Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors

Edward Krupat,[1] Jolie Wormwood,[2] Richard M Schwartzstein[1] & Jeremy B Richards[1]

CONTEXT Early studies suggested that experienced clinicians simply generate more accurate diagnoses than those less experienced. However, more recent studies indicate that experienced clinicians may be subject to biases in formulating and confirming hypotheses that lead to inaccuracy.

OBJECTIVES The goal of this study was to identify factors associated with the ability to process information in ways that overcome premature closure and result in accurate diagnosis using a set of vignettes in which inconsistent information was introduced midway.

METHODS Seventy-five participants (25 Year 3 medical students, 25 internal medicine residents in their second year of residency and 25 internal medicine faculty) were recruited to solve each of four complex clinical vignettes. In each vignette, the first four rounds of information pointed toward a narrowing range of diagnostic possibilities, but patient information presented in and after the fifth round was inconsistent with prior findings. In addition to accuracy, outcome measures were length of differential diagnosis, certainty of diagnosis, persistence in data collection and tendency to switch diagnoses.

RESULTS There were no significant differences in diagnostic accuracy across the three groups, each of which differed in level of training. However, across experience levels, diagnostic accuracy was associated with the mean number of items in the differential, tendency to persist (e.g. to request a greater number of rounds of information), and openness to switch (e.g. to change the most likely diagnosis on receipt of disconfirming information).

CONCLUSIONS Level of training (i.e. clinical experience) was not associated with accuracy on this task. As faculty clinicians certainly have more knowledge than their junior counterparts, it is important to identify ways in which cognitive factors can lead to more or less persistence and openness, and to teach clinicians how to overcome tendencies associated with error.

Medical Education 2017: 51: 1127–1137
doi: 10.1111/medu.13382

[1]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[2]Department of Psychology, Northeastern University, Boston, Massachusetts, USA

Correspondence: Edward Krupat, Beth Israel Deaconess Medical Center, Center for Educaton, 330 Brookline Ave, Boston, MA 02215 USA; E-mail: edkrupat@gmail.com

INTRODUCTION

Studies of medical decision making vary considerably in terms of the focus of the predictive and outcome variables investigated. Early studies focused on differences in levels of experience and expertise, and sought to establish whether practising physicians are more successful in the diagnostic process than students or trainees.[1–4] Rather than focusing on those making the diagnoses, other studies have investigated the content and context of the cases encountered, asking, for instance, whether the likelihood of error depends on case complexity and the consistency and order of the information presented.[5–8] Using lessons learned about processing, others have introduced various interventions aimed at influencing clinicians' cognitive processing and have attempted to determine their efficacy.[9–12]

Increasingly, studies have focused not simply on the extent to which correct versus incorrect inferences are made, but on how different approaches to formulating, confirming and changing hypotheses lead to more positive outcomes.[13–15] The present study compares the responses of medical students, residents and faculty physicians, respectively, when presented with a series of cases in which disconfirming information is delivered midway through a case, in an attempt to further investigate the extent to which processing differences mediate and explain the relationship between experience and accuracy.

In differentiating between experience and expertise, a number of perspectives have been offered.[16–18] Bereiter and Scardamalia[18] noted that 'experienced non-experts' do well on tasks requiring pattern recognition, but many experienced clinicians fail to monitor the external environment for unexpected or unanticipated cues. Joseph and Patel[19] observed that when physicians deal with areas in which they have expertise, they are often subject to biases such as that manifested by searching for confirming evidence. More generally, Mylopoulos and colleagues[20,21] noted that 'routine expertise' is applied when clinicians seek rapid solutions to familiar problems, whereas 'adaptive expertise' is needed to generate reflective and flexible solutions to complex challenges.

Moulton et al.[22] noted that not all physicians are capable of 'slowing down when [they] should'. In a study of surgeons, some were observed to actually stop during an operation to consider an unanticipated complication, whereas others were not as capable of monitoring or responding to the environment to deal with new and conflicting information, and continued the surgical procedure without pause, sometimes with negative consequences.[23]

Conceptually, these differences have been discussed in the context of System 1 versus System 2 cognitive processing.[12,24–26] System 1 processing is described as automatic, intuitive and low in effort. In System 1 mode, people attempt to incorporate new information within pre-existing patterns, move quickly toward resolution, and may be subject to bias. System 2 processing, which is described as logical and effortful, involves more careful and deliberative analysis. Environmental circumstances such as fatigue and time pressures may push for System 1 processing; however, clinicians operating in System 2 mode are more likely to avoid heuristic thinking and to be less subject to bias.

In an attempt to clarify questions about the ability to overcome the tendency to make decisions before sufficient information has been gathered, referred to as 'premature closure',[27,28] this study focuses exclusively on those circumstances in which clinicians are confronted with information that is inconsistent or disconfirming. Kruglanski and Webster describe the problem of premature closure as one of 'seizing' and 'freezing'.[29] Operating in a System 1-like manner, this process occurs in two steps. Initially, the clinician may 'seize' upon a first diagnostic solution that appears plausible and, with that solution in mind, may 'freeze' on it even upon the receipt of information that should cause the clinician to re-evaluate and entertain alternative diagnostic possibilities. Eva and colleagues used a different metaphor by asking whether clinicians are capable of 'swapping horses midstream'.[30] Several factors have been associated with openness to new information. In one study, younger, less experienced physicians were found to have greater openness to change, but no differences by experience were detected in clinicians' certainty in an initial diagnosis or in the breadth and type of diagnostic hypotheses generated.[31]

To study the diagnostic process, we recruited three groups of participants: medical students; residents, and faculty physicians. Each participant individually attempted to diagnose a set of clinical vignettes in which disconfirming information was introduced midway through each case. In addition to data on

the accuracy of the diagnoses generated by participants, we also collected data on their differential diagnoses, their levels of certainty, their persistence in collecting patient data, and the extent to which they showed an openness to changing diagnoses. This allowed us to investigate not only person-level variables (e.g. experience) associated with accuracy, but also those characteristics of the diagnostic process associated with accuracy across individuals who differed in experience. We predicted that characteristics of diagnostic style would be more important determinants of accuracy than experience, and expected that in this context in which confirmation biases might exist, experienced clinicians would not demonstrate greater accuracy than those with less experience.

METHODS

Participants and recruitment

Three groups of participants were recruited with the goal of enrolling 25 volunteers per group. These included: (i) medical students who had just completed their core clinical (third) year at one of the teaching hospitals affiliated with Harvard Medical School (HMS); (ii) residents in internal medicine who were completing their second year of training at the same hospital, and (iii) faculty physicians in internal medicine at the same hospital. All participants were recruited by e-mail.

The content of the recruitment message, which was identical for all three groups, asked for participation in an online study of clinical reasoning. The message indicated that participation would involve being sent one clinical vignette per month via e-mail for four consecutive months. To avoid the possibility of order effects, participants received the vignettes in different orders. At the end of the final vignette, all participants completed two well-validated instruments for measuring disposition: Curiosity (Intrinsic [I]- and Deprivation [D]-type)[32] and Need for Cognition.[33] The former was selected because of speculation that clinicians' trait-based curiosity may be associated with an interest in and an openness to new information. The latter, which assesses the extent to which people are inclined toward effortful cognitive activity, was chosen because it has been associated with persistence at intellectual tasks. Participants were compensated for each completed vignette. This project was approved via expedited review by the HMS Institutional Review Board.

For each group, the first 25 individuals who agreed to participate were included. One participant in each group dropped out midway through the study and was not replaced. Had every participant responded to each vignette, the number of responses available for analysis would have been 300 (three levels of experience × 25 participants per level × four vignettes). However, because three individuals dropped out, a total of 288 served as the *n* for all analyses at the level of vignette.

Vignettes and procedure

In consultation with several clinicians, we created four internal medicine clinical vignettes for this study. As all four were derived from real clinical cases, a correct diagnosis could be discerned for each. The medical conditions varied considerably, with correct diagnoses ranging from mesothelioma to multiple sclerosis. Participants were sent an e-mail with a link to an online web-based platform created for this study; each link led the participant to the specific vignette to which he or she was to respond that month.

In this study, we introduced an instrument that had not been used previously, which we called the Decision Certainty Analysis Tool (DCAT). Using the DCAT, participants encountered four vignettes, each intentionally written to be complex, and all with an identical structure. Each began with a two- or three-sentence description of the patient, including the patient's name, relevant demographic information and a statement of his or her presenting symptoms. At this point, subjects were asked to list as many items as they desired as possibilities in their differential diagnosis. For each item in the differential diagnosis, subjects then indicated their certainty using a scale extending from 1 (not at all certain) to 100 (completely certain).

Participants were instructed that their objective was to provide a diagnosis for each vignette, with the eventual goal of announcing the diagnosis to the patient ('Mr/Ms X, it is my opinion that…') and providing a diagnosis in their own words (rather than choosing from a set of multiple-choice-type options). The diagnosis offered was to be described in more specific terms than, for example, 'cardiac in nature' or 'sounds like a pulmonary problem' and thus more like 'viral pneumonia' or 'aortic aneurysm'.

When subjects had completed the listing of their differential diagnoses and had offered a certainty

rating for each diagnosis included in the first step, they were given the choice of announcing their diagnosis to the patient or asking for more information. If the participant selected the former, a page appeared in which he or she was asked to state the diagnosis. If the participant chose the latter, a second cue appeared (Step 2), of approximately similar length, offering additional relevant information (e.g. a more detailed description of symptoms, physical findings or relevant history). At this point, participants were able to modify their differential diagnoses based on the new information by adding or deleting diagnoses, and adjusting the level of certainty (as desired) for each diagnosis included in this step.

Although the procedure was designed to allow participants to receive up to 10 sets of cues (10 steps), in order to bring the diagnostic situation as close to reality as possible (i.e. where the process simply continues until the clinician has enough certainty in his or her diagnosis to stop), no mention was made of the amount of potential information available, no explicit reference was made to time, and there were no incentives or disincentives to stop early or late in the process. This procedure for Step 1 was followed for each of the following steps until the participant chose to declare a diagnosis, or until 10 steps had been completed, at which point participants were asked to offer a single final diagnosis.

The information given to participants in Steps 1–4, although of varying levels of diagnostic value, was designed to make participants increasingly confident in a narrowing set of possibilities. However, Step 5 (known as the 'disconfirming information step' [DIS]) contained information that was intentionally inconsistent with the prior data such that it might cause the participant to pause and feel the need to reformulate. For example, in one of the cases, Steps 1–4 presented information that could be interpreted as consistent with coronary artery disease or congestive heart failure; however, at the DIS, the information provided was intended to raise the possibility of chronic pulmonary emboli. The information provided in Steps 6–10 was generally consistent with that contained in the DIS.

This procedure, with one exception, was followed for every vignette. The exception occurred if a participant wished to announce his or her diagnosis before reaching the DIS (Step 5). In order to ensure that every participant was exposed to the disconfirming information, those participants who wanted to announce a diagnosis earlier were asked to read one more piece of information (the disconfirming information given in Step 5) before completing the case. After reading the DIS, participants who had requested to stop before Step 5 could choose to announce their final diagnosis at that point, or they could opt to continue to receive additional information as they had in previous steps.

Assessment of accuracy

The critical outcome was whether the final diagnosis for each vignette was correct or not; this was coded by two physicians working independently. A response was coded as 'correct' if it stated the diagnosis accurately and with an acceptable degree of specificity, or if it was considered sufficiently close to the precisely correct answer to lead quickly to the proper treatment and to be unlikely to compromise patient safety. A response was rated as 'incorrect' if it was sufficiently at variance with the actual diagnosis such that treatment based on the response would not have been effective. Instances of disagreement between coders were rare and all were resolved after brief discussion.

Analysis plan

The analysis of the data had several components. Firstly, we performed generalisability (G) studies[34,35] to assess the reliability of participants' responses across vignettes. Secondly, we examined the results according to experience level, Curiosity, and Need for Cognition to determine whether these person-level variables were associated with making the correct diagnosis. Thirdly, we sought to identify whether differences in 'processing style' served as useful indicators of whether the diagnosis was correct. Finally, we performed an informal analysis of outliers, or participants whose levels of accuracy across the four cases had been either very good or very poor. To do this, we compared data for the 27 participants who diagnosed all four cases correctly with data for the 12 participants whose diagnoses were inaccurate in three or four of the cases in order to test whether any of the outcome variables differentiated between the most and least accurate diagnosticians.

Generalisability

Variance components were independently computed at the level of participant, vignette and

order of vignettes, using either the total number of diagnoses considered or diagnostic accuracy (correct versus incorrect) as the primary outcome. Subsequently, G coefficients were determined using participant, vignette, order of vignettes, participant within vignette, participant within order, and participant within vignette and order. Variance components were determined by minimum norm quadratic unbiased estimation (MINIQUE) with diagnostic accuracy as the dependent variable, and participant, vignette and order as random, independent variables.

Person-level variables

Hierarchical linear modelling (HLM [described in more detail below]) was used to compare the extent to which experience level, Curiosity and Need for Cognition could each account for differences in rates of accuracy. Secondary analyses were conducted to indicate whether participants at the three levels of experience varied in the manner in which they processed and utilised the diagnostic information in the vignettes.

Process-level variables

The process variables collected fell into three categories: (i) characteristics of the differential (e.g. number of diagnoses listed, per step or on the first differential; breadth of differential, indicated by the number of systems considered) and certainty of the diagnosis (e.g. certainty of the most likely diagnosis on the initial differential); (ii) persistence or, more accurately, lack of persistence (e.g. intent to stop before the DIS; stopping after the DIS; number of steps completed), and (iii) openness to disconfirmation (e.g. switch of the most likely diagnosis from the first to the final step taken; switch of the most likely diagnosis immediately after the DIS).

Analytic approach for person- and process-level variables

Rather than aggregate across vignettes, as is typical, we chose to analyse the data using ratings for each individual for each vignette as the unit of analysis. We chose this analytic strategy because our primary research question focused on the variables associated with a correct diagnosis for any given vignette. This approach allows us to identify specific processes that tend to promote correct diagnosis even among participants with poor accuracy overall, or processes that are problematic even when used by participants with better accuracy overall.

Because the data from the individual vignettes were not independent (i.e. each subject responded to four vignettes), we employed HLM,[36] which is designed to deal with data that is nested (i.e. non-independent). Hierarchical linear modelling has advantages over many alternative methods of analysing repeated-measures data because it allows for the simultaneous estimation of within-subject and between-subject variance, more efficient estimation of effects, and lower rates of type I error.[36] Moreover, because HLM provides information on the consistency of relationships across individuals, this approach aids in the identification of processes with benefits or decrements that should generalise to any clinician in terms of his or her diagnostic accuracy, even one who has not actually demonstrated the use of a given process.

Because the key outcome variable, diagnostic accuracy, is binary, we also utilised an extension of HLM, hierarchical generalised linear modelling (HGLM),[36,37] for all models with this variable (or any other binary variable). All models yield coefficients (*B*s) for the model intercept and for each predictor variable. *B* coefficients for each predictor variable can be interpreted as average unstandardised regression coefficients; the model provides a *t*-ratio and p-value for each *B*, that indicate whether the effect of each predictor variable is significant. We report effect sizes as odds ratios with 95% confidence intervals. All analyses were carried out using HLM 7 (Scientific Software International, Inc. (Lincolnwood, IL, USA).

RESULTS

Generalisability

Individual vignette and order of the vignettes were not significantly associated with variability in participants' responses, based on the G coefficients and the calculated variance components. With regard to the number of diagnoses considered, individual participants accounted for 42.3% of total variability between cases, whereas vignette content and vignette order accounted for only 3.5% and 0.5% of variance, respectively. Using variance from participants, case content and case order resulted in a G coefficient of 0.82, which indicates adequate reliability between cases and order of cases. Similar analyses for correctness of the final diagnosis showed that variance between respondents was mostly attributable to individual participants within each vignette (41.0% of variability) rather than

Table 1 Process-level predictors of accuracy in diagnosis

	B	OR (95% CI)	p-value
Aspects of the differential			
Number of diagnoses considered on first step	0.12	1.13 (0.99–1.30)	0.08
Average number of diagnoses considered	0.44	1.55 (1.28–1.87)	<0.001
Total number of systems considered	0.07	1.07 (0.89–1.30)	0.46
Certainty of most likely diagnosis at first step	−0.01	0.99 (0.97–1.00)	0.11
Persistence			
Number of steps completed	0.74	2.09 (1.78–2.47)	<0.001
Attempted to quit before Step 5 (Yes/No)	−2.10	0.12 (0.06–0.24)	<0.001
Quit at Step 5 (Yes/No)	−3.19	0.04 (0.02–0.11)	<0.001
Openness to disconfirmation			
Switch diagnosis between first and last step (Yes/No)	2.96	19.29 (9.55–38.94)	<0.001
Switch diagnosis at Step 5 (Yes/No)	0.74	2.09 (0.80–5.42)	0.03

B represents the average population regression coefficient generated by hierarchical generalised linear modelling.
CI = confidence interval; OR = odds ratio.

individual participant (14.8% of variability), content of the vignette (1.0% variability) or participants within order of vignettes (0% of variability). The G coefficient for correctness of the final diagnosis was 0.60, indicating adequate reliability of results between vignettes regardless of the order in which they were considered by participants.

Person-level predictors of accuracy

Analyses at the participant level indicated no significant differences according to level of experience. Faculty physicians were correct in 67% of cases, residents were correct in 77% of cases and medical students were correct in 70% of cases (all two-way comparisons, $p \geq 0.20$). In addition, differences in accuracy were not predicted by participants' level of Curiosity ($p = 0.15$) or Need for Cognition ($p = 0.35$).

Process-level predictors of accuracy

The findings in Table 1 are based on HLM analyses and organised according to the relationship between accuracy and: (i) factors associated with the differential; (ii) factors associated with lack of persistence, and (iii) factors associated with openness to disconfirmation. Firstly, only one aspect of the participants' differential, the average number of items considered, predicted accuracy of diagnosis: the larger the mean number of items on the differential, the more likely the diagnosis was correct ($p < 0.001$). In addition, the length of the differential diagnosis on the initial step tended to relate to accuracy in the same way ($p = 0.08$). Breadth of the differential (i.e. number of different systems considered) and certainty of the most likely diagnosis on the initial differential did not predict accuracy ($p = 0.46$ and $p = 0.11$, respectively).

Secondly, the tendency to persist, operationalised in several ways, predicted accuracy (Table 1). Accuracy of diagnosis was associated with completing a greater number of steps, a tendency not to want to stop early (i.e. before the disconfirming information was presented), and a tendency not to quit immediately after reading the disconfirming information (all: $p < 0.001$). For the most part, participants did show persistence, based on simple calculations using the raw data. A request to stop before the DIS occurred in only 15% (44/288) of instances. In those instances, 61% (27/44) of participants who had wanted to stop before the DIS were sufficiently influenced by the content of the DIS to choose to continue; these participants reached an accurate diagnosis in 44% (12/27) of cases. However, in all of the 17 instances in which the participant wanted to stop before the DIS and then actually did stop after seeing the disconfirming information, the final diagnosis was incorrect.

Concerning openness to change, when participants switched their most likely diagnosis from Step 1 to

Table 2 Process-level predictors of openness to disconfirmation (switching diagnosis between first and last step)

	B	OR (95% CI)	p-value
Aspects of the differential			
Number of diagnoses considered on first step	0.15	1.16 (0.99–1.37)	0.07
Average number of diagnoses considered	0.45	1.57 (1.19–2.07)	0.002
Total number of systems considered	0.33	1.39 (1.13–1.70)	0.002
Certainty of most likely diagnosis at first step	−0.03	0.97 (0.95–0.99)	0.01
Persistence			
Number of steps completed	0.72	2.06 (1.71–2.48)	<0.001
Attempted to quit before Step 5 (Yes/No)	−1.64	0.19 (0.10–0.37)	<0.001
Quit at Step 5 (Yes/No)	−2.82	0.06 (0.02–0.15)	<0.001

B represents the average population regression coefficient generated by hierarchical generalised linear modelling.
CI = confidence interval; OR = odds ratio.

their final diagnosis or switched their most likely diagnosis immediately after reading the disconfirming information, the final diagnosis was more often accurate ($p < 0.001$ and $p = 0.03$, respectively). Calculations using the raw data showed a switch in the most likely diagnosis from first to last in 83% (239/288) of cases. In instances in which the most certain diagnosis was the same at the beginning and end, the diagnosis was accurate in only 16% (8/49) of cases, whereas in instances in which a switch was made from beginning to end, the diagnosis was correct in 82% (197/239) of cases.

Process-level predictors of openness to disconfirmation

As the vignettes used were built around disconfirmation, which implied a strong need to switch from one diagnosis to another, we also studied switching of the most likely diagnosis between the first and final steps as an additional outcome. Based on the HLM analyses, Table 2 indicates that several factors predicted whether participants switched diagnoses. When the average length of the differential (number of diagnoses considered) and breadth of the differential (number of total systems considered) were larger, participants were more likely to switch (both: $p = 0.002$). Switching was also predicted by lesser certainty on the first step ($p = 0.01$). Each measure of persistence (more steps completed, not wanting to quit before the disconfirming information, and not quitting immediately after the disconfirming information) was also significantly associated with willingness to switch from an initial diagnosis to a new one in response to the disconfirming information (all: $p < 0.001$).

Differences in diagnostic style among participants

Finally, although we did not find significant differences in the diagnostic *accuracy* of participants at differing levels of experience, we assessed whether the three groups demonstrated *stylistic* (i.e. process) differences. Analyses by HLM indicated a pattern of difference in the differentials of the three groups as faculty physicians listed significantly fewer items on Step 1 than did residents ($p = 0.01$) and medical students ($p = 0.01$), an average of fewer items per step than residents ($p = 0.02$) and medical students ($p = 0.002$), and included fewer systems (i.e. narrower differentials) on Step 1 than residents ($p < 0.001$) and medical students ($p = 0.002$). Residents and medical students did not differ from one another on any of these variables. Of the three groups, medical students most often wanted to stop before the DIS (22% of instances), which was significantly more often than residents, who wanted to stop prematurely in 9% of cases ($p = 0.01$). Faculty physicians wanted to stop prior to the DIS at an intermediate level, in 16% of instances (not significantly different from either of the other two groups).

Most and least accurate clinicians

Of the 24 participants in each group, eight medical students, 12 residents and seven faculty physicians

achieved an accurate diagnosis in all four vignettes. Conversely, five medical students, five faculty physicians and 12 residents produced the least accurate set of diagnoses and were wrong on three or four of the vignettes. The most consistently accurate clinicians differed from their less accurate counterparts in the following ways: (i) they included more items in their diagnostic lists ($p < 0.003$); (ii) they considered more systems in total ($p < 0.002$); (iii) they completed more steps ($p < 0.001$); (iv) they less frequently quit before or at the DIS (both: $p < 0.001$), and (v) they more often switched diagnoses during the course of the vignette ($p < 0.001$).

DISCUSSION

As predicted, we found no direct relationship between experience and diagnostic accuracy in this set of complex vignettes. In addition, several factors relating to the manner in which the clinicians processed the information referred to diagnostic accuracy. Our findings support Bereiter and Scardamalia's[18] notion of 'experienced non-experts' and reinforce the suggestion that not all expertise is *adaptive* expertise.[20] These results illustrate the pitfalls of low-effort, quick-to-conclusion System 1 thinking, characterised by a lack of persistence in data collection and resistance to change.

Specifically, we found that a larger differential was associated with greater accuracy. This suggests that initially formulating a wider range of possible explanations may contribute to accuracy by allowing for the consideration of many hypotheses. Although we cannot make cause-and-effect inferences from our data, we can nonetheless ask whether the faculty physicians, who were certainly more knowledgeable than the medical students, might have achieved better results had they entertained differentials that were larger or broader. Based on our findings, we propose that the experienced physicians, who were more prone than residents and students to rely initially on System 1 thinking, produced shorter, 'intuitive' differential diagnoses and were more resistant to change in the face of subsequent conflicting information. In essence, their process may have shown a tendency toward 'seizing and freezing'.[29]

Another factor associated with accuracy was tendency to persist in the acquisition of data rather than stopping quickly. Although it is impossible for clinicians to know in advance whether or not subsequent information will reinforce or disconfirm their initial diagnosis, or will merely distract them from the essential elements of a case, it is important to recognise that continuing to collect information may reveal new wrinkles or discordant information that will require problem reformulation. When the search for information is shut off quickly, *in some cases* critical information will not be obtained and certainly not processed.

In this study, the ratio between the levels of accuracy achieved by those who did not want to quit before the DIS (accurate in 79% of instances) and those who requested to stop before the DIS (accurate in 30% of instances) was > 2.5 : 1. Although it is impossible to make generalisations about how much information is too little or too much to achieve diagnostic accuracy, those participants who stayed with the vignettes longer were considerably more likely to achieve the correct diagnosis, and those clinicians who were most consistently accurate almost always went to the end (completing an average of 9.78 steps) compared with those who were least accurate (who completed a mean of only 6.25 steps). Although some have equated 'quick' and 'slow' thinking with time on the clock[6] (parenthetically, we found no association between accuracy and time spent per case), we prefer the metaphoric interpretation that suggests that slowing down, which represents a willingness to continue the work of gathering information in order to assure certainty and accuracy, is a characteristic associated with reaching the right diagnosis.

Finally, participants' openness and willingness to 'change horses' was a key factor in diagnostic accuracy. Starting with a large differential and collecting more information are of no benefit unless one is willing to act on that information when all the pieces of the puzzle do not fit neatly together. In this study, there were 17 instances in which a participant wanted to quit before Step 5 (i.e. was not very persistent) and then did quit immediately after seeing the disconfirming information in the DIS. Although the number of instances in which this occurred is far too small to support firm generalisations, this 'toxic' combination of lack of persistence and lack of openness to change was associated with making a correct diagnosis in none of the 17 instances, which is particularly notable when we recognise that the overall base rate for accuracy in these vignettes was 71%.

The findings reported here, although suggestive, are subject to significant limitations. This is the first use of the DCAT, and our results are based on only four vignettes, all of which were structured in a highly specific way (to imply the need for reformulation midway through the case), which may not be typical in clinical practice. There are many examples in clinical medicine in which discordant information is merely a distractor and does not, in fact, lead to an alternative correct diagnosis. In fact, it is possible that the characteristics noted here (i.e. persistence and willingness to change horses) might in some situations lead to unnecessary testing and diminished diagnostic accuracy.

Additional limitations include the fact that the data came from a relatively small sample at one medical school and associated teaching hospital. As the cases were completed online, we cannot be certain that participants were fully attentive or of the amount of effort they made, although a median time of almost 17 minutes per vignette suggests that participants did take them seriously. Although the G analyses produced satisfactory results bearing on the reliability of the findings, we recognise the need to address concerns about content specificity by using a broad range of cases with a variety of clinicians in further data collection involving the DCAT. Another important way of validating this new method would be to compare the results of DCAT studies against findings using established measures of clinical reasoning such as the script concordance test,[38,39] especially utilising the 'written think-aloud' approach of Power et al.[40]

Subject to further testing, we believe the DCAT has potential to add both methodological and educational value. The many and sometimes conflicting findings produced in the domain of diagnostic reasoning raise the question of whether method variance, the fact that each study introduces information in different ways and asks participants to take on different tasks, may account in part for the differences reported. Given the great variety of approaches, the DCAT might serve as a standardised tool with which to investigate the process of diagnostic reasoning, using cases written for different disciplines and at different levels of complexity, while allowing for the collection of a number of different types of outcome data.

Educationally, the DCAT can be used in a variety of ways. It might serve as a training instrument for use in medical students and residents, allowing them, in small groups, to compare their differentials and levels of certainty, and offering a teaching platform for instruction and discussion about dealing with inconsistent information and bias. The DCAT can also serve as a mechanism with which to identify clinicians at any level of experience whose diagnostic skills are weak, and may be used for instruction, remediation, formative assessment and, once sufficiently validated, possibly even summative assessment.

CONCLUSIONS

Clinical experience was not associated with accuracy on this task. Rather, factors such as the number of diagnostic possibilities entertained, persistence in collecting information and openness to change were of greatest importance in overcoming the confirmatory bias, avoiding premature closure and achieving a correct diagnosis. As faculty clinicians certainly have more knowledge than their junior counterparts, it is important to identify ways in which cognitive factors interact with experience, and to introduce knowledge about this into medical training at all levels, including in continuing education. This would enable clinicians to avoid becoming over-reliant on premature, intuitive diagnoses identified via a process of pattern recognition, and to overcome cognitive tendencies and processing shortcuts that are most closely associated with diagnostic error.

Contributors: EK and JBR contributed to the conception and design of the work, and the acquisition, analysis and interpretation of data. JW contributed to the analysis and interpretation of data. RMS contributed to the conception of the work, and to the analysis and interpretation of data. All authors contributed to the drafting and critical revision of the paper, and approved the final manuscript for publication. All authors have agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
Acknowledgements: The authors acknowledge the significant contributions of David Roberts, Department of Medicine, Beth Israel Deaconess Medical Center, and Jannelle Yopchick, who was a doctoral student in the Department of Psychology, Northeastern University, at the time of her participation to this work in its earlier stages.
Funding: This study was funded by the President's Office, Harvard University.
Conflicts of interest: None.
Ethical approval: This study was approved by the Institutional Review Board of Harvard Medical School.

REFERENCES

1 Arocha JF, Patel VL, Patel YC. Hypothesis generation and the coordination of theory and evidence in novice diagnostic reasoning. *Med Decis Making* 1993;**13** (3):198–211.
2 Norman GR, Muzzin LJ, Rosenthal D. Expert–novice differences in perception and categorisation in dermatology. Presented at the Annual Meeting of the American Educational Research Association, 31 March to 4 April 1985, Chicago, IL.
3 Sisson JC, Donnelly MB, Hess GE, Wooliscroft JO. The characteristics of early diagnostic hypotheses generated by physicians (experts) and students (novices) at one medical school. *Acad Med* 1991;**66** (10):607–12.
4 Benbassat J, Bachar-Bassan E. A comparison of initial diagnostic hypotheses of medical students and internists. *J Med Educ* 1984;**59** (12):951–6.
5 Mamede S, von Gogh T, van den Berge K, van Saase JL, Schmidt HG. Why do doctors make mistakes? A study of the role of salient distracting features. *Acad Med* 2014;**89** (1):1–7.
6 Cavalcanti RB, Sibbald M. Am I right when I am sure? Data consistency influences the relationship between diagnostic accuracy and certainty. *Acad Med* 2014;**89** (1):107–13.
7 Cunnington JP, Turnbull JM, Regehr G, Marriott M, Norman GR. The effect of presentation order in clinical decision making. *Acad Med* 1997;**72** (10 Suppl 1):40–2.
8 Bergus GR, Chapman GB, Levy BT, Ely JW, Oppliger RA. Clinical diagnosis and the order of information. *Med Decis Making* 1998;**18** (4):412–17.
9 Corderre S, Wright B, McLaughlin K. To think is good: querying an initial hypothesis reduces diagnostic error in medical students. *Acad Med* 2010;**85** (7):1125–9.
10 Norman G, Young M, Brooks L. Non-analytic models of clinical reasoning: the role of experience. *Med Educ* 2007;**41** (12):1140–5.
11 Corderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. *Med Educ* 2003;**37** (8):695–703.
12 Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Adv Health Sci Educ Theory Pract* 2009;**14** (Suppl 1):27–35.
13 Ark TK, Brooks LR, Eva KW. The benefits of flexibility: the pedagogical value of instructions to adopt multi-faceted diagnostic reasoning strategies. *Med Educ* 2007;**41** (3):281–7.
14 Hess BJ, Lippner RS, Thompson V, Homboe ES, Graber ML. Blink or think: can further reflection improve initial diagnostic impressions? *Acad Med* 2015;**90** (1):112–18.
15 Norman G, Monteiro S, Sherbino J. Reflecting upon reflection in diagnostic reasoning. *Acad Med* 2014;**89** (9):1195.
16 Weinger MB. Experience does not equal expertise. Can simulation be used to tell the difference? *Anesthesiology* 2007;**107** (5):691–4.
17 Murray DJ, Boulet JR, Avidan M, Kras JF, Heinrichs B, Woodhouse J, Evers AS. Performance of residents and anaesthesiologists in a simulation-based skill assessment. *Anesthesiology* 2007;**107** (5):705–13.
18 Bereiter C, Scardamalia M. *Surpassing Ourselves: an Inquiry into the Nature and Implications of Expertise*. Chicago, IL: Open Court 1993;1.
19 Joseph GM, Patel VL. Domain knowledge and hypothesis generation in diagnostic reasoning. *Med Decis Making* 1990;**10** (1):31–46.
20 Mylopoulos M, Regehr G. Cognitive metaphors of expertise and knowledge: prospects and limitations for medical education. *Med Educ* 2007;**41** (12): 1159–65.
21 Mylopoulos M, Woods NN. When I say … adaptive expertise. *Med Educ* 2017;**51**(7):685–6.
22 Moulton CE, Regehr G, Mylopoulos M, McRae HM. Slowing down when you should: a new model of expert judgement. *Acad Med* 2007;**82** (10 Suppl):109–16.
23 Moulton CE, Regehr G, Lingard L, Merritt C, MacRae H. Slowing down to stay out of trouble in the operating room: remaining attentive to automaticity. *Acad Med* 2010;**85** (10):1571–7.
24 Kahneman D. *Thinking Fast and Slow*. New York, NY: Farrar, Straus & Giroux 2011.
25 Croskerry P. A universal model of diagnostic reasoning. *Acad Med* 2009;**84** (8):1022–8.
26 Tay SW, Ryan P, Ryan CA. Systems 1 and 2 thinking processes and cognitive reflection testing in medical students. *Can Med Educ J* 2016;**7** (2):e97–103.
27 McSherry D. Avoiding premature closure in sequential diagnosis. *Artif Intell Med* 1997;**10** (3):269–83.
28 Kumar B, Kanna B, Kumar S. The pitfalls of premature closure: clinical decision-making in a case of aortic dissection. *BMJ Case Rep* 2011. https://doi.org/10.1136/bcr.08.2011.4594.
29 Kruglanski AW, Webster DM. Motivated closing of the mind: 'seizing' and 'freezing'. *Psychol Rev* 1996;**103** (2):263–83.
30 Eva KW, Link CL, Lutfey KE, McKinlay JB. Swapping horses midstream: factors related to physicians' changing their minds about a diagnosis. *Acad Med* 2010;**85** (7):1112–17.
31 Eva KW, Cunnington JP. The difficulty with experience: does practice increase susceptibility to premature closure? *J Contin Educ Health Prof* 2006;**26** (3):192–8.
32 Litman JA. Relationships between measures of I- and D-type curiosity, ambiguity tolerance, and need for closure: an initial test of the wanting–liking model of information seeking. *Pers Indiv Diff* 2010;**48** (4):397–402.
33 Cacioppo JT, Petty RE, Feinstein JA, Jarvis WBG. Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychol Bull* 1996;**119** (2):197–253.
34 Tavakol M, Brennan RL. Medication education assessment: a brief overview of concepts in generalisability theory. *Int J Med Educ* 2013;**4**:221–2.

35 Eva KW, Reiter HI, Rosenfeld J, Norman GR. The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Acad Med* 2004;**79** (6):602–9.
36 Raudenbush S, Bryk A, Cheong YF, Congdon R, Du Toit M. *HLM 7*. Lincolnwood, IL: Scientific Software International, Inc 2011.
37 Kenny DA, Korchmaros JD, Bolger N. Lower level mediation in multilevel models. *Psychol Methods* 2003;**8** (2):115–28.
38 Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten CP. The script concordance test: a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12** (4):189–95.
39 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;**45** (4):329–38.
40 Power A, Lemay JF, Cooke S. Justify your answer: the role of written think aloud in script concordance testing. *Teach Learn Med* 2017;**29** (1):59–67.

Received 21 December 2016; editorial comments to author 3 March 2017, accepted for publication 24 May 2017