative data.

At the end of the experiment, the researcher administered a semi-structured interview to better understand what the participants feel their diagnostic reasoning approach tends to be. The main questions are indicated below, each with a corresponding follow-up question in case they are not answered by responses to the main questions:

1. What's your general approach to making diagnoses?

Follow-Up: Do you have those cognitive aids or frameworks you use?

1. Do you tend to keep a broad set of differentials in mind?

Follow-Up: Are there particular situations where having a narrower set would be more useful?

1. How do you decide what information or tests to get on a patient?

Follow-Up: Would you say you tend to seek information to confirm or to rule out differentials that you have in mind?

1. How similar was your diagnostic reasoning on this task versus how you would approach diagnosis in real life?

Follow-Up: Was there anything that prevented you from approaching the task as you would in real life?

**Data Analysis**

Once the audio recordings had been transcribed, the transcriptions for specific categories of utterances. Hence, this study adopted a theory-driven semantic thematic analysis (as per definitions detailed by Braun and Clarke, 2006), given that we use pre-existing theory to produce codes a priori to be applied to our data. This kind of thematic analysis is also suitable given that our qualitative data is from a structured experiment, rather than a dataset with a looser structure (e.g. interview recordings).

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

***Differential Evaluation***: any time that the participant (each of the following is considered a separate subcode):

* Mentions a new condition that they are considering
* Rules out or eliminates a condition from consideration
* Mention of increased likelihood of a previously mentioned condition, or that information seems to correspond with a condition
* Mention of decreased likelihood of a previously mentioned condition, or that information seems to contradict with a condition

***Information Seeking Strategies:*** any time the participant expresses why they may or may not request a particular piece of information in relation to ruling out or confirming a condition.

***Uncertainty Expression:*** any time the participant expresses being unsure about their diagnosis or understanding of the patient.

We also define a group of codes that indicate reasoning strategies, which pertains to how the participants generate and considered differentials. These codes are based around the work of Coderre et al (2003), which had also used a think-aloud protocol for a diagnostic reasoning task. Their paper defines three different diagnostic reasoning strategies: hypothetico-deductive reasoning, scheme-inductive reasoning and pattern recognition. Adapted from their definitions, we define each as follows:

**Hypothetico-Deductive Reasoning** - prior to selecting the most likely diagnosis, the participant analysed any alternative differentials one by one through something akin to a process of elimination.

**Scheme Inductive Reasoning** - participant structures their diagnosis by anatomical systems or categories of conditions (e.g infective vs cardiovascular causes) to determine root causes of patient symptoms rather than focusing on specific conditions.

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

**Hybrid Strategy** – participants uses a reasoning strategy that is an amalgamation of at least two of the preceding strategies.

We first code specific utterances within each case that suggested one of these strategies (aside from the Hybrid Strategy), and then determined which strategy was most prevalent or influential for cases as a whole such that each case was categorised under one of these strategies. In addition to coding each case under one of these strategies, we also code participants on an overall level based on their subjective perception of how they make diagnostic decisions. This is based on responses provided during the debrief interview (as described in the Procedure section). Hence, reasoning strategy codes are at the case level and also at the participant level. Finally, we apply thematic analysis to the responses from the debrief interview to find other themes that seem to recur in the data. As questions were fairly closely related to diagnostic decision making, we could ensure that participants gave an account of their decisional process.

Although we do not record differentials in the same way as in Study 1 (in a list with corresponding likelihood and severity ratings), we do obtain the other variables from Study 1. Namely, we record confidence at each stage of information seeking and data around the information sought by participants.

**Results**