Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation

Matt Sibbald[1,2] · Jonathan Sherbino[2] · Jonathan S. Ilgen[3] · Laura Zwaan[4] · Sarah Blissett[5] · Sandra Monteiro[1,2,6] · Geoffrey Norman[2,6]

Abstract

There is an ongoing debate regarding the cause of diagnostic errors. One view is that errors result from unconscious application of cognitive heuristics; the alternative is that errors are a consequence of knowledge deficits. The objective of this study was to compare the effectiveness of checklists that (a) identify and address cognitive biases or (b) promote knowledge retrieval, as a means to reduce errors in ECG interpretation. Novice postgraduate year (PGY) 1 emergency medicine and internal medicine residents (n=40) and experienced cardiology fellows (PGY 4–6) (n=21) were randomly allocated to three conditions: a debiasing checklist, a content (knowledge) checklist, or control (no checklist) to be used while interpreting 20 ECGs. Half of the ECGs were deliberately engineered to predispose to bias. Diagnostic performance under either checklist intervention was not significantly better than the control. As expected, more errors occurred when cases were designed to induce bias (F=96.9, $p<0.0001$). There was no significant interaction between the instructional condition and level of learner. Checklists attempting to help learners identify cognitive bias or mobilize domain-specific knowledge did not have an overall effect in reducing diagnostic errors in ECG interpretation, although they may help novices. Even when cognitive biases are deliberately inserted in cases, cognitive debiasing checklists did not improve participants' performance.

✉ Matt Sibbald
  Sibbald@mcmaster.ca

1   Department of Medicine, Centre for Simulation Based Learning, McMaster University, 1200 Main St W, Hamilton, ON, Canada

2   McMaster Faculty of Health Sciences Education Research, Innovation and Theory Program, McMaster University, Hamilton, Canada

3   Department of Emergency Medicine and Center for Leadership and Innovation in Medical Education, University of Washington, Seattle, USA

4   Institute of Medical Education Research Rotterdam, Erasmus MC, Rotterdam, The Netherlands

5   University of California, San Francisco, USA

6   Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Canada

Introduction

The past decade has seen increasing interest in instructional or system-based interventions to reduce diagnostic errors. As reviewed by Graber, a number of strategies have been attempted, with varying success (Graber et al. 2012). One of the earliest strategies used technology to ameliorate errors using computer-based "decision support systems." These systems were based on an "expert-system" framework that employs explicit rules extracted from experts. Such systems were cumbersome, and not particularly accurate, listing as many as 30–40 diagnoses and with overall accuracy that rarely exceeded 40% (Elstein et al. 1996). Not surprisingly, large scale trials (Berner et al. 1999; Friedman et al. 1999) showed these types of decision-support interventions had relatively small impact on errors. Integration of artificial intelligence to improve the usability and functionality of these systems is appealing, and may allow machine-assisted diagnostic accuracy which surpasses that of clinicians (Haenssle et al. 2018), however even the most robust technology still requires human oversight (Obermeyer and Emanuel 2016).

A second popular strategy was based on the assumption that most errors derive from cognitive biases suggesting that interventions designed to educate clinicians about biases should reduce errors (Croskerry 2003a, b, Croskerry et al. 2013a, b). The underlying assumption have been challenged (Norman and Eva 2009), some studies have shown that while students can learn to identify biases (Reilly et al. 2013), such knowledge does not reduce the rate of errors (Sherbino et al. 2011, 2014).

A third strategy incorporates knowledge mobilization techniques, following a theory that errors are a consequence of knowledge deficits (Kahneman and Egan 2011). Zwaan et al. (2017) conducted a retrospective chart review of successive cases of COPD—both with, and without, errors—and found that insufficient knowledge was the primary cause of "suboptimal clinical acts". Strategies to mobilize knowledge have been investigated by Mamede and Schmidt (2008, 2014) who showed that a structured "reflection" intervention designed to encourage clinicians to systematically explore their knowledge of relationships between features and diagnoses, demonstrated small, but reasonably consistent effects in reducing diagnostic error. However the process is very resource intensive, is ineffective when the clinician selects the cases to review (Monteiro et al. 2015) and is only effective when the original case is available in written form for review (Schmidt, personal communication).

One strategy that may improve diagnostic accuracy is the use of checklists. Checklists have been proposed as safety interventions in many domains of medicine, with improved clinical outcomes related to central line infections (Marsteller et al. 2012) and peri-operative communication (Haynes et al. 2009). Ely et al. (2011) have advocated for the use of checklists to reduce diagnostic errors. They point out that such checklists can be designed to achieve different purposes. They describe three kinds of checklists, directing the clinician to:

a) Identify cognitive biases using general questions like "Could I be anchoring on the wrong diagnosis? (Availability, anchoring)", "Have I selectively elicited data to confirm my diagnosis?" "(confirmation bias), "Did I consider the flaws of heuristic thinking?" (all biases)
b) Consider alternative diagnoses within the domain
c) Critically examine the process of data gathering/interpretation.

Each of these types of checklists interacts with a different component of the diagnostic reasoning process (see Fig. 1), which is most commonly typified as an interplay between automatic pattern recognizing system 1 processes and more cognitively intensive system 2 processes which are variably involved in data gathering and interpretation around the illness presentation, arriving at a diagnosis and ratifying it through a calibration step (Croskerry 2009).

While there is no evidence showing that checklists to identify cognitive bias can reduce errors, Sibbald et al. (2013, 2014) demonstrated that the checklists directed at interpretation of ECG components can result in a small but consistent reduction in errors at all levels, although the benefit was larger for novices (Sibbald et al. 2014). Critically, only one study has examined a differential diagnosis checklist in a head-to head comparison. Shimizu et al. (2013) compared a differential diagnosis checklist to a cognitive bias checklist and found that the former reduced errors (statistically significant) but the latter did not. It is possible that the differential diagnosis checklist acted as a retrieval aid or knowledge mobilization tool. However, the differences were very small and inconsistent. In one group the difference emerged on easy cases but not difficult; in a second group the reverse was true. Critically, the study involved medical students and written clinical cases, and may not generalize to more experienced clinicians. The current study re-examines this comparison in a visual diagnostic task with clinicians of different levels of expertise.

To understand the potential impact of checklists to improve diagnostic performance requires a direct comparison of error reduction strategies based on identification of
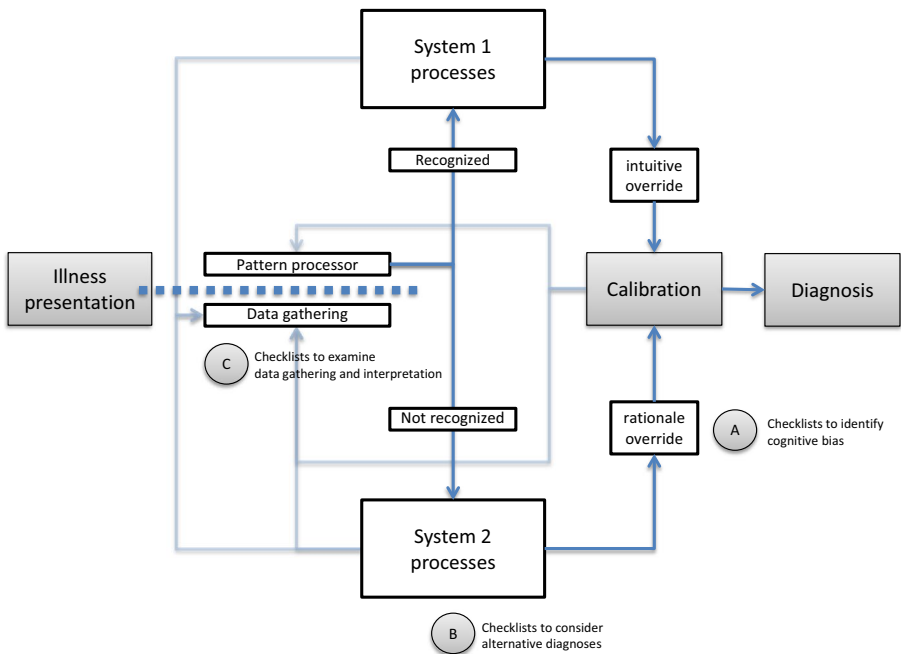
Model adapted from Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic medicine*, *84*(8), 1022-1028.

Fig. 1 3 different types of checklists in a model of diagnostic decision making

cognitive biases with alternative strategies based on mobilizing condition-specific knowledge and no intervention.

Research goal

The primary goal of the current study is to examine the whether or not use of checklists could reduce diagnostic error. We trial two different types of checklists; the first focused on identifying and addressing cognitive bias (debiasing checklists) and the second directed towards a systematic review of the ECG features (content checklists). To gauge the impact of these checklist-based interventions on reducing diagnostic error in ECG interpretation, we compared performance on these conditions to a no-intervention control.

As a secondary goal, we examined the relative effectiveness of these checklist interventions in novice (postgraduate year 1) and experienced trainees (postgraduate year 4 or 5). We also examined the relative effectiveness of these strategies in cases that are specifically designed to contain cognitive biases compared to cases that were not. The rationale for this design feature is that most of the experimental literature that implicates cognitive bias as a cause of diagnostic error uses written cases or experimental strategies that are designed to induce bias (Christensen et al. 1991; Hatala et al. 1999; Mamede et al. 2010). We anticipated that experts would be less susceptible to bias, and therefore benefit less from either intervention. We also anticipated that participants using a checklist focused on identifying cognitive bias would have lower error rates when interpreting content engineered to induce biases.

Methods

Study design

The study was designed to test the relative effectiveness of three educational interventions designed to reduce errors in ECG interpretations described in Table 1, following the method described by Sibbald et al. (2013). The conditions were as follows:

1. A 'control' condition containing only general instructions to review their interpretation carefully for diagnostic errors.
2. A 'debiasing' condition involving instruction on identifying common biases coupled with a checklist that prompts learners to identify any cognitive biases that might lead to errors in interpretation, based on a checklist used in prior study (Shimizu et al. 2013).
3. A 'content' checklist condition involving instruction on how to systematically interpret an ECG, then a content-specific checklist drawing attention to specific features of the ECG, based on a checklist used in prior study (Sibbald et al. 2013, 2014).

Study flow

The study flow consisted of four steps (see Fig. 1). Following informed consent, all participants watched a review session on ECG diagnosis with worked examples and embedded instruction on the underlying concepts of ECG reading. All participants received the same instructions. The main goal of this component was to ensure that all participants

Table 1 Checklists for each condition

Condition	Control condition	Debiasing condition (based on Shimizu et al. 2013)	Content checklist (based on Sibbald et al. 2013)
	Please check your interpretation carefully and amend it as required	Please check your interpretation carefully considering each of the following: 1. Was I comprehensive? 2. Did I consider the inherent flaws of heuristic thinking? 3. Was my judgment affected by any cognitive bias? 4. Were any of the following biases present (anchoring, availability, confirmation, search satisficing, framing)? 5. What is the worst case scenario?	Please check your interpretation carefully considering each of the following: 1. Rate 2. Rhythm 3. Axis 4. Hypertrophy 5. Ischemia 6. Intervals

have, to some extent, a common working knowledge of principles of ECG interpretation. The instructional material used was from a popular online ECG teaching site (ECG Made Simple 2017). Next, participants watched a slide show on the prevalence and importance of diagnostic errors, their potential origins in system or cognitive processes, reasons why clinical decision-making might be faulty as well as the value of checking diagnostic decisions.

Participants then received instruction customized to their random group assignment, as follows:

1. In the control condition, participants were given general instructions to review their diagnoses carefully.
2. For the debiasing condition, participants were given instructions in identifying specific cognitive biases adapted from the online materials recommended by the Society to Improve Diagnosis in Medicine. We provided definitions and examples for anchoring bias, availability bias, confirmation bias, framing effect and search satisficing. These biases are all identified as important in the Institute of Medicine report (Donaldson et al. 2000), have been described in detail previously and have some evidential support (Blumenthal-Barby and Krieger 2015).
3. For the content checklist condition, participants received instruction on use of the content checklist (Table 1) which drew attention to various aspects of the ECG, including rate, rhythm, axis, hypertrophy, ischemia, and intervals.

Finally, all groups practiced interpreting ECGs using the same set of 6 ECGs with their condition-specific checklist (or nothing else for control condition). Table 1 includes the checklists. All practice ECGs contained clinical features that could induce a cognitive bias as described below. The correct answer was provided for each ECG.

Instruction and testing was completed on a secure web-based testing platform (LimeSurvey, Hamburg, Germany).

Testing

During the testing phase, all participants were given the same 20 ECGs, each consisting of a brief clinical history and an ECG. They were instructed to provide a written interpretation in a free text box for each ECG, and then advanced to a separate screen and asked to check their interpretation using with their condition-specific instruction. They were allowed to revise their interpretation in a free text box.

The 20 ECGs were selected from an online case bank (www.ecgmadesimple.com, Table 2) and presented in random order. Ten of these ECGs included case stems that were designed to suggest 2 of the 6 cognitive biases described in the debiasing checklist. The two biases were selected for their ease of adoption in the ECG format:

1. *Search satisficing*, which occurs when search for a diagnosis stops after one cause or explanation is found, even though multiple problems are present. The ECG cases designed to promote search satisficing had more than one important diagnosis (e.g., rapid atrial fibrillation AND inferior ST elevation myocardial infarction).
2. *Confirmation Bias* describes the tendency to look for confirming data to support a diagnosis rather than look for disconfirming evidence to refute it, although the latter may be more persuasive and definitive. The ECG cases designed to promote confirmation bias

Table 2 ECG testing material

ECG #	Description	Bias	Answer
1	45 year old presenting for routine physical	None	Sinus rhythm, left ventricular hypertrophy
2	58 year old presents with ankle pain in the emergency department	None	Sinus rhythm with Mobitz 1 s degree heart block
3	59 year old with palpitations	None	Supraventricular tachycardia
4	86 year old woman transferred from a nursing home with severe diarrhea post- amoxicillin therapy.	None	Ventricular tachycardia
5	38 year old man admitted to hospital with flank pain	None	Hyperkalemia
6	73 year old retired lawyer with hypertension	None	Sinus, Right bundle branch block
7	44 year old scheduled to have rectal tumor removed	None	Sinus, left anterior fasicular block
8	48 year old with severe chest pain	None	Sinus arrhythmia, acute inferior ST elevation MI
9	56 year old with dyspnea and palpitations	None	Atrial fibrillation with rapid ventricular response
10	68 year old witnessed collapse in the emergency room	None	Ventricular fibrillation
11	39 year old man with chest pain and diaphoresis	Satisficing	sinus rhythm, right axis deviation, S1Q3T3, anterior T inversions
12	78 year old complaining of tiredness and fatigue	Satisficing	2:1 Atrial flutter, left ventricular hypertrophy
13	56 year old presenting with chest pain	Satisficing	Sinus rhythm with second degree Mobitz! Heart block, inferior ST elevation MI
14	39 year old 1 day after resection of a myxoma	Satisficing	Complete heart block, Pericarditis
15	75 year old for follow up post bypass cardiology clinic	Satisficing	Atrial fibrillation, junctional escape with complete heart block, setpal MI
16	84 year old with irregular heart rate, triage nurse concerned about atrial fibrillation	Confirmation	Sinus rhythm with atrial ectopy, left ventricular hypertrophy
17	47 year old man with pleuritic pain, concerned about recurrent pericarditis	Confirmation	Sinus rhythm with anterior ST elevation MI
18	23 year old referred for abnormal ECG. Query left bundle branch block	Confirmation	Sinus rhythm with pre-excitation
19	59 year old after mitral valve repair. Rule out ventricular tachycardia	Confirmation	Ventricular pacing
20	77 year old with COPD. Respirologist concerned about right ventricular hypertrophy	Confirmation	Sinus rhythm with right bundle branch block

included a plausible but incorrect diagnosis in the stem (e.g., "rule out atrial fibrillation" when the ECG was sinus rhythm with multiple premature atrial contractions).

ECGs were selected by MS, reviewed by JS, SM and GN to identify their fit to suggest the 2 cognitive biases, and pilot tested with two practicing cardiologists. Answers were adopted from the online case bank that publishes each ECG's interpretation and a justification for those interpretations.

Participants

In order to ensure a baseline working knowledge of ECG interpretation, we enrolled novice emergency medicine and internal medicine residents in their first year of postgraduate training. Experienced participants were cardiology fellows or residents in their fourth, fifth or sixth year of postgraduate training. Participants were recruited from Canada (McMaster University, Western University), United States (University of Washington), and the Netherlands (Rotterdam University). Participants were paid a small honorarium for participating. We planned a sample size of at least 20 novice and 20 expert learners, giving a power of 0.80 to detect a 10% difference between all three groups, accepting a 5% type I error rate (http://powerandsamplesize.com).

Scoring

Consistent with the approach of Sibbald et al. (2013), errors in each ECG interpretation before and after the condition-specific instruction were counted. Errors included missing correct diagnoses, as well as additionally provided incorrect diagnoses. Interpretation time and the additional time taken following the condition specific instruction were also collected as secondary endpoints.

Analysis

Data were analyzed using repeated measures ANOVA, with condition (control, debiasing checklist and content checklist) and learner level (novice, experienced) as between-subject factors, and case (20 levels) as a repeated measure. Initial analysis looked at the effect of expertise on accuracy and time to diagnosis, as a form of validation.

The primary analysis focused on the number of errors following revision after the use of the two checklist and control conditions. Secondary analysis examined how often the diagnosis was revised under different instructional conditions. To determine whether the effect of instructional condition differed by educational level, we examined the interaction between instruction condition and level.

We did not look at initial diagnoses prior to application of the checklist intervention for each case, since it may be influenced to an unknown degree by the intervention based on recall of the checklist for prior cases.

Finally, we examined performance on specific cases to determine whether the debiasing instruction was more effective than the content checklist or control condition for those cases that included a specific cognitive bias. All analyses were conducted with SPSS version 23 (IBM, Redmond United States).

Table 3 Participant demographics

	Control condition	Debiasing checklist	Content checklist	Total
Level				
Novice learners	11	14	15	40
Expert learners	8	7	6	21
Gender				
Female	5	8	6	19
Male	14	13	15	42
Age				
25–29 years	8	11	10	29
30–34 years	10	6	8	24
35–39 years	1	4	3	8
Nationality				
Canada	4	4	3	11
United States	6	6	6	18
Netherlands	11	11	10	32

Table 4 Errors in ECG interpretation by condition, learner level and content

Instructional condition		Control condition		Debiasing checklist condition		Content checklist condition	
Learner level	Case type	Mean	SD	Mean	SD	Mean	SD
Novice	No bias	1.13	1.09	.87	1.01	1.04	1.08
	Confirmation	1.84	1.41	1.75	1.54	1.34	1.31
	Satisficing	2.60	1.57	2.25	1.48	2.36	1.33
Expert	No bias	.30	.68	.32	.75	.36	.70
	Confirmation	.40	.55	.57	1.22	.60	1.01
	Satisficing	1.23	1.12	1.20	1.47	1.29	1.45

Results

Sixty-one participants were recruited, 40 novice learners and 21 experienced learners. Participant demographics are reported in Table 3.

Effect of educational level

Overall, there were 1.19 ± 1.34 errors per ECG. Experienced learners made fewer errors than novice learners (0.60 vs.. 1.52, F=66.0, $p < 0.0001$). There were fewer errors in the two checklist conditions compared to controls (Table 4), although the differences were small and not significant (F=0.61, $p = 0.55$).

Although novice residents appeared to take longer to enter their preliminary diagnosis (276 s vs. 164 s) these time differences were not significantly different (F=2.14, $p = .15$).

In examining revision after use of the checklist, juniors took longer (31 s vs. 11 s), which was significant (F = 4.60, p = .04).

The substantial difference between the novice and experienced groups constitutes some evidence of the appropriateness of the materials and task.

Effect of instructional condition on errors and time

The error rates for junior and senior learners by condition are shown in Fig. 2. The control condition had slightly more errors among juniors, but the main effect of condition was not significant (F = 0.26, p = .77). and the interaction between condition and trainee experience was not significant (F = 0.82, p = .44). Participants working under the content checklist condition took longer, on average, than the debias and control condition (checklist 276 s, debias 225 s, control 223 s) but these differences were not significant (F = 0.206, p = .81).

The process of revision

Of the 1220 interpretations made on the ECGs, there were only 97 instances where participants changed their original interpretation in the context of their 'second look' (e.g. interpreting the ECG in the context of one the two checklists, or no additional information in the control condition). The mean number of errors for these changed interpretations was 1.87 before the intervention and 1.86 after (F = .000, p = .997). Novice trainees made 1.92 errors before and 2.06 after the intervention while more experienced trainees made 1.79 before and 1.65 after; the interaction between condition and experience on change in accuracy was not significant (F = 0.64, p = .42).

Specific effect of debiasing intervention on cases with cognitive biases

As shown in Table 4, cases engineered to induce specific biases resulted in significantly more errors for both juniors and seniors (0.67 overall for non-bias-inducing cases compared
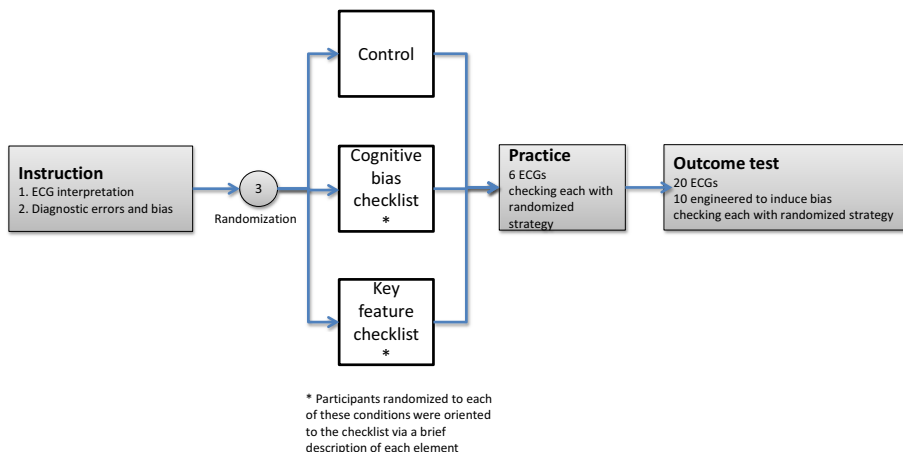

Fig. 2 Study process

to 1.08 and 1.82 for cases with confirmation bias or search satisficing bias, respectively (F = 96.9, $p < 0.0001$). There was no evidence that specific debiasing instructions resulted in fewer errors on the biased cases, for either novice or experienced residents (F = 0.25, $p = .78$). There was, however, a significant interaction between the case type and learner level with novice participants being more susceptible to confirmation bias (F = 5.59, $p = 0.001$).

Discussion

This study systematically compared the strategy of debiasing checklists with knowledge retrieval checklists to reduce errors in ECG interpretation. While we found significant effects of expertise on diagnostic accuracy, we did not demonstrate any advantage for checklists, either those directed at specific aspects of the ECG or those designed to identify cognitive biases. Moreover, even when cases contained specific cognitive biases related to the instruction and checklist, there was no increase in accuracy related to use of the debiasing checklist.

One obvious conclusion is that is that the checklists and instruction were inadequate. However, we specifically attempted to use "state of the art" materials (Sibbald et al. 2013, 2014; Society to Improve Diagnosis in Medicine). The testing content was taken from a well-rated ECG education site, and carefully engineered to promote bias, a manipulation that successfully resulted in a higher error rate both among novice and expert learners. Learners were taken from multiple institutions in multiple countries to minimize systematic influences of individual training contexts.

Despite all this careful effort, little impact on diagnostic errors was seen using either of these strategies. Certainly the systematic application of a checklist after the initial diagnosis resulted in relatively few changes, and no clear improvement. To be fair, novice learners experienced a small reduction in errors in both the debiasing and content checklist conditions, however, the effect size was small and only present in a post hoc analysis of the novice group, whereas the interaction between expertise and condition was not significant. It is not clear whether this observation reflects the specific content of these instructional conditions or is a more general phenomenon of increased instructional scaffolding translating into better novice performance.

The results contrast with an earlier study of ECG diagnosis using the same content checklist (Sibbald et al. 2013). It may be that the present study was relatively underpowered; however, one clear distinction was that the Sibbald et al. (2013) study used more complex cases, with almost three times as many errors per case as the present study. It also showed a significant number of error corrections, particularly among novices.

While the low frequency of changed diagnosis and a resulting minimal effect on accuracy as a consequence of a systematic intervention after an initial diagnosis is a negative result from the present study, it is not a finding unique to this study. Monteiro, using written clinical cases, found a similar 8% change in diagnoses after revision, with similarly small overall improvement (Monteiro et al. 2015).

This may in turn suggest to some that using decision supports after an initial diagnosis may be futile, and students should be taught to be systematic and thorough from the outset, using checklists similar to those in this study to guide systematic search and reduce premature closure and confirmation bias. Surprisingly a direct test of this approach

of "hypothesis-free" systematic inquiry using feature checklists and ECG cases led to the opposite conclusion; such approaches actually increased diagnostic errors (Norman et al. 1999).

Importantly, teaching about bias and providing checklists to learners to identify biases did not mitigate the increase in errors in testing content engineered to promote bias. This can be viewed as a 'best case' scenario, to the extent that the cases clearly exhibited the defined cognitive bias and participants were actively reminded of the biases during the diagnostic process. However, the study by Zwaan et al. (2017) raises concern that even those with substantial experience with cognitive biases have difficulty identifying and agreeing on which bias is present. Moreover, the process of identification of a bias does not guarantee error correction unless the clinician is aware of the correct alternative (Dhaliwal 2017).

This study has some important limitations. We chose to use ECG diagnosis as an example of visual test interpretation. While visual test interpretation is frequently used for studies on diagnostic error, contexts in which data are not automatically provided, but must be collected by a healthcare professional may be more prone to bias. Second, this study included learners at the beginning and near the end of their postgraduate training. We cannot comment on more novice learners (e.g. medical students) or more expert clinicians (e.g. practicing cardiologists). However, given the range of diagnostic performance captured in the two groups in this study, and the minimal impact of both interventions, it is unlikely larger effects will be seen in other groups. Finally, our intervention represents a relatively blunt checklist instrument to tackle a nuanced problem of bias infiltrating diagnostic error. Our findings do not preclude the efficacy of more selective application of checklists for more targeted use.

What are the implications of this study? The effects of these interventions are disappointingly small. Educators might reduce curriculum time in these types of interventions until further work shows larger impact, either through empirically refined tools or more targeted application of these ideas. It is difficult to justify taking ECG instructional time to cover the interventions used in this study given their minimal effect. It is becoming increasingly clear that interventions that rely on encouraging clinicians to "think harder", however that is defined e.g. "take more time, slow down" (Sherbino et al. 2012); "be aware of biases" (Sherbino et al. 2014); "pay attention to specific aspects" (Sibbald et al. 2013, 2014) or "reflect" (Mamede and Schmidt 2014; Mamede et al. 2008, 2010), but which rely entirely on existing knowledge have little or no impact on errors (Dhaliwal 2017). Certainly, one recurrent issue is that clinicians are unaware that they are committing an error, so are unlikely to take efforts to correct it (Eva and Regehr 2011). While such errors may appear obvious in hindsight, this retrospective process has its own problems (Zwaan et al. 2017).

It is difficult to escape the conclusion that, however serious the problem of diagnostic errors, there are unlikely to be any 'quick fixes' to improve diagnostic performance based on general, real time reminders that clinicians broadly apply in practice.

Compliance with ethical standards

References

Berner, E. S., Maisiak, R. S., Cobbs, C. G., & Taunton, O. D. (1999). Effects of a decision support system on physicians' diagnostic performance. *Journal of the American Medical Informatics Association, 6*(5), 420–427. https://doi.org/10.1136/jamia.1999.0060420.

Blumenthal-Barby, J. S., & Krieger, H. (2015). Cognitive biases and heuristics in medical decision making. *Medical Decision Making, 35*(4), 539–557. https://doi.org/10.1177/0272989X14547740.

Christensen, C., Heckerling, P. S., Mackesy, M. E., Bernstein, L. M., & Elstein, A. S. (1991). Framing bias among expert and novice physicians. *Academic Medicine, 66*(9), S76–S78. https://doi.org/10.1097/00001888-199109000-00047.

Croskerry, P. (2003a). Cognitive forcing strategies in clinical decisionmaking. *Annals of Emergency Medicine, 41*(1), 110–120. https://doi.org/10.1067/MEM.2003.22.

Croskerry, P. (2003b). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*(8). Retrieved from https://journals.lww.com/academicmedicine/Fulltext/2003/08000/The_Importance_of_Cognitive_Errors_in_Diagnosis.3.aspx

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703.

Croskerry, P., Singhal, G., & Mamede, S. (2013a). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*, *22*(Suppl 2), ii58 LP–ii64. Retrieved from http://qualitysafety.bmj.com/content/22/Suppl_2/ii58.abstract

Croskerry, P., Singhal, G., & Mamede, S. (2013b). Cognitive debiasing 2: Impediments to and strategies for change. *BMJ Quality & Safety*, *22*(Suppl 2), ii65 LP–ii72. Retrieved from http://qualitysafety.bmj.com/content/22/Suppl_2/ii65.abstract

Dhaliwal, G. (2017). Premature closure? Not so fast. *BMJ Quality & Safety, 26*(2), 87–89. https://doi.org/10.1136/bmjqs-2016-005267.

Donaldson, M. S., Corrigan, J. M., & Kohn, L. T. (2000). *To err is human: Building a safer health system*. Washington: National Academies Press.

ECG Made Simple. (2017). *Stepwise videos.* Retrieved from: http://www.ecgmadesimple.com/stepwise-vids/story_html5.html. Accessed 16 Sept 2017.

Elstein, A. S., Friedman, C. P., Wolf, F. M., Murphy, G., Miller, J., Fine, P., et al. (1996). Effects of a decision support system on the diagnostic accuracy of users: A preliminary report. *Journal of the American Medical Informatics Association, 3*(6), 422–428. https://doi.org/10.1136/jamia.1996.97084515.

Ely, J. W., Graber, M. L., & Croskerry, P. (2011). Checklists to reduce diagnostic errors. *Academic Medicine, 86*(3), 307–313. https://doi.org/10.1097/ACM.0b013e31820824cd.

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education, 16*(3), 311–329. https://doi.org/10.1007/s10459-010-9263-2.

Friedman, C. P., Elstein, A. S., Wolf, F. M., Murphy, G. C., Franz, T. M., Heckerling, P. S., et al. (1999). Enhancement of clinicians' diagnostic reasoning by computer-based consultation. *JAMA, 282*(19), 1851. https://doi.org/10.1001/jama.282.19.1851.

Graber, M. L., Kissam, S., Payne, V. L., Meyer, A. N. D., Sorensen, A., Lenfestey, N., et al. (2012). Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ Quality & Safety*, *21*(7), 535 LP–557. Retrieved from http://qualitysafety.bmj.com/content/21/7/535.abstract.

Haenssle, H. A., et al. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology, 29*(8), 1836–1842.

Hatala, R., Norman, G. R., & Brooks, L. R. (1999). Impact of a clinical scenario on accuracy of electrocardiogram interpretation. *Journal of General Internal Medicine, 14*(2), 126–129. https://doi.org/10.1046/j.1525-1497.1999.00298.x.

Haynes, A. B., Weiser, T. G., Berry, W. R., Lipsitz, S. R., Breizat, A.-H. S., Dellinger, E. P., et al. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine, 360*(5), 491–499. https://doi.org/10.1056/NEJMsa0810119.

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.

Mamede, S., & Schmidt, H. G. (2014). Reflection in diagnostic reasoning. *Academic Medicine, 89*(7), 959–960. https://doi.org/10.1097/ACM.0000000000000306.

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education, 42*(5), 468–475. https://doi.org/10.1111/j.1365-2923.2008.03030.x.

Mamede, S., van Gog, T., van den Berge, K., Rikers, R. M. J. P., van Saase, J. L. C. M., van Guldener, C., et al. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA, 304*(11), 1198. https://doi.org/10.1001/jama.2010.1276.

Marsteller, J. A., Sexton, J. B., Hsu, Y.-J., Hsiao, C.-J., Holzmueller, C. G., Pronovost, P. J., et al. (2012). A multicenter, phased, cluster-randomized controlled trial to reduce central line-associated bloodstream infections in intensive care units*. *Critical Care Medicine, 40*(11), 2933–2939. https://doi.org/10.1097/CCM.0b013e31825fd4d8.

Monteiro, S. D., Sherbino, J., Patel, A., Mazzetti, I., Norman, G. R., & Howey, E. (2015). Reflecting on diagnostic errors: Taking a second look is not enough. *Journal of General Internal Medicine, 30*(9), 1270–1274. https://doi.org/10.1007/s11606-015-3369-4.

Norman, G. R., Brooks, L. R., Colle, C. L., & Hatala, R. M. (1999). The benefit of diagnostic hypotheses in clinical reasoning: Experimental study of an instructional intervention for forward and backward reasoning. *Cognition and Instruction, 17*(4), 433–448. https://doi.org/10.1207/S1532690XCI1704_3.

Norman, G. R., & Eva, K. W. (2009). Diagnostic error and clinical reasoning. *Medical Education, 44*(1), 94–100. https://doi.org/10.1111/j.1365-2923.2009.03507.x.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *The New England Journal of Medicine, 375*(13), 1216.

Reilly, J. B., Ogdie, A. R., Von Feldt, J. M., & Myers, J. S. (2013). Teaching about how doctors think: A longitudinal curriculum in cognitive bias and diagnostic error for residents. *BMJ Quality & Safety*, *22*(12), 1044 LP–1050. Retrieved from http://qualitysafety.bmj.com/content/22/12/1044.abstract

Sherbino, J., Dore, K. L., Siu, E., & Norman, G. R. (2011). The effectiveness of cognitive forcing strategies to decrease diagnostic error: An exploratory study. *Teaching and Learning in Medicine, 23*(1), 78–84. https://doi.org/10.1080/10401334.2011.536897.

Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Kreuger, S., et al. (2012). The relationship between response time and diagnostic accuracy. *Academic Medicine, 87*(6), 785–791. https://doi.org/10.1097/ACM.0b013e318253acbd.

Sherbino, J., Kulasegaram, K., Howey, E., & Norman, G. (2014). Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: A controlled trial. *CJEM, 16*(01), 34–40. https://doi.org/10.2310/8000.2013.130860.

Shimizu, T., Matsumoto, K., & Tokuda, Y. (2013). Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. *Medical Teacher, 35*(6), e1218–e1229. https://doi.org/10.3109/0142159X.2012.742493.

Sibbald, M., de Bruin, A. B. H., & van Merrienboer, J. J. G. (2013). Checklists improve experts' diagnostic decisions. *Medical Education, 47*(3), 301–308. https://doi.org/10.1111/medu.12080.

Sibbald, M., De Bruin, A. B. H., & van Merrienboer, J. J. G. (2014). Finding and fixing mistakes: Do checklists work for clinicians with different levels of experience? *Advances in Health Sciences Education, 19*(1), 43–51. https://doi.org/10.1007/s10459-013-9459-3.

Zwaan, L., Monteiro, S., Sherbino, J., Ilgen, J., Howey, B., & Norman, G. (2017). Is bias in the eye of the beholder? A vignette study to assess recognition of cognitive biases in clinical case workups. *BMJ Quality & Safety*, *26*(2), 104 LP–110. Retrieved from http://qualitysafety.bmj.com/content/26/2/104.abstract.