Is bias in the eye of the beholder? A vignette study to assess recognition of cognitive biases in clinical case workups

Laura Zwaan,[1,2] Sandra Monteiro,[3] Jonathan Sherbino,[4] Jonathan Ilgen,[5] Betty Howey,[6] Geoffrey Norman[3]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2015-005014).

For numbered affiliations see end of article.

Correspondence to
Dr Laura Zwaan, Institute of Medical Education Research Rotterdam, Erasmus MC, Wytemaweg 80, Rotterdam, 3015 CN, The Netherlands; l.zwaan@erasmusmc.nl

Received 5 November 2015
Revised 8 January 2016
Accepted 11 January 2016
Published Online First 29 January 2016

▸ http://dx.doi.org/10.1136/bmjqs-2016-005267

To cite: Zwaan L, Monteiro S, Sherbino J, *et al*. *BMJ Qual Saf* 2017;**26**:104–110.

ABSTRACT

Background Many authors have implicated cognitive biases as a primary cause of diagnostic error. If this is so, then physicians already familiar with common cognitive biases should consistently identify biases present in a clinical workup. The aim of this paper is to determine whether physicians agree on the presence or absence of particular biases in a clinical case workup and how case outcome knowledge affects bias identification.

Methods We conducted a web survey of 37 physicians. Each participant read eight cases and listed which biases were present from a list provided. In half the cases the outcome implied a correct diagnosis; in the other half, it implied an incorrect diagnosis. We compared the number of biases identified when the outcome implied a correct or incorrect primary diagnosis. Additionally, the agreement among participants about presence or absence of specific biases was assessed.

Results When the case outcome implied a correct diagnosis, an average of 1.75 cognitive biases were reported; when incorrect, 3.45 biases ($F=71.3$, $p<0.00001$). Individual biases were reported from 73% to 125% more often when an incorrect diagnosis was implied. There was no agreement on presence or absence of individual biases, with κ ranging from 0.000 to 0.044.

Interpretation Individual physicians are unable to agree on the presence or absence of individual cognitive biases. Their judgements are heavily influenced by hindsight bias; when the outcome implies a diagnostic error, twice as many biases are identified. The results present challenges for current error reduction strategies based on identification of cognitive biases.

INTRODUCTION

Diagnostic errors have serious consequences for the health of a nation. The Institute of Medicine (IoM) in the USA recently released a comprehensive report, 'Improving Diagnosis in Health Care',[1] warning that 'most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences'. The report makes several recommendations to reduce the prevalence of diagnostic error, covering many aspects of healthcare, from research funding to legal liability to patient advocacy.

Central to the report is the recognition that 'understanding the clinical reasoning process and the factors that can impact it are important to improving diagnosis'.[2] One chapter is devoted to summarising research on the psychology of clinical reasoning, described by a 'dual process' theory[2 3] in which two very different cognitive processes are at play: system 1, which is rapid, subconscious and relies heavily on cognitive shortcuts or 'heuristics' which may lead to a bias, and system 2, which is slow, conscious and analytical. While errors may arise in both systems, it is presumed that most errors arise from cognitive biases.

The notion that cognitive biases are a primary cause of diagnostic error has been prevalent in the medical literature for over 30 years.[2 4–10] The approach taken by this diversity of authors is remarkably consistent: they begin with a description of a selection of the heuristics and biases originally identified in the Tversky and Kahneman[11] research pro-

gramme in the 1970s, and then provide an illustrative example of how this might arise in clinical practice and what one might do to avoid it.

While these papers make a formidable argument that the biases described in the literature *might* cause a diagnostic error, empirical evidence that any of these biases *actually* causes diagnostic errors is sparse. A recent systematic review of 213 studies[12] cited evidence for 18 cognitive biases in medicine, yet upon our careful inspection of this review, only 15 of the 213 studies (7%) examined linkages between biases and diagnostic error by health professionals, and these studies examined only 7 biases.

Two broad strategies have been used to identify and study biases, each with benefits and limitations. Many experimental studies use written cases designed to intentionally invoke a bias. These methods offer the advantages of case consistency across participants, more definitive knowledge of the diagnoses being assessed (since investigators typically create these cases themselves) and more controlled manipulation of variables that may impact diagnostic performance.[13 14] Yet the ecological validity of these techniques remains a concern, as Blumenthal-Barby[12] states:

> Most of the studies on biases and heuristics in medical decision-making are based on hypothetical vignettes, raising concerns about the applicability of these findings to actual decision making.

An alternative approach is to examine the occurrence of cognitive biases in practice by reviewing instances where errors arose, then inferring the presence of cognitive biases as a causal factor in these errors.[15] This ostensibly captures 'real world' performance, though several limitations should be highlighted. First, this method typically presumes that the individuals reviewing these cases can determine what the clinician may have been thinking from the information available in the record. This charting may not reflect the depth or variety of thinking processes that actually occurred. Second, reviewers using this approach are aware that an error was committed, and may thus be vulnerable to 'hindsight bias'.[16] Finally, ascribing errors to just cognitive biases may oversimplify what is actually happening in these clinicians' minds.[17–19]

Nevertheless, if cognitive biases are indeed a primary source of diagnostic errors, then presumably errors can be reduced by ensuring that students learn to recognise the more common biases and apply 'cognitive forcing strategies'[20] to improve accuracy. However, a prerequisite for the success of these strategies is the demonstration that biases can be reliably identified in the reasoning process.

The purpose of this study is therefore to examine the reliability of bias identification among clinicians with a special interest in diagnostic error, and to determine the extent to which the identification of these biases may be influenced by knowledge of the outcome—hindsight bias.[16]

METHODS

Subjects

Recruitment and inclusion criteria

Individuals on the Society to Improve Diagnosis in Medicine (SIDM) listserv were contacted by email using a recruitment letter. This society was formed half a decade ago, with the explicit purpose of reducing the incidence of diagnostic error through research and education. The society holds annual conferences, sponsors a journal and conducts various other activities to raise awareness. The letter invited members to participate in a brief recruitment survey to determine their eligibility and willingness to participate. Members were asked for their status as a physician, medical specialty and consent to participate. Their willingness to participate implied consent to be involved in the main survey of bias. They were informed that the survey would take 20 min to complete. The study received ethics approval from the McMaster Research Ethics Board, no. 11–409.

One hundred and thirteen people responded to the recruitment email. Of those, 71 met the inclusion criteria (ie, were practicing physicians, comfortable communicating in English and consented to participate in the survey). Emails were sent based on these responses and 37 volunteers (52%) completed the survey. Two reminders were sent to all those who consented and the survey was closed after 6 weeks.

Materials

The study materials were based on 12 general medicine cases used in previous research in clinical reasoning.[21] These cases were constructed to contain ambiguity, with two equally probable diagnoses. To do this, the frequency of signs and symptoms suggestive of one diagnosis was purposefully balanced with findings that suggested the alternative diagnosis. Vignettes were presented such that patients were seen in either a clinic or an emergency department setting, with equal frequency of these two settings.

In order to examine the influence of hindsight bias, an experimental manipulation was built into the last part of the case description, so that the workup concluded with the clinician ordering a test or initiating a treatment plan specific to one of the two diagnoses, with information suggesting that the patient encounter had ended (eg, the patient was sent home with a prescription for antibiotics). The other disease suggested by the case was not investigated. Finally, each case then revealed the patient outcome, which was experimentally manipulated to reveal one of two end-points. For the 'consistent' version of the case, the patient follow-up revealed improvement of the condition (eg, evidence that the shortness of breath presumably due to pneumonia resolved with antibiotics), suggesting

that the clinician made the correct diagnosis. For the 'inconsistent' version, the details of the patient follow-up indicated that the patient had not recovered or had worsening symptoms (eg, antibiotics did not resolve the shortness of breath which was subsequently identified as due to pulmonary embolism), indicating that provisional diagnosis pursued by the clinician was incorrect. Critically, the case stem, including patient history, preliminary lab results and primary complaint, were identical up to when the patient outcome was manipulated to include 'consistent' or 'inconsistent' details. A sample case is shown in online supplementary appendix 1.

There was no attempt to ensure that a cognitive bias was present (and no attempt to exclude biases). The only guiding principle for the cases was that they were ambiguous, with features of two possible diagnoses that were approximately equiprobable. In other words, contrary to previous studies that have explicitly manipulated the case to illustrate particular biases,[13 14] the current study made no attempt to design cognitive biases into the cases.

To pretest the original collection of 12 cases, they were given to a sample of 9 emergency medicine specialists at the University of Washington, who were all full-time providers, board certified and in practice for an average of 7.0 years. They were asked to list the diagnoses they were considering, as well as estimates of likelihood. Eight of the 12 cases showed an expected '50/50' pattern with two clear primary diagnoses (table 1). The remaining four cases showed a clear primary diagnosis and were excluded from further testing.

Procedure

The cases and questions were hosted on Lime Survey (http://www.limesurvey.org). Eight counterbalanced versions of the survey were created such that participants saw a total of eight cases; four cases appeared with 'inconsistent' follow-up results and four appeared with 'consistent' follow-up results. A randomised cross-over design was employed so that an equal number of volunteers received each version of the survey.

The survey began with a review of six common cognitive biases: anchoring, availability, base rate neglect, confirmation bias, premature closure and representativeness. These biases were chosen in part because they have some empirical support, either in medicine or psychology. Definitions for the six biases are presented in online supplementary appendix 2, and were taken directly from a frequently cited review on this topic.[2] (Five of the six are also described in the IoM report.)[1]

Participants were instructed to read each case and then decide (1) if it was likely that a diagnostic error was committed, and (2) if cognitive bias played a role in the physician's decision-making process. Participants could select any of the six defined biases as well as enter any additional biases in an open field labelled 'other'. To assist participants in their task, they were provided with the definitions at the bottom of each case. There was also space for a write-in.

ANALYSIS

The primary analysis examined the influence of the case outcome (ie, consistent or inconsistent with the diagnosis initially pursued) on (a) the total number of biases identified, and (b) the response to the question 'Is it likely that a diagnostic error was committed?' Secondary analysis examined the number of biases identified following a positive or negative response about the likelihood of diagnostic error; the increase in the presence of each specific bias when the test was consistent or inconsistent with the diagnosis; and the inter-rater reliability (κ) for each bias.

Because of the incomplete crossover design, it was not possible to run a full repeated measures analysis of variance (ANOVA) including cases as a factor, so the first two questions were addressed using a two-way ANOVA with case (eight levels) and follow-up outcome (consistent or inconsistent) as the two factors. The same analysis was conducted on individual biases, where the dependent variable was simply whether a bias was present (yes/no). An additional analysis by subject was conducted on the average number of biases each participant identified in the four consistent and four inconsistent cases. The

Table 1 Average subjective probability of two primary diagnoses based on pretest with nine emergency physicians in the state of Washington

Diagnosis A	Diagnosis B	Probability (A)	Probability (B)	Probability other
Pneumonia	Pulmonary embolism	0.44	0.30	0.26
Acute MI	Type A aortic dissection	0.53	0.33	0.14
Tubo-ovarian abscess	Appendicitis	0.46	0.39	0.15
Subarachnoid haemorrhage	Meningitis	0.36	0.57	0.07
Kidney stone	Type B aortic dissection	0.49	0.38	0.13
Pyelonephritis	Abdominal aortic aneurism	0.52	0.34	0.14
Pancreatitis	Cholecystitis	0.42	0.58	0.00
Cellulitis	Deep vein thrombosis	0.56	0.43	0.01

MI, myocardial infarction.

relation between number of biases and response to the 'error' question was also conducted as a two-way ANOVA, with question response (yes/no) and case as the two factors.

Finally, the analysis of agreement used the fact that κ and the intraclass correlation are mathematically identical[22] to estimate an average κ across all raters by computing variance due to cases, raters and error, and then calculating an intraclass correlation coefficient (ICC). κ Values were classified as poor (ICC values: 0–0.2), fair (ICC values: 0.21–0.4), moderate (ICC values: 0.41–0.6) or substantial (ICC values: 0.61–0.8) based upon definitions outlined by Landis and Koch.[23]

RESULTS

As shown in table 2, the majority of participants had training in primary care, mostly internal medicine and emergency medicine. They had been in practice for anywhere from 1 to 45 years (mean 27 years), and about three-quarters of them saw patients more than 25% of their work time.

The primary analysis by case examined the number of cognitive biases identified when the follow-up outcome was consistent or inconsistent with the clinician's initial diagnostic plan. When the follow-up result was inconsistent (ie, supported the diagnosis that the clinician had *not* pursued), there were twice as many biases identified as when it was consistent (3.44 (1.53) vs 1.75 (1.80)): $F(1,280)=73.3$, $p<0.0001$. The equivalent analysis of the average number of biases by subject yielded an $F(1,36)$ of 104.3, $p<0.00001$. As shown in table 3, an excess number of biases in the presence of an inconsistent follow-up result was present for all cases.

In response to the question, 'Was it likely that a diagnostic error was present?' again there was a large difference depending on whether the follow-up result was consistent or inconsistent, with 8% of respondents saying an error was present when the outcome was consistent with the initial diagnostic plan versus 60% when it was inconsistent ($\chi^2=80.8$, $p<0.0001$). When the participant judged that a diagnostic error was present, there was a greater number of biases identified: 3.22 (1.50) vs 1.27 (1.82) ($F(1,280)=89.6$, $p<0.0001$).

We examined the prevalence of individual biases in the two conditions and the agreement among respondents regarding the presence or absence of each bias in each case, as shown in table 4. The prevalence of individual biases was 73%–125% greater when the test result was inconsistent. The greatest increase occurred with availability and premature closure.

Finally we examined agreement among raters regarding the presence or absence of individual biases on each case, expressed as κ coefficients. The calculated κ coefficients were consistently poor (ranging from 0 to 0.063).

Table 2 Demographics of the sample

Specialty		
Internal medicine	13	35%
Emergency medicine	10	27%
Family medicine	4	11%
Other	10	27%
Number of years in practice		
Mean	26	
SD	12.2	
Range	1–45	
Per cent of time seeing patients in current position		
<25	9	24%
25–49	5	14%
50–75	13	35%
76–100	10	27%

Table 3 Number of biases by case and consistent/inconsistent

	Inconsistent			Consistent		
Case	Mean	SD	Range	Mean	SD	Range
Pneumonia	3.40	1.40	1–5	2.63	1.67	0–6
Acute MI	3.72	1.35	1–5	1.23	1.48	0–5
Tubo-ovarian abscess	3.74	1.60	1–6	1.50	1.40	0–4
Subarachnoid haemorrhage	3.54	1/72	1–6	1.08	1.80	0–6
Kidney stone	3.45	1.57	0–5	1.88	1.90	0–6
Pyelonephritis	3.85	1.27	1–6	1.59	2.03	0–6

MI, myocardial infarction.

DISCUSSION

The current study investigated the ability of generalist physicians with special interest in diagnostic error to reliably detect cognitive biases. We examined agreement among raters about presence or absence of specific biases, and the extent to which the case outcome (consistent or inconsistent with the primary diagnosis pursued) affected the number of biases identified. The presence of an inconsistent outcome resulted in identification of twice as many biases, despite the identical case description to that point. This suggests that

Table 4 Per cent of cases where individual biases were present by consistent/inconsistent test result and inter-rater reliability

	Consistent (%)	Inconsistent (%)	Relative increase (%)	κ Value
Anchoring	35	70	100	0.0
Availability	25	55	120	0.025
Base rate neglect	11	28	150	0.063
Confirmation	35	62	77	0.024
Premature closure	39	88	125	0.046
Representativeness	26	45	73	0.044

reviewers exhibited a strong 'hindsight' bias,[16] where knowledge of a delayed diagnosis or an adverse event led to a more intensive search for possible causes. Further, an inconsistent outcome was far more likely to be considered a diagnostic error than a consistent outcome. However, not all inconsistent outcomes were considered diagnostic errors and some of the consistent outcomes were considered diagnostic errors. It may be that reviewers were considering different definitions of diagnostic error or different clinical problem-solving strategies.

Further, the inter-rater agreement about the presence or absence of specific biases was poor. This lack of agreement is surprising, although careful examination of prior studies that identified cognitive biases retrospectively did not identify any studies where reliability was reported. One study[24] coded biases as part of a larger text coding scheme and reported agreement of 80%–95% on a single pilot transcript. It is unclear how these numbers inform the issue of identifying specific biases. A second retrospective study of diagnostic errors in primary care[25] computed inter-rater reliability for 'process' errors but did not specifically identify any cognitive biases.

There are several limitations to this study. It may be that the individuals who agreed to participate in the study were not sufficiently expert in recognising cognitive biases. Though they are all members of the SIDM listserv, thus having expressed an interest in diagnostic error and presumably read discussions on the listserv, we could have tested their understanding. Still, we presume they have at least similar knowledge of biases to physicians who are participating as reviewers of real world cases of diagnostic error. It must also be recognised that an assumption of much of the literature on cognitive bias in diagnostic error is that straightforward instruction about biases at the undergraduate or residency level will be sufficient to reduce errors; the fact that these individuals, who likely have both more interest and more experience, were unable to agree on biases may serve as a cautionary note.

It is not clear how one could find a more appropriate cohort. On the other hand, a strength of the study was the enrolment of a geographically diverse cohort of experienced generalist physicians, whose clinical skills are directed toward the management of undifferentiated patient complaints where multiple diagnoses are possible.

The sample size was relatively small, although the results were highly significant. Perhaps more serious is the relatively low response rate. From the nature of the sampling strategy, we have no way to identify response bias by comparing with non-responders; however, a higher response rate may well have led to even stronger effects, since presumably those who volunteered were confident in their understanding of cognitive biases. In any case, the observed difference was large. A sensitivity analysis suggests that it would require an additional 166 participants who showed no evidence of outcome bias to make the overall effect *not* significant.

Finally, the study could be faulted for using written cases, which Blumenthal-Barby[12] identified as potential threat to validity. However, there is a critical difference; participants in the studies that identified bias through written cases were required to assume the role of a clinician working up a case. In our study, participants were asked to act as reviewers of a case workup. The review format was not unlike what might be encountered in any retrospective case review.

It may be argued that the cases lacked sufficient detail to permit identification of cognitive biases. However, again, we expect that the details are similar to those available in chart reviews. The cases are also very close to the kinds of cases that have been used in studies of cognitive bias[11 13 14] with the critical difference that the present cases were not deliberately designed to induce specific bias.

The imprecision of the definitions may be viewed as a limitation. However, it should be emphasised that these were not chosen by the authors, but rather are taken directly from a published source[2] and in fact several are nearly identical to definitions in the IoM report.[1] Certainly, the brief definitions suggest that the biases are overlapping and not distinct, and may simply be impossible to identify unambiguously from a retrospective report. For example, it could be argued that all case workups showed 'premature closure' since, although there were two possible diagnoses suggested by the case, the 'physician' pursued only one. Since both diagnoses are roughly equiprobable, this would be true regardless of whether the test was consistent or inconsistent. However, hindsight bias seems to greatly impact this attribution: premature closure was cited as a bias in only 39% of consistent cases, whereas 88% of inconsistent cases were deemed to demonstrate this bias. Similarly, it could be argued that 'base rate neglect' is present any time an error is committed (unless both diagnoses have exactly the same prevalence), since either the clinician picked the common diagnosis when the rare one was correct or picked the rare diagnosis when the common one was correct.

Conversely, 'availability' bias—which amounts to overestimating the likelihood of a particular diagnosis because past salient cases cause clinicians to overrepresent the probability of a given diagnosis—cannot be inferred from our experimental case descriptions because respondents had no access to the 'internal narrative' of the treating clinician, and there was no information given about these clinicians' past experiences. Despite this, availability bias was cited frequently by participants, and again appears vulnerable to hindsight, implicated in 25% of consistent cases and 55% of inconsistent cases.

These findings have consequences for the application of procedures to identify cognitive biases as a strategy for diagnostic error reduction. While many authors have advocated teaching students to prospectively recognise cognitive bias in their own reasoning so they can apply error-correcting 'cognitive forcing strategies',[2] the fact that even relative experts are unable to distinguish among various biases suggests that this may not be possible. This may explain why the few studies teaching instructional strategies to combat errors by consciously overcoming cognitive biases[26–28] have consistently demonstrated little benefit.

The finding of a very strong 'hindsight bias' has been identified previously in a medical context.[29] However, in that study, the judgement was simply whether care was adequate or not. Diagnostic error was not a focus of this investigation and, critically, the methods did not explore the causes underlying the judgement of adequate or inadequate care. In the present study, we have critically examined the purported role of cognitive bias in diagnostic error and shown that these causal judgements are strongly influenced by case outcomes.

The findings have implications for retrospective chart review of diagnostic errors. As Wears[16] described, hindsight bias can result in a simplistic misunderstanding of the cause of errors. Two strategies that might help are: (a) inclusion of cases in which no errors were detected, and (b) interviewing the clinician involved. One of us (LZ) used precisely this approach, and found that many of the errors that appeared to be a consequence of cognitive biases were revealed to be knowledge gaps.[19] Nevertheless, interviews are not a panacea; they depend on fallible memory and reconstruction, and prevent the types of experimental outcome manipulations that a study of this kind permits. And if, as some propose,[2] diagnostic errors arise in Type 1 processes, these are unconscious processes and hence not available to introspection.

CONCLUSIONS

This study has shown that observers with an interest in diagnostic error are unable to agree on the presence of specific cognitive biases in clinical case workups. Further, they are strongly influenced by 'hindsight bias, where knowledge of the case outcome strongly influences the detection of biases and error in the process. While cognitive bias may play a role in diagnostic error, and there may be potential to reduce error by better understanding cognitive bias, this study suggests that our current approaches to recognising and reducing cognitive biases are unlikely to have an impact on the prevalence of diagnostic errors.

Author affiliations
[1]Institute of Medical Education Research Rotterdam, Erasmus MC, Rotterdam, The Netherlands
[2]Department of Public and Occupational Health, VU University Medical Center/EMGO Institute, Amsterdam, The Netherlands
[3]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[4]Department of Medicine, McMaster University, Hamilton, Ontario, Canada
[5]Department of Medicine, University of Washington, Seattle, Washington, USA
[6]Program for Educational Research and Development, McMaster University, Hamilton, Ontario, Canada

Correction notice This article has been edited since it first published Online First. Minor edits have been made to κ coefficient values in the Results section and table 4 has been updated.

Twitter Follow Laura Zwaan at @laurazwaan81

Acknowledgements The authors wish to acknowledge the support of Mark Graber and the executive board of the Society for Improving Diagnosis in Medicine. We also wish to thank the participants who willingly volunteered their time to complete the survey.

Contributors All authors made substantial contributions to the conception or design of the work and to the acquisition and interpretation of data. All authors were involved in either drafting the work or revising it critically for important intellectual content. All gave their final approval for the version to be published. All authors agree to be accountable for all aspects of the work and for ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests None declared.

Ethics approval McMaster Research Ethics Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Balogh EP, Miller BT, Ball JR. *Improving diagnosis in medicine*. Washington: National Academy of Sciences, 2015.
2 Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med* 2003;78:775–80.
3 Norman G. Research in clinical reasoning: past history and current trends. *Med Educ* 2005;39:418–27.
4 Klein JG. Five pitfalls in decisions about diagnosis and prescribing. *BMJ* 2005;330:781–3.
5 Detmer DE, Fryback DG, Gassner K. Heuristics and biases in medical decision-making. *J Med Educ* 1978;53:682–3.
6 Redelmeier DA, Ferris LE, Tu JV, *et al*. Problems for clinical judgement: introducing cognitive psychology as one more basic science. *Can Med Assoc J* 2001;164:358–60.
7 Redelmeier DA. The cognitive psychology of missed diagnoses. *Ann Int Med* 2005;142:115–20.
8 Elstein AS. Heuristics and biases: selected errors in clinical reasoning. *Acad Med* 1999;74:791–4.
9 Schmitt BP, Elstein AS. Patient management problems: heuristics and biases. *Med Decis Making* 1988;8:224–5.
10 Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Qual Saf* 2013;22 (Suppl 2):ii58–64.
11 Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science* 1974;185:1124–31.
12 Blumenthal-Barby JS, Krieger H. Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy. *Med Decis Making* 2015;35:539–57.
13 Mamede S, van Gog T, van den Berge K, *et al*. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA* 2010;304:1198–203.

14 Hatala R, Norman GR, Brooks LR. Impact of a clinical scenario on accuracy of electrocardiogram interpretation. *J Gen Int Med* 1999;14:126–9.
15 Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. *Arch Intern Med* 2005;165:1493–9.
16 Wears RL, Nemeth CP. Replacing hindsight with insight: toward better understanding of diagnostic failures. *Ann Emerg Med* 2007;49:206–9.
17 Kogan JR, Hess BJ, Conforti LN, *et al*. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med* 2010;85(Suppl 10): S25–28.
18 Jacoby LL. A process dissociation framework: Separating automatic from intentional uses of memory. *J Memory Language* 1991;30:513–41.
19 Zwaan L, Thijs A, Wagner C, *et al*. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Acad Med* 2012;87:149–56.
20 Croskerry P. ED cognition: and decision by anyone at anytime. *Can J Emerg Med* 2014;16:13–19.
21 Eva KW, Cunnington JP. The difficulty with experience: does practice increase susceptibility to premature closure? *J Contin Educ Health Prof* 2006;26:192–8.
22 Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Ed Psych Measur* 1973;33:6113–19.
23 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
24 Ogdie AR, Reilly JB, Pang WG, *et al*. Seen through their eyes: Residents' reflections on the cognitive and contextual components of diagnostic errors in medicine. *Acad Med* 2012;87:1361–7.
25 Singh H, Giardina TD, Meyer AN, *et al*. Types and origins of diagnostic errors in primary care settings. *JAMA Intern Med* 2013;173:418–25.
26 Reilly JB, Ogdie AR, Von Feldt JM, *et al*. Teaching about how doctors think: a longitudinal curriculum in cognitive bias and diagnostic error for residents. *BMJ Qual Saf* 2013;22:1044–50.
27 Sherbino J, Dore KL, Siu E, *et al*. The effectiveness of cognitive forcing strategies to decrease diagnostic error: an exploratory study. *Teach Lrn Med* 2011;23:78–84.
28 Sherbino J, Kulasegaram K, Howey E, *et al*. Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. *Can J Emerg Med* 2014;16:34–40.
29 Caplan RA, Posner KL, Cheney FW. Effect of outcome on physician judgments of appropriateness of care. *JAMA* 1991;265:1957–60.

Appendix 1:
Example case: showing the consistent and inconsistent version

History of Present Illness:
A 43-year old woman was brought to the Emergency Department by her husband at 0200 in the morning because of shortness of breath. The dyspnea occurred suddenly at 1100 pm and awoke the patient from sleep. This dyspnea was accompanied by retrosternal chest pain, which was worse on deep breathing. She also reports that she had awoken with chest tightness the prior night, but this resolved after a short while. The patient reports that she has been feeling unwell for about 4 days, with throat and sinus congestion, fever and chills, and vomited a small amount of bile. She has also had a cough for several days, and had coughed up small amounts of blood. The patient complained of nausea and vomited a small amount of bile during the triage interview. She has had no recent surgery.

Past Medical History
tubal ligation, 8 years ago
Pneumonia, 2 years ago
No recent surgery

Social History
Prior smoking, stopped 2 years previously.

Medications
None.

Physical examination
Her temp was 37.4, pulse 96, BP 110/96, RR 30.
The chest was clear to auscultation.
The heart sounds were normal as was the abdominal exam.
There was some left calf tenderness without swelling.

Further Testing and Imaging
Her WBC count was elevated (13,0 x 10^9 /L).
Her hemoglobin level was normal.
The ECG demonstrates non-specific ST depression in V3-V6.

A Chest X-ray was ordered to diagnose pneumonia.

Consistent case version:
This demonstrated an infiltrate in the lingula of her left lung field consistent with pneumonia.

Inconsistent case version:
This demonstrated a wedge shaped, pleural-based consolidation in the patient's left lower lobe (Hampton's hump), suggestive of a pulmonary embolism.

Appendix 2:

Definitions of Biases presented to Participants

Bias Definitions

Anchoring

The tendency to perceptually lock into salient features in the patient's initial presentation too early in the diagnostic process, and failing to adjust this initial impression in the light of later information.

Availability Heuristic

The disposition to judge things as being more likely or frequently occurring, if they readily come to mind. Thus recent experience with a disease may inflate the likelihood of its being diagnosed. Conversely, if a disease has not been seen for a long time (i.e. is less available) it may be underdiagnosed.

Base Rate Neglect

The tendency to ignore the true prevalence of a disease, either inflating or reducing its base rate, and distorting Bayesian reasoning.

Confirmation Bias

The tendency to look for confirming data to support a diagnosis rather than look for disconfirming evidence to refute it, despite the latter often being more persuasive and definitive.

Premature Closure

The tendency to apply premature closure to the decision-making process, accepting a diagnosis before it has been fully verified. The consequences of the bias are reflected in the maxim – "when the diagnosis is made, the thinking stops."

Representativeness Bias

The tendency to look for a prototypical manifestations of disease. Restraining decision-making along pattern recognition lines leads to atypical variants being missed.