Reflecting on Diagnostic Errors: Taking a Second Look is Not Enough

Sandra D. Monteiro, PhD[1,4], Jonathan Sherbino, MD[2], Ameen Patel, MD[2], Ian Mazzetti, MD[2], Geoffrey R. Norman, PhD[1], and Elizabeth Howey, BSc[3]

[1]Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; [2]Department of Medicine, McMaster University, Hamilton, Ontario, Canada; [3]Program for Education Research and Development, McMaster University, Hamilton, Ontario, Canada; [4]Program for Education Research and Development, McMaster University, Hamilton, Ontario, Canada.

BACKGROUND: An experimenter controlled form of reflection has been shown to improve the detection and correction of diagnostic errors in some situations; however, the benefits of participant-controlled reflection have not been assessed.

OBJECTIVE: The goal of the current study is to examine how experience and a self-directed decision to reflect affect the accuracy of revised diagnoses.

DESIGN: Medical residents diagnosed 16 medical cases (pass 1). Participants were then given the opportunity to reflect on each case and revise their diagnoses (pass 2).

PARTICIPANTS: Forty-seven medical Residents in postgraduate year (PGY) 1, 2 and 3 were recruited from Hamilton Health Care Centres.

MAIN MEASURES: Diagnoses were scored as 0 (incorrect), 1 (partially correct) and 2 (correct). Accuracies and response times in pass 1 were analyzed using an ANOVA with three factors—PGY, Decision to revise yes/no, and Case 1–16, averaged across residents. The extent to which additional reflection affected accuracy was examined by analyzing only those cases that were revised, using a repeated measures ANOVA, with pass 1 or 2 as a within subject factor, and PGY and Case or Resident as a between-subject factor.

KEY RESULTS: The mean score at pass 1 for each level was PGY1, 1.17 (SE 0.50); PGY2, 1.35 (SE 0.67) and PGY3, 1.27 (SE 0.94). While there was a trend for increased accuracy with level, this did not achieve significance. The number of residents at each level who revised at least one diagnosis was 12/19 PGY1 (63 %), 9/11 PGY2 (82 %) and 8/17 PGY3 (47 %). Only 8 % of diagnoses were revised resulting in a small but significant increase in scores from Pass 1 to 2, from 1.20/2 to 1.22 /2 (t=2.15, p=0.03).

CONCLUSIONS: Participants did engage in self-directed reflection for incorrect diagnoses; however, this strategy provided minimal benefits compared to knowing the correct answer. Education strategies should be directed at improving formal and experiential knowledge.

KEY WORDS: clinical reasoning; diagnostic error; reflection.
J Gen Intern Med 30(9):1270–4
DOI: 10.1007/s11606-015-3369-4

Electronic supplementary material *The online version of this article (doi:10.1007/s11606-015-3369-4) contains supplementary material, which is available to authorized users.*

Published online July 15, 2015

INTRODUCTION

One popular dual process model of reasoning[1–5] proposes that errors in reasoning are a consequence of cognitive biases that originate in System 1 processes: a rapid error-prone memory retrieval system. Importantly, System 1 is contrasted with System 2, which functions to both detect and correct the biases that lead to error.[1–5] The applied interpretations of this model have had a significant impact on strategies for error reduction[6–13]; specifically that additional cognitive processing can prevent diagnostic errors.[6] Proponents of this dual process model of reasoning argue that improved metacognitive skills (i.e., cognitive forcing strategies[7]) and a slower, cautious approach can increase reliance on System 2, resulting in more errors being detected and corrected.[2,3,8–10] However, the available scientific evidence does not support this strategy.[14–19] Critically, a series of experiments demonstrated that instructing residents to proceed slowly and cautiously when diagnosing written medical cases did not improve diagnostic accuracy compared to being instructed to be more rapid.[16–19] In another study, participants had improved diagnostic accuracy after receiving a warning about the extreme difficulty of written medical cases compared to participants who did not receive the same warning; however, the authors were not able to replicate this effect with new cases.[20]

Still others have suggested that physicians should consider their initial diagnosis to be incorrect and take a critical 'second look'.[8–10] In one study of this reflective strategy, participants were asked to list only their initial diagnosis for half the cases.[21] For the other half of cases, they were required to evaluate their initial diagnosis using a series of time-consuming steps that critically appraised the evidence in the case and identified alternate diagnoses.[21] In that study, participants required two hours to fully diagnose 16 written cases, and there was still no overall improvement in accuracy for cases diagnosed using the reflective strategy.[21] In another study, medical residents were able to detect and correct some diagnostic errors after taking a second look at the case;[22] however, the majority of errors were artificially induced, experimenters controlled which cases the participants were allowed to revise and participants were given access to all the case details to assist in their reflective strategy.

Thus far, investigations of strategies to reduce diagnostic error have been limited to a prescribed form of error detection

or reflection,[16,19–22] with mixed results. It is not clear if any benefit will remain when 1) the decision to reflect is left to the clinician and 2) there is no prescription about how reflection should proceed. Further research is required to determine if reflection is an appropriate and worthwhile strategy for improving accuracy at differing levels of expertise if participants select which cases to revise. In the present study, we addressed the following questions:

1) How will physicians choose to use an opportunity to reflect?
2) Can reflection help detect and correct errors?

Identification of difficult cases should result in increased time during initial diagnosis. As well, if reflection is beneficial, revisions should increase accuracy significantly.

We also had a number of secondary research questions:

1) Will access to case details during reflection improve performance compared to having limited access?
2) Does response time predict a decision to revise a diagnosis?
3) Are more senior residents more accurate overall?

METHOD

Design

The study was a randomized mixed design, comparing between subject effects due to access to case details and within-subject effects resulting from decisions to reflect again and revise a prior case diagnosis.

Setting

Residents doing a medicine rotation in the teaching hospitals associated with McMaster University in Hamilton were invited to participate. The test sites were the Juravinski Cancer Centre, St. Joseph's Hospital and McMaster Children's Hospital. The study was conducted by SM using laptops set up in conference rooms within each of the test sites.

Participants

Recruitment. Selected residents were informed of the study by e-mail and invited to participate during the hour before morning rounds or during the lunch hour by co-authors AP and IM. Recruitment continued for several months, until we acquired a sufficient sample size. This study includes a total of 47 residents; 19 in post-graduate year (PGY) 1, 11 in PGY 2 and 17 in PGY 3. The study was approved by the McMaster Integrated Research Ethics Board HIREB 11-409.

Materials

Participants were presented with 16 general medicine cases that were a randomly selected subset of cases, some of which were used in Sherbino et al.,[16] Norman et al.[17] and Monteiro et al.[18] These cases were created by a panel of two experienced Emergency Medicine physicians and two experienced Internal Medicine physicians.[16] All cases were reviewed by the panel to ensure that there was only one correct diagnosis. All cases followed the same structure, presenting the patient's primary complaint and a representative patient photograph, followed by the history, tests ordered and a diagnostic image (e.g., CT scan, rhythm strip, etc.). These images were not critical to the diagnosis, but only supported the results reported in the text. Cases were matched for word count and reading time, but not difficulty. The level of case difficulty ranged qualitatively from rare and difficult to straightforward acute medical conditions. Diagnostic performance for these cases ranged from 21 to 82 %.[16–18] A sample case is shown in Appendix 1. In previous studies, performance on this sample case was 82 %.[16–18]

All participants reviewed the same set of cases, but in randomized order. Cases were presented on laptop computers using RunTime Revolution (version 2.8.1; Edinburgh Scotland) software. Case processing time and case diagnoses were recorded by the software and exported as text.

Procedure

Participants were told they would be asked to diagnose 16 general medicine cases. Once the experiment started, participants saw a welcome screen on the computer and they entered basic information: name, program name and program year. The primary instructions, including general tips for navigating through the program and entering responses as well as a description of how the case information would be presented were delivered in written form on screen. In the first phase (Pass 1) of the experiment, participants were encouraged to proceed quickly and enter only a single diagnosis for each case. Their instructions were:

> "You will be asked to read and diagnose several cases in 20 min. Each case description includes a brief description of the patient and vital statistics, as well as a photograph of the patient and an accompanying diagnostic image when available (e.g., x-ray, ECG, etc.)…Remember that you will not be able to go back to the case file once you have advanced to the diagnosis screen. Read the case information completely, but remember to use your time carefully as you only have 20 min."

After a first pass through the cases, participants were then given an opportunity for further reflection and asked to review all 16 cases again and choose between retaining or revising their previous diagnosis (Pass 2) . The message presented before proceeding with the review was:

> "Thank you for assessing these cases quickly. We would now like you to carefully reconsider every

diagnosis. Please re-consider all the evidence, before confirming or changing your initial diagnosis…"

Through random assignment, half the participants were able to review the full case details and half the participants only saw the primary complaint and patient photograph during Pass 2.

Scoring

All responses were scored for accuracy on a three-point system. This system was created by consensus from an expert panel of two experienced Emergency Medicine physicians and two experienced Internal Medicine physicians.[16] The panel created a list of correct, partially correct and incorrect potential diagnoses for each of the 16 cases. While all cases only had a single correct diagnosis, the list included acceptable synonyms for correct diagnoses. The list also included synonyms for incorrect and partially correct diagnoses. These cases have been used in a number of previous studies,[16–18] and the list of diagnoses was continually revised to include scoring for new responses that arose from those studies. The scoring rubric for the sample case used in the current study is provided in Appendix 2.

All participant responses were scored using this list. Incorrect diagnoses received a score of 0, partially correct responses received a score of 1 and correct diagnoses were assigned a 2. Diagnoses for the current study were scored and tallied by the author (SM) who was blind to condition. We report accuracy as average scores out of two and the standard error of the mean (SEM), percent correct and also as a count for incorrect, partially correct and correct diagnoses. We also report response times in seconds (sec) and standard deviation (SD).

Primary Analysis. The analysis focused on 1) the conditions under which a decision was made to review a case, and 2) the consequences of that decision. As the decision arose on a case-by-case basis, the unit of analysis was Case. A complete analysis would examine the accuracy and time taken to reach a diagnosis for each participant, case, and pass 1 or 2 and all interactions. However, as discussed in the next section, relatively few cases were revised, so there would be large amounts of missing data (many cases would have no second pass). Instead, the first question, the extent to which clinicians are aware of their errors, was addressed by examining the accuracies and response times on the first pass using an ANOVA with three factors—PGY, Decision to revise yes/no, and Case 1–16—averaged across residents, The analysis was then replicated using Resident, averaged across cases. All results are cited for the analysis using Case; the Resident analysis led to the same conclusions.

The second question, the extent to which additional reflection resulted in increased accuracy, was examined by analyzing only those cases that were revised, using a repeated measures ANOVA, with Pass 1 or 2 as a within-subject factor, and PGY and Case or Resident as a between-subject factor. All analyses were calculated using IBM SPSS Statistics, Version 22.0.

RESULTS

Summary

All participants completed the initial diagnosis of all 16 cases; average time was 26 min (range=12 to 38). Less than 0.5 % of all responses (7/752) were not recorded due to participant error (i.e., incorrectly advancing to the next screen). A total of 745 diagnoses were scored, resulting in 309 correct and 436 partially or completely incorrect diagnoses in Pass 1. In Pass 2, only 60 diagnoses were revised overall, resulting in 322 correct and 423 partially or completely incorrect diagnoses. The mean score at Pass 1 for each level was PGY1, 1.17 (SEM 0.50); PGY2, 1.35 (SEM 0.67) and PGY3, 1.27 (SEM 0.94). While there was a trend for increased accuracy with level, this did not achieve significance. Average time to diagnosis was 110 s per case for PGY1, 86 s for PGY2 and 101 s for PGY3, which although significant, did not show a consistent trend. There was a significant overall effect of case difficulty (F=2.91, p<0.001), but no interaction with postgraduate level or decision to revise.

a) How did physicians use the opportunity to reflect?
When residents were offered the opportunity to review each case again, only 8 % (60 out of 745) of all diagnoses were revised, suggesting that residents were generally confident in their initial diagnosis, despite the fact their accuracy was only 58 to 64 % on average. On average, residents took 97 s (SD 24) to read a case in Pass 1 and only 17 s (SD 12) to read a case in Pass 2. Fourteen residents revised only one diagnosis, seven revised two, and seven revised more than two cases. The proportion of cases revised was 10 % for PGY1 residents, 9 % for PGY2 and 5 % for PGY3 [X^2 (2) = 6.59, p=0.04]. The number of residents at each level who revised at least one diagnosis was 12/19 PGY1 (63 %), 9/11 PGY2 (82 %) and 8/17 PGY3 (47 %). Availability of the case resulted in a higher rate of revisions [38 vs. 22 %, X^2 (2) = 4.1, p=0.04], but did not affect accuracy.

b) Did reflection help detect and correct errors?
Diagnoses that were revised were significantly less accurate initially than those that were not (0.64/2 vs. 1.25/2; F(1,671 = 17.7, p<0.001). There was no significant interaction with level. Further, diagnoses that were eventually revised took about 10 sec longer than those that were not revised; however, this was not significant. Repeated ANOVA measures showed that scores for the revised diagnoses increased significantly from 0.64 to 0.90, F(1,28) = 4.26, p=0.05. The table examines the relation between initial and revised accuracy in detail and shows that 28 of 158 (18 %) completely incorrect (i.e., score of 0) diagnoses were revised, and the average final score of

the revised diagnoses was 0.50 out of two in Pass 2 (Table 1). Similarly, 28 of 279 (10 %) partially correct (i.e., score of 1) diagnoses were revised, and this resulted in an increase of 0.14 in score. Conversely, the few diagnoses (6) that received a 2.0 score in Pass 1 and were revised after reflection experienced a drop in score of about 0.8.

Because so few diagnoses were revised, the impact on overall accuracy was small, resulting in an increase in scores from Pass 1 to 2, from 1.20/2 to 1.22 /2 (t=2.15, p=0.03), respectively. Therefore, although residents were, to some degree, able to identify their own mistakes and made attempts to correct them, the impact of revisions on diagnostic accuracy was minimal.

DISCUSSION

Studies to date of the effect of instruction to slow down, be systematic or reflect have been of two forms—a parallel groups design where one group proceeds quickly and the other slowly,[16,17] and a longitudinal design where participants initially proceed quickly then go through the cases more intensively.[19–22] The latter studies have shown some success; however, they involve an intensive "reflection" intervention in which the clinician creates comprehensive matrices of signs and symptoms against diagnoses. Further, "reflection" is mandatory, and not under clinician control. Finally, the method involves reviewing the original case protocol. In the present study, we focused on the longitudinal design to assess the impact of revisiting a diagnosis in a more ecologically valid fashion, in which 1) no instruction about how to be reflective was given, 2) participants could choose to review a case or not, and 3) the effect of presence or absence of the case description was examined experimentally. Examining the overall performance under various conditions, we showed that 1) unstructured reflection on a review of the cases provided some benefit on individual cases, but the overall effect was small. Relatively few of the incorrect diagnoses were revised, and overall accuracy only increased by 2 %, and 2) to some degree, participants were able to recognize diagnostic errors and correct them, and this was associated with slightly longer reading times both initially and on case revision, replicating and extending previous work.[16]

Why were rates of revision so low? One possibility is that participants were unsuccessful at improving their scores because of limits in their knowledge or experience,[23] so that they had insufficient knowledge to recognize their errors. As a consequence, additional reflection resulted in only minimal improvement in accuracy. Outside of medical education, undergraduate psychology students were far more accurate (66 %) on general knowledge questions they answered immediately than for questions they deferred and revised (4 %), consistent with the suggestion that people make quick judgements about their knowledge and only reflect when they are uncertain or do not have the knowledge.[24] In the present study, participants with the knowledge to diagnose a medical case correctly the first time did not need to reflect further, while participants without the required knowledge could not benefit from further reflection. This suggests that diagnostic performance is not modulated by reasoning skills, added reflection or identification of cognitive biases, but by experience and knowledge.[18]

The results of the present study also provide information on the ability of physicians to self-assess and identify possible errors. We did demonstrate that if physicians are aware of their diagnostic mistakes, they will attempt to correct them by trying to revise incorrect diagnoses. However, their ability to detect and correct errors is far from perfect; only 18 % of incorrect diagnoses were revised correctly. The overall accuracy of revised diagnoses remained much lower than Pass 1 diagnoses that were not revised in pass 2; most errors remained undetected.

Table 1 Count (Percent) of Diagnoses that Maintained or Changed in Accuracy from Pass 1 to Pass 2

		Second pass		
		Incorrect (0 or 1)	Correct (2)	Total
First pass	Incorrect (0 or 1)	420 (56.4 %)	16 (2.1 %)	436
	Correct (2)	3 (0.4 %)	306 (41.0 %)	309
		423	322	745

LIMITATIONS

One clear concern is that the study's findings are based on written cases, which obviously leave out important aspects of the dynamics of clinical reasoning. But the question is not whether the study is a good representation of the "real world" (it is not), but whether the constraints of the study methods invalidate the findings. Evidence from other studies[25] indicates that students learn clinical reasoning as well from written cases as from videos and from live standardized patients.

A second limitation is that the reflection phase was clearly constrained in time, and participants had no opportunity to seek additional knowledge. Further research could expand this step to permit such strategies as internet searches and measure the impact on accuracy.

Additionally, the range of expertise was constrained. We have not examined whether expert clinicians are equally vulnerable to errors, although other evidence suggests that they differ more in degree than in kind.[26]

CONCLUSION

There remains little consistent evidence that strategies that focus on improved reasoning and added reflection are reliable. In some retrospective reports, the rate of diagnostic errors linked to knowledge deficits is quite low compared to the rate

of errors linked to reasoning deficits.[9,10] However, interventions directed at reducing errors by making clinicians aware of cognitive biases have been negative.[14,15] It may well be that there is no "quick fix" to reduce errors, and strategies should be directed at improving formal and experiential knowledge.

Acknowledgements: Contributors: *There are no other contributors to acknowledge.*

Funders: *This research was funded by a Canada Research Chair awarded to Dr. Norman.*

Prior Presentations: *The abstract for this study appeared in the Research in Medical Education Conference Program, (RIME) 2013. This study was also presented as an electronic poster at The Hodges Education Scholarship International Symposium (THESIS) organized by the Wilson Centre, University of Toronto in 2014.*

Conflict of Interest: *The authors do not have any conflicts of interest to declare.*

Corresponding Author: *Sandra D. Monteiro, PhD; Program for Education Research and DevelopmentMcMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4L1, Canada (e-mail: monteisd@mcmaster.ca).*

REFERENCES

1. **Evans JSB.** Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol. 2008;59:255–278.
2. **Evans JSB, Stanovich KE.** Dual-process theories of higher cognition advancing the debate. Perspect Psychol Sci. 2013;8(3):223–241.
3. **Kahneman D.** Thinking, fast and slow. New York: Farrar, Straus and Giroux; 2011.
4. **Kahneman D, Frederick S.** Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich T, Griffin D, Kahneman D, eds. Heuristics and biases: the psychology of intuitive judgment. Cambridge: Cambridge University Press; 2002:49–81.
5. **Stanovich KE, West RF.** On the relative independence of thinking biases and cognitive ability. J Pers Soc Psychol. 2008;94(4):672.
6. **Croskerry P, Sinclair D.** Emergency medicine: a practice prone to error. CJEM. 2001;3(4):271–276.
7. **Croskerry P.** Cognitive forcing strategies in clinical decision making. Ann Emerg Med. 2003;41(1):110–120.
8. **Croskerry P.** Clinical cognition and diagnostic error: applications of a dual process theory of reasoning. Adv Health Sci Educ. 2009;14(1):27–35.
9. **Chisholm CD, Weaver CS, Whenmouth L, Giles B.** A task analysis of emergency physician activities in academic and community settings. Ann Emerg Med. 2011;58(2):117–122.
10. **Graber ML, Gordon R, Franklin N.** Reducing diagnostic errors in medicine: what's the goal? Acad Med. 2002;77(10):981–992.
11. **Schiff GD, Bates DW.** Can electronic clinical documentation help prevent diagnostic errors? N Engl J Med. 2010;362(12):1066–1069.
12. **Klein JG.** Five pitfalls in decisions about diagnosis and prescribing. BMJ. 2005;330(7494):781.
13. **Redelmeier DA.** The cognitive psychology of missed diagnoses. Ann Intern Med. 2005;142(2):115–120.
14. **Sherbino J, Dore KL, Siu E, Norman GR.** The effectiveness of cognitive forcing strategies to decrease diagnostic error: an exploratory study. Teach Learn Med. 2011;23(1):78–84.
15. **Sherbino J, Kulasegaram K, Howey E, Norman G.** Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. CJEM. 2012;15:1–7.
16. **Sherbino J, Dore KL, Wood TJ, Young ME, Gaissmeier W, Krueger S, Norman GR.** The relationship between response time and diagnostic accuracy. Acad Med. 2011;87:785–791.
17. **Norman GR, Sherbino J, Dore K, Young M, Gaissmeier W, Monteiro S, Kreuger S.** The etiology of diagnostic errors: a controlled trial of system 1 vs. system 2 reasoning. Acad Med. 2014;89(2):277–284.
18. **Monteiro SD, Sherbino JD, Ilgen JS, Dore KL, Wood TJ, Young ME, Bandiera G, et al.** Disrupting diagnostic reasoning: do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents and emergency physicians? Acad Med. 2015;89(2):277–284.
19. **Ilgen JS, Bowen JL, McIntyre LA, Banh KV, Barnes D, Coates WC, Druck J, Fix ML, Rimple D, Yarris LM, Eva KW.** The impact of instruction to use first impressions or directed search on candidate diagnostic performance and the utility of vignette-based assessment. Acad Med. 2013;88:535–541.
20. **Mamede S, Schmidt HG, Rikers RMJP, Penaforte JC, Coelho-Filho JM.** Influence of perceived difficulty of cases on physicians' diagnostic reasoning. Influence of perceived difficulty of cases on physicians' diagnostic reasoning. Acad Med. 2008;83(12):1210–1216.
21. **Mamede S, Schmidt H, Penaforte JC.** Effects of reflective practice on the accuracy of medical diagnosis. Med Educ. 2008;42:468–475.
22. **Mamede S, van Gog T, van den Berge K, Rikers RM, van Saase JL, van Guldener C, Schmidt HG.** Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. JAMA. 2010;304(11):1198–1203.
23. **Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DR.** Relating faults in diagnostic reasoning with diagnostic errors and patient harm. Acad Med. 2012;87(2):149–156.
24. **Eva KW, Cunnington JP, Reiter HI, Keane DR, Norman GR.** How can I know what I don't know? Poor self assessment in a well-defined domain. Adv Health Sci Educ. 2004;9(3):211–224.
25. **La Rochelle JS, Durning SJ, Pangaro LN, Artino AR, van der Vleuten CP, Schuwirth L.** Authenticity of instruction and student performance: a prospective randomised trial. Med Educ. 2011;45(8):807–817.
26. **Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling, et al.** Do physicians know when their diagnoses are correct? J Gen Intern Med. 2005;20(4):334–339.