UNRAVELING CLINICAL REASONING AND DIAGNOSTIC ERROR:
Mechanisms and interventions

Justine Staal

Staal, Justine
Unraveling clinical reasoning and diagnostic error: mechanisms and interventions
Thesis, Erasmus University Rotterdam, the Netherlands

ISBN: 978-94-6483-276-1
Cover design: Annemieke van Harten
Layout design: Publiss | www.publiss.nl
Printed by: Ridderprint | www.ridderprint.nl

Unraveling Clinical Reasoning and Diagnostic Error
Mechanisms and Interventions

Het ontrafelen van klinisch redeneren en diagnosefouten
Mechanismen en interventies

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof. dr. A.L. Bredenoord

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

woensdag 1 november 2023 om 13.00 uur

door

Justine Staal
geboren te Goes.

Erasmus University Rotterdam

Promotiecommissie:

Promotoren:	prof. dr. W.W. van den Broek
	prof. dr. M.A. Frens
Overige leden:	dr. F. van Kooten
	prof. dr. ir. A. Burdorf
	prof. dr. D.R.M. Timmermans
Copromotoren:	dr. L. Zwaan
	dr. J. Alsma

Contents

1. Introduction

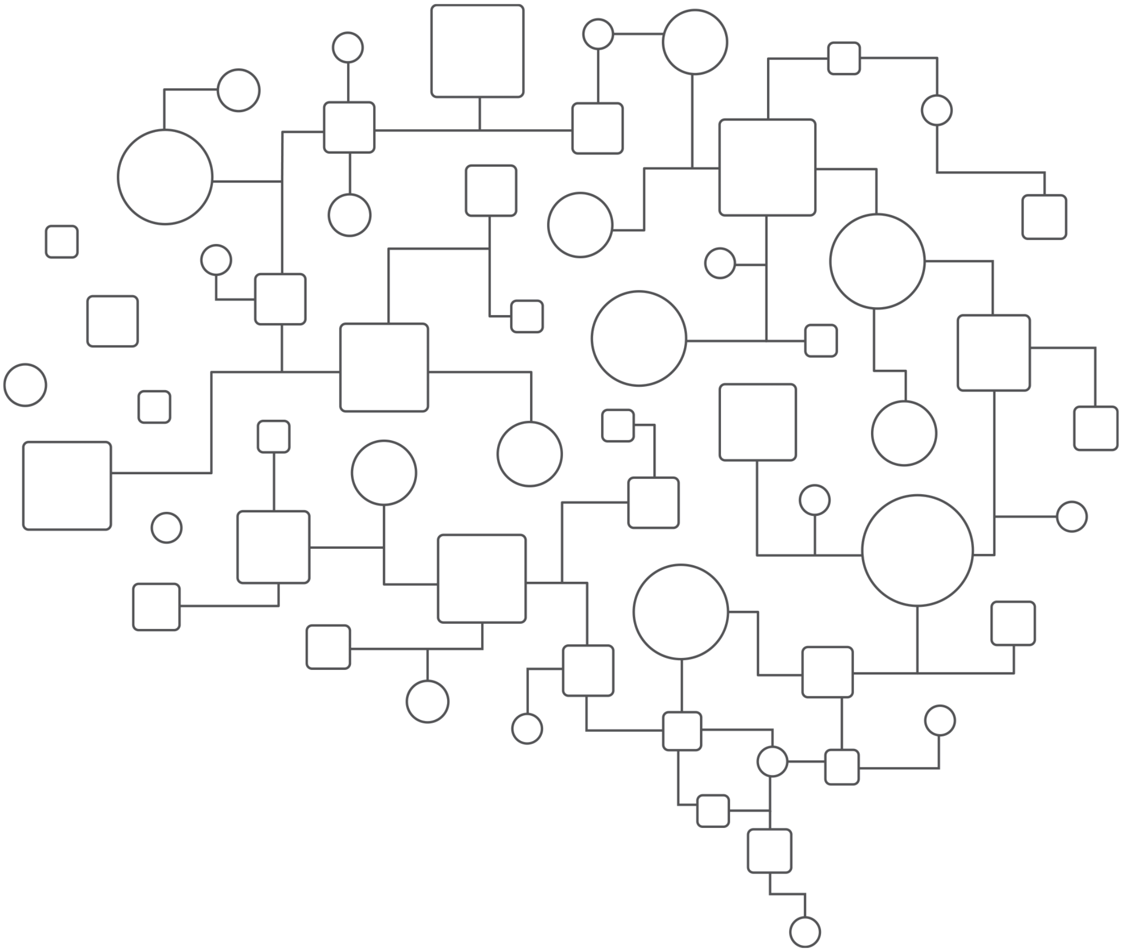
2. Cognitive processes underlying diagnostic errors

3. Interventions to prevent diagnostic errors

4. Discussion and summary

INTRODUCTION

I



CHAPTER 1

General introduction

General introduction

The medical field has gone through tremendous advancements through the decades and huge strides have been made in improving healthcare. For example, new diseases and treatments have been discovered and advanced techniques have been implemented. In 1999, attention was called to healthcare quality and safety when the National Academies of Sciences, Engineering, and Medicine (NASEM) reported on the prevalence and consequences of medical errors.(1) Such errors are defined as the "failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim". Deaths due to medical errors in the United States alone exceeded the combined fatalities from road accidents, breast cancer, and AIDS.(1) A significant portion of medical errors is attributed to flaws in the diagnostic process, yet these errors remained underemphasized (2-7) despite their high incidence, highly preventable nature, and severe consequences.(8) It was not until 2015 that the NASEM raised awareness by publishing *Improving diagnosis in healthcare*, a report focused specifically on diagnostic errors.(9) In this report it was estimated that "most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences." The report concluded that "improving the diagnostic process is not only possible, but also represents a moral, professional, and public health imperative". It is therefore vital to understand the causes of diagnostic errors and to develop strategies to reduce them.

Diagnostic errors are defined as "the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient".(9) Generally, diagnostic errors can be divided into three categories.(10) First, no-fault errors are mistakes that could not have been prevented, for example due to unusual disease presentation. Second, system errors occur due to technical or organizational circumstances, such as breakdowns in communication or equipment failure. Last, cognitive errors result from breakdowns in clinicians' diagnostic processes, such as faulty data gathering or interpretation. Causes of diagnostic errors are often multifactorial (11, 12): an observational study by Graber et al. (10) showed that cognitive errors occurred in 74% of cases, and often in conjunction with system errors, which occurred in 65% of cases. Cognitive errors are therefore seen as major contributors to diagnostic errors.(3)

Much is still unclear about the mechanisms underlying cognitive errors and the types of interventions that could counteract them. In the current literature, this debate centers on whether cognitive biases (i.e., predispositions to think in a way that leads to systematic failures in judgement (13)) (10, 14-18) or knowledge deficits (8, 18-21) are the main cause for errors. Bias errors can also be caused by knowledge deficits instead of faulty reasoning: for instance, when a clinician settles on a wrong diagnosis before considering other possibilities (premature closure bias) this could be explained as the clinician using a heuristic to arrive at their diagnosis and not considering alternatives, or as the clinician not knowing the correct diagnosis at all.(19) Furthermore, our current understanding is limited because diagnostic errors are primarily studied in retrospective studies (8, 10, 12) and experimental laboratory studies. Retrospective studies are vulnerable to hindsight bias or outcome bias (22, 23) and only focus on cases in which errors occurred. This leaves it unclear whether the processes in faulty reasoning, such as the use of biased heuristics, are limited to error cases or also occur in effective reasoning. Experiments can avoid this by prospectively inducing errors but exactly this controlled nature also leads to reduced validity for clinical practice. More research is necessary to determine the mechanisms of cognitive errors and the effectiveness of interventions because empirical evidence is scarce.(3, 24) Therefore, the aim of this thesis is to increase our understanding of the cognitive mechanisms underlying diagnostic errors and to determine the effectiveness of interventions to counteract cognitive errors. A brief overview of the literature concerning causes and prevention of cognitive errors will precede the chapters of this thesis.

Diagnostic process

Diagnostic errors are best explained within the context of the diagnostic reasoning process. This process is often conceptualized as both a classification scheme, because it involves labeling a pattern of signs and symptoms as a specific disease, and a process.(25) The process starts as soon as a health problem is detected and the patient interacts with the healthcare system.(9) Following this, an iterative cycle starts where healthcare professionals gather information on the patient's health problem, integrate this information, and interpret it. The cycle is repeated until sufficient information has been collected to decide on a diagnosis and a course of action for the patient, which is then communicated to the patient. The planned action is taken and finally, the outcomes for the patient are evaluated. Despite its apparent simplicity, the diagnostic reasoning process becomes rather complex in practice. For example, although the cycle of gathering and interpreting information should theoretically continue until one correct diagnosis can be confidently assigned to the patient, in reality diagnostic uncertainty is inherent to the process and a diagnosis can never be established with complete certainty.(26) In addition, factors such as organizational influences, technologies, the physical work environment, and the members of a diagnostic team can all impact the outcomes of the diagnostic process, both independently and in interaction with other factors.(9, 27, 28)

Accurately detecting and measuring diagnostic errors within the already complex diagnostic reasoning process is a challenging task. Available data is scarce and often not reliable due to gaps and variability in both the amount and quality of the data.(9, 29, 30) This variability emerges because data on diagnostic errors originate from a variety of sources (e.g., autopsies, malpractice claims, clinician surveys) and a variety of settings (e.g., primary or specialty care, different specialisms such as radiology or cardiology that partially rely on visual diagnosis, compared to internal medicine or intensive care). All these sources give insights in different groups and facets of diagnostic error, but are difficult to aggregate and provide a poor overview of diagnostic error in general.

Interpretation of these data sources is further influenced by complexities in the diagnostic process itself. A notable example is the evolution of a disease over time, which complicates the determination of whether or not an error occurred.(9, 25) Diseases are often already present before the symptoms surface and it might take even longer before these symptoms are sufficient to be recognized and diagnosed. Even if symptom patterns appear, these can differ substantially between different diseases and patients, and the patterns might not be recognizable or be obscured by comorbidities.(9) This is further compounded by the evolution of the diagnostic process itself over time: information is gathered in cyclical stages rather than all at once. In the time between these stages the available information and symptoms may change and lead to different conclusions at different moments. At which stage in this evolution does a mistake become a diagnostic error rather than a no-fault error? Moreover, the measurement of diagnostic error is unduly influenced by the outcomes of the diagnostic process.(25) When determining whether or not an error occurred, observers are influenced by hindsight bias or outcome bias and rate the likelihood that the correct diagnosis should have emerged as a plausible solution at the time of diagnosis higher than it actually was.(22, 23) This influences the detection of diagnostic errors, as cases with an incorrect diagnostic outcome are labeled as diagnostic errors, whereas possible errors or breakdowns in cases with a correct final outcome are overlooked. But when considering factors such as the evolution of the disease and the diagnostic process, even if the diagnostic outcome was incorrect, the diagnostic process might not necessarily need to be labeled as erroneous. In summary, the detection and measurement of errors is a daunting task that requires a better understanding of diagnostic errors.

Dual process theory

The quality of a diagnosis depends on the clinician's competency in clinical reasoning, defined as "the cognitive processes necessary to evaluate and manage a patient's medical problems". (31) Clinical reasoning is fundamentally a cognitive process and cognitive psychology is considered the basis to understanding it. Currently, the dual process theory (13) is the most common framework used to explain clinical reasoning (9, 32) and its proposed mechanisms underlying cognitive errors are central to this thesis.

Dual process theory is an influential decision making model that originated from the field of psychology (33-35) and has since been applied to many tasks that require decision making, including clinical reasoning.(9, 15) Dual process theory generally proposes that cognition consists of two systems, often referred to as System 1 and System 2.(13, 33, 36-39) System 1 is defined as an intuitive system which relies on heuristics (i.e., mental shortcuts) (13) to achieve fast and automatic processing. System 1 is thought to work best for routine problems, e.g., when a clinician encounters a patient with readily recognizable symptom patterns. On the other hand, System 2 is seen as a slow system that uses deliberate reasoning, which places a burden on working memory. This type of reasoning is analytical, with a known process and outcome. It is thought to follow the principles of hypothetico-deductive reasoning, where falsifiable hypotheses are generated, tested, and then accepted or rejected.(40) System 2 reasoning is most useful for new, complex, or non-routine problems, e.g., when a patient's symptoms are not recognized. The existence of two cognitive systems is corroborated by several psychological and neuropsychological studies, although these cannot be mapped directly to the theoretical constructs of System 1 and System 2.(39, 41) Despite the seemingly strict division, many reasoning tasks contain a certain measure of both non-analytical and analytical processes and therefore tasks are often considered along a continuum instead of being categorized as either pure System 1 or System 2 reasoning.(32, 42)

The interaction between System 1 and System 2 is conceptualized mainly in three models.(43) First, the parallel model proposes that both processes occur simultaneously: both System 1 and 2 offer a solution to the problem at hand and if these solutions are conflicting, the conflict is resolved to arrive at an answer.(44) Second, the serial, or default-interventionist, model (34, 45, 46) suggests that System 1 always provides default responses to a problem and that System 2 monitors these responses, only overriding them when necessary or possible. Third, a hybrid model combining aspects of the parallel and serial models has been proposed.(47) This model incorporates both the idea that System 1 and System 2 are activated sequentially, and that System 2 activation only occurs when necessary. Additionally, the hybrid model also proposes parallel activation of multiple automatic responses in System 1, which will compete as well. Our current understanding of cognitive diagnostic errors and interventions is primarily based in this default-interventionist perspective (48), although the interaction between System 1 and System 2 remains debated.(49)

Heuristics and biases perspective

Within the dual processing framework, cognitive errors are generally explained as a result of incorrect heuristics, or mental shortcuts, often associated with nonanalytical System 1 reasoning.(13, 19, 50) These shortcuts are traditionally thought to trade accuracy for efficiency. In the heuristics and biases perspective, errors occur when flawed heuristics lead to cognitive biases (i.e., predispositions to think in a way that leads to systematic failures in judgement).(13-16, 32, 51-57) Examples of cognitive biases are confirmation biases (i.e., the tendency to interpret new information as confirming one's beliefs) or availability biases (i.e., the tendency to think that examples that come easily to mind are more representative than they actually are), but many other biases have been identified.(14) In medicine, these biases can occur in the diagnostic process: for example, new symptoms can be interpreted in light of a previous diagnosis even if they are unrelated; or if a clinician encounters patients with similar symptoms but different diagnoses, the diagnosis of the first patient might more easily come to mind and bias the diagnosis of the other patients. These biases are considered to be highly preventable.(58)

Although biases are a popular and widely accepted explanation of diagnostic errors (19, 59) and empirical evidence has shown that biases can indeed cause such errors (3, 17, 60), the origins of cognitive diagnostic errors cannot fully be explained by biases in nonanalytical reasoning. For instance, System 1 is not actually more prone to bias than System 2; fast reasoning is not necessarily associated with more errors (45, 61); slow reasoning does not guarantee a correct diagnosis (9); and biases can also occur due to the limited cognitive capacity of System 2 reasoning.(62, 63) Nonanalytical reasoning is actually an important component of clinical reasoning. It allows clinicians to make decisions even when information is missing, making it very effective in the real world.(64-68) They are often accurate and can even outperform decisions made via analytical reasoning.(64-67) The value and effectiveness of nonanalytical reasoning are shown by experts, who can make accurate decisions based off of limited information.(18, 69, 70)

Knowledge perspective

The knowledge perspective offers an alternative explanation of what could cause diagnostic errors.(18, 68, 71) In this view, clinical reasoning is defined as a categorization or classification task where knowledge on diseases and their accompanying symptoms are encoded. The exact method of encoding differs between theories: some propose the existence of illness scripts (72), exemplars, or protoypes (73), which all store different types of information and to a different extent. However, all these forms of encoded knowledge play the same role in diagnosis. When encountering a patient, knowledge on diseases similar to the features of the patient are activated and the best matching one is selected as the diagnosis.(9) Dual processing theory can also be viewed through the lens of the knowledge perspective. In this case, the two systems differ in whether they retrieve knowledge automatically or nonautomatically.(74) The nonanalytical System 1 operates using experiential knowledge that is formed between disease knowledge and specific features through experience. Automatic pattern recognition matches a patient's features to disease knowledge.(9) The analytical System 2 uses the encoded formal knowledge, which is the learned knowledge clinicians have of the clinical and patient features representing a disease.(9, 18) As opposed to the heuristics and biases perspective, knowledge is central to successful clinical reasoning instead of processing strategies.(9, 18) After all, if a clinician does not possess the necessary knowledge to arrive at the correct diagnosis, no amount of analytical reasoning can provide the answer.(37, 75-78)

Cognitive error interventions

Both the heuristics and biases perspective and the knowledge perspective offer different approaches for counteracting cognitive errors. Interventions aimed at preventing cognitive biases are called debiasing strategies or cognitive forcing strategies (14) and primarily focus on improving or supporting clinician's reasoning processes. These interventions operate on the principle of metacognition, which is one's awareness and understanding of their own reasoning processes. Debiasing strategies attempt to engage metacognitive thinking in clinicians by having them reflect on their reasoning processes, for instance by asking them to slow down and consider whether they missed anything, or to ask themselves whether any biases influenced their reasoning.(14, 18, 24) In terms of dual process theory, debiasing strategies aim to mobilize analytical thinking to reduce errors that arise from nonanalytical thinking. As long as a clinician is aware of potential pitfalls and of their own reasoning processes, they could be taught to detect and prevent their own mistakes. Debiasing strategies are often presented as mnemonics or checklists, or clinicians are taught to incorporate metacognitive questions into their usual reasoning processes.

On the other hand, error interventions based on remedying knowledge deficits focus on increasing knowledge and expertise, enhancing recall, or organizing knowledge.(19, 79, 80) Metacognition remains an important component of such strategies, because clinicians have to be aware of their knowledge deficits before they will seek to remedy them; however, the focus of these interventions is not on the metacognitive principle but on acquiring or triggering the appropriate knowledge. Examples of such interventions are educational strategies (81-83), feedback interventions (84-86), deliberate reflection strategies (80, 87), checklists (88, 89), or clinical decision support algorithms.(90)

Although both frameworks provide many suggestions for reducing cognitive errors, empirical evidence is relatively scarce (3, 24), especially evidence in practical settings.(91, 92) In the current literature, little evidence exists concerning debiasing strategies (92): several studies report improvements in diagnostic accuracy of self-assessments (79) but other studies conclude that debiasing is ineffective.(59, 93, 94) The overall evidence is too limited to conclude that debiasing strategies are effective. Of the interventions focused on knowledge deficits or content, most empirical evidence relates to deliberate reflection, a method that asks clinicians to consider alternative diagnoses and assess the case features that fit or do not fit with these diagnoses. This information must then be weighted and the clinician is asked to rank the diagnoses. Deliberate reflection strategies are found to lead to small but consistent improvements.(69, 79-81, 95, 96) Other knowledge-based interventions have not been studied to this extent, but especially clinical decision support (24, 90), checklists (97) and feedback interventions (95, 98) show small overall effects and are seen as promising. However, much still remains unknown about exactly in which settings and subgroups (e.g., dependent on clinician's level of experience or case difficulty) these interventions are most effective.(97) Overall, more research is necessary to determine how effective these strategies will be in preventing errors in clinical practice.

Research questions and thesis outline

In summary, even though cognitive flaws such as biases are considered a main source of error, the mechanisms underlying it are incompletely understood.(59, 99) This leaves open questions concerning the origins of cognitive diagnostic errors and how they are best prevented. Additional research is necessary to advance our understanding of diagnostic errors. This thesis aims to provide further insight in the cognitive mechanisms underlying diagnostic errors and possible error interventions using quantitative methods. This translates to the main research questions of this thesis, which are discussed in the following section.

The studies in Chapter 2 through 5 are concerned with increasing our understanding of causes of diagnostic errors by examining how several cognitive factors relate to the occurrence of diagnostic errors.

In Chapter 2, a multi-center laboratory experiment was conducted to examine how time to diagnose related to diagnostic accuracy in a within-subjects design. Using Mamede et al.'s (87) methodology, we prospectively induced availability bias in internal medicine residents. Residents were asked to judge the likelihood of a suggested working diagnosis to be correct for several cases, after which they diagnosed clinical cases that resembled the previous cases but in truth had another correct diagnosis. Resident's diagnostic accuracy and time taken to diagnose were measured to answer the main research question. In addition, confidence, perceived case complexity, mental effort, resource use, and confidence-accuracy calibration were measured to examine how induced bias impacted diagnostic performance besides accuracy.

The laboratory experiment in Chapter 3 presents a within-subjects study aimed at assessing how diagnostic suggestions influenced the diagnostic performance of medical interns. Interns were asked to provide a most likely diagnosis for clinical cases in the form of a general practitioner's referral letter to the emergency department. The referral letters contained either no suggestion (as the control condition), a correct suggestion, or an incorrect suggestion. The suggestion was meant to prospectively induce confirmation bias. We measured intern's diagnostic accuracy, number of differential diagnoses, confidence, and time taken to diagnose.

Chapter 4 presents a laboratory experiment that measured resident's eye movements to determine how the type of information they engaged with during diagnosis affected diagnostic accuracy. Based on the results from Chapter 2, the methodology to prospectively induce bias was modified to induce confirmation bias instead of availability bias. Residents were asked to diagnose clinical cases with a suggested working diagnosis. Half of the cases contained a correct suggestion, the other half of the suggestions was incorrect. Residents determined whether the suggested diagnosis was correct, and if not, what would be the most likely diagnosis instead. Each case contained regions of interest, which were defined as information units necessary to distinguish between the correct and the incorrect diagnosis. We measured resident's diagnostic accuracy, confidence, and the total time their eyes were fixated on the regions of interest. Gaze fixation time is a measure of engaging with and processing information, so the longer someone fixates on information, the more they are thought to use it in their reasoning process.(100) Additionally, we measured time taken to diagnose and resident's confidence-accuracy calibration.

In Chapter 5, an observational study investigated how medical students used clinical information (i.e., patient history, physical examination, investigations) during diagnosis. Previous research showed that knowledge deficits can cause diagnostic errors (20), but it remains unknown how well clinical information is used in student's diagnostic process. First and second year medical students were asked to diagnose clinical cases in an online learning environment, TeachingMedicine.com. For each case, students filled out their differential diagnosis and identified clinical information that would reduce or increase the likelihood of each differential diagnosis. We measured how many relevant differential diagnoses students included and how well they identified clinical information relevant to each included diagnosis. Their ability to correctly assign information was compared for the patient's history, the physical examination, and further investigations. Additionally, we compared whether the assignment of clinical information was performed similarly for information that increased or decreased the likelihood of a specific diagnosis. Lastly, we assessed whether student's ability to assign information predicted their performance on ordering appropriate investigations.

The studies in Chapter 6 through 8 examined how effective several diagnostic error interventions were in improving diagnostic accuracy.

Chapter 6 concerns a mixed design laboratory experiment that compared the effectiveness of a debiasing checklist and a content-specific checklist for normal and abnormal electrocardiogram (ECG) diagnosis. Residents enrolled in the general practice program were asked to diagnose ECGs in two sessions. In the first session, they diagnosed the ECGs without a checklist and in the second phase a week later, half of the residents diagnosed the same ECGs with a debiasing checklist and the other half diagnosed the same ECGs with a content-specific checklist. The ECGs could have no abnormalities (i.e., normal ECGs), an easily recognizable abnormality (i.e., atrial fibrillation), or a complex abnormality (i.e., ischemia). We measured resident's diagnostic accuracy, confidence, patient management, time taken to diagnose, and confidence-accuracy calibration. These measures were then compared between the first and the second session to determine the impact of checklist use on the diagnosis of normal and abnormal ECGs.

In Chapter 7, the effectiveness of performance feedback and information feedback on the diagnostic performance of medical students was assessed. Students were asked to diagnose chest X-rays in two phases. In the first phase, the learning phase, students provided an initial most likely diagnosis and then received feedback in one of three formats. In the control condition, students were only asked to inspect the X-ray again and were not given the correct answer; in the performance feedback condition, students were shown the X-ray again and were informed of the correct diagnosis; and lastly, in the information feedback condition, students were shown the X-ray again with the correct diagnosis and an indication of where the abnormality could be seen, along with a short explanation of how they could have recognized this abnormality. After this, students again diagnosed X-rays without feedback in the test phase. We measured student's diagnostic accuracy, confidence, time taken to diagnose, and their confidence-accuracy calibration.

Chapter 8 concerns a systematic review and meta-analysis on the effectiveness of workplace-oriented cognitive reasoning tools (i.e., interventions aimed at improving clinician's clinical reasoning processes) on improving diagnostic accuracy. This review included experimental studies that compared medical students and professionals diagnostic performance with and without a tool on a diagnostic task. Additionally, the study aimed to identify factors associated with greater tool effectiveness. We aimed to assess the effectiveness of workplace-oriented tools alone, because previous reviews aggregated the effects of workplace-oriented and education-oriented studies. A random-effects meta-analysis was used to quantify tool effectiveness.

Finally, Chapter 9 provides an overview and discussion of the main findings of this thesis and additionally considers implications and opportunities for future research.

References

1. Kohn LT, Corrigan JM, Donaldson M. To Err is Human: Building a Safer Health System (Institute of Medicine, National Academy Press, Washington, DC). 1999.
2. Berner ES. Diagnostic error in medicine: introduction. Springer; 2009. p. 1-5.
3. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. European journal of internal medicine. 2013;24(6):525-9.
4. Wachter RM. Why diagnostic errors don't get any respect—and what can be done about them. Health Affairs. 2010;29(9):1605-10.
5. Graber ML, Carlson B. Diagnostic error: the hidden epidemic. Physician executive. 2011;37(6):12-8.
6. Thammasitboon S, Thammasitboon S, Singhal G. Diagnosing diagnostic error. Current Problems in Pediatric and Adolescent Health Care. 2013;43(9):227-31.
7. Zwaan L, Schiff GD, Singh H. Advancing the research agenda for diagnostic error reduction. BMJ quality & safety. 2013;22(Suppl 2):ii52-ii7.
8. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
9. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
10. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.
11. Schiff G, Volodarskaya M, Nieva HR, Singh H, Wright A, editors. Diagnostic Pitfalls: A New Paradigm to Understand and Prevent Diagnostic Error. Journal of general internal medicine; 2016: SPRINGER ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATES.
12. Schiff GD, Hasan O, Kim S, Abrams R, Cosby K, Lambert BL, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. Archives of internal medicine. 2009;169(20):1881-7.
13. Kahneman D. Thinking, fast and slow: Macmillan; 2011.
14. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic medicine. 2003;78(8):775-80.
15. Croskerry P. The cognitive imperative thinking about how we think. Academic Emergency Medicine. 2000;7(11):1223-31.
16. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. BMJ quality & safety. 2013;22(Suppl 2):ii58-ii64.
17. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203.
18. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. Journal of Evaluation in Clinical Practice. 2018;24(3):666-73.
19. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Medical education. 2010;44(1):94-100.
20. Mamede S, Goeijenbier M, Schuit SCE, de Carvalho Filho MA, Staal J, Zwaan L, et al. Specific disease knowledge as predictor of susceptibility to availability bias in diagnostic reasoning: a randomized controlled experiment. Journal of general internal medicine. 2021;36(3):640-6.

21. Braun LT, Zwaan L, Kiesewetter J, Fischer MR, Schmidmaier R. Diagnostic errors by medical students: results of a prospective qualitative study. BMC medical education. 2017;17(1):1-7.
22. Zwaan L, Monteiro S, Sherbino J, Ilgen J, Howey B, Norman G. Is bias in the eye of the beholder? A vignette study to assess recognition of cognitive biases in clinical case workups. BMJ quality & safety. 2017;26(2):104-10.
23. Wears RL, Nemeth CP. Replacing hindsight with insight: toward better understanding of diagnostic failures. Annals of emergency medicine. 2007;49(2):206-9.
24. Graber ML, Kissam S, Payne VL, Meyer AND, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ quality & safety. 2012;21(7):535-57.
25. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis. 2015;2(2):97-103.
26. Kassirer JP. Our stubborn quest for diagnostic certainty. Mass Medical Soc; 1989. p. 1489-91.
27. Carayon P, Hundt AS, Karsh BT, Gurses AP, Alvarado CJ, Smith M, et al. Work system design for patient safety: the SEIPS model. BMJ Quality & Safety. 2006;15(suppl 1):i50-i8.
28. Smith MJ, Sainfort PC. A balance theory of job design for stress reduction. International journal of industrial ergonomics. 1989;4(1):67-79.
29. McGlynn EA, McDonald KM, Cassel CK. Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the Institute of Medicine. Jama. 2015;314(23):2501-2.
30. Singh H, Sittig DF. Advancing the science of measurement of diagnostic errors in healthcare: the Safer Dx framework. BMJ Quality & Safety. 2015;24(2):103-10.
31. Barrows HS, Tamblyn RM. Problem-based learning: An approach to medical education: Springer Publishing Company; 1980.
32. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education. 2009;14(1):27-35.
33. Evans JSBT. In two minds: dual-process accounts of reasoning. Trends in cognitive sciences. 2003;7(10):454-9.
34. Evans JSBT, Stanovich KE. Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science. 2013;8(3):223-41.
35. Osman M. An evaluation of dual-process theories of reasoning. Psychonomic bulletin & review. 2004;11(6):988-1010.
36. Kahneman D, Frederick S. A model of heuristic judgment: Cambridge University Press; 2005.
37. Stanovich K. Rationality and the reflective mind: Oxford University Press; 2011.
38. Frankish K. Dual-process and dual-system theories of reasoning. Philosophy Compass. 2010;5(10):914-26.
39. Smith ER, DeCoster J. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and social psychology review. 2000;4(2):108-31.
40. Evans C, Kakas AC, editors. Hypothetico-deductive Reasoning. Fgcs; 1992.
41. Rotgans JI, Schmidt HG, Rosby LV, Tan GJS, Mamede S, Zwaan L, et al. Evidence supporting dual-process theory of medical diagnosis: a functional near-infrared spectroscopy study. Medical education. 2019;53(2):143-52.
42. Norman G, Monteiro S, Sherbino J. Is clinical cognition binary or continuous? Academic Medicine. 2013;88(8):1058-60.

43. Evans JSBT. On the resolution of conflict in dual process theories of reasoning. Thinking & Reasoning. 2007;13(4):321-39.

44. Sloman SA. The empirical case for two systems of reasoning. Psychological bulletin. 1996;119(1):3.

45. Bago B, De Neys W. Fast logic?: Examining the time course assumption of dual process theory. Cognition. 2017;158:90-109.

46. De Neys W, Glumicic T. Conflict monitoring in dual process theories of thinking. Cognition. 2008;106(3):1248-99.

47. De Neys W. Bias, conflict, and fast logic: Towards a hybrid dual process future? Dual process theory 20: Routledge; 2017. p. 47-65.

48. Monteiro SM, Norman G. Diagnostic reasoning: where we've been, where we're going. Teaching and learning in medicine. 2013;25(sup1):S26-S32.

49. Gronchi G, Giovannelli F. Dual process theory of thought and default mode network: A possible neural foundation of fast thinking. Frontiers in psychology. 2018;9:1237.

50. Norman G, Sherbino J, Dore K, Wood T, Young M, Gaissmaier W, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. Academic Medicine. 2014;89(2):277-84.

51. Croskerry P. From mindless to mindful practice—cognitive bias and clinical decision making. N Engl J Med. 2013;368(26):2445-8.

52. Elia F, Apra F, Verhovez A, Crupi V. "First, know thyself": cognition and error in medicine. Acta Diabetologica. 2016;53(2):169-75.

53. Mithoowani S, Mulloy A, Toma A, Patel A. To err is human: A case-based review of cognitive bias and its role in clinical decision making. Canadian Journal of General Internal Medicine. 2017;12(2).

54. Phua DH, Tan NC. Cognitive aspect of diagnostic errors. Ann Acad Med Singapore. 2013;42(1):33-41.

55. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC medical informatics and decision making. 2016;16(1):1-14.

56. Vickrey BG, Samuels MA, Ropper AH. How neurologists think: a cognitive psychology perspective on missed diagnoses. Annals of neurology. 2010;67(4):425-33.

57. Kempainen RR, Migeon MB, Wolf FM. Understanding our mistakes: a primer on errors in clinical reasoning. Medical teacher. 2003;25(2):177-81.

58. Croskerry P. Achieving quality in clinical decision making: cognitive strategies and detection of bias. Academic Emergency Medicine. 2002;9(11):1184-204.

59. O'Sullivan ED, Schofield S. Cognitive bias in clinical medicine. Journal of the Royal College of Physicians of Edinburgh. 2018;48(3):225-31.

60. Schmidt HG, Mamede S, Van Den Berge K, Van Gog T, Van Saase JLCM, Rikers RMJP. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. Academic Medicine. 2014;89(2):285-91.

61. Sherbino J, Dore KL, Wood TJ, Young ME, Gaissmaier W, Kreuger S, et al. The relationship between response time and diagnostic accuracy. Academic Medicine. 2012;87(6):785-91.

62. Norman G. Dual processing and diagnostic errors. Advances in Health Sciences Education. 2009;14(1):37-49.

63. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine. 2017;92(1):23-30.

64. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annual review of psychology. 2011;62(1):451-82.

65. Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: models of bounded rationality. Psychological review. 1996;103(4):650.

66. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. Dialogues in clinical neuroscience. 2022.

67. McLaughlin K, Eva KW, Norman GR. Reexamining our bias against heuristics. Advances in Health Sciences Education. 2014;19(3):457-64.

68. Norman G, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. Medical education. 2007;41(12):1140-5.

69. Hess BJ, Lipner RS, Thompson V, Holmboe ES, Graber ML. Blink or think: can further reflection improve initial diagnostic impressions? Academic Medicine. 2015;90(1):112-8.

70. Brush Jr JE, Sherbino J, Norman GR. How expert clinicians intuitively recognize a medical diagnosis. The American journal of medicine. 2017;130(6):629-34.

71. Eva KW, Norman GR. Heuristics and biases− a biased perspective on clinical reasoning. Medical education. 2005;39(9):870-2.

72. Schmidt HG, Boshuizen H. On acquiring expertise in medicine. Educational psychology review. 1993;5(3):205-21.

73. Nosofsky RM. Exemplars, prototypes, and similarity rules. Essays in honor of William K Estes. 1992;1:149-67.

74. Logan GD. Toward an instance theory of automatization. Psychological review. 1988;95(4):492.

75. Aczel B, Bago B, Szollosi A, Foldes A, Lukacs B. Is it time for studying real-life debiasing? Evaluation of the effectiveness of an analogical intervention technique. Frontiers in psychology. 2015;6:1120.

76. Aron AR. Progress in executive-function research: From tasks to functions to regions to networks. Current directions in psychological science. 2008;17(2):124-9.

77. Best JR, Miller PH, Jones LL. Executive functions after age 5: Changes and correlates. Developmental review. 2009;29(3):180-200.

78. Zelazo PD. The development of conscious control in childhood. Trends in cognitive sciences. 2004;8(1):12-7.

79. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ quality & safety. 2016;25(10):808-20.

80. Mamede S, Schmidt HG. Reflection in medical diagnosis: a literature review. Health Professions Education. 2017;3(1):15-25.

81. Griffith PB, Doherty C, Smeltzer SC, Mariani B. Education initiatives in cognitive debiasing to improve diagnostic accuracy in student providers: A scoping review. Journal of the American Association of Nurse Practitioners. 2021;33(11):862-71.

82. Cooper N, Bartlett M, Gay S, Hammond A, Lillicrap M, Matthan J, et al. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Medical Teacher. 2021;43(2):152-9.

83. Graber ML. Educational strategies to reduce diagnostic error: can you teach this stuff? Advances in health sciences education. 2009;14(1):63-9.

84. Schiff GD. Minimizing diagnostic error: the importance of follow-up and feedback. The American journal of medicine. 2008;121(5):S38-S42.

85. Branson CF, Williams M, Chan TM, Graber ML, Lane KP, Grieser S, et al. Improving diagnostic performance through feedback: the Diagnosis Learning Cycle. BMJ quality & safety. 2021;30(12):1002-9.

86. Kuhn J, van den Berg P, Mamede S, Zwaan L, Bindels P, van Gog T. Improving medical residents' self-assessment of their diagnostic accuracy: does feedback help? Advances in Health Sciences Education. 2021:1-12.

87. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Medical education. 2008;42(5):468-75.

88. Sibbald M, de Bruin ABH, van Merrienboer JJG. Checklists improve experts' diagnostic decisions. Medical education. 2013;47(3):301-8.

89. Sibbald M, Sherbino J, Ilgen JS, Zwaan L, Blissett S, Monteiro S, et al. Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. Advances in Health Sciences Education. 2019;24(3):427-40.

90. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. bmj. 2020;370.

91. Dave N, Bui S, Morgan C, Hickey S, Paul CL. Interventions targeted at reducing diagnostic error: systematic review. BMJ quality & safety. 2022;31(4):297-307.

92. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: Still no easy answers. Western Journal of Emergency Medicine. 2020;21(6):125.

93. Sherbino J, Kulasegaram K, Howey E, Norman G. Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. Canadian Journal of Emergency Medicine. 2014;16(1):34-40.

94. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. Medical teacher. 2013;35(6):e1218-e29.

95. Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. Medical Teacher. 2019;41(5):517-24.

96. Richmond A, Cooper N, Gay S, Atiomo W, Patel R. The student is key: a realist review of educational interventions to develop analytical and non-analytical clinical reasoning ability. Medical Education. 2020;54(8):709-19.

97. Zwaan L, Staal J. Evidence on Use of Clinical Reasoning Checklists for Diagnostic Error Reduction. AHRQ Papers on Diagnostic Safety Topics [Internet]. 2020; (3).

98. Chamberland M, Setrakian J, St-Onge C, Bergeron L, Mamede S, Schmidt HG. Does providing the correct diagnosis as feedback after self-explanation improve medical students diagnostic performance? BMC medical education. 2019;19(1):1-8.

99. Janssen EM, Velinga SB, de Neys W, Van Gog T. Recognizing biased reasoning: Conflict detection during decision-making and decision-evaluation. Acta Psychologica. 2021;217:103322.

100. Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nature neuroscience. 2010;13(10):1292-8.

COGNITIVE PROCESSES UNDERLYING DIAGNOSTIC ERRORS

II

CHAPTER 2

The relationship between time to diagnose and diagnostic accuracy among internal medicine residents: a randomized experiment

Staal, J., Alsma, J., Mamede, S., Olson, A.P.J., Prins-van Gilst, G., Geerlings, S.E., Plesac, M., Sundberg, M.A., Frens, M.A., Schmidt, H.G., Van den Broek, W.W., & Zwaan, L.

Published in *BMC Medical Education* 2021; 21: 1-9.

DOI: 10.1186/s12909-021-02671-2.

Abstract

Background: Diagnostic errors have been attributed to cognitive biases (reasoning shortcuts), which are thought to result from fast reasoning. Suggested solutions include slowing down the reasoning process. However, slower reasoning is not necessarily more accurate than faster reasoning. In this study, we studied the relationship between time to diagnose and diagnostic accuracy.

Methods: We conducted a multi-center within-subjects experiment where we prospectively induced availability bias (using Mamede et al.'s methodology) in 117 internal medicine residents. Subsequently, residents diagnosed cases that resembled those bias cases but had another correct diagnosis. We determined whether residents were correct, incorrect due to bias (i.e. they provided the diagnosis induced by availability bias) or due to other causes (i.e. they provided another incorrect diagnosis) and compared time to diagnose.

Results: We did not successfully induce bias.: no significant effect of availability bias was found. Therefore, we compared correct diagnoses to all incorrect diagnoses. Residents reached correct diagnoses faster than incorrect diagnoses (115s vs. 129s, $p < .001$). Exploratory analyses of cases where bias was induced showed a trend of time to diagnose for bias diagnoses to be more similar to correct diagnoses (115s vs 115s, $p = .971$) than to other errors (115s vs 136s, $p = .082$).

Conclusions: We showed that correct diagnoses were made faster than incorrect diagnoses, even within subjects. Errors due to availability bias may be different: exploratory analyses suggest a trend that biased cases were diagnosed faster than incorrect diagnoses. The hypothesis that fast reasoning leads to diagnostic errors should be revisited, but more research into the characteristics of cognitive biases is important because they may be different from other causes of diagnostic errors.

Keywords: cognitive bias, decision making, diagnostic error, patient safety

Background

Diagnostic errors are a serious patient safety concern that went largely unrecognized (1) until the National Academies of Sciences, Engineering, and Medicine (NASEM) published the report 'Improving Diagnosis in Healthcare' in 2015.(2) Understanding the underlying causes of diagnostic errors is a crucial step towards reducing those errors. Research findings of a variety of studies (3-6) have led to the consensus that cognitive flaws are a major cause of diagnostic errors.(7-12) However, researchers disagree about the type of cognitive flaw that is the main cause.(13, 14) The discussion is centered around the question whether cognitive biases or other cognitive flaws, such as knowledge deficits, are the most common cause of error.(13). In the diagnostic error literature, a common explanation is that errors are caused by cognitive biases due to fast reasoning and that slowing down and taking more time can prevent these errors.(8, 10, 15, 16) Contributing to clarifying the influence of time taken to diagnosis on the likelihood of making mistakes is of the utmost importance in determining what strategies may be effective in decreasing diagnostic errors.

Diagnostic reasoning is frequently described by dual process theory (DPT), an influential theory on decision-making in the field of psychology.(17, 18) DPT describes that reasoning consists of two systems, called System 1 and System 2.(18) System 1 relies on heuristics (mental shortcuts) and on fast and automatic reasoning. We are only conscious of the final product of System 1 reasoning and therefore it is called non-analytical reasoning. On the other hand, System 2 is slow, sequential, and allows for deliberate reasoning, although the system is limited by the capacity of our working memory. System 2 reasoning is regulated: we are conscious of both the process and the result, and therefore it is called analytical reasoning.(16-19) The separation of System 1 and System 2 is primarily relevant in theory, as non-analytic and analytic processes tend to blend together in practice.

The shortcuts in non-analytical reasoning can introduce cognitive biases (predispositions to think in a way that leads to systematic failures in judgement (17)). An example is availability bias, where people rely on examples that come to mind easily; e.g. clinicians are more likely to diagnose a patient with the same condition as in a recently seen patient.(6) Based on this rationale, non-analytical (and therefore, fast) reasoning is purported to be a major cause of bias-induced diagnostic errors.(7, 12, 15, 16, 20, 21)

To prevent such errors, many interventions stimulate slower, more analytical reasoning. However, this idea is contradicted by the studies of Sherbino et al.(22) and Norman et al.,(23) who showed that faster diagnoses were more often or just as often correct as slower diagnoses. This implies that fast (or faster) reasoning cannot be equated to faulty reasoning and actually may lead to excellent diagnostic performance. It has also been suggested to only slow down when necessary to make sure that correct diagnostic processes are not disrupted; however, it seems that clinicians often do not know when they would require extra time or help. This was shown in a study by Meyer et al.(24) where clinicians' confidence and their intention to request for help (e.g. from a colleague) did not correctly reflect their diagnostic accuracy.

Despite these arguments, diagnostic errors are still primarily attributed to fast diagnostic reasoning (10, 15, 16, 20, 21) and the overall view of diagnostic errors has not shifted much. An important limitation of the studies showing that faster diagnoses were just as often correct as slower diagnoses is that they used a between subjects design and therefore can alternatively be explained by assuming that faster participants were just better diagnosticians than slower participants.(22, 23) Additionally, these studies only focused on correct versus incorrect diagnoses and did not examine how bias-induced diagnoses related to reaction times.

To determine how time to diagnose relates to diagnostic error within subjects, we induced availability bias (by using Mamede et al.'s methodological procedure for bias-induction (6)). First residents evaluated the accuracy of simple cases and subsequently diagnosed a similar case with a different diagnosis. If they would provide the same diagnosis as they had evaluated before, this was considered an error due to availability bias. If they provided another incorrect answer, this was considered a diagnostic error due to other reasons. We compared their time to diagnose and confidence when they were correct, incorrect due to bias or incorrect for other reasons. Furthermore, we explored perceived case complexity and mental effort invested in diagnosis, and determined residents' confidence-accuracy calibration and resource use to study how these measures would be affected by bias, the effect of which was not examined by Meyer et al. (24) .

We expected to replicate Sherbino et al. (22) and Norman et al. (23),'s findings, but now in a within-subjects design, and to show that faster reasoning was not necessarily related to diagnostic errors. Specifically, we expected that both bias-induced diagnostic errors and correct diagnoses would be diagnosed faster than other errors. Furthermore, we expected that confidence would be lower for both bias errors and other errors than for correct diagnoses.

Methods

Design

The study was a two-phase computer-based experiment with a within-subjects design (Fig. 1),based on a study by Mamede et al. (6) where availability bias was induced. All methods were carried out in accordance with the relevant guidelines and regulations. The experiment consisted of two phases, with no time-lag between the phases:

1) Bias phase: Residents were randomly divided into two groups, who each evaluated 6 clinical cases with a provisional diagnosis. Both groups saw four filler cases (cases meant to create a diverse case mix and to distract from the bias cases) and two biasing cases. The biasing cases were different for each group: residents in group 1 saw biasing cases A (pneumonia) and B (hypercapnia) and residents in group 2 saw biasing cases C (Hodgkin's lymphoma) and D (ileus) (Fig. 1). This way, the two groups were biased towards different cases and acted as each other's controls in the test phase. Additionally, creating two groups allowed us to correct for case complexity and increase generalizability.

2) Test phase: Residents diagnosed 8 clinical cases. Half of the cases were similar to the biasing cases shown to group 1; the other half were similar to the biasing cases shown to group 2 (Fig. 1). Thus, residents diagnosed four cases for which they saw the similar case in Phase 1 and four for which they did not, resulting in four cases that were exposed to bias and four cases that were not exposed to bias for each resident.

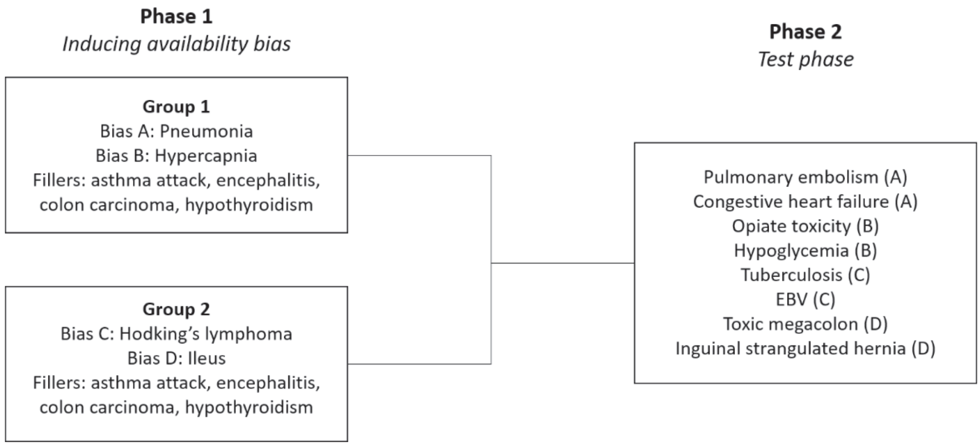

Phase 1
Inducing availability bias
Phase 2
Test phase
Group 1
Bias A: Pneumonia
Bias B: Hypercapnia
Fillers: asthma attack, encephalitis, colon carcinoma, hypothyroidism
Group 2
Bias C: Hodking's lymphoma
Bias D: Ileus
Fillers: asthma attack, encephalitis, colon carcinoma, hypothyroidism
Pulmonary embolism (A)
Congestive heart failure (A)
Opiate toxicity (B)
Hypoglycemia (B)
Tuberculosis (C)
EBV (C)
Toxic megacolon (D)
Inguinal strangulated hernia (D)

Figure 1. Study design and clinical cases shown in each phase.

Participants

In total, 117 Internal Medicine residents in their 1st to 6th year of training participated (Table 1). Group 1 and 2 consisted of 57 residents and 60 residents respectively. Residents were in training at one of the three participating academic medical centers: two in the Netherlands and one in the USA. Residents from the Dutch academic centers were recruited during their monthly educational day; residents from the American academic center were recruited individually (by APJO, MAS, and MP).

Sample size was prospectively estimated in G-power (25). We calculated sample size for an ANCOVA (analysis of covariance) with a medium effect size, a power of 80%, an α of 0.05, 2 groups and 2 covariates. This estimation indicated that 128 participants would be required.

Table 1. *Participant demographics.*

Hospital	***N***	***Age (SD)***	***N(%) Female***	***Years as resident (SD)***
Erasmus Medical Centre (Rotterdam, Netherlands)	26	31 (3.5)	14 (54%)	2.2 (1.1)
University Medical Center Amsterdam (Amsterdam, Netherlands)	69	35 (2.5)	47 (69%)	2.5 (1.3)
University of Minnesota (Minnesota, U.S.A.)	23	29 (2.0)	12 (52%)	1.6 (1.1)

Materials

Sixteen written cases (Fig. 1) were developed by one internist and diagnosed and confirmed by another internist who was not aware of the diagnoses of the first internist (JA and GP). Cases consisted of a short history of a fictional patient, combined with test results (Appendix A). Cases were designed in sets with the same presenting symptom, but each case had a different final diagnosis. Cases in each set were matched by superficial details such as patient gender and age. All cases were piloted (N = 10) to ensure appropriate level of difficulty. All materials were available in Dutch and English. An online questionnaire (Appendix B) was prepared in Qualtrics (an online survey tool).

Procedure

Residents received an information letter and were asked to sign informed consent. They were told that the goal of the study was to examine information processing during diagnosis when evaluating diagnoses, and when diagnosing cases themselves.

In the first phase (bias induction), residents estimated (on a scale from 0-100%) the likelihood that a provided provisional diagnosis was correct. All diagnoses were in fact correct. This was followed by a test phase in which residents were given 8 clinical cases for which they had to provide the most likely diagnosis as a free text response.

After diagnosing all cases, residents were shown the history of each case again and were then asked to provide for each case the confidence in their diagnosis, their perceived complexity, and their invested mental effort in diagnosing the case. We also measured residents' confidence-accuracy calibration by correlating their average confidence and accuracy ratings. Lastly, we asked if they had wanted to use additional resources to diagnose the case.

Finally, we provided feedback by showing the cases, the diagnosis the resident had provided, and the correct diagnosis. For cases with a provisional diagnosis, we showed the residents' indicated likelihood of the diagnosis being correct and told them that all provisional diagnoses had been correct.

Outcome measures

The independent variable was the type of bias exposure: participants were biased to either cases A/B or to cases C/D (Fig. 1). The main dependent variable was the final diagnosis, which was defined as correct, bias error, or other error. A bias error occurred when the diagnosis from Phase 1 was given; other errors occurred when another incorrect diagnosis was given(other error). A diagnosis could only be defined as a bias error if residents saw the corresponding bias case in Phase 1 of the study; otherwise their diagnosis was labelled ”other error”. Additionally, we calculated the frequency with which residents mentioned the bias diagnosis of a case in the control condition (when they did not see the bias case), which had to be significantly lower than in the bias condition. Otherwise, the 'bias' diagnosis could also be a probable differential diagnosis, which prevented us from concluding the error was made due to bias. This was scored by two internists (JA and GP), who independently assessed and assigned a score to all diagnoses. A score of 0 was given for incorrect diagnoses; a score of 0.5 was given for partially correct diagnoses (e.g. the participant answered sepsis, but the diagnosis was pneumonia with sepsis); a score of 1 was given for fully correct diagnoses. After the first ratings, their responses were compared and discrepancies were resolved through discussion.

We measured time to diagnose in seconds spent on each clinical case and confidence on a scale from 0-100%. We additionally measured case complexity (26) and mental effort (27, 28), also on a scale from 0-100%. The confidence-accuracy calibration was expressed by a goodness-of-fit (R^2) measure through a scatterplot of average confidence and accuracy per resident. Finally, resource use was measured as the percentage of residents who wanted to use extra resources.

Statistical analysis

First, we examined whether the bias induction was successful by comparing if the frequency with which residents mentioned the bias diagnosis in the control condition (when they did not see the bias case) was significantly lower than in the bias condition. This determined which comparisons we could analyze.

We then calculated the mean for time to diagnose, confidence, complexity, and mental effort over all cases for each error type. The time to diagnose variable was scaled prior to the calculation of the mean to correct for differences due to case length. This was done by calculating a grand mean from the individual means of all 8 cases and subtracting the grand mean from the individual means for time to diagnose. This indicated the number by which every individual time would have to be corrected and resulted in the scaled times to diagnose. Furthermore, per analysis we excluded residents for whom a mean could not be calculated due to missing values.

Statistical tests. We compared residents' correct diagnoses, bias errors, and other errors on time to diagnose, confidence, complexity, mental effort, using two-sided repeated measures t-tests. The originally planned ANCOVA was not performed because we did not induce bias. We used three tests to compare these types of diagnoses instead of one encompassing test because such a test would unnecessarily exclude residents due to listwise exclusion. For each instance of multiple testing, the alpha level was corrected to $\alpha = .017$ (.05/3) using a Bonferroni correction. Analyses were performed in Spyder (Python 3.7). Additionally, for each significant result we calculated the Cohen's d (29) and the 95% confidence interval around the mean difference. The relation between resource use and diagnostic accuracy was evaluated using a repeated measures binomial logistic regression in Rstudio (version 1.2.5003), for which we calculated the odds ratio.

Results

Bias induction

A one-way analysis of variance (ANOVA) showed no significant difference in the frequency with which the bias diagnosis was given on cases that were exposed to bias and cases that were not exposed to bias ($p > .05$). Additionally, out of the 117 residents, residents infrequently mentioned the bias diagnosis (0-20 times for any case). Because bias induction was unsuccessful, we could not analyze bias error as a separate error type. Therefore, we merged bias errors and other errors into one category in the main analyses.

Main analyses

Residents were faster to reach a correct diagnosis than an incorrect diagnosis, $t(112) = 4.51$, $p < .001$, 95% CI [4.11 23.89], $d = 0.37$ (Fig. 2). Residents' confidence was higher for correct diagnoses than for incorrect diagnoses, $t(112) = 8.75$, $p < .001$, 95% CI [8.48 15.52], $d = 0.89$ (Fig. 3).

Exploratory analyses

Case complexity and mental effort

Residents found correct diagnoses less complex than incorrect diagnoses, $t(113) = 7.51$, $p < .001$, 95% CI [5.49 12.51], $d = 0.67$, and invested less effort in correct diagnoses as opposed to incorrect diagnoses, $t(113) = 8.52$, $p < .001$, 95% CI [7.23 14.77], $d = 0.81$ (Fig. 3).

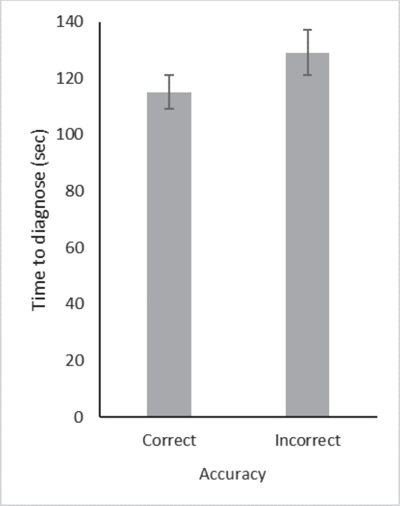


Figure 2. Mean time to diagnose (adjusted for case length) for correct and all incorrect diagnoses (N = 113). Bars indicate the 95% confidence interval.

Confidence-accuracy calibration

Residents' confidence-accuracy calibration trend line (Fig. 4) for average accuracy and confidence achieved a goodness-of-fit of $R^2 = 0.03$, indicating that most residents were not well calibrated and that confidence-accuracy calibration varied widely between residents.

Resources

Residents indicated they wanted to consult one or more additional resources during diagnosis in 63% of the cases. We performed a repeated measures binomial logistic regression in RStudio, using the glmer package (30), to assess whether diagnostic accuracy was a predictor for resource use. We corrected for participant and case repetitions. The model showed no significant difference in how often residents indicated they wanted to use resources when they were correct (59%) versus when they were incorrect (68%), b = -.204, SE = 0.18, OR = 0.82, $p > .05$.

Bias diagnoses

Despite the overall unsuccessful bias induction, in several cases (opiate intoxication, hypoglycemia, tuberculosis, toxic megacolon) the bias diagnosis was given more frequently (although not significantly) on cases that were exposed to bias. Average time to diagnose and confidence (Table 2) were calculated in the same way as for correct diagnoses and other errors. We performed independent measures t-tests for these analyses, because the low numbers of bias responses would cause many data points to be excluded in a repeated measures test.

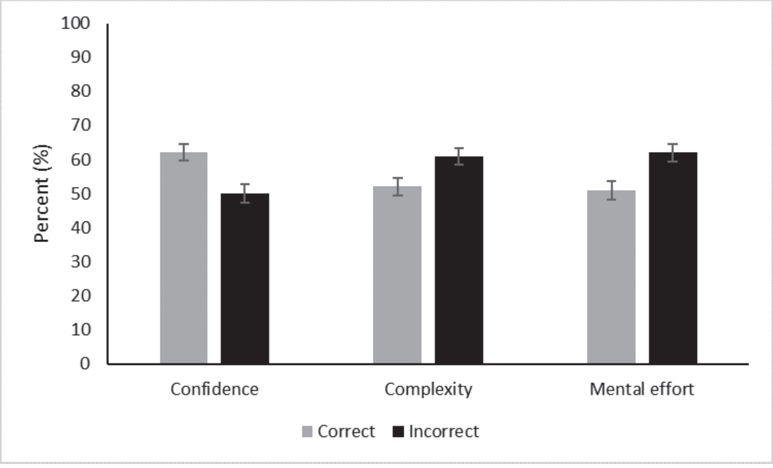


Figure 3. Mean confidence, complexity, and mental effort for correct and all incorrect diagnoses (N = 113). Bars indicate the 95% confidence interval.

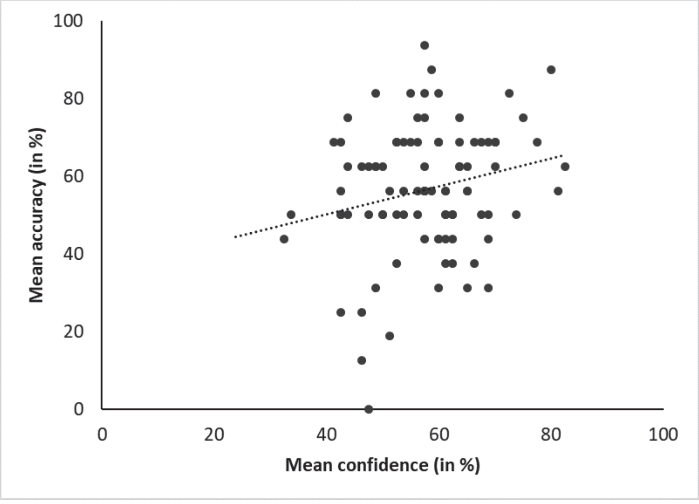


Figure 4. The relationship (linear trend line) between mean accuracy and mean confidence over all cases.

Time to diagnose did not differ between bias errors and correct diagnoses, t(122) = -0.03, p = .971, but a trend was present towards significance showing that bias errors were diagnosed faster than other errors, t(92) = 1.75, p = .082. Conversely, confidence showed a trend towards significance for residents to be less confident in bias errors than in correct diagnoses, t(122) = 2.07, p = .041, 95% CI [1.00 17.00], but no difference in confidence between bias errors and other errors, t(92) = 1.53, p = .130).

Table 2. *Descriptive statistics for the time to diagnose (adjusted for case length) and confidence.*

	Time (sec)			***Confidence (%)***		
Error type	M	SD	95% CI	M	SD	95% CI
Correct (*n* = 96)	115	49.89	105-125	63	18.58	59-66
Bias (*n* = 28)	115	45.23	98-133	54	19.28	47-62
Other (*n* = 64)	136	54.85	122-150	47	20.37	42-52

Discussion

In this study we examined how time to diagnose related to diagnostic error. Because bias induction was unsuccessful we could not analyse bias errors and other errors separately. In line with our hypotheses, we found that even within subjects, residents took less time when they were correct and had more confidence in correct diagnoses. With this increased confidence, we also saw residents found correct cases were less complex and invested less effort in correct diagnoses. Additional analyses showed that residents' confidence-accuracy calibration was poor and that accuracy did not influence how often residents requested resources. Further exploratory analyses of the bias errors were performed on the cases with a (non-significant) effect of bias. Although the results should be interpreted with caution, it was interesting that the results were in line with our hypotheses about correct diagnoses versus bias errors. Residents took equal amounts of time to diagnose correct and bias diagnoses (Table 2) and we saw a trend for bias errors to be reached faster than other errors. Contrary to our hypotheses, we found that confidence was similar between bias errors and other errors (Table 2) and that there was a trend for confidence to be lower for bias errors than for correct diagnoses.

Our findings regarding time to diagnose support and expand on the work of Norman et al. (23) and Sherbino et al.,(22) who showed that physicians who diagnosed cases quickly were equally or more often correct than those who diagnosed cases more slowly. We have now shown that this applies on an individual level as well, i.e. physicians were faster when they were correct compared to incorrect, and that this cannot just be attributed to faster physicians being better diagnosticians. Further interesting insights come from the exploratory analyses where bias-induced errors showed a trend to be diagnosed faster than incorrect diagnoses (Table 2). These fast reaction times suggest that bias errors might differ from other types of errors.

This study and others show that fast diagnoses are not necessarily wrong and that correct diagnoses are not necessarily slow. The difference in time to diagnose between correct and incorrect diagnoses could partially be explained by the differences in relative difficulty of the cases: physicians could find some cases easier than other cases and might solve those cases quickly and correctly. The cases where they had more doubts would take longer. Although it is likely this occurred in some cases, the fact that residents were poor judges of their performance, which was evidenced by their poor confidence-accuracy calibration and their reluctance to use resources when necessary, speaks against this explanation for all cases taken together. This makes it unlikely that they consistently sped up or slowed down for cases where they were correct or incorrect. It is therefore less likely that time to diagnose for correct and incorrect cases can on average fully be explained by differences in case difficulty. Other causes of diagnostic errors need to be explored to gain better understanding of the diagnostic process. One such example would be knowledge deficits,(13) which have also been shown to reduce cognitive biases.(31)

Although our finding that correct diagnoses are made faster than incorrect diagnoses is not novel in itself, there is still a need to demonstrate and emphasize this finding: partially due to the pervasive notion that fast reasoning is primarily a cause for errors despite the findings of previous studies, and partially due to the limitations of these previous studies, which are in part overcome by the within-subjects design of the current study. Moreover, even though it seems logical that fast diagnosis is also a crucial part of the diagnostic process, many interventions focused on reducing errors in diagnostic reasoning still recommend stimulating analytical reasoning and slowing down the diagnostic reasoning process. Research that tested such interventions and educational strategies (such as the SLOW tool (32), general debiasing checklists (33) and cognitive forcing training (34)) did not show improved diagnostic accuracy.(35) Therefore, these interventions could result in harm because they would target both bias errors and correct diagnoses. It could be that reconsidering correctly diagnosed cases would result in more diagnostic tests and consequently overdiagnosis, which could also be harmful for patients (36).

The obvious solution would be to slow down only when necessary. However, this study confirms Meyer et al.,(24)'s finding that clinicians' confidence is not well calibrated with accuracy and they do not ask for additional resources when necessary, whether they are residents or experts. Further, correct diagnoses and bias errors were similar, which makes it hard to differentiate between them. This suggests it would be difficult to use the concept of fast versus slow reasoning to detect diagnostic errors. Additional research is necessary to identify means to improve clinicians' calibration, for example through feedback.(37)

This study has several strengths and limitations. Strengths are that our study is a multi-center study with a randomized within-subjects design, which made the residents their own control and reduced variance between subjects. We additionally induced bias prospectively instead of assessing it retrospectively, which avoids issues like hindsight bias.(38) However, not all residents were vulnerable to bias and because we ended up with a small number of bias errors we were unable to replicate the induction of availability bias in Mamede et al.'s study.(6) This limited the analyses we could perform, because residents could only be biased to 4 cases at most, so the computed means for time to diagnose and confidence sometimes contain only one value for a resident, making the exploratory analyses less robust. However, we thought it best to be strict in our definition and selection of bias responses in order to approximate errors due to bias as closely as possible. It is unclear why bias induction was unsuccessful. One explanation is that the cases we developed to induce bias had many possible underlying diseases: this could have resulted in there being many possible differential diagnoses, which may have induced some analytical reasoning.

A further limitation is that our sample included a relatively large range of years of experience. It could be that the effects of time to diagnose and confidence are different for different levels of experience. This should be studied in a follow-up study. A final limitation is the use of written case vignettes: these limit the ecological validity of the study and do not allow residents to look up extra information while diagnosing the case. However, written cases provided the best way to prospectively induce bias and have been shown to offer a good approximation of real clinician performance.(39, 40)

In conclusion, this study shows that correct diagnoses are reached faster than incorrect diagnoses and that this is not due to faster physicians being better diagnosticians. This indicates that fast diagnostic reasoning underlies correct diagnoses and does not necessarily lead to diagnostic errors. Exploratory analyses indicate that this might be different for diagnostic errors caused by cognitive biases, although more research into the characteristics of cognitive biases would be necessary to determine this. Both diagnostic error interventions and educational strategies should not promote focusing on slowing down to reduce errors and the common view of fast reasoning primarily being a cause for errors should be reconsidered.

Acknowledgements

We would like to thank the residents who participated in the study.

References

1. Wachter RM. Why diagnostic errors don't get any respect—and what can be done about them. Health Affairs. 2010;29(9):1605-10.
2. National Academies of Sciences E, and Medicine. Improving diagnosis in health care: National Academies Press; 2015.
3. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
4. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DRM. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. Academic Medicine. 2012;87(2):149-56.
5. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.
6. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203. Available from: http://dx.doi.org/10.1001/jama.2010.1276.
7. Phua DH, Tan NC. Cognitive aspect of diagnostic errors. Ann Acad Med Singapore. 2013;42(1):33-41.
8. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education. 2009;14(1):27-35.
9. Croskerry P. Diagnostic failure: a cognitive and affective approach. Advances in patient safety: from research to implementation. 2005;2:241-54. Available from: 10.1037/e448242006-001.
10. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic medicine. 2003;78(8):775-80.
11. Croskerry P. The cognitive imperative thinking about how we think. Academic Emergency Medicine. 2000;7(11):1223-31.
12. Elia F, Apra F, Verhovez A, Crupi V. "First, know thyself": cognition and error in medicine. Acta Diabetologica. 2016;53(2):169-75.
13. Norman, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine. 2017;92(1):23-30. Available from: http://dx.doi.org/10.1097/ACM.0000000000001421.
14. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. Journal of Evaluation in Clinical Practice. 2018;24(3):666-73.
15. Mithoowani S, Mulloy A, Toma A, Patel A. To err is human: A case-based review of cognitive bias and its role in clinical decision making. Canadian Journal of General Internal Medicine. 2017;12(2).
16. Frankish K. Dual-process and dual-system theories of reasoning. Philosophy Compass. 2010;5(10):914-26.
17. Kahneman D, Egan P. Thinking, fast and slow: Farrar, Straus and Giroux New York; 2011.
18. Evans JSBT. In two minds: dual-process accounts of reasoning. Trends in cognitive sciences. 2003;7(10):454-9.
19. Smith ER, DeCoster J. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and social psychology review. 2000;4(2):108-31.

20. Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Ann Emerg Med. 2003;41(1):110-20. Available from: 10.1067/mem.2003.22.

21. Elstein AS. Heuristics and biases: selected errors in clinical reasoning. Academic Medicine. 1999. Available from: http://dx.doi.org/10.1097/00001888-199907000-00012.

22. Sherbino J, Dore KL, Wood TJ, Young ME, Gaissmaier W, Kreuger S, et al. The relationship between response time and diagnostic accuracy. Academic Medicine. 2012;87(6):785-91.

23. Norman, Sherbino J, Dore K, Wood T, Young M, Gaissmaier W, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. Academic Medicine. 2014;89(2):277-84. Available from: http://dx.doi.org/10.1097/ACM.0000000000000105.

24. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine. 2013;173(21):1952-8.

25. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91.

26. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Medical education. 2008;42(5):468-75.

27. Robinson MD, Johnson JT, Herndon F. Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. Journal of Applied Psychology. 1997;82(3):416. Available from: http://dx.doi.org/10.1037/0021-9010.82.3.416.

28. Franssens S, De Neys W. The effortless nature of conflict detection during thinking. Thinking & Reasoning. 2009;15(2):105-28. Available from: http://dx.doi.org/10.1080/13546780802711185.

29. Cohen J. Statistical power analysis for the behavioral sciences, 2nd edn. Á/L. Erbaum Press, Hillsdale, NJ, USA; 1988.

30. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:14065823. 2014.

31. Mamede S, de Carvalho-Filho MA, de Faria RMD, Franci D, Nunes MdPT, Ribeiro LMC, et al. 'Immunising'physicians against availability bias in diagnostic reasoning: a randomised controlled experiment. BMJ Quality & Safety. 2020. Available from: http://dx.doi.org/10.1136/bmjqs-2019-010079.

32. O'Sullivan ED, Schofield SJ. A cognitive forcing tool to mitigate cognitive bias–a randomised control trial. BMC medical education. 2019;19(1):12.

33. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. Medical teacher. 2013;35(6):e1218-e29.

34. Sherbino J, Kulasegaram K, Howey E, Norman G. Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. Canadian Journal of Emergency Medicine. 2014;16(1):34-40.

35. Graber ML, Kissam S, Payne VL, Meyer AND, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ Qual Saf. 2012;21(7):535-57.

36. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis. 2015;2(2):97-103.

37. Zwaan L, Hautz WE. Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. BMJ Publishing Group Ltd; 2019.

38. Zwaan L, Monteiro S, Sherbino J, Ilgen J, Howey B, Norman G. Is bias in the eye of the beholder? A vignette study to assess recognition of cognitive biases in clinical case workups. BMJ quality & safety. 2017;26(2):104-10.

39. Mohan D, Fischhoff B, Farris C, Switzer GE, Rosengart MR, Yealy DM, et al. Validating a vignette-based instrument to study physician decision making in trauma triage. Medical decision making. 2014;34(2):242-52. Available from: http://dx.doi.org/10.1177/0272989X13508007.

40. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. Jama. 2000;283(13):1715-22. Available from: http://dx.doi.org/10.1001/jama.283.13.1715.

Appendix A – Example of a clinical case (biasing case A)

History of present illness

A 68-year old man is referred to the emergency room because of shortness of breath. He has had flu symptoms for two weeks. His wife has had the same symptoms, and she has since recovered well. The last two days he has started coughing up sputum. The sputum was green at first, but since this morning it has been rust brown. In addition, he suddenly became much sicker today. His wife thinks he is less alert than usual. Deep breathing causes pain in his left chest half. He does not smoke, and he drinks alcohol in moderation.

Past medical history

Hypertension – managed by PCP with Perindopril.

Hypercholesterolemia – treated with a statin.

Physical examination

Temperature 101.3°F. Oxygen saturation 94% (94-100%) on room air. Blood pressure 120/80 mmHg. Pulse 95 bpm. Moderately sick appearing man. JVD: not elevated.

Cardiac exam: normal auscultation without murmurs.

Lung exam: auscultation with crepitation in left posterior basal lung field.

The remainder of the physical examination was normal.

Laboratory testing

Leukocytes 12x109/L (<10 x 109/L), CRP 76 mg/L (<10 mg/L). Electrolytes and liver chemistry are within normal ranges.

Additional testing

CXR: good expiratory view. Left basal consolidation. No pleural effusion. ECG: sinus rhythm.

Diagnosis: pneumonia.

Appendix B – Survey questions

Diagnosing cases:

First, the participants will be asked to evaluate 6 clinical cases. For each case the working diagnosis has been provided. For each case the participants is asked the following questions:

1. Indicate in percentages (%) the probability that the working diagnosis is correct:
2. Indicate how difficult it was to diagnose this case: (scale from 0-10, easy to difficult)

Then the participants will see 8 clinical cases without a working diagnosis. They are asked to diagnose each case themselves. For each case the participants will be asked the following questions:

1. What is the most likely diagnosis?
2. How confident are you in your diagnosis? (scale from 0-10, little confidence to a lot of confidence)
3. How difficult did you find it to diagnose the case? (scale from 0-10, easy to difficult)

Relevant personal information:

1. How old are you?
2. What is your gender?
3. Are you dyslexic?
4. In which year did you finish your study in Medicine?
5. How many years have you been a resident of internal medicine?
6. Which subspecialism do you want to, or will you, practice after residency?
7. How many years of experience do you have in clinical practice?
8. How many years of this experience did you acquire outside of your residency?

Feedback on the study:

1. What do you think the goal of this study is?

Finally, we ask the participants to provide some general feedback (if they have more to remark than what was asked in the previous questions).

1. Describe additional points that caught your attention during the study (e.g. mistakes in terminology, difficulty of the cases etc.) and, if necessary, elaborate on the answers provided on the previous feedback points.

The last optional question is whether participants would like to receive information about the study and its outcomes (and their own performance) when study 1 and 2 have concluded. If so, they can leave their email address.

CHAPTER 3

Does a suggested diagnosis in a general practitioners' referral letter impact diagnostic reasoning: an experimental study

Staal, J.*, Speelman, M.*, Brand, R., Alsma, J., Zwaan, L.

*J. Staal and M. Speelman are shared first authors; they contributed equally to the work.

Published in *BMC Medical Education* 2022; 22:256.

DOI: 10.1186/s12909-022-03325-7

Abstract

Background: Diagnostic errors are a major cause of preventable patient harm. Studies suggest that presenting inaccurate diagnostic suggestions can cause errors in physicians' diagnostic reasoning processes. It is common practice for general practitioners (GPs) to suggest a diagnosis when referring a patient to secondary care. However, it remains unclear via which underlying processes this practice can impact diagnostic performance. This study therefore examined the effect of a diagnostic suggestion in a GP's referral letter to the emergency department on the diagnostic performance of medical interns.

Methods: Medical interns diagnosed six clinical cases formatted as GP referral letters in a randomized within-subjects experiment. They diagnosed two referral letters stating a main complaint without a diagnostic suggestion (control), two stating a correct suggestion, and two stating an incorrect suggestion. The referral question and case order were randomized. We analysed the effect of the referral question on interns' diagnostic accuracy, number of differential diagnoses, confidence, and time taken to diagnose.

Results: 44 medical interns participated. Interns considered more diagnoses in their differential without a suggested diagnosis (M = 1.85, SD = 1.09) than with a suggested diagnosis, independent of whether this suggestion was correct (M = 1.52, SD = 0.96, d = 0.32) or incorrect ((M = 1.42, SD = 0.97, d = 0.41), $\chi^2(2)$ =7.6, p = 0.022). The diagnostic suggestion did not influence diagnostic accuracy ($\chi^2(2)$ = 1.446, p = 0.486), confidence, ($\chi^2(2)$= 0.058, p = 0.971) or time to diagnose ($\chi^2(2)$= 3.128, p = 0.209).

Conclusions: A diagnostic suggestion in a GPs referral letter did not influence subsequent diagnostic accuracy, confidence, or time to diagnose for medical interns. However, a correct or incorrect suggestion reduced the number of diagnoses considered. It is important for healthcare providers and teachers to be aware of this phenomenon, as fostering a broad differential could support learning. Future research is necessary to examine whether these findings generalize to other healthcare workers, such as more experienced specialists or triage nurses, whose decisions might affect the diagnostic process later on.

Trial registration: The study protocol was preregistered and is available online at Open Science Framework (https://osf.io/7de5g).

Keywords: diagnostic error, clinical reasoning, cognitive bias, patient safety

Introduction

Diagnostic errors are a large burden on patient safety. It is estimated that a majority of patients will suffer at least one diagnostic error during their lifetime, sometimes with devastating consequences.(1-3) Diagnostic errors are defined as "the failure to establish and/or communicate an accurate and timely explanation of the patient's health problem(s)".(1) Most of these errors are thought to be preventable.(1, 4) In order to develop successful interventions, it is crucial to understand the underlying causes of diagnostic errors.

Physicians working in the ED often use clinical information (e.g., symptoms, examination findings, or test results) from patient referral letters in diagnostic decisions. The referral process is vulnerable to breakdowns in the process itself (5-7) and can also be influenced by flaws in the cognitive processes of the involved physicians. Flawed cognitive processes are seen as an important cause of diagnostic errors. These cognitive errors are often explained using dual process theory (DPT), which hypothesizes that reasoning consists of a non-analytical and fast System 1, and an analytical and more deliberate System 2.(8, 9) Errors in System 1 are often ascribed to cognitive biases (10), which are introduced into the reasoning process because of incorrect assumptions or missed information. Errors in System 2, on the other hand, are often ascribed to knowledge deficits.(11, 12) In a clinical context, cognitive errors could cause physicians to be influenced by incorrect information from another physician or to incorrectly interpret clinical information, which could ultimately result in diagnostic errors. Especially emergency medicine physicians are prone to such errors, due to domain specific factors such as complex decision making under time pressure and high uncertainty.(13, 14)

Previous studies show that clinical information can indeed influence diagnostic accuracy. For example, accurate clinical information improved physicians' true positive rates in radiology and test reading (15-17), whereas inaccurate clinical information reduced diagnostic accuracy (18) and even biased physicians' diagnostic reasoning towards incorrect working diagnoses suggested by the clinical information.(19) This effect was found for medical students as well as for experienced physicians.(20) However, it remains unclear via which underlying processes clinical information can impact diagnostic accuracy. For example, accuracy could decrease due to overconfidence, a limited differential diagnosis, or because physicians do not spend enough time on a case.

In this experimental study, we examined the effect of a general practitioner's (GPs) suggested diagnosis when they refer a patient from primary care (i.e., general practice) to secondary care (i.e., the ED) on the diagnostic performance of medical interns. The suggested diagnosis in a GPs referral question could be correct or incorrect, or did not contain a diagnostic suggestion at all (control group). We studied diagnostic performance in terms of diagnostic accuracy, and expanded on previous research by adding measures of differential diagnosis, confidence, and time spent on a case.

We expected that the suggested diagnosis in a GPs referral letter would cause interns to more often follow the suggested diagnosis than when no suggested diagnosis was provided (control condition). We hypothesized that this would also be true if the suggestion was incorrect. Furthermore, we hypothesized that both a correctly and an incorrectly suggested diagnosis would reduce the number of differential diagnoses considered and decrease the time spent to diagnose compared to the control group. Lastly, we expected that confidence in the most likely diagnosis would increase relative to the control group.

Methods

Participants

Medical interns associated with the Erasmus University Rotterdam (EUR) and the Erasmus University Medical Center (Erasmus MC) were invited to participate. Participants were eligible if they had completed their clinical rotation in internal medicine. Using G-power 3.1.9.7 (21) a sample size of 36 participants was estimated for a repeated measures analysis of variance (ANOVA) with a medium effect size based on Meyer et al. (20), a power of 0.95, and an alpha level of 0.05.

Design

A randomized within-subjects experiment was conducted in which each participant diagnosed six cases in three conditions. Participants were presented with two cases stating the patient's main complaint without a diagnostic suggestion, two cases with a correct diagnostic suggestion, and two cases with an incorrect diagnostic suggestion. Case order and condition were randomised through partial counterbalancing using a Latin square (Appendix A).

Materials

Six fictional cases were developed by an expert internist (JA), a medical doctor (RB) and a medical student (MS) and were piloted by 6 medical doctors specialized in primary care or internal medicine. Each case had one correct diagnosis and one plausible (but incorrect) alternative diagnosis (Table 1). All cases were formatted to look like genuine referral letters from a primary care physician (Appendix B) and were presented in Dutch. Participants used their own device (laptop or mobile phone) to access the survey in which the cases were presented (Appendix C).

Table 1. *Overview of the primary complaint, the alternative (incorrect) diagnostic suggestion and the correct diagnostic suggestion.*

	No suggestion	Correct suggestion	Incorrect suggestion
Case 1	Abdominal pain	Ovarian torsion	Appendicitis
Case 2	Painful leg	Erysipelas	Deep vein thrombosis
Case 3	Dyspnea	Heart failure	Pneumonia
Case 4	Epigastric pain	Pancreatitis	Cholecystitis
Case 5	Retrosternal pain	Peptic ulcer	Myocardial infarction
Case 6	Colic	Gallstone	Kidney stone

Procedure

Participants read an information letter and signed informed consent before participation. In order to study the effect of the manipulated referral question, the study's purpose was not fully disclosed to participants in advance. Instead, participants were told that we wanted to pilot the difficulty level of several clinical cases that were to be used for education. Participants diagnosed the six cases and after every case, they were asked to provide their most likely diagnosis (free text) and to rate their confidence in this diagnosis (0 = no confidence, 10 = very confident). The time that participants took to complete each case was registered upon submitting the diagnosis. After diagnosing all cases, participants were shown the case again and were asked to provide differential diagnoses for each of the cases. The differential diagnosis was elicited after all six cases were diagnosed to prevent the possible induction of reflective reasoning, which could reduce the effect of our manipulation.(22) Finally, they were asked to provide their demographic information and what they thought the real goal of the study was.

Outcome measures

Diagnostic accuracy was quantified by scoring the most likely diagnosis as either correct (1 point), partially correct (0.5 points) or incorrect (0 points). A diagnosis was scored as correct if participants mentioned the correct diagnosis or a different term for the same diagnosis. Closely adjacent diagnoses were also given full points (i.e., the correct diagnosis was pancreatitis and the participant mentioned acute pancreatitis). A diagnosis was scored as partially accurate if the participant captured an element of the diagnosis, but left out another core element (i.e., the correct diagnosis was peptic ulcer and the participant mentioned only ulcer). Any other diagnoses were scored as incorrect and did not receive any points. Scoring was performed independently by a medical doctor (RB) and a medical student (MS). If there was a discrepancy, this was solved via discussion with an expert internist (JA) as the third rater. Confidence in the most likely diagnosis was measured on a scale from 0 to 10 as self-reported by the participant. Time spent to diagnose was measured in seconds and automatically recorded by the survey software (Qualtrics). Based on the time taken to diagnose in the pilot, any entrees that took less than 25 seconds were considered not realistic and therefore excluded. Lastly, differential diagnosis was measured as the number (count) of alternative diagnoses given in a free text box.

Demographics

We measured the following demographic information: age, sex, months spent in the clinical phase, current internship, and specialism of interest. Additionally, we performed a manipulation check by asking participants to guess the study's goal.

Statistical analysis

According to the Kolmogorov-Smirnov test, the data were not normally distributed. Therefore, a within-subjects Friedman's ANOVA was performed to test if the referral question (within-subjects factor) impacted students' diagnostic performance. Separate Friedman's ANOVAs were performed for mean diagnostic accuracy, differential diagnosis, confidence, and time to diagnose a case, which were averaged per participant per condition. Additionally, differential diagnosis, confidence, and time to diagnose for correct and incorrect most likely diagnoses were compared using the Wilcoxon signed rank test. If a Friedman's ANOVA was significant, post-hoc tests were performed using individual Wilcoxon signed rank tests. A p-value of < 0.05 was considered statistically significant. Statistical analyses were performed using SPSS statistical software, version 25 for Windows (IBM Corp., Armonk, New York).

Results

44 out of the total 97 participants (45%) completed the experiment, 5 (5%) quit halfway through the study and 48 (50%) did not get past the initial instructions. Of the 44 students who completed the study, five participants were excluded based on the cut-off value for time to diagnose ($<$ 25 seconds), leaving 39 participants in the main analysis. For the analysis of the differential diagnosis, an additional five students were excluded because they did not provide a differential diagnosis for any of the cases. Demographics were available for 38

participants. 31 participants (82%) were female. On average, participants were 24 years (SD = 1) old and had spent 21 months (SD = 8) in the clinical phase. Age, sex, and months in the clinical phase did not moderate accuracy, number of differential diagnoses, confidence or time to diagnose (all $p > 0.05$) and thus, did not need to be corrected for.

Manipulation check

Seven out of the 39 participants (17.94%) correctly identified the study's goal. Despite this, their performance (diagnostic accuracy: M = 0.50, SD = 0.32) was similar to participants who did not identify the study's goal (diagnostic accuracy: M = 0.51, SD = 0.32). Therefore, all participants were analysed as one group.

Main analysis

The diagnostic suggestion did not influence diagnostic accuracy, $\chi^2(2) = 1.45$, $p = 0.486$, but did impact the number of differential diagnoses generated, $\chi^2(2) = 7.60$, $p = 0.022$. Interns considered significantly more diagnoses when they did not receive a diagnostic suggestion compared to when they did, which resulted in a small effect size compared to both correct suggestions (d = 0.32) and incorrect suggestions (d = 0.41). Confidence, $\chi^2(2) = 0.06$, $p = 0.971$ and time to diagnose, $\chi^2(2) = 3.13$, $p = 0.209$, did not differ significantly depending on the referral question. Descriptive data are reported in Table 2.

Table 2. *Mean (M) and standard deviation (SD) for accuracy, differential diagnosis, confidence and time to diagnose.*

	N	No suggestion	Correct suggestion	Incorrect suggestion
Accuracy, M (SD)	39	0.46 (0.34)	0.57 (0.27)	0.51 (0.35)
Differential diagnosis, M (SD)	33	1.85 (1.09)	1.52 (0.96)	1.42 (0.97)
Confidence, M (SD)	39	6.23 (1.23)	6.21 (1.00)	6.36 (1.13)
Time to diagnose, M (SD)	39	120.21 (55.87)	133.79 (81.87)	140. 40 (74.78)

Exploratory analyses

Accuracy per case

The effect of diagnostic suggestion on diagnostic accuracy was not significant overall, but there was substantial variation between the cases used (Table 3). Notably, accuracy was descriptively higher for a correct diagnostic suggestion in case 1 (50%) and case 5

(63.64%) compared to an incorrect diagnostic suggestion (case 1: 13.33%; case 5: 53.85%) or no diagnostic suggestion (case 1: 16.67%; case 5: 26.67%). Conversely, accuracy was descriptively lower for the correct diagnostic suggestion for case 3 (68.76%) and case 6 (43.75%) compared to an incorrect diagnostic suggestion (case 3: 90.91%; case 6: 75%) or no diagnostic suggestion (case 3: 100%; case 6: 72.72%).

Table 3. *The number of responses and percentage (%) of correct responses per case.*

Condition	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**	**Case 6**	**Total**
No suggestion	2/12 (16.67%)	4 /13 (30.77%)	12/12 (100%)	7/15 (46.67%)	4/15 (26.67%)	8/11 (72.73%)	37/78 (47.43%)
Correct suggestion	6/12 (50%)	6/11 (54.55%)	11/16 (68.76%)	9/12 (75%)	7/11 (63.64%)	7/16 (43.75%)	46/78 (58.97%)
Incorrect suggestion	2/15 (13.33%)	11/15 (73.33%)	10/11 (90.91%)	5/12 (41.67%)	7/13 (53.85%)	9/12 (75%)	44/78 (56.41 %)
Total correct per case	10/39 (25.64%)	21/39 (53.85%)	33/39 (84.62%)	21/39 (53.85%)	18/39 (46.15%)	24/39 (61.54%)	127/234 (54.27%)

Correct and incorrect diagnosis

The number of diagnoses considered in the differential diagnosis did not differ between participants who gave a correct diagnosis (M = 1.54, SD = 1.22) and participants who gave an incorrect diagnosis (M = 1.59, SD = 1.22), T = 1407.00, p = 0.767. The time that participants spent to diagnose cases also did not differ between correct and incorrect diagnoses (correct: M = 129.29, SD = 104.07; incorrect: M = 136.57, SD = 88.27, T = 2620.00 p = 0.322).

Confidence

Participants were more confident when their most likely diagnosis was correct (M = 6.51, SD = 0.97), compared to when it was incorrect (M = 6.00, SD = 1.03), T = 1592.00, p = 0.006, d = 0.51. This did not differ based on the diagnostic suggestion, χ^2 (3) = 4.29, p = 0.232 (Table 2).

Discussion

This study examined the effect of clinical information in the form of a diagnostic suggestion in a GPs referral letter on the diagnostic performance of medical interns. Contrary to our hypotheses, we found no effect of diagnostic suggestion on accuracy, confidence, or time taken to diagnose. Diagnostic suggestions did, however, affect the number of diagnoses participants considered in their differential diagnosis. They considered more diagnoses when the referral letter did not contain any suggestion compared to when either a correct or incorrect suggestion was presented. Exploratory analyses further suggested a positive correlation between accuracy and confidence.

Research on the effect of clinical information on test reading has shown that diagnostic suggestions can bias physicians towards the suggested diagnosis, decreasing diagnostic accuracy if the suggestion was incorrect.(18, 19) The interns in the current study, however, were able to overcome the potential bias of an incorrect suggestion, as their accuracy did not decrease. This contrast to previous studies might be explained by the relative inexperience of our participants. It is suggested that inexperienced physicians rely more on analytical reasoning than on non-analytical reasoning, as they have not accumulated enough previous experiences to rely on pattern recognition.(23) Reliance on analytical thinking could result in a more conscious approach to diagnosis, possibly making our participants more vigilant for information in the case that conflicted with the suggestion. Such an approach would make participants less likely to be biased by the suggestion as analytical approaches such as deliberate reflection have been shown to reduce diagnostic errors due to biases.(22) This possibility is supported by our finding that confidence was higher when participants were correct: they seemed capable of estimating how valid their diagnoses were, which fits the profile of analytical reasoning.

Although overall diagnostic accuracy was not affected by the type of diagnostic suggestion, exploratory analyses suggested there were differences at case-level. Specifically, our findings indicated that depending on the case, correct diagnostic suggestions could either be beneficial or detrimental to accuracy (Table 3). These differences were descriptive and not statistically significant, but provide considerations for future research. In two cases where less than 50% of participants were correct when receiving no diagnostic suggestion, accuracy improved when they received the correct suggestion. In this scenario, the correct suggestion could possibly compensate for gaps in knowledge by suggesting a diagnosis that the participant otherwise would not have considered.(11, 12) For example, in the first case interns were likely more familiar with appendicitis (the alternative incorrect diagnosis) than with ovarian torsion (the correct diagnosis). The correct suggestion might have prevented them from missing the less prevalent diagnosis and allowed them to suggest the correct diagnosis instead. However, in two other cases accuracy descriptively decreased when a correct diagnostic suggestion was considered. This could indicate that knowledge gaps should be acknowledged before a suggestion can be beneficial. For example, if the incorrect diagnosis seems likely, participants might still choose to reject the correct suggestion. All in all, perhaps the effect of diagnostic suggestions depends on the case diagnosis, participant's prior knowledge, and their willingness to consider suggestions.

The type of diagnostic suggestion did impact interns' differential diagnosis: just providing a suggested diagnosis, either correct or incorrect, reduced the number of diagnoses considered. This is consistent with Meyer et al. (20) who showed that an a priori diagnosis, regardless of whether this diagnosis was correct or not, led to fewer questions asked during history taking and a less systematic assessment of differential diagnoses. Failure to consider the correct diagnosis is an important cause of diagnostic error.(24) It is vital that the correct diagnosis is at least considered in the differential diagnosis, even if it is not considered as the most likely diagnosis. The importance of the differential diagnosis is associated with the dynamic nature of diagnostic reasoning. If the course of a disease changes, it will be easier to consider another diagnosis that is already included in the differential diagnosis. But although our diagnostic suggestions did reduce the number of differential diagnoses considered, they did not decrease diagnostic accuracy. Future research should examine whether this reduction in differential diagnoses results in a qualitatively worse differential diagnosis, or conversely if it produces a more specific and efficient differential diagnosis without a reduction in accuracy. Though it is difficult to make practical recommendations based on the current results, we suggest it might be valuable for education to have interns practice diagnosing cases without a diagnostic suggestion as this can allow them to foster a broader differential diagnosis. Additionally, educators could vary between using cases with and without diagnostic suggestions, so that interns can practice with both scenarios and might perhaps learn to overcome possible negative influences or benefit from possible positive influences of suggested diagnoses. For example, interns could be trained using methods such as deliberate reflection, which promote the generation of multiple differential diagnoses and considering information that increases or reduces the likelihood of the differential diagnoses.(25)

The current study had several strengths and limitations. Because of the experimental within-subjects design with randomized presentation of the cases and diagnostic suggestions, it was possible to isolate the effect of the diagnostic suggestion. Furthermore, this study had a high power due to the within-subjects design. However, the experimental design also poses a limitation, as we could not replicate the time constraints and high level of uncertainty present in clinical practice. Additionally, the current findings are limited in their generalizability to practice, as we included relatively inexperienced interns. Future research should investigate how diagnostic suggestions affect primary to secondary care referral in clinical practice and in more experienced physicians. Lastly, this study did not consider the impact of diagnostic suggestions on some steps in the diagnostic process, such as ordering and interpreting investigations, due to practical considerations. Future studies should also consider how diagnostic suggestions impact other steps in the diagnostic process.

In conclusion, diagnostic suggestions can reduce the number of diagnoses considered in the differential diagnosis of medical interns. Other aspects of diagnostic performance, namely interns' diagnostic accuracy, confidence, and time to diagnose, were not affected. Healthcare providers should be aware of this phenomenon in order to limit unwanted effects. When training medical students in clinical reasoning, one could avoid diagnostic suggestions in order to train students in broad differential thinking. Considering the fact that various professionals are involved with the work-up in the ED, future research should repeat the experiment in other groups of professionals, such as medical specialists and triage nurses.

Acknowledgements

We would like to thank the residents who participated in the study.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9. Available from: http://dx.doi.org/10.1001/archinte.165.13.1493.
3. Gunderson CG, Bilan VP, Holleck JL, Nickerson P, Cherry BM, Chui P, et al. Prevalence of harmful diagnostic errors in hospitalised adults: a systematic review and meta-analysis. BMJ quality & safety. 2020;29(12):1008-18. Available from: http://dx.doi.org/10.1136/bmjqs-2019-010822.
4. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21. Available from: http://dx.doi.org/10.1001/archinternmed.2010.146.
5. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA internal medicine. 2013;173(6):418-25. Available from: http://dx.doi.org/10.1001/jamainternmed.2013.2777.
6. Manser T, Foster S. Effective handover communication: an overview of research and improvement efforts. Best practice & research Clinical anaesthesiology. 2011;25(2):181-91. Available from: http://dx.doi.org/10.1016/j.bpa.2011.02.006.
7. van Heesch G, Frenkel J, Kollen W, Zwaan L, Mamede S, Schmidt H, et al. Improving Handoff by Deliberate Cognitive Processing: Results from a Randomized Controlled Experimental Study. The Joint Commission Journal on Quality and Patient Safety. 2021;47(4):234-41. Available from: http://dx.doi.org/10.1016/j.jcjq.2020.11.008.
8. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education. 2009;14(1):27-35. Available from: http://dx.doi.org/10.1007/s10459-009-9182-2.
9. Kahneman D, Egan P. Thinking, fast and slow: Farrar, Straus and Giroux New York; 2011.
10. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic medicine. 2003;78(8):775-80. Available from: http://dx.doi.org/10.1097/00001888-200308000-00003.
11. Mamede S, Goeijenbier M, Schuit SCE, de Carvalho Filho MA, Staal J, Zwaan L, et al. Specific Disease Knowledge as Predictor of Susceptibility to Availability Bias in Diagnostic Reasoning: a Randomized Controlled Experiment. Journal of general internal medicine. 2021;36(3):640-6. Available from: http://dx.doi.org/10.1007/s11606-020-06182-6.
12. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine. 2017;92(1):23-30. Available from: http://dx.doi.org/10.1097/ACM.0000000000001421.
13. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. European journal of internal medicine. 2013;24(6):525-9. Available from: http://dx.doi.org/10.1016/j.ejim.2013.03.006
14. Zwaan L, Hautz WE. Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. BMJ Publishing Group Ltd; 2019. p. 352-5.
15. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. American Journal of Roentgenology. 1981;137(5):1055-8. Available from: http://dx.doi.org/10.2214/ajr.137.5.1055
16. Leslie A, Jones AJ, Goddard PR. The influence of clinical information on the reporting of CT by radiologists. The British journal of radiology. 2000;73(874):1052-5. Available from: http://dx.doi.org/10.1259/bjr.73.874.11271897

17. Song KS, Song HH, Park SH, Ahn KJ, Yang IK, Byun JY, et al. Impact of clinical history on film interpretation. Yonsei medical journal. 1992;33(2):168-72. Available from: http://dx.doi.org/10.3349/ymj.1992.33.2.168

18. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. Jama. 2004;292(13):1602-9. Available from: http://dx.doi.org/10.1001/jama.292.13.1602

19. LeBlanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. Academic Medicine. 2002;77(10):S67-S9. Available from: http://dx.doi.org/10.1097/00001888-200210001-00022

20. Meyer FML, Filipovic MG, Balestra GM, Tisljar K, Sellmann T, Marsch S. Diagnostic Errors Induced by a Wrong a Priori Diagnosis: A Prospective Randomized Simulator-Based Trial. Journal of Clinical Medicine. 2021;10(4):826. Available from: http://dx.doi.org/10.3390/jcm10040826

21. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91.

22. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203. Available from: http://dx.doi.org/10.1001/jama.2010.1276.

23. Coderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. Medical education. 2003;37(8):695-703.

24. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. Academic Medicine. 2011;86(3):307-13. Available from: http://dx.doi.org/10.1097/ACM.0b013e31820824cd

25. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Medical education. 2008;42(5):468-75. Available from: http://dx.doi.org/10.1111/j.1365-2923.2008.03030.x.

Appendix A – Randomisation procedure

Table 1. *Partial randomisation of referral questions and clinical cases using a Latin square.*

Order	Broad	Broad	Specific congruent	Specific congruent	Specific incongruent	Specific incongruent
A	Case 1	Case 2	Case 6	Case 3	Case 5	Case 4
B	Case 2	Case 3	Case 1	Case 4	Case 6	Case 5
C	Case 3	Case 4	Case 2	Case 5	Case 1	Case 6
D	Case 4	Case 5	Case 3	Case 6	Case 2	Case 1
E	Case 5	Case 6	Case 4	Case 1	Case 3	Case 2
F	Case 6	Case 1	Case 5	Case 2	Case 4	Case 3

Appendix B – Example case 1: correct referral question.

Referral letter for Emergency Department

Sender		Patient	
Name:	[redacted]	Name:	[redacted]
AGB-code:	[redacted]	Date of birth:	[redacted]
Organisation:	[redacted]	Citizen service nr.:	[redacted]
	[redacted]	Address:	[redacted]
		City:	[redacted]
Org. AGB-code:	[redacted]	Phone number:	[redacted]
Address:	[redacted]	Health insurer:	[redacted]
City:	[redacted]		
Phone:	[redacted]	Patient ID:	[redacted]
Peer consultation:	[redacted]	Healthcare institution:	[redacted]

Referral			
Date:	[redacted]	Name of product:	[redacted]
ZD-number:	[redacted]	Waiting time:	[redacted]
Organisation:	[redacted]	Care question:	[redacted]
		Address:	[redacted]
		Residence:	[redacted]

Core part 25-09-2020 14:32

Dear colleague,

Reason for referral

I hereby refer the following patient (details below) with complaints of pain in the abdomen, ovarian torsion?

Journal

Patient contact 23-05-2019

Patient has pain in the abdomen since yesterday evening, mostly pain in the lower abdomen. Also a bit nauseous. She didn't vomit until now. Last stool was yesterday afternoon. No pain during movement.

Last menstruation was two weeks ago. No trauma prior to this pain.

PE: very uncomfortable and painful patient, temp. 37.2(ear), heartrate 103.

Calm peristalsis, normal tympany, active defense, possible to palpate the abdomen with deep breathing, tenderness in particular in the right lower quadrant of the abdomen.

Further investigation:

HCG negative

CRP 35

Medical history

Medical history

11-02-19 Insertion of IUD – ultrasound: good position, benign adnexal enlargement 5 cm

15-10-18 Allergic reaction/allergy - dust mite, cat, dog

12-03-15 Mononucleosis infectiosa

22-09-13 Eczema – Dermovate lotion

Actual medication

Dermovate lotion

Kind regards,

[redacted]

Appendix C – Survey questions

First, participants are asked to evaluate 6 clinical cases. For two cases, no diagnostic suggestion is provided (but only the patients' main complaint); for two cases, a correct diagnostic suggestion is provided; and for two cases, an incorrect diagnostic suggestion is provided. For each case participants are asked to answer the following questions:

1. What is the most likely diagnosis?
2. How confident are you in your diagnosis? (on a scale form 0-10, no confidence to very confident)

After completing all cases, participants are again shown the case and are asked:

1. Have you considered any other diagnosis? If yes, please list these diagnoses.

Then, participants are asked for relevant demographic information:

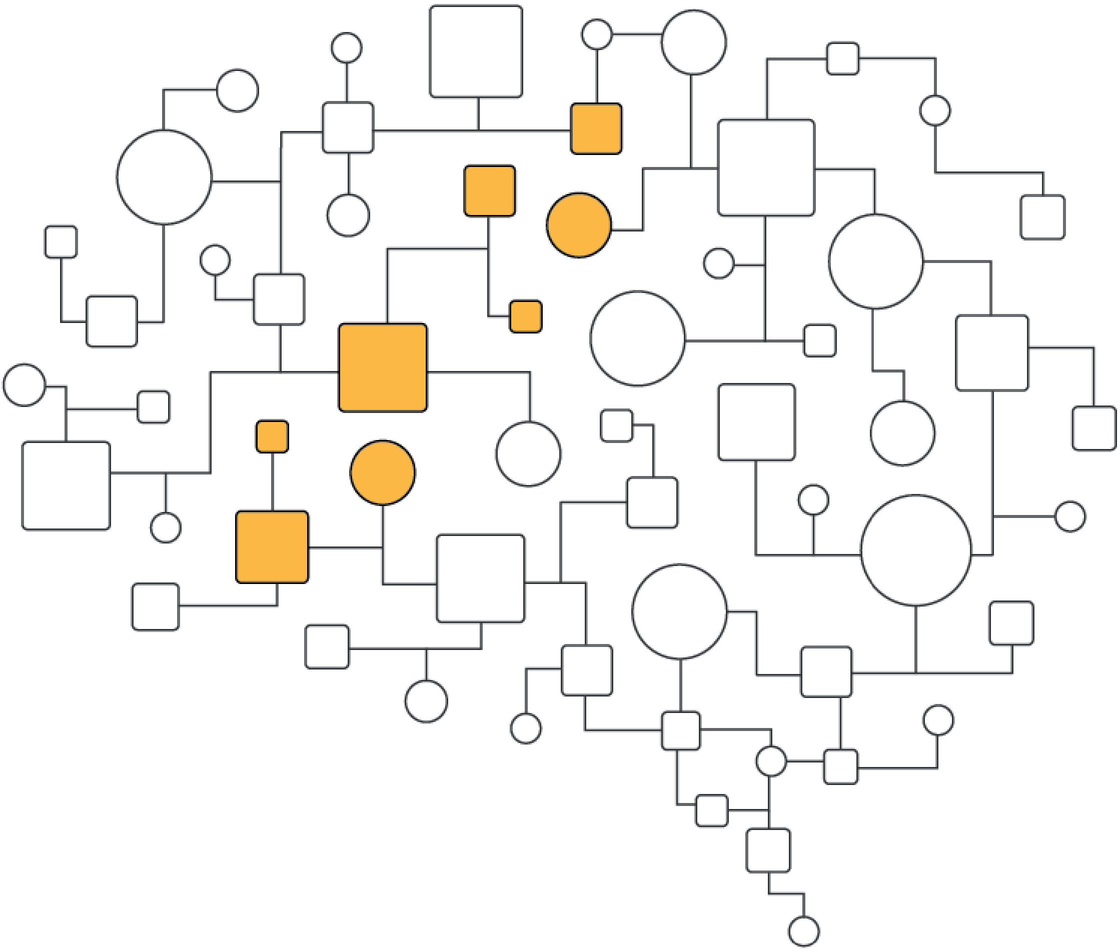
1. How old are you?
2. What is your sex?
3. How many months have you spent in your clinical phase?
4. What is the department of your current internship? In case you are currently not following an internship, please name your last completed internship.
5. Which specialism do you want to, or will you, practice after residency?

And we asked participants to guess the goal of the study to check our manipulation:

1. At the start of this study, we informed you that we aimed to evaluate the included cases as exam materials. In addition to that, we also have a secondary goal. Do you have any idea what this goal could be?

Finally, participants are asked to leave their e-mail address so that they can receive information on the study's outcomes after the study has been completed.

1. If you would like to receive information about the study and its outcomes (and your own performance) when the study has concluded, please leave your email address.



CHAPTER 4

Selectivity in information processing in correct and incorrect diagnoses: a randomized controlled eye-tracking experiment

Staal, J., Alsma, J., Van der Geest, J., Mamede, S.,
Jansen, E., Frens, M.A., Van den Broek, W.W., Zwaan, L.

Submitted for publication.

Abstract

Background: Diagnostic errors are often attributed to erroneous selection and interpretation of patients' clinical information, either due to cognitive biases or knowledge deficits. We hypothesized that the type of information processed during clinical reasoning distinguishes between correct and incorrect diagnoses.

Methods: In this within-subjects eye-tracking experiment, 19 internal and emergency medicine residents diagnosed 12 written case vignettes. Half the cases contained a correct diagnostic suggestion and the other half an incorrect suggestion. We measured how often (i.e., number of fixations) and how long (i.e., relative dwell time) residents attended to clinical information crucial for either the correct diagnosis or the incorrect suggestion. Additionally, we measured confidence and time to diagnose in each case.

Results: No main effects of diagnostic suggestion or final diagnostic accuracy were observed. However, an interaction was observed where residents fixated more often on the information relevant to the correct diagnosis when they received an incorrect suggestion, but overcame that bias to arrive at the correct final diagnosis (M: 25 fixations, SD: 20; compared to an average of 16-17 fixations in other conditions). This interaction was not significant for relative dwell time. Confidence (range: 64 - 67%) or time to diagnose (range: 68 – 86 sec) did not differ depending on residents' accuracy or the diagnostic suggestion.

Conclusions: Selectivity in information processing was not directly associated with an increase in diagnostic errors but rather seemed related to recognizing and rejecting a biased suggestion in favor of the correct diagnosis. This could indicate an important role for case-specific knowledge in avoiding biases and diagnostic errors. Future research should examine information processing for other types of clinical information.

Introduction

Diagnostic errors, defined as missed, wrong, or delayed diagnoses, are a large burden on patient safety. It is estimated most of people will experience a diagnostic error during their lifetime, possibly with devastating consequences.(1) Preventing diagnostic errors is challenging, as clinical reasoning is a complex process that encompasses many different cognitive skills.(2) Flaws in these cognitive skills are thought to be among the main causes of diagnostic errors and are viewed as highly preventable.(3, 4) However, assessing what the cause of the error is, is difficult. Clinical reasoning occurs rapidly and clinicians may not always be aware of the reasoning steps that occurred, much less of whether or not these steps were flawed.(5, 6) Therefore, it is crucial to gain more insight in the cognitive processes underlying clinical reasoning to reduce diagnostic errors.

To understand how errors due to cognitive flaws can be prevented, it should first be understood how clinicians make a diagnosis. Clinical reasoning is mostly explained using the dual process theory (7), which proposes that reasoning occurs via two systems. System 1 facilitates fast and automatic reasoning and is primarily active in routine situations, whereas System 2 is associated with more deliberate and conscious decision making which is useful in new or complex situations. Although many variations on dual process theory exist (8-10), the general consensus is that System 1 uses heuristics, or mental shortcuts, to quickly arrive at a solution. These heuristics are often efficient and useful (11-13) but can also result in cognitive biases (systematic reasoning errors that occur when not all relevant information is considered (7)) and subsequently, in diagnostic errors.(7, 14, 15) For example, a premature closure bias can occur when a clinician does not continue considering likely alternative diagnoses after an initial diagnosis was reached.(16) The literature furthermore suggests that knowledge deficits could be a significant cause of errors.(17, 18) From this perspective, errors can occur because clinicians do not have sufficient knowledge, or cannot access it, when diagnosing a patient. The exact mechanisms behind how cognitive biases and knowledge deficits lead to errors have yet to be elucidated. One possibility, however, is that inappropriate use of available clinical information could be an underlying factor. Cognitive biases suggest a failure in the clinical reasoning process itself and knowledge deficits assume a deficiency in pre-existing knowledge. Both result in the erroneous selection and interpretation of a patients' clinical information. This selectivity in information processing could provide further insight in the causes of diagnostic errors.

Previous research supports an association between faulty information gathering or integration and diagnostic errors. Graber et al. (16) found that faulty information synthesis was the most common cause of errors. Later studies analyzing error cases similarly reported that clinician assessment errors, such as not considering diagnoses or incorrectly weighing competing diagnoses, occurred frequently.(19, 20) Zwaan et al. (21, 22) furthermore showed in record review studies that often, insufficient information was gathered and that the follow-up on relevant findings was lacking. This selectivity was also associated with an increase in diagnostic errors and patient harm.(21) Mamede et al. (23) found that salient distracting features (i.e., pieces of information that grab attention because they are strongly associated to a certain diagnosis despite not being relevant to the correct diagnosis) caused diagnostic errors. Additionally, the ability to appropriately select information was a distinguishing factor between experts and novices (24) and all steps from information search to the final integration of information improve as expertise develops.(25) Though these studies differ in their goals and designs, they all give indications that selectivity in information processing could be a cause of diagnostic errors.

Several limitations should be kept in mind when interpreting previous research. First, previous studies often exclusively examined error cases and did not determine to what extent selectivity, or other processes that might have led to errors, occurred in correctly diagnosed cases.(16, 19, 20) The comparison between error cases and correctly diagnosed cases is important to understand how certain processes can lead to errors, and to what extent they are just part of the standard clinical reasoning process. Second, a majority of studies retrospectively assessed cases (16, 19-22), a process that is susceptible to hindsight bias (i.e., when knowing the outcome of a case influences the judgement of the case) (26). Assessors might overestimate the likelihood that the clinician involved in the error could have made the diagnosis at that point in time.(27) Prospective studies circumvent both issues, but to our knowledge, current prospective studies did not specifically examine selectivity in information processing in written case vignettes.

The current study aimed to prospectively investigate information processing during diagnosis in both error cases and correctly diagnosed cases. Information processing was measured using eye-tracking, an increasingly popular technique in clinical reasoning research (28), primarily for visual diagnostic specialisms such as radiology.(29-31) Eye-tracking allows a more objective and “live” observation of clinicians’ reasoning processes and does not rely on self-report. Eye-tracking assumes that information that receives more attention (i.e., is looked at longer or more often) is processed more (32), and indirectly relates this to information processing.(33) Residents diagnosed written clinical case vignettes while wearing a head mounted eye-tracker. Each case contained either a correct or an incorrect diagnostic suggestion, which was meant to induce confirmation bias. This way, we could examine whether biased residents do indeed cut corners, or ignore relevant information. Information processing was measured as what information residents looked at, and how long and how often they looked at that information. Additionally, we measured residents’ final diagnostic accuracy, their confidence in that diagnosis, and total time spent on the case. We hypothesized that residents would look longer and more often at clinical information necessary to arrive at their final diagnosis, regardless of the diagnostic suggestion. Furthermore, we expected that residents would be more confident if their final diagnosis matched the diagnostic suggestion, regardless of their accuracy. We expected no differences in the time spent on each case depending on the diagnostic suggestion or their final accuracy.

Methods

Design

The study was a single-phase eye-tracking experiment with a within-subjects design. The study was approved by the medical ethical committee of the Erasmus University Medical Center (MEC-2018-1571). All participants gave informed consent. All methods were carried out in accordance with the relevant guidelines and regulations. Residents' eye movements were measured while they diagnosed 12 written clinical case vignettes, accompanied by a diagnostic suggestion designed to induce confirmation bias, in a random order. Each case contained a suggested provisional diagnosis, which was correct in six of the cases and incorrect in the other six. Whether the suggested diagnosis was correct or incorrect was randomized.

Participants

Nineteen residents in their 1st to 6th year of training participated between November 2020 and August 2022 (Table 1). Residents were in training at the Erasmus University Medical Center Rotterdam in the Netherlands, either for the internal medicine or emergency medicine department. They were recruited individually through mail and phone contact (by JS, EJ, and JA). Residents were excluded if they could not read the case vignettes on the monitor at 60cm distance. Glasses or contact lenses were allowed if they did not distort the eye-tracker signal. All participants provided written consent.

Sample size was calculated a priori using G-power.(34) We estimated the sample size for a repeated measures ANOVA with within factors. Sample size was calculated for a small effect (0.20), a power of 0.80, an alpha of 0.05, 1 group, and 12 measurements. This estimation indicated that 19 participants would be required.

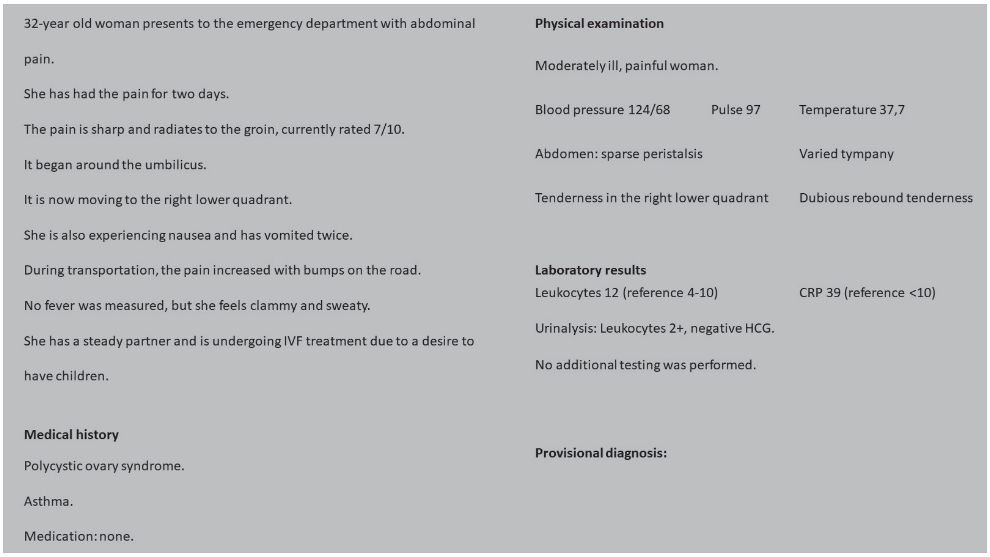
Table 1. *Participant demographics.*

Medical specialty	*N*	*Age (SD)*	*Sex N (% Female)*	*Years as resident (SD)*
Internal medicine	16	32 (2)	10 (62%)	3.1 (2)
Emergency medicine	3	30 (6)	2 (66%)	3.7 (3)

Materials

Cases

Twelve written clinical case vignettes were developed by one internist (JA) and independently diagnosed and confirmed by one emergency physician (EJ) (Table 2). Cases concerned a variety of internal medicine and emergency medicine diagnosis that all junior doctors should be expected to recognize based on their teaching curriculum (Table 2). Each case consisted of the history, medication details, physical examination findings, and test results of fictional patients (Figure 1). Cases were designed to have one correct diagnosis and one plausible, but incorrect, alternative suggestion. Clinical information in the cases included several distinguishing features that fit with either the correct or the incorrect diagnosis. When all considered together, these features provided all information necessary to prefer the correct diagnosis over the incorrect suggestion. The cases were piloted by third year residents and an emergency physician (N = 5).



32-year old woman presents to the emergency department with abdominal pain.

She has had the pain for two days.

The pain is sharp and radiates to the groin, currently rated 7/10.

It began around the umbilicus.

It is now moving to the right lower quadrant.

She is also experiencing nausea and has vomited twice.

During transportation, the pain increased with bumps on the road.

No fever was measured, but she feels clammy and sweaty.

She has a steady partner and is undergoing IVF treatment due to a desire to have children.

Medical history

Polycystic ovary syndrome.

Asthma.

Medication: none.

Physical examination

Moderately ill, painful woman.

Blood pressure 124/68 Pulse 97 Temperature 37,7

Abdomen: sparse peristalsis Varied tympany

Tenderness in the right lower quadrant Dubious rebound tenderness

Laboratory results

Leukocytes 12 (reference 4-10) CRP 39 (reference <10)

Urinalysis: Leukocytes 2+, negative HCG.

No additional testing was performed.

Provisional diagnosis:

Figure 1. Example of a clinical case as presented during eye-tracking (Case 1). The case was presented in Dutch. An overview of all cases in English is presented in Appendix A.

Regions of interest

Regions of interest reflected the distinguishing features of each case and were defined a priori. These features could be any information in the case, such as symptoms or test results. Regions of interest were classified based on whether the information they contained was relevant only to the correct diagnosis, or only to the incorrect diagnosis. Regions of interest were defined by one internist (JA) and one emergency physician (EJ), who independently marked the regions of interest and resolved discrepancies via discussion. All areas of a case that were not within a region of interest were designated as “background” and were not considered in the analyses. The regions of interest are indicated in Appendix A.

Case presentation

he cases were scaled to fit a 1920x1080px monitor. All information for one case was shown immediately on the same screen. The cases were presented on a light-grey background to prevent strain on residents during the eye-tracking procedure. All cases had a font size of 18 and double line spacing. No text was placed in the center of the screen, to prevent accidental overlap between residents’ gaze starting position and the regions of interest. All cases were written in Dutch.

Eye-tracker

The head-mounted EyeLink II (SR-Research, Canada) was used to record residents’ eye movements during diagnosis. The EyeLink II tracks the corneal reflection and the pupil using infra-red light at 500Hz. We tracked the right eye, unless the resident indicated sight was worse in this eye. A 9-point grid was used for calibration. Each calibration was validated a second time. The calibration was accepted if the inaccuracy between the gaze and the measurement was less than 4 degrees. Residents were asked to keep their chin in a chin rest during the experiment, which kept their head stable and at approximately 60cm distance from the monitor.

Voice recorder

Residents were asked to state their most likely diagnosis out loud. This was recorded using a Basic voice recorder Premium. After the diagnoses were transcribed by the first author (JS), the voice files were deleted permanently.

Survey

Additional outcome measures not directly related to the eye-tracking measurement were acquired via Qualtrics, an online survey tool. Feedback was also provided via Qualtrics.

Procedure

Residents received an information letter and signed informed consent. They were informed that the goal of the study was to investigate the cognitive processes underlying clinical reasoning, in terms of speed and information processing. They were not informed that the cases were designed to induce confirmation bias until after the experiment had concluded. The experiment took approximately 20 to 30 minutes.

The experiment started with setting up the eye-tracker. The head-mounted eye-tracker and seat were adjusted to allow a comfortable position in the chin rest. Residents were then asked to diagnose 12 cases and indicate whether they agreed with the provisional diagnostic suggestion, which was presented as the suggested diagnosis from a colleague. If they did not agree, they were asked to provide their most likely diagnosis. The eye-tracker was then calibrated: residents saw a black fixation cross in the center of the screen before each case, to correct for possible drift in their position and to ensure that the gaze starting position did not overlap with the case text. After this initial fixation cross, the case was shown. There was no time limit for diagnosis and residents had to click after reading the case to indicate they wanted to provide a diagnosis. After clicking, a blank screen appeared and residents' most likely diagnosis was recorded using a voice recorder. The researcher then proceeded to the next case.

After all cases had been diagnosed, the eye-tracker was removed and participants were asked to fill out a final set of questions in a Qualtrics survey. They were shown the history of each case again and were asked to indicate how confident they were in their diagnosis for that case. They additionally provided demographic information. Finally, we performed a manipulation check by asking whether they suspected the goal of the study and then residents were shown the correct expert diagnosis for each case. They were debriefed of the goal of the experiment after they had read the feedback.

Outcome measures

The independent variable was the diagnostic suggestion: this suggestion could either reflect the correct diagnosis or the plausible, but incorrect, alternative diagnosis. The main dependent variables were diagnostic accuracy and the eye-tracking measures reflecting information processing. Diagnostic accuracy was scored by one internist (JA) and one intensivist (EJ), who independently assessed and assigned a score of 0 to incorrect diagnoses and of 1 to correct diagnoses. Discrepancies in scoring were resolved through discussion. The measures of information processing were relative dwell time, or the percentage of total fixations, and the number of fixations on the regions of interest in each case. Additionally, residents' confidence (0: not confident to 10: very confident) and total time on task (in seconds) were measured.

Lastly, we asked residents to provide their medical specialty, age, sex, and years of experience as a resident as demographic information.

Statistical analysis

First, bias induction was checked by comparing diagnostic accuracy in cases where a correct suggestion was provided with cases where an incorrect suggestion was provided, using an independent measures *t*-test. The main analysis compared the dependent variables for regions of interest for the correct diagnosis with regions of interest for the incorrect suggestion in a 2 (diagnostic suggestion) x 2 (diagnostic accuracy) repeated measures ANOVA. The dependent variables were averaged per participant over all cases. All tests were performed in IBM SPSS Statistics for Windows (Version 26.0). All tests were considered significant at the α = .05 level.

Table 2. *Correct diagnoses and diagnostic suggestions, and average diagnostic accuracy (N = 19) for each case.*

Case	Correct diagnosis	Incorrect suggestion	Diagnostic accuracy Mean (SD)
1	Ovarian torsion	Appendicitis	0.21 (0.42)
2	Nephrotic syndrome	Heart failure	0.42 (0.51)
3	Viral pericarditis	Pulmonary embolism	0.63 (0.50)
4	Giardia lamblia infection	Coeliac disease	0.45 (0.44)
5	Thrombotic thrombocytopenic purpura (TTP)	Immune thrombocytopenic purpura (ITP)	0.39 (0.47)
6	Sarcoidosis	Metastatic prostate cancer	0.24 (0.42)
7	Epstein-Barr virus (EBV) infection	Lymphoma	0.47 (0.50)
8	Toxic megacolon	Ileus	0.58 (0.51)
9	Hypoglycemia	Benzodiazepine intoxication	0.26 (0.45)
10	Alcoholic hepatitis	Pancreatic cancer	0.08 (0.25)
11	Gout	Cellulitis	0.21 (0.42)
12	Obstructive sleep apnea syndrome (OSAS)	Primary hyperaldosteronism	0.38 (0.49)

Funding

The funding body was not involved in the design of the study and the collection, analysis, and interpretation of data and in writing the manuscript.

Results

The 12 cases showed a considerable range in average diagnostic accuracy, varying from 8% to 63% accuracy. Analyzing our outcome measures per case was not possible, however, because too few observations populated each combination of diagnostic suggestion and diagnostic accuracy to perform the ANOVA. Therefore, we opted to take all observations together and changed our analysis method to a between-subjects ANOVA with case and participant number as covariates, to correct for differences between cases and participants. Treating these within-subjects observations as between-subjects could lead to an underestimation of the study effect, as between-subjects analyses account for multiple possible sources of variation.

In 8 of the 12 cases, data from all participants could be used. In the other cases, some data was not usable because the eye was not correctly captured. This led to the removal of 6 case recordings spread across 4 cases.

Bias induction

Diagnostic accuracy was lower in cases with an incorrect suggestion (20%, 95% CI: 12-27%) than in cases with a correct suggestion (57%, 95% CI: 48-66%), $p < 0.001$, indicating that confirmation bias was successfully induced.

Manipulation check

None of the participants correctly guessed the true goal of the study.

Main analysis

The descriptive statistics of all outcome measures are shown in Table 3.

Information processing.

Relative dwell time

Relative dwell time on either the regions of interest relevant for the correct diagnosis ($p = 0.231$) or the incorrect diagnosis ($p = 0.237$) did not differ depending on diagnostic accuracy. Whether or not the diagnostic suggestion was correct did not affect relative dwell time on the regions of interest for the correct diagnosis ($p = 0.345$), but the relative dwell time was reduced in the regions of interest for the suggested diagnosis ($p = 0.014$) when the suggestion was incorrect. The interaction between the diagnostic suggestion and diagnostic accuracy was not significant for either type of region of interest (correct: $p = 0.541$; incorrect: $p = 0.818$). The covariates case and participant did not explain these effects ($p > 0.050$).

Number of fixations

The number of fixations was higher for the regions of interest relevant for the correct diagnosis ($p = 0.043$) but not for the incorrect diagnosis ($p = 0.799$) if the final diagnosis was correct. Similarly, the number of fixations was higher for regions of interest relevant for the correct diagnosis ($p = 0.020$) but not for the incorrect diagnosis ($p = 0.176$) if the suggestion was incorrect. The interaction between the diagnostic suggestion and diagnostic accuracy was significant for the regions of interest for the regions of interest for the correct diagnosis (p = 0.038, Figure 2), though not for the regions of interest for the incorrect suggestion (p = 0.830). The interaction showed that residents fixated more on information relevant to the correct diagnosis when they arrived at the final correct diagnosis despite receiving the incorrect diagnosis, compared to when they did not arrive at the correct diagnosis or when they received a correct suggestion. The covariates case and participant did not explain these effects (p > 0.050).

Confidence

Residents' confidence did not differ depending on the diagnostic suggestion (p = 0.953) or their diagnostic accuracy (p = 0.839). There was no interaction (p = 0.189).

Time to diagnose

Residents did not take longer to diagnose a case depending on the diagnostic suggestion (p = 0.083) or their diagnostic accuracy (p = 0.142). There was no interaction (p = 0.239).

Table 3. *Descriptive statistics of confidence and total time spent on diagnosis depending on diagnostic suggestion and diagnostic accuracy. Relative dwell time and number of fixations are further divided between the type of region of interest (ROI).*

Condition (number of observations)	**Confidence Mean (SD)***	**Total time Mean (SD)***		**Relative dwell time Mean (SD)**	**Number of fixations Mean (SD)**
Correct suggestion Correct diagnosis (N = 64)	67 (14)	69 (33)	ROI – Correct diagnosis	7.0 (5)	16 (11)
			ROI – Incorrect diagnosis	5.3 (4)	14 (12)
Correct suggestion Incorrect diagnosis (N = 48)	64 (18)	68 (36)	ROI - Correct	6.4 (4)	16 (12)
			ROI - Suggestion	5.8 (4)	13 (11)
Incorrect suggestion Correct diagnosis (N = 22)	64 (17)	86 (47)	ROI - Correct	8.0 (4)	25 (20)
			ROI - Suggestion	3.6 (2)	11 (14)
Incorrect suggestion Incorrect diagnosis (N = 88)	67 (19)	71 (39)	ROI - Correct	7.0 (5)	17 (14)
			ROI - Suggestion	6.0 (4)	11 (10)

*Note: Confidence and time were calculated for each of the four conditions over both regions of interest related to the correct diagnosis and to the incorrect suggestion.

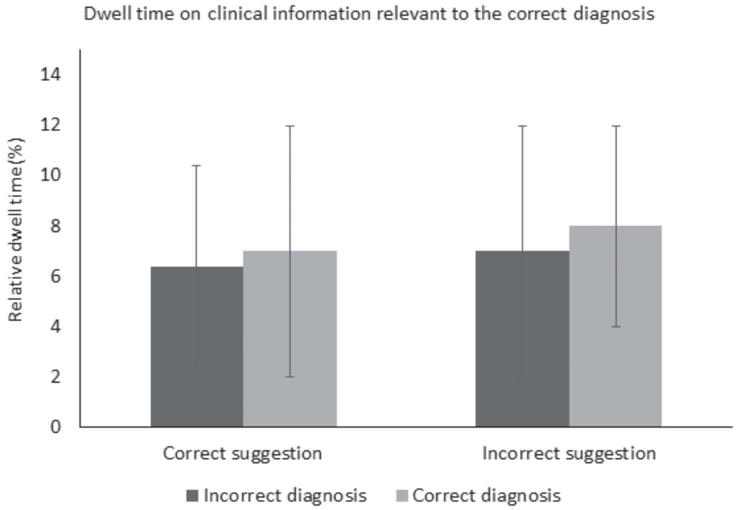


Figure 2. Average number of fixations on areas of interest containing clinical information relevant to the correct diagnosis. The figure shows an interaction between the diagnostic suggestion and the final diagnosis.

Discussion

The current study examined how residents' selectivity in information processing during diagnosis differed between cases where an error occurred compared to correctly diagnosed cases. No differences in selectivity were observed between error cases and correct cases, or between cases with an incorrect or a correct diagnostic suggestion. However, an interaction between diagnostic accuracy and the diagnostic suggestion was observed: residents looked more often at information necessary to make to the correct diagnosis if they made the correct diagnosis despite being biased by an incorrect diagnostic suggestion. If residents did not make the correct diagnosis at all, independent of the suggestion, or if they made the correct diagnosis in a case with a correct diagnostic suggestion, no differences were observed. This interaction showed a similar trend for both how often and how long residents looked at certain information (Figure 2) but the interaction was only significant for how often they looked (i.e., the number of fixations). These findings were partially in line with our hypotheses: selectivity in information processing was not necessarily related to residents' final diagnosis, unless they overcame a biased suggestion in order to arrive at that diagnosis. Residents' confidence in their diagnosis or their total time to diagnose did not differ depending on the diagnostic suggestion or their final accuracy. This was unexpected for confidence, which we hypothesized would vary depending on the match between the diagnostic suggestion and the final diagnosis, but expected for time to diagnose.

Our findings do not fully substantiate the hypothesis that selective information processing is associated with diagnostic errors. In fact, the selective processing of pertinent information was only observed when residents were able to overcome the biased diagnostic suggestion. This can be linked to Mamede et al.'s (35) finding that clinicians with higher knowledge (as measured by how many distinguishing disease features they recognized) were less susceptible to cognitive biases. So perhaps selectivity is guided by the proper prior knowledge to know what to look for, rather than an information processing behavior that can cause errors. Similar findings are reported in eye-tracking studies that examined clinical reasoning in visual diagnostic tasks. Generally, these studies show that compared to novices, experts in the relevant clinical domain spent more time looking at areas in the diagnostic image that were clinically relevant (36-39) and were faster at identifying abnormalities in these areas.(37, 40) Selectivity in information processing, specifically on areas that might provide pertinent information, might therefore be an indication of expertise and one's ability to make the correct diagnosis. Of course, the flipside of our main finding is the interpretation that in cases where the biased suggestion was not overcome, or where an incorrect diagnosis was made, that information relevant to the correct diagnosis was attended to less often, which can also be taken as a form of selectivity in information processing related to diagnostic errors. However, within error cases we did not observe any specific selectivity favoring either information relevant to the correct diagnosis or the biased suggestion, as we did in cases where the biased suggestion was corrected.

This interaction suggested that residents' information processing was not solely dictated by one's final diagnosis but also by a measure of effort expended in reaching that diagnosis, or awareness of the bias in the diagnostic suggestion. Refuting an incorrect suggestion would require more effort than simply agreeing with a correct suggestion, even if both final diagnoses were correct. This might alternatively explain why selectivity in information was only observed when the suggestion was corrected. Eye-tracking measures are also known to give an indication of the difficulty someone had with processing the presented information (33), and longer fixation times might also indicate more effort. "Correcting" a correct diagnostic suggestion with an incorrect final diagnosis, on the other hand, would likely be a similar process to refuting an incorrect suggestion, even though that did not result in differences in selectivity. Just detection of a bias or effort expended on a case can therefore not fully explain our findings, although these factors likely play a part in the measured eye-tracking behaviors.

A question of causality does remain for the interaction we describe. Did residents who spent more time looking at relevant information arrive at the correct diagnosis because they processed this relevant information more? Or did residents with the appropriate knowledge to make the correct diagnosis know which information they needed to look for to distinguish between the biased suggestion and the correct diagnosis? Based on the previous studies in visual diagnosis (36-40), the latter explanation might be more likely: experts, similar to those with higher knowledge, seem to know what information is clinically relevant and are thus able to process the necessary information and make the correct diagnosis by focusing more on those areas. In terms of our study, this could mean that residents who overcame the bias might have been more “expert” on a certain diagnosis than their peers, which might have translated in the more “expert” search pattern we observed. The differences that emerged in accuracy between cases might further indicate that case-specific knowledge could play a role. Additionally, our comparison was made within participants. Residents’ time spent looking at certain information under specific circumstances, which would not be expected in the same participant unless something occurred to make them look longer. This is, however, speculation based on the observed patterns, as our experiment could not differentiate between the possible underlying mechanisms of information processing. Future research might further specify and confirm or falsify expectations for patterns in information processing or visual search behavior, which might elucidate possible mechanisms.

Some strengths and limitations of this study should be considered when interpreting the results. This study improves on previous studies into clinicians’ reasoning processes by prospectively inducing diagnostic errors to avoid hindsight bias. Additionally, we assessed information processing in both correct and incorrect diagnoses. The use of eye-tracking furthermore allowed a more objective observation of clinicians’ reasoning than self-report measures. We also included 12 observations per participant and our participants were relatively experienced, which allowed us insight in the diagnostic reasoning process as it might occur in practice. However, comparisons to practice were limited on other aspects. The clinical case vignettes we used presented all necessary information simultaneously and the usual cyclicality of the diagnostic process and its progression over time were not incorporated. Residents could also not request extra information, so we could not measure whether follow-up was neglected or whether residents would gather all relevant information themselves. We did make sure to not only include relevant information in the case but also additional and not strictly necessary information, that was not relevant for either the correct or the bias diagnosis. We could therefore still observe whether residents were able to pick out the useful information when diagnosing a case. Furthermore, while the eye-tracking methodology is an objective measure of attention and information processing, it only accounts for overt attention, or the information that participants were directly (and probably consciously) focusing on. Any covert attention or peripherally observed information cannot be accounted for using eye-tracking.(41, 42) Perhaps this is not as problematic for written case vignettes as it might be for visual diagnostic material, however, people can also absorb information from quickly scanning text, even without explicit focus. If residents would then have to determine which differential diagnosis was the most likely, they would probably have to explicitly focus on the relevant information anyway, but it should be noted that eye-tracking does not cover all information processing that occurs. This is further the case in our experiment, where we solely focused on residents' ability to use information that distinguished between the correct and the incorrect diagnosis. Other information was not assessed. Eye-tracking experiments to assess information processing during diagnosis could be designed in numerous ways, for example by inducing a bias other than confirmation bias or by selecting different regions of interest, or cases of other levels of difficulty. Examining the effects of such factors could be valuable for future research. Finally, though correcting for participant and case number did not affect the results, individual differences or differences in case difficulty might have unobserved effects, as we aggregated over all observations. These factors might be further examined in future research.

In conclusion, selectivity in information processing likely plays a more complex role than simply being a direct cause of errors. The current results suggest that appropriate selectivity was associated with refuting incorrect diagnoses, whereas no specific differences were observed between correct and incorrect diagnoses in general. Selectivity in information processing might be a marker of cognitive processes underlying diagnostic errors rather than a cause of such errors. Eye-tracking is a valuable method to more objectively and precisely test hypotheses regarding information processing and it could be useful in refuting or confirming assumptions about information processing that could distinguish between the role of knowledge and biases in reasoning. Future research should include more participants to assess individual differences and should explore the influence of many more factors, such as different types of information or cognitive biases on clinicians' information processing. Better understanding how information processing occurs and what assumptions can be made about information processing during diagnosis will require more research.

Acknowledgements

We would like to thank Jonathan van Leeuwen for his valuable help and input during data analysis and Jacky Hooftman for her help with collecting data. Furthermore, we thank the residents who participated in our study.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. Royce CS, Hayes MM, Schwartzstein RM. Teaching critical thinking: a case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. Academic Medicine. 2019;94(2):187-94.
3. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic medicine. 2003;78(8):775-80.
4. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
5. Croskerry P. From mindless to mindful practice—cognitive bias and clinical decision making. N Engl J Med. 2013;368(26):2445-8.
6. Wachter RM. Why diagnostic errors don’t get any respect—and what can be done about them. Health Affairs. 2010;29(9):1605-10.
7. Kahneman D. Thinking, fast and slow: Macmillan; 2011.
8. Evans JSBT. In two minds: dual-process accounts of reasoning. Trends in cognitive sciences. 2003;7(10):454-9.
9. Osman M. An evaluation of dual-process theories of reasoning. Psychonomic bulletin & review. 2004;11(6):988-1010.
10. De Neys W. Bias, conflict, and fast logic: Towards a hybrid dual process future? Dual process theory 20: Routledge; 2017. p. 47-65.
11. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annual review of psychology. 2011;62(1):451-82.
12. Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: models of bounded rationality. Psychological review. 1996;103(4):650.
13. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. Dialogues in clinical neuroscience. 2022.
14. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education. 2009;14(1):27-35.
15. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203.
16. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.
17. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. Journal of Evaluation in Clinical Practice. 2018;24(3):666-73.
18. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine. 2017;92(1):23-30.
19. Schiff GD, Hasan O, Kim S, Abrams R, Cosby K, Lambert BL, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. Archives of internal medicine. 2009;169(20):1881-7.

20. Baartmans MC, Hooftman J, Zwaan L, van Schoten SM, Erwich JJHM, Wagner C. What Can We Learn From In-Depth Analysis of Human Errors Resulting in Diagnostic Errors in the Emergency Department: An Analysis of Serious Adverse Event Reports. Journal of Patient Safety. 2022:10.1097.

21. Zwaan L, Thijs A, Wagner C, Timmermans DRM. Does inappropriate selectivity in information use relate to diagnostic errors and patient harm? The diagnosis of patients with dyspnea. Social science & medicine. 2013;91:32-8.

22. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DRM. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. Academic Medicine. 2012;87(2):149-56.

23. Mamede S, Van Gog T, Van Den Berge K, Van Saase JLCM, Schmidt HG. Why do doctors make mistakes? A study of the role of salient distracting clinical features. Academic Medicine. 2014;89(1):114-20.

24. Kumar B, Ferguson K, Swee M, Suneja M. Diagnostic Reasoning by Expert Clinicians: What Distinguishes Them From Their Peers? Cureus. 2021;13(11).

25. Crowley RS, Naus GJ, Stewart Iii J, Friedman CP. Development of visual diagnostic expertise in pathology: an information-processing study. Journal of the American Medical Informatics Association. 2003;10(1):39-51.

26. Zwaan L, Monteiro S, Sherbino J, Ilgen J, Howey B, Norman G. Is bias in the eye of the beholder? A vignette study to assess recognition of cognitive biases in clinical case workups. BMJ quality & safety. 2017;26(2):104-10.

27. Zwaan L, Schiff GD, Singh H. Advancing the research agenda for diagnostic error reduction. BMJ quality & safety. 2013;22(Suppl 2):ii52-ii7.

28. Blondon KS, Wipfli R, Lovis C, editors. Use of eye-tracking technology in clinical reasoning: a systematic review. Mie; 2015.

29. Brunyé TT, Drew T, Weaver DL, Elmore JG. A review of eye tracking for understanding and improving diagnostic interpretation. Cognitive research: principles and implications. 2019;4(1):1-16.

30. Sqalli MT, Al-Thani D, Elshazly MB, Al-Hijji M, Alahmadi A, Houssaini YS. Understanding Cardiology Practitioners' Interpretations of Electrocardiograms: An Eye-Tracking Study. JMIR human factors. 2022;9(1):e34058.

31. Al-Moteri MO, Symmons M, Plummer V, Cooper S. Eye tracking to investigate cue processing in medical decision-making: A scoping review. Computers in Human Behavior. 2017;66:52-66.

32. Just MA, Carpenter PA. A theory of reading: from eye fixations to comprehension. Psychological review. 1980;87(4):329.

33. Findlay JM, Gilchrist ID. Active vision: The psychology of looking and seeing: Oxford University Press; 2003.

34. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91.

35. Mamede S, Goeijenbier M, Schuit SCE, de Carvalho Filho MA, Staal J, Zwaan L, et al. Specific disease knowledge as predictor of susceptibility to availability bias in diagnostic reasoning: a randomized controlled experiment. Journal of general internal medicine. 2021;36(3):640-6.

36. Giovinco NA, Sutton SM, Miller JD, Rankin TM, Gonzalez GW, Najafi B, et al. A passing glance? Differences in eye tracking and gaze patterns between trainees and experts reading plain film bunion radiographs. The Journal of Foot and Ankle Surgery. 2015;54(3):382-91.

37. Krupinski EA, Tillack AA, Richter L, Henderson JT, Bhattacharyya AK, Scott KM, et al. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. Human pathology. 2006;37(12):1543-56.

38. Matsumoto H, Terao Y, Yugeta A, Fukuda H, Emoto M, Furubayashi T, et al. Where do neurologists look when viewing brain CT images? An eye-tracking study involving stroke cases. PloS one. 2011;6(12):e28928.

39. Wood G, Knapp KM, Rock B, Cousens C, Roobottom C, Wilson MR. Visual expertise in detecting and diagnosing skeletal fractures. Skeletal radiology. 2013;42(2):165-72.

40. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology. 2007;242(2):396-402.

41. Posner MI. Orienting of attention. Quarterly journal of experimental psychology. 1980;32(1):3-25.

42. Belopolsky AV, Theeuwes J. When are attention and saccade preparation dissociated? Psychological Science. 2009;20(11):1340-7.

Appendix A – Clinical cases

Below the cases used in the experiment are presented, both with their correct diagnosis and the incorrect diagnostic suggestion. All cases were presented to participants in Dutch, but were translated to English.

Furthermore, the used regions of interest are indicated in each case. The colour blue indicates a region of interest containing clinical information necessary to make the correct diagnosis, which is not relevant for the incorrect diagnostic suggestion. The colour green indicates a region of interest containing clinical information necessary to make the incorrect suggested diagnosis, which is not relevant for the correct diagnosis.

Case 1

A 32-year-old woman presents to the Emergency Department with abdominal pain. She has had the pain for two days. The pain is sharp and radiates to the groin, currently rated 7 out of 10. It began around the umbilicus and is now moving to the right lower quadrant. She is also experiencing nausea and has vomited twice. During transportation, the pain increased with bumps on the road. No fever was measured, but she feels clammy and sweaty. She has a steady partner and is undergoing IVF treatment due to a desire to have children.

Medical history: polycystic ovary syndrome, asthma. Medication: none.

Physical examination: Moderately ill, painful woman. Blood pressure 124/68. Pulse 97 Temperature 37.7°C.
Abdomen: sparse peristalsis. Varied tympany. Tenderness in the right lower quadrant. Dubious rebound tenderness.

Laboratory results: Leukocytes 12 (reference 4-10); CRP 39 (reference <10).
Urinalysis: 2+ leukocytes, negative HCG.
No additional testing was performed.

Provisional diagnosis: ovarian torsion (correct) or appendicitis (incorrect).

Case 2

A 73-year-old man presents to the Emergency Department with dyspnea. He has had swelling in his legs for the past two weeks. He also feels more tired, sleeps poorly, and has difficulty lying flat. He has been waking up 2-3 times a night to urinate. He has experienced shortness of breath with exertion for the past week, but has not had any chest pain. His weight has increased, despite normal eating and drinking habits. The patient smokes 2 packs of cigarettes per week and drinks alcohol moderately.

Medical history: Type 2 diabetes since 1998. Myocardial infarction in 2017. Under evaluation by urologist for elevated PSA, no diagnosis yet. Medications: Metformin, Gliclazide, aspirin, Prasugrel, Simvastatin, Bisoprolol, Perindopril.

Physical examination: Blood pressure 165/87. Pulse 82. Saturation 95%. Elevated central venous pressure.
Lungs: crackles heard bilaterally at the bases.
Heart: normal sounds, no murmurs.
Extremities: pitting edema up to the knees.

Laboratory results: Hemoglobin 8.2 g/dL (reference 8.6-10.5); Creatinine 109 µmol/L (reference 65-115); Cardiac enzymes negative; Albumin 22 g/L (reference 35-50).
Urinalysis: 4+ protein, 2+ leukocytes, 1+ erythrocytes.
ECG: Sinus rhythm, normal conduction intervals. Pathologic Q waves in leads II, III, and aVF.

Provisional diagnosis: nephrotic syndrome (correct) or heart failure (incorrect).

Case 3

A 24-year-old female presents to the Emergency Department with thoracic pain complaints. She had a recent episode of gastroenteritis in the preceding week, with fever, nausea, and diarrhea. As a result, she spent a lot of time in bed. Since one day, she has been experiencing stabbing chest pain that is retrosternal and worsens with breathing. She also feels feverish. She has never had this pain before.

Medical history: Past medical history is unremarkable except for the use of oral contraceptives.

Physical examination: Painful woman. Pulse 109/min. Temperature 38.1.
Lungs: vesicular breath sounds bilaterally.
Heart: S1S2, no murmurs.
Extremities: supple calves.

Laboratory results: Leukocyte count 7.1 (reference 4-10); CRP 52 (reference < 10); D-dimer 0.85 (reference < 0.50).
ECG: Sinus tachycardia, normal conduction times. Diffuse subtle ST elevations.

Provisional diagnosis: viral pericarditis (correct) or pulmonary embolism (incorrect).

Case 4

A 30-year-old man presents to the Emergency Department with concerns about weight loss. He had traveled around the world last year and lost 7kg in a few months after returning home (from 85kg to 78kg). His dietary intake is normal, and he has no difficulties with digestion. He tried a low FODMAP diet without effect. He reports feeling bloated with excessive flatulence. His stool consistency is slightly softer than usual. He also feels tired and lacking in energy. He does not smoke but consumes alcohol socially.

Medical history: Past medical history includes recurrent respiratory infections. He recently tested positive for H. influenza.

Physical examination A man who is not acutely ill. Abdomen: No abnormalities.

Laboratory results: Hemoglobin level 7.2 (reference 8.6-10.5); Mean corpuscular volume (MCV) 103 (reference 80-100); Vitamin B12 120 (reference 145-637).

Provisional diagnosis: giardia lamblia infection (correct) or coeliac disease (incorrect).

Case 5

A 72-year-old man presents to the Emergency Department with pain in his back, radiating to the left groin. The pain started quite suddenly this morning, and he rates it as 7/10 on the visual analogue scale. He feels nauseous but has not vomited. He has no difficulty with transportation and no fever. He feels restless and cannot find a comfortable position. He had a normal diet and bowel movement yesterday and no dysuria. He stopped smoking 5 years ago and drinks 2-3 units daily.

Medical history: hypertension, type 2 diabetes, kidney stones, and gout. Medications: metformin, amlodipine, and allopurinol.

Physical examination Painful man, sweaty, slightly pale. BP 194/112. HR 92/min. Temperature 37.2. Sat 95% on room air, Respiratory rate 24/min. Heart/lungs normal.
Abdomen: soft, obese, no palpable abnormalities, slight tenderness in upper abdomen radiating to the back.

Laboratory results: Hb 9.1 (reference 8.6-10.5); WBC 11.6 (reference 4-10); Creatinine 97 (reference 65-115); Na 136 (reference 135-145); K 5.1 (reference 3.8-5.0); CRP 21 (reference <10).
Urine: leukocytes negative, erythrocytes negative, nitrite negative, protein 1+.
Renal ultrasound: left kidney has calculi, no hydronephrosis or dilated ureters.

Provisional diagnosis: thrombotic thrombocytopenic purpura (TTP) (correct) or immune thrombocytopenic purpura (ITP) (incorrect).

Case 6

62-year-old man presents to the emergency department with general malaise and abdominal pain. He has been experiencing pain in his abdomen, particularly in the epigastrium, for about a month, with a slow progressive course. Additionally, he has been experiencing nausea, vomiting, and reduced food intake, and for the past week, he has only been able to consume liquids. He has lost 10kg of weight. He has not had a fever and his bowel movements have been normal to firm. He has been experiencing difficulty with urination for some time and today noticed hematuria.

Medical history: hypertension and type 2 diabetes mellitus.

Physical examination A slightly weakened man. Blood pressure 155/94. Pulse 86/min. No fever.
Abdomen: lively peristalsis, changing tympany, supple. Tenderness in epigastrium.

Laboratory results: Hb 6.5 (reference 8.6-10.5);Tr 250 (reference 150-370); Kreat 295 (reference 65-115); Calcium 3.40 (reference 2.20-2.65); Phosphate 1.65 (reference 0.8-1.4); PSA 12 (reference < 4.5); 25-hydroxyvitamin D 142 ng/mL (reference 50-120); 1,25-dihydroxyvitamin D 170 pg/mL (reference 55.2-223.2).
Additional tests: X-ray; bi-hilar lymphadenopathy.

Provisional diagnosis: sarcoidosis (correct) or metastatic prostate cancer (incorrect).

Case 7

A 25-year-old man comes to the outpatient clinic with swollen lymph nodes in his neck. He noticed the swelling three weeks ago and went to see his general practitioner when it didn't go away. The lymph nodes appeared during a period of sore throat. Currently, he only has flu-like symptoms, has lost 3 kg, and sweats more at night, having to change the sheets. He feels somewhat feverish but has not measured it.

Medical history: Past medical history is unremarkable.

Physical examination: Non-acutely ill man. Slightly swollen eyes. Temperature: 36.8 °C. No abnormalities over heart and lungs. Multiple submandibular and cervical lymph nodes. Slightly tender on palpation. Some tenderness in the liver region.

Laboratory results: Hb 8.6 (reference 8.6-10.5); MCV 87 (reference 82-98); Leukocytes 13.5 x 10^9 (reference <10 x 10^9/L); with atypical lymphocytes in the differential count; ASAT 215 (reference <45); ALAT 310 (reference <50); LD 430 (reference 135-225); Bili 12 (reference <17); CRP 45 (reference <10).
Chest X-ray: No abnormalities.

Provisional diagnosis: Epstein-Barr virus (EBV) infection (correct) or lymphoma (incorrect).

Case 8

63-year-old woman comes to the emergency department with abdominal pain. She was recently treated for urinary tract infection. A week ago, she developed a bloated abdomen, followed by multiple episodes of diarrhea and abdominal pain. The diarrhea is greenish and later had blood in it. She used loperamide, which seemed to improve the symptoms. Due to the medication, she has less frequent bowel movements. Last night, she became sicker, with fever and chills. She has a distended abdomen and severe abdominal pain.

Medical history: Recurrent urinary tract infections. Two previous cesarean sections. Laparoscopic hysterectomy due to myomas. COPD GOLD I, for which she uses inhalation medication.

Physical examination: Ill woman. Blood pressure 86/55. Pulse 120. Temperature 38.2 °C. Respiratory rate 30/min. No abnormalities over heart and lungs.
Abdomen: Status post multiple surgeries. Distended abdomen. Hyperresonant percussion. Almost no peristalsis with occasional tinkling. Tender on palpation. No rebound tenderness.

Laboratory results: Na+ 121 (reference <10); CRP 175 (reference <10); Urea 8.4 (reference 2.5-7.5); Creatinine 134 (reference 45-100); Hb 6.8 (reference 7.5-9.5); MCV 83 (reference 82-98).
Clostridium toxin: positive.
No additional tests were performed.

Provisional diagnosis: toxic megacolon (correct) or ileus (incorrect).

Case 9

A 72-year-old woman presents to the emergency department with decreased level of consciousness. She did not wake up well in the morning. She had diarrhea and episodes of nausea in the past few days. She has been eating and drinking less. Her general practitioner visited her a day ago. Antihypertensive medication was temporarily stopped. It was agreed to take stool cultures for the diarrhea. Other medication could be continued. Temazepam was prescribed for sleeping.

Medical history: Depressive disorder. COPD Gold II. Type 2 diabetes mellitus, treated with two types of oral medication. Patient does not want insulin despite high HbA1C levels. Hypertension, treated with enalapril and hydrochlorothiazide. Dyslipidemia, treated with a statin.

Physical examination: E2M4V2. Patient has snoring breath sounds. AHF 16 times per minute. Temperature 36.7 °C. BP 124/80 mmHg. Pulse 102/min. Pupils 4+/4+. No neck stiffness. No lateralization on painful stimulus.

Laboratory results: Na 132 (reference 135-145); K 4.9 (reference 3.8-5.0); Creatinine 300 (reference 55-90) (last known value 75); Venous blood gas: pH 7.36; pCO2 6 (reference 45); Base excess -2 / HCO3- 21.
CT brain: no abnormalities except for some age-related atrophy.

Provisional diagnosis: hypoglycemia (correct) or benzodiazepine intoxication (incorrect).

Case 10

68-year-old man presents to the outpatient clinic due to yellow vision. His wife noticed that he has yellow sclerae. He is surprised that his general practitioner referred him as he has no complaints himself. Upon questioning, he mentions that he has lost some weight. No clearly reduced appetite. Stool is lighter in color. Urine is darker in color. Intoxications: Smokes 3 packs of shag/week. 2-6 alcoholic drinks per day.

Medical history: Laparoscopic appendectomy. Head trauma after falling from a bicycle.

Physical examination: Unwell man. Occasionally incoherent.
Chest: normal heart sounds, no murmurs.
Abdomen: slightly distended. Sparse peristalsis. Variable tympany. Soft. Tenderness in the right upper abdomen. No other abnormalities on physical examination.

Laboratory results: ASAT 294 (reference <45); ALAT 160 (reference <50); AF 150 (reference <15); gGT 580 (reference 5-50); Bilirubin 78 (reference <17).

Provisional diagnosis: alcoholic hepatitis (correct) or pancreatic cancer (incorrect).

Case 11

72-year-old man is referred to the outpatient clinic with a painful leg. He has had acute onset of pain in his left ankle and foot for 3 days. The leg also feels warmer and is slightly swollen. The foot is slightly reddish. The leg is painful to touch and stand on. He has not had a fever. He has never experienced this before. The patient does not smoke and consumes 2-3 alcoholic drinks per day.

Medical history: hypertension, type 2 diabetes mellitus, TIA, AF. Medications: Enalapril, Hydrochlorothiazide, Clopidogrel, Metformin, Apixaban.

Physical examination: Temperature: 37.3°C. Heart rate: 82. Blood pressure: 152/76. BMI: 31.
No abnormalities found over the heart, lungs, or abdomen
Extremities: Some edema and redness in the forefoot and ankle. The leg is warm and very painful to palpation and movement. Left leg is 2cm > right leg. Tangential pressure pain in the foot.

Laboratory results: Leukocytes 9 x 109 (reference <10x109); CRP: 34 (reference <10); Creatinine: 134 (reference 65-115).

Provisional diagnosis: gout (correct) or cellulitis (incorrect).

Case 12

A 58-year-old woman presents to the emergency department due to vomiting and watery diarrhea since this morning. She feels nauseous and vomits during intake. Upon standing, she feels lightheaded. She has been slightly congested in the past few days. No sick contacts. She ate normally until yesterday and craved more salt.

Medical history: Appendectomy Hypertension 2021. Melanoma, treated with nivolumab and ipilimumab. Medication: Metformin, Amlodipine, Allopurinol.

Physical examination: Ill-appearing woman, somewhat drowsy. Blood pressure 95/55. Heart rate 120. Temperature 37.2°C. Respiratory rate 24/min.
Abdomen: lively peristalsis, varying tympany, supple.

Laboratory results: Leukocytes 7 x 10^9/L (reference 4-10); CRP 12 mg/L (reference <10); Sodium 132 mmol/L (reference 135-145); Potassium 4.1 mmol/L (reference 3.8-5.0); Creatinine 54 µmol/L (reference 65-115); TSH 5.4 mU/L (reference 0.4-4.0).

Provisional diagnosis: obstructive sleep apnea syndrome (correct) or primary hyperaldosteronism (incorrect).

CHAPTER 5

Deliberate practice of diagnostic clinical reasoning reveals low performance and improvement of diagnostic justification in pre-clerkship students

Staal, J., Waechter, J., Allen, J., Hee Lee, C., Zwaan, L.

Accepted in BMC Medical Education (in press).

Abstract

Background: Diagnostic errors are a large burden on patient safety and improving clinical reasoning (CR) education could contribute to reducing these errors. To this end, calls have been made to implement CR training as early as the first year of medical school. However, much is still unknown about pre-clerkship students' reasoning processes. The current study aimed to observe how pre-clerkship students use clinical information during the diagnostic process.

Methods: In a prospective observational study, pre-clerkship medical students completed 10-11 self-directed online simulated CR diagnostic cases. CR skills assessed included: creation of the differential diagnosis (Ddx), diagnostic justification (DxJ), ordering investigations, and identifying the most probable diagnosis. Student performances were compared to expert-created scorecards and students received detailed individualized formative feedback for every case.

Results: 121 of 133 (91%) first- and second-year medical students consented to the research project. Students scored much lower for DxJ compared to scores obtained for creation of the Ddx, ordering tests, and identifying the correct diagnosis, (30-48% lower, $p < 0.001$). Specifically, students underutilized physical exam data ($p < 0.001$) and underutilized data that decreased the probability of incorrect diagnoses ($p < 0.001$). We observed that DxJ scores increased 40% after 10-11 practice cases ($p < 0.001$).

Conclusions: We implemented deliberate practice with formative feedback for CR starting in the first year of medical school. Students underperformed in DxJ, particularly with analyzing the physical exam data and pertinent negative data. We observed significant improvement in DxJ performance with increased practice.

Keywords: clinical reasoning, deliberate practice, diagnostic justification, pre-clerkship students

Introduction

Diagnostic errors, defined as missed, wrong, or delayed diagnoses, pose a significant burden on patient safety: most patients will likely experience one during their lifetime, sometimes with devastating consequences.(1) Flaws in clinical reasoning (CR), such as cognitive errors (2-8) or knowledge deficits, (9-12) are thought to be the main causes of diagnostic error. (10-13) CR encompasses many complex cognitive skills (14, 15) and is a core competency for graduating medical students.(16) Therefore, improving CR training could contribute to reducing diagnostic errors.(2, 13, 16, 17)

While *teaching* students about CR starts early in medical school, practice opportunities focused on *training* CR skill development does not start until clerkship, typically via observing expert clinicians and performing assessments on real patients.(18, 19) It is generally expected that students' CR skills will improve markedly and sufficiently during clerkships; however, this is contradicted by research showing that students' improvements are about similar to, or even less than their improvements in the pre-clerkship years.(20) This indicates that current CR training remains suboptimal, likely due to limitations in the methods for training and assessment of CR skills throughout medical school.(17, 27) Outside the workplace, both training and assessment are restricted by the time, funding, and manpower resources required to collect and analyze CR relevant data. Additionally, the current methods of assessment, such as using students' final diagnostic accuracy, have been doubted in their sensitivity to truly measure CR.(21)

One proposed solution includes beginning CR training for pre-clerkship students in first year medical school and throughout all phases of undergraduate medical education. (1, 19, 22-24) This will increase opportunities for formative feedback and allow students to start developing diagnostic skills prior to clinical rotations. Deliberate practice, the iterative process of repeated practicing and receiving formative feedback with simulation has also been proposed as an effective strategy for training CR.(23, 25, 26) Key aspects of CR that should be incorporated into medical school training and assessment include: building a Ddx, ordering tests, choosing a most probable diagnosis, and importantly, diagnostic justification. (22, 25-27)

Diagnostic justification (DxJ) is the process of identifying clinical data that increases or decreases the probability that a diagnosis is the correct diagnosis (or alternatively, is not the correct diagnosis). DxJ performance was observed to be below expectations in medical students and differentiates experts from novices.(25, 26, 29-31) Novices made errors because they had difficulty recognizing or interpreting relevant information, (32, 33) had limited knowledge of pertinent information (33) and underreported both positive, and to a larger extent, negative pertinent information.(34-36) These findings have primarily been observed in medical students during or after their clerkship training and much remains unknown about the reasoning processes of pre-clerkship students. When included as a component of assessment, DxJ was found to be the most predictive of graduate competency exam performance, have the highest item discrimination and increased assessment reliability.(26)

The current study aimed to determine how pre-clerkship medical students utilized clinical information in diagnostic cases. Our research questions (RQ) focused on overall processes of CR:

1. How do students perform at: creating a Ddx, performing DxJ, ordering and using investigations, and determining the correct Dx, and does this performance change with increased practice?
2. Within DxJ, are there differences in scores among the different categories of data (history, physical exam, and investigation results)?
3. Within DxJ, how do students assign data as increases versus decreases probability to the diagnoses in their Ddx and does this change with correct versus incorrect diagnoses?
4. How do first year students perform on all research questions compared to second year students?

We expected students would improve scores for all CR skills using deliberate practice (RQ1). RQ2 was observational and our null hypothesis was that there would be no differences. For RQ3 we expected more data to be assigned as “increases” than “decreases” for all diagnoses but we hypothesized that more data would be assigned to “decreases probability” for incorrect diagnoses than for correct diagnoses. Finally, we hypothesized that second year students would outperform first year students on all research questions (RQ4).

Methods

This was a prospective single-site observational study, approved by the Conjoint Health Research Ethics Board at the University of Calgary (REB19-0065) and the University of North Dakota (IRB00001300).

Participants

First and second year medical students were recruited from the University of North Dakota; the average age of the first-year medical students was 24. Simulated CR cases on teachingmedicine.com/dx were integrated in the mandatory curriculum and 133 students completed multiple cases throughout the school year. Students provided informed consent for their data to be included in research and did not receive compensation for participating.

Case Creation

Eleven text-based case vignettes were created for training. An 8-member committee representing internal medicine, critical care, brainstormed a minimum of three common diagnoses from 15 different systems/categories (total 58 diagnoses); 11 cases with typical presentations were created based on diagnoses from this list. The order of case presentation targeted 25% overlap with the currently taught organ system and 75% overlap with previously taught organ systems.

Each clinical scenario contained four stages: a one sentence introduction, the history, the physical exam, and investigations. The student could create a Ddx of up to 5 diagnoses and assign data from the history and physical exam to each diagnosis, indicating whether this data increases or decreases the probability of the diagnosis being correct. Students could do the same with investigation results. No investigation results were provided unless ordered by the student. Students could navigate back and forth between stages as needed. Further details about the CR software are described in a recent Innovation Report.[37]

Case completion by student

Students registered an online account on teachingmedicine.com, for which they provided their name, email, and a password. Cases were provided one at a time for the first-year students and one or two at a time for the second-year students. First year students completed 11 cases, one case per month and started the first case in the first month of their first year; second year students completed 10 cases spaced out over six months and started early in the second year. All students were given two to four weeks to complete each case and completed them during self-study. Students were encouraged to work in groups and to use internet searches and textbooks as needed. Case order was different between 1st and 2nd year students, but was the same for all students in a given year.

Students were provided with individualized formative feedback for each completed case. This feedback was based on a comparison between the over 100 data points collected per case and the scorecard (see *Scoring* below). Students' iterative cycle of practice and feedback with each case comprises deliberate practice. The feedback contained quantitative and qualitative information on the correct and incorrect choices they made when building their differential diagnosis, performing DxJ, ordering and using the results of investigations, and identifying the most probable diagnosis at the end of the case.

Students also attended a whole-class 1 hour video-conference review of each case, during which a faculty member demonstrated a "think aloud" demonstration of navigating the case, followed by a review of whole-class performance statistics and an informal online survey of the students. The survey results were collected and displayed to students and faculty during the review session and were used for curriculum improvement but were not included for research analysis.

Scoring

A scorecard was created for each clinical scenario (see Appendix for example). Specific diagnoses were designated as "appropriate" for the case; if an appropriate diagnosis was added to the Ddx by the student, a point was earned for building their Ddx. Data from the history, physical exam, and investigations were coded as "required", "neutral", or "wrong" for each category of "increases" or "decreases" probability for every Dx submitted by all students. A point was: earned if data was assigned where "required", missed if not assigned where "required" and half point deducted if assigned where "wrong". Investigations were coded as "required" or "inappropriate": the learner earned points for ordering appropriate tests and lost points for inappropriate tests. If the correct diagnosis was chosen as the most probable Dx for the case, the user scored 100% for this section. There were algorithms to score partial marks if the user did not assign the most probable Dx for the case as most probable, but instead, assigned it as less probable. Scores out of 10 were calculated for each of: Ddx, DxJ, investigations, and final diagnosis.

Statistical analysis

Performance data and scoring were collected and calculated using teachingmedicine.com; the data were then de-identified, exported, and analyzed using the programming language and statistical environment R-4.0.1 (R Core Team). The Kruskal-Wallis test was used to compare: 1) students' mean scores on creating a Ddx, performing DxJ, ordering and using investigations, and determining the most probable correct Dx; 2) how students performed on DxJ on the history, physical exam, and investigation results; and 3) how students assign pertinent positive and pertinent negative information to diagnoses. A linear mixed-effects model was used to explore changes in scores across cases. Finally, a Wilcoxon test and *t*-test were used to compare performance between first year and second year students.

Results

121 of 133 (91%) first- and second-year medical students consented to the research project and completed 10 and 11 clinical cases respectively. All cases were completed between August 2021 and May 2022.

RQ1: Student performance on building Ddx, DxJ, investigations, and final diagnosis

Figure 1 shows students' mean scores across all cases and all students. Students' mean scores differed significantly between building the Ddx (8.21, SD = 2.0), performing DxJ (3.9, SD = 1.6), investigations (5.65, SD = 2.6), and the final diagnosis (7.45, SD = 4.2) ($p < 0.001$). .

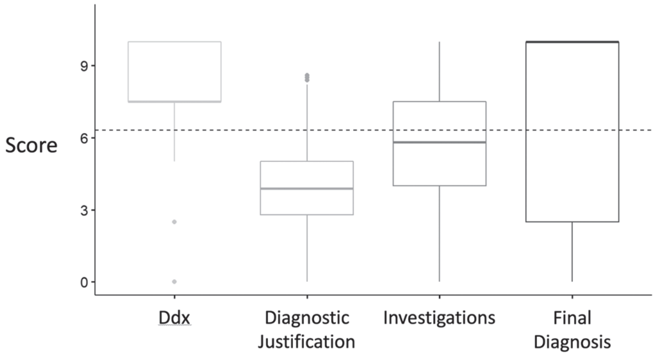


Figure 1. Mean score of all 4 scores for all cases completed by all students; all scores are statistically different from each other (p < 0.001) with Dx justification notably having the lowest score. The maximum score possible was 10 for each score. All scores are calculated independently from each other.

DxJ scores were 30% lower than scores for building the Ddx and 48% lower than for investigations. We investigated if these scores increased with deliberate practice (Figure 2). Only DxJ scores increased over 10-11 cases, from 3.13 to 4.40, showing an increase of 40% improvement (*p* < 0.001)

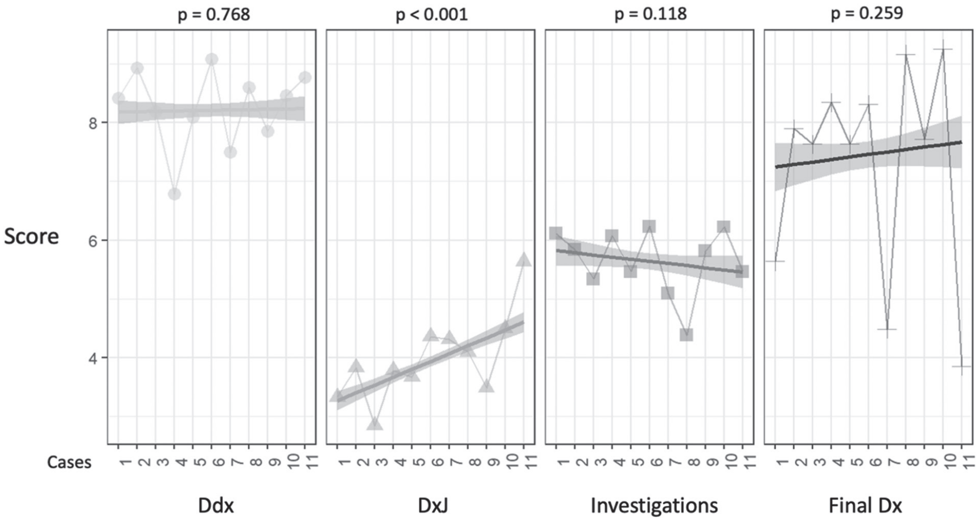


Figure 2. The X axis shows completed cases 1 through 11 for both 1st year (10 cases) and 2nd year (11 cases) students; the Y axis shows the scores for each of Ddx, Dx justification, investigations, and final Dx. Scores for Dx justification increased by 40% (p < 0.001); no statistical changes were observed for the scores for Ddx, investigations, and final Dx across the cases.

RQ4: First year versus second year student performance

Figure 3 shows that second year students scored higher than first year students for Ddx (7.88 vs 8.56), DxJ (3.67 vs. 4.14), and investigations (5.39 vs. 5.92), (*p* < 0.001 for all) but not for final diagnosis (7.32 vs. 7.58, *p* = 0.24).

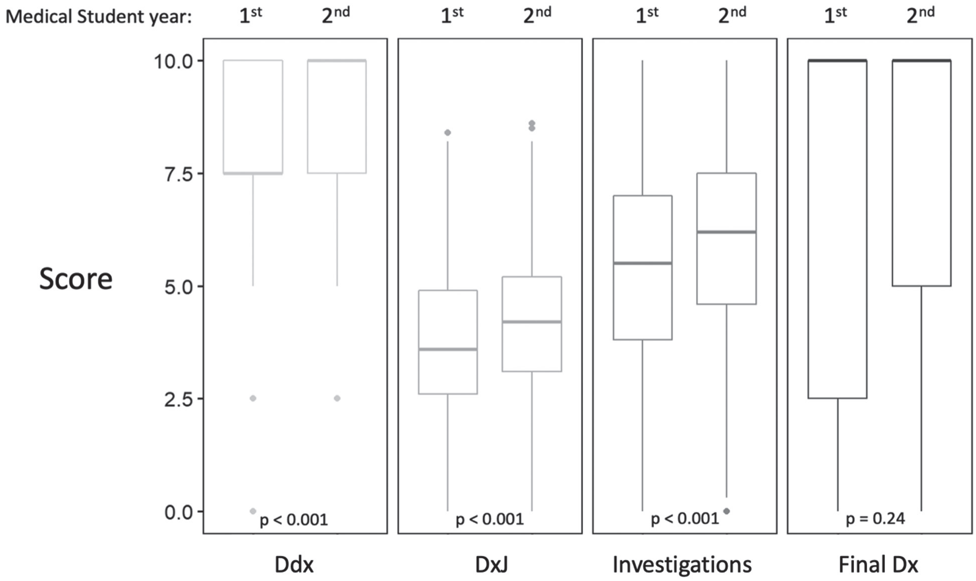


Figure 3. Comparison of 1st and 2nd year students. There were statistical differences between 1st and 2nd year students for all scores except the final dx (p < 0.05).

RQ2: Student DxJ performance in history, physical exam, and investigations

Figure 4 displays students' mean scores for DxJ in the history (M = 0.47, SD = 0.2), physical exam (M = 0.38, SD = 0.2), and investigations (M = 0.49, SD = 0.2). The maximum possible scaled value is 1.0. The physical exam score was significantly lower than the history and investigations scores. (*p* < 0.001); the difference between history and investigations was not statistically different (p = 0.058).

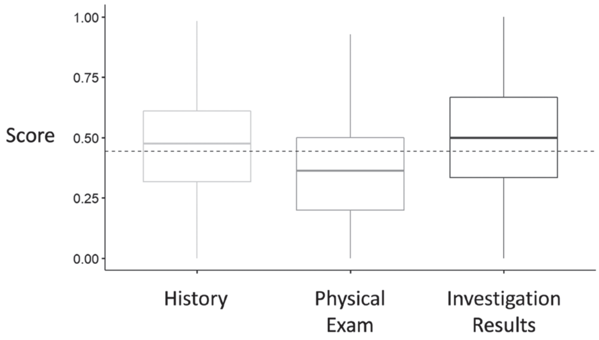


Figure 4. Mean Dx justification scores when classified by history, physical exam, and investigation results. Scores are scaled to a maximum of 1. Scores for the physical exam were significantly lower than both history and investigation results.

RQ4: First year versus second year student performance

Figure 5 shows that second year students scored higher DxJ scores for history (p = 0.001) and physical exam ($p < 0.001$), but not for investigations (p = 0.12).

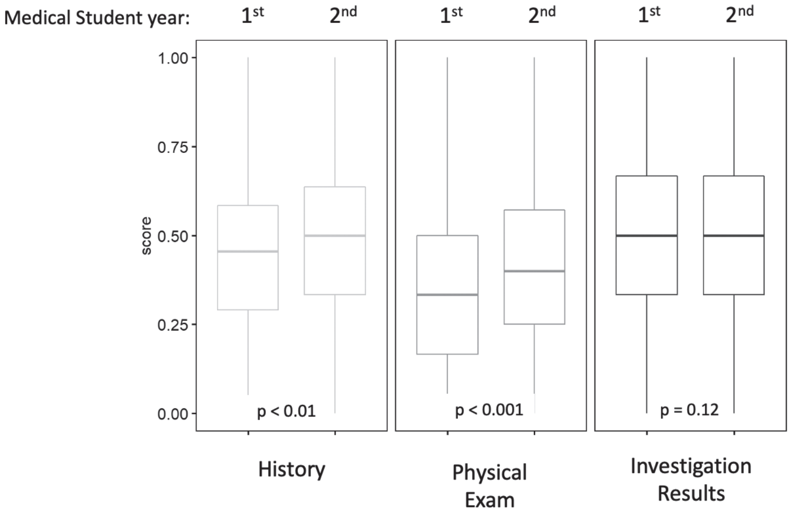


Figure 5. Comparison of scores for 1st versus 2nd year students. 2nd year students scored higher than 1st year students for Dx justification using history and physical exam data, but there was no observed difference for analyzing the investigation results.

RQ3: Student DxJ performance in assign data as increases versus decreases probability

Figure 6 shows that students assigned significantly more data to the “increase probability” category than to the “decrease probability” category (7.7 times more, $p < 0.001$) for both correct and incorrect diagnoses. We predicted, but did not observe, that the ratio of “decreases to increases” probability data would be higher for incorrect diagnoses compared to correct diagnoses ($p = 0.41$).

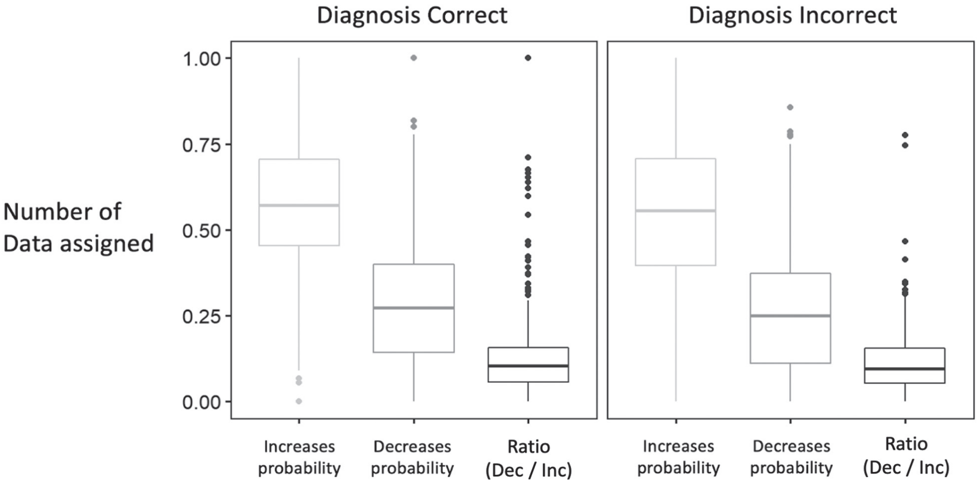


Figure 6. The ratio of decreases to increases data, compared for correct diagnoses and incorrect diagnoses, was measured using the number of data assigned as “increases probability” and “decreases probability”. Data is not weighted with respect to importance or magnitude of impact on probabilities. We predicted and observed that data was much more frequently assigned as “increases” compared to “decreases” probability for both correct and incorrect diagnoses. We predicted the ratio of “decreases to increases” data to increase for incorrect diagnoses but we observed no statistical difference between the ratios (value = 0.13 for both, $p = 0.41$).

RQ4: First year versus second year student performance

There were no differences observed for RQ3 between 1st and 2nd year medical students.

Discussion

We provided deliberate practice and formative feedback for diagnostic clinical reasoning to pre-clerkship medical students starting in the first month of first year medical school. We observed and analyzed students’ CR performance for the following activities: building a Ddx, performing DxJ, ordering investigations, and selecting a final dx. We made five important observations. First, students scored: well on constructing the Ddx and choosing the correct diagnosis; moderately on ordering investigations; and poorly with diagnostic justification. Second, after diagnosing 10-11 cases with formative feedback, performance on DxJ increased by 40%; the other scores remained unchanged. Third, clinical data was rarely assigned as "decreasing probability" for diagnoses, even when the diagnoses were incorrect. Fourth, DxJ scores were lower for physical exam data than for history and investigation data. Fifth, second year students performed better than first year students on most measures; notably however, they did not outperform first year students on analyzing data from the investigations section nor on identifying the correct diagnosis.

DxJ skills are crucial for medical students to develop before they are ready for clinical practice.(25) Once a Ddx has been created, the process of DxJ provides the structure and evidence to demonstrate that both the correct diagnosis is indeed present and that all the incorrect diagnoses within the Ddx are concurrently absent. We believe that DxJ is the most important part of CR; the *absence* of DxJ is no better than assuming or guessing. If DxJ is not being performed, then what process *is* being used to determine the probabilities of the diagnoses within the Ddx?

Overall, the current findings suggest that DxJ is a skill that is difficult to learn and requires a lot of practice or experience to develop. The low DxJ scores we observed in pre-clerkship students accord with previous studies reporting that students' DxJ scores were lower than expected *despite* exhibiting high final diagnostic accuracy.(25, 28, 38) One study showed that absences or major deficiencies in DxJ were observed in up to 48% of third year students.(26)

Our study showed that DxJ scores improved 40% with one year of deliberate practice. This suggests that DxJ is indeed a skill that can be developed and improved with formative feedback. This observation supports Hayden et al. (38), who showed a dose-response relationship of increased DxJ scores with increased attendance of CR simulation. However, 1 year of practice is likely inadequate, as scores remained low (4.4 out of 10) after 1 year of training.

Within DxJ processes, we predicted and observed that students assigned more data as "increases" than "decreases" probability to their Ddx. However, contrary to our prediction, we did not observe more data assigned as "decreases" probability for the incorrect diagnoses compared to correct diagnoses. This finding complements previous studies reporting that novices neglected to use pertinent negative information.(33-36) Another finding within DxJ processes that is important was that scores for physical exam data analysis were lower than for history or investigations data; to our knowledge, this observation has not been previously reported. It is possible that inexperienced students tend to ignore or under-appreciate data from the physical exam and similarly, data that decreases the probability of diagnoses. It is also possible that current methods of instruction do not sufficiently train students to appropriately analyze these clinical data. These findings should receive attention in future CR training methods; we have upgraded our curriculum accordingly to explicitly highlight this feedback to students based on these findings.

We observed that scoring for analyzing investigation results and identifying the final diagnosis were not different between 1st and 2nd year students; this observation could suggest that these metrics are invalid or ineffective methods of assessing CR. Other publications have similarly concluded that final diagnostic accuracy is an insensitive CR assessment tool because students can get to the right diagnosis even if they score low on CR processes.(21, 25, 28) DxJ is likely a better indicator of the quality of underlying reasoning processes; when DxJ is incorporated into assessment of CR, it has been observed to be the most predictive of graduate competency exam performance, have the highest item discrimination and increases assessment reliability.(26)

Strengths of this study include the large sample size, the large volume of data collected, and the highly organized storage structure of the data, making it easily analyzed at scale. The absence of multiple assessors of performance eliminates inter-observer variability since all scores were generated from a single unique scorecard for each case. We performed longitudinal data collection with 10-11 data collection events over a period of 6 to 10 months which we believe to be superior to a single assessment event. Furthermore, our data provides initial validity evidence based on observations that 1) second year students consistently outperformed first year students on most measures; 2) performance improved with practice; and 3) we replicated patterns found in previous studies on DxJ.

Limitations of the study include the generalizability of the data, the study design, and the creation of the scorecard. Generalizability was limited because all data were collected at a single site and on a limited number of cases. The study design was observational and does not allow us to draw causal conclusions. Lastly, given that this was the first year of deployment of this curriculum, we had two, but only sometimes three experts to review the scorecards of each case. We expect to discover and correct minor scorecard errors with increased expert review of the data in the scorecards.

Future research will extend our observations over 2 years and 20 cases for the same cohorts of learners. We are currently collecting data for a multi-centered project. We have upgraded our software to collect data to not only identify when and where the misdiagnoses occur, but to also inform why these misdiagnoses are occurring; this analysis will be included in our upcoming multi-site project.

In conclusion, pre -clerkship 1st and 2nd year medical students completed 10 and 11 deliberate practice CR cases respectively with formative feedback over one year in this single site study. Students scored particularly low for DxJ processes, especially related to physical exam data, and assigning data as “decreases” probability. We did observe, however, that DxJ scores improved by 40% within 1 year of deliberate practice. DxJ is a key component of CR and deliberate practice of CR starting in pre-clerkship students has now been shown to be a feasible and effective strategy to improve diagnostic CR skills.

Acknowledgements

We thank all students we participated in this study.

References

1. National Academies of Sciences, Engineering, and Medicine. 2015. Improving diagnosis in health care. Washington DC: The National Academies Press.
2. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic medicine. 2003;78(8):775-80.
3. Croskerry P. The cognitive imperative thinking about how we think. Academic Emergency Medicine. 2000;7(11):1223-31.
4. Evans JSBT. In two minds: dual-process accounts of reasoning. Trends in cognitive sciences. 2003;7(10):454-9.
5. Kahneman D. Thinking, fast and slow: Macmillan; 2011.
6. Frankish K. Dual-process and dual-system theories of reasoning. Philosophy Compass. 2010;5(10):914-26.
7. Elia F, Apra F, Verhovez A, Crupi V. “First, know thyself”: Cognition and Error in Medicine. Acta Diabetologica. 2016;53(2):169-75.
8. Phua DH, Tan NC. Cognitive aspect of diagnostic errors. Ann Acad Med Singapore. 2013;42(1):33-41.
9. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
10. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Medical education. 2010;44(1):94-100.
11. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. Journal of Evaluation in Clinical Practice. 2018;24(3):666-73.
12. Mamede S, de Carvalho-Filho MA, de Faria RMD, Franci D, Nunes MdPT, Ribeiro LMC, et al. ‘Immunising’ physicians against availability bias in diagnostic reasoning: a randomised controlled experiment. BMJ quality & safety. 2020;29(7):550-9.
13. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.
14. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. Bmj. 2002;324(7339):729-32.
15. Elstein AS, Shulman LS, Sprafka SH, Sprafka SA. 1978. Medical problem solving an analysis of clinical reasoning. Cambridge: Harvard University Press.

16.x Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. Academic Medicine. 2020;95(8):1166-71.

17. Graber ML, Rencic J, Rusz D, Papa F, Croskerry P, Zierler B, et al. Improving diagnosis by improving education: a policy brief on education in healthcare professions. Diagnosis. 2018;5(3):107-18.
18. Parsons AS. Harvard Macy Institute. 2020. [cited 2022]. Available from: https://harvardmacy.org/index.php/hmi/teaching-clinical-reasoning.
19. Rencic J, Trowbridge RL, Fagan M, Szauter K, Durning S. Clinical reasoning education at US medical schools: results from a national survey of internal medicine clerkship directors. Journal of general internal medicine. 2017;32(11):1242-6.
20. Williams RG, Klamen DL, White CB, Petrusa E, Fincher R-ME, Whitfield CF, et al. Tracking development of clinical reasoning ability across five medical schools using a progress test. Academic Medicine. 2011;86(9):1148-54.

21. Hege I, Kononowicz AA, Kiesewetter J, Foster-Johnson L. Uncovering the relation between clinical reasoning and diagnostic accuracy–an analysis of learner's clinical reasoning processes in virtual patients. PloS one. 2018;13(10):e0204900.

22. Cooper N, Bartlett M, Gay S, Hammond A, Lillicrap M, Matthan J, et al. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Medical Teacher. 2021;43(2):152-9.

23. Khin-Htun S, Kushairi A. Twelve tips for developing clinical reasoning skills in the pre-clinical and clinical stages of medical school. Medical teacher. 2019;41(9):1007-11.

24. Medicine TStIDi. Coalition for improved diagnosis.: The Society to Improve Diagnosis in Medicine; 2021 [updated December 15 2021]. Available from: https://www.improvediagnosis.org/coalition/.

25. Williams RG, Klamen DL. Examining the diagnostic justification abilities of fourth-year medical students. Academic Medicine. 2012;87(8):1008-14.

26. Yudkowsky R, Park YS, Hyderi A, Bordage G. Characteristics and implications of diagnostic justification scores based on the new patient note format of the USMLE Step 2 CS exam. Academic Medicine. 2015;90(11):S56-S62.

27. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. Academic Medicine. 2019;94(6):902-12.

28. Williams RG, Klamen DL, Markwell SJ, Cianciolo AT, Colliver JA, Verhulst SJ. Variations in senior medical student diagnostic justification ability. Academic Medicine. 2014;89(5):790-8.

29. Nendaz MR, Gut AM, Perrier A, Louis-Simonet M, Blondon-Choa K, Herrmann FR, et al. Brief report: beyond clinical experience: features of data collection and interpretation that contribute to diagnostic accuracy. Journal of general internal medicine. 2006;21(12):1302-5.

30. Hobu PPM, Schmidt HG, Boshuizen HPA, Patel VL. Contextual factors in the activation of first diagnostic hypotheses: expert-novice differences. Medical education. 1987;21(6):471-6.

31. Crespo KE, Torres JE, Recio ME. Reasoning process characteristics in the diagnostic skills of beginner, competent, and expert dentists. Journal of dental education. 2004;68(12):1235-44.

32. Brenner P. From novice to expert. American Journal of Nursing. 1982;82(3):402-7.

33. Gilliland S. Clinical reasoning in first-and third-year physical therapist students. Journal of Physical Therapy Education. 2014;28(3):64-80.

34. Corbett A, Sandholdt C, Bakerjian D, editors. Learning analytics with virtual patient data reveals subgroup of students who miss pertinent findings. Innovate Learning Summit; 2020: Association for the Advancement of Computing in Education (AACE).

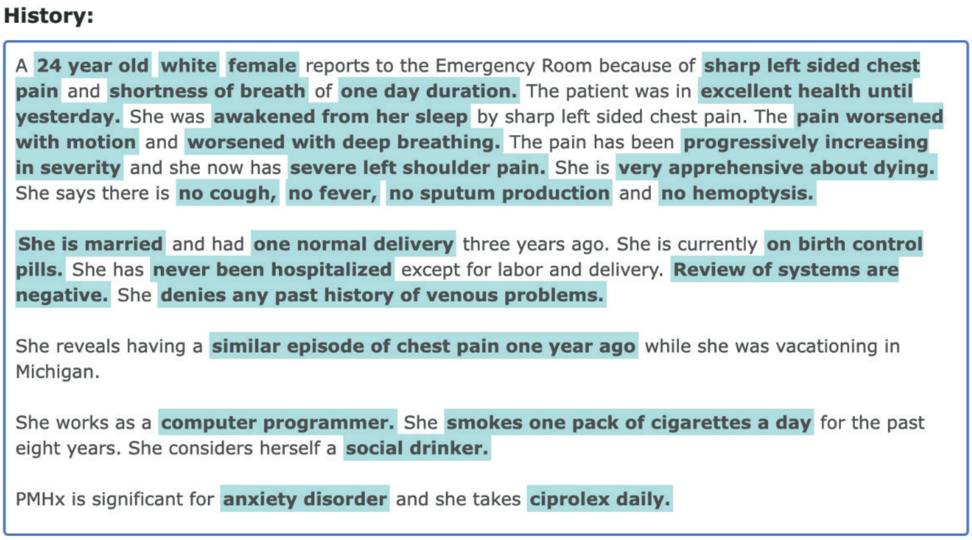
35. Walling A, Moser SE, Dickson G, Zackula RE. Are students less likely to report pertinent negatives in post-encounter notes? Family Medicine-Kansas City. 2012;44(1):22.

36. Kim J. Necessity of introducing post encounter note describing history and physical examination at clinical performance examination in Korea. Korean Journal of Medical Education. 2014;26(2):107-15.

37. Waechter J, Allen J, Lee CH, Zwaan L. Development and Pilot Testing of a Data-Rich Clinical Reasoning Training and Assessment Tool. Academic Medicine. 2022;97(10):1484-8.

38. Hayden EM, Petrusa E, Sherman A, Feinstein DM, Khoury K, Krupat E, et al. Association of Simulation Participation With Diagnostic Reasoning Scores in Preclinical Students. Simulation in Healthcare. 2022;17(1):35-41.

Supplementary Material A. Examples of the TeachingMedicine.com scorecard.



History:

A **24 year old** **white** **female** reports to the Emergency Room because of **sharp left sided chest pain** and **shortness of breath** of **one day duration.** The patient was in **excellent health until yesterday.** She was **awakened from her sleep** by sharp left sided chest pain. The **pain worsened with motion** and **worsened with deep breathing.** The pain has been **progressively increasing in severity** and she now has **severe left shoulder pain.** She is **very apprehensive about dying.** She says there is **no cough,** **no fever,** **no sputum production** and **no hemoptysis.**

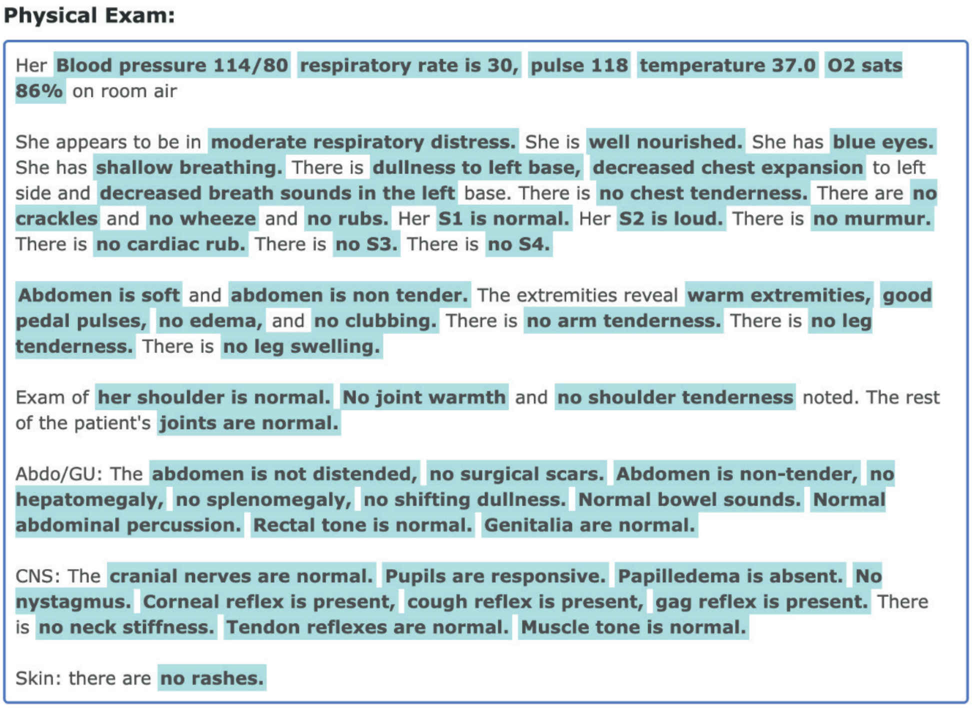
She is married and had **one normal delivery** three years ago. She is currently **on birth control pills.** She has **never been hospitalized** except for labor and delivery. **Review of systems are negative.** She **denies any past history of venous problems.**

She reveals having a **similar episode of chest pain one year ago** while she was vacationing in Michigan.

She works as a **computer programmer.** She **smokes one pack of cigarettes a day** for the past eight years. She considers herself a **social drinker.**

PMHx is significant for **anxiety disorder** and she takes **ciprolex daily.**

Figure S1: Example of clinical information students could select from the case history.



Physical Exam:

Her **Blood pressure 114/80** **respiratory rate is 30,** **pulse 118** **temperature 37.0** **O2 sats 86%** on room air

She appears to be in **moderate respiratory distress.** She is **well nourished.** She has **blue eyes.** She has **shallow breathing.** There is **dullness to left base,** **decreased chest expansion** to left side and **decreased breath sounds in the left** base. There is **no chest tenderness.** There are **no crackles** and **no wheeze** and **no rubs.** Her **S1 is normal.** Her **S2 is loud.** There is **no murmur.** There is **no cardiac rub.** There is **no S3.** There is **no S4.**

Abdomen is soft and **abdomen is non tender.** The extremities reveal **warm extremities,** **good pedal pulses,** **no edema,** and **no clubbing.** There is **no arm tenderness.** There is **no leg tenderness.** There is **no leg swelling.**

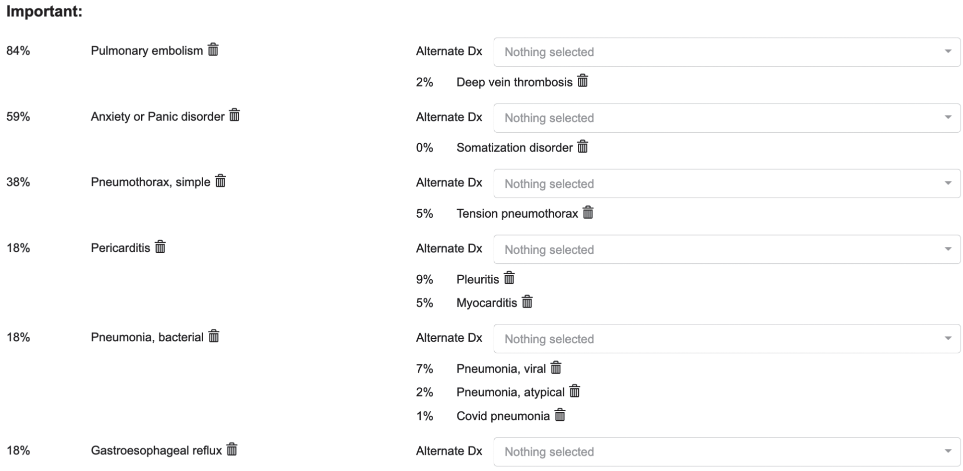
Exam of **her shoulder is normal.** **No joint warmth** and **no shoulder tenderness** noted. The rest of the patient's **joints are normal.**

Abdo/GU: The **abdomen is not distended,** **no surgical scars.** **Abdomen is non-tender,** **no hepatomegaly,** **no splenomegaly,** **no shifting dullness.** **Normal bowel sounds.** **Normal abdominal percussion.** **Rectal tone is normal.** **Genitalia are normal.**

CNS: The **cranial nerves are normal.** **Pupils are responsive.** **Papilledema is absent.** **No nystagmus.** **Corneal reflex is present,** **cough reflex is present,** **gag reflex is present.** There is **no neck stiffness.** **Tendon reflexes are normal.** **Muscle tone is normal.**

Skin: there are **no rashes.**

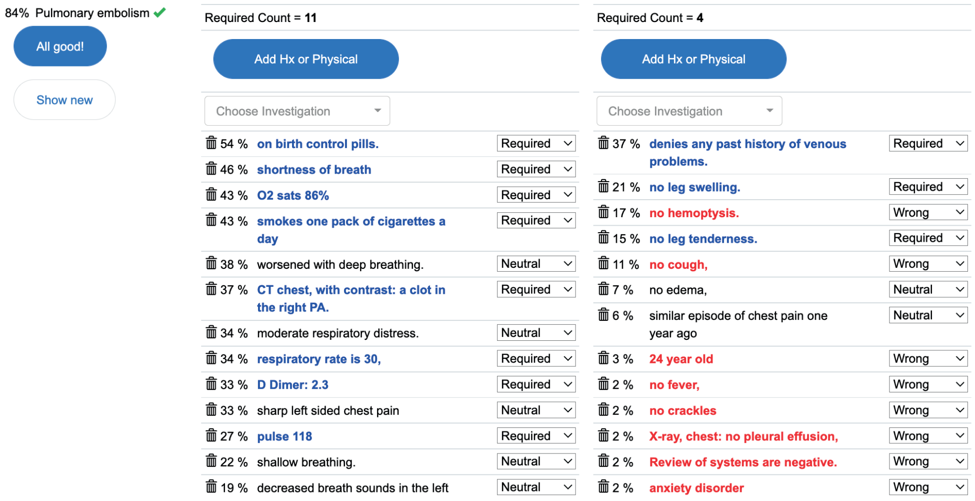
Figure S2: Example of clinical information students could select from the case physical exam.



Important:

84%	Pulmonary embolism 🗑	Alternate Dx	Nothing selected
		2%	Deep vein thrombosis 🗑
59%	Anxiety or Panic disorder 🗑	Alternate Dx	Nothing selected
		0%	Somatization disorder 🗑
38%	Pneumothorax, simple 🗑	Alternate Dx	Nothing selected
		5%	Tension pneumothorax 🗑
18%	Pericarditis 🗑	Alternate Dx	Nothing selected
		9%	Pleuritis 🗑
		5%	Myocarditis 🗑
18%	Pneumonia, bacterial 🗑	Alternate Dx	Nothing selected
		7%	Pneumonia, viral 🗑
		2%	Pneumonia, atypical 🗑
		1%	Covid pneumonia 🗑
18%	Gastroesophageal reflux 🗑	Alternate Dx	Nothing selected

Figure S3. Scorecard for building the differential diagnosis (Ddx).



84% Pulmonary embolism ✓

Required Count = **11**

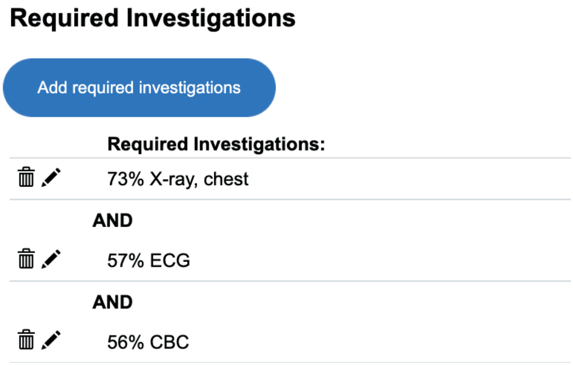
54 %	**on birth control pills.**	Required
46 %	**shortness of breath**	Required
43 %	**O2 sats 86%**	Required
43 %	**smokes one pack of cigarettes a day**	Required
38 %	worsened with deep breathing.	Neutral
37 %	**CT chest, with contrast: a clot in the right PA.**	Required
34 %	moderate respiratory distress.	Neutral
34 %	**respiratory rate is 30,**	Required
33 %	**D Dimer: 2.3**	Required
33 %	sharp left sided chest pain	Neutral
27 %	**pulse 118**	Required
22 %	shallow breathing.	Neutral
19 %	decreased breath sounds in the left	Neutral

Required Count = **4**

37 %	**denies any past history of venous problems.**	Required
21 %	**no leg swelling.**	Required
17 %	**no hemoptysis.**	Wrong
15 %	**no leg tenderness.**	Required
11 %	**no cough,**	Wrong
7 %	no edema,	Neutral
6 %	similar episode of chest pain one year ago	Neutral
3 %	**24 year old**	Wrong
2 %	**no fever,**	Wrong
2 %	**no crackles**	Wrong
2 %	**X-ray, chest: no pleural effusion,**	Wrong
2 %	**Review of systems are negative.**	Wrong
2 %	**anxiety disorder**	Wrong

Figure S4. Scorecard for Diagnostic Justification.

S5a

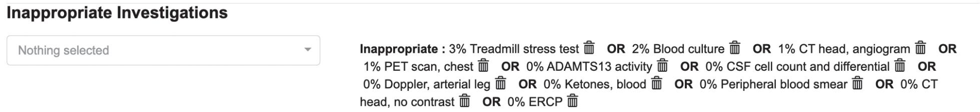


Required Investigations

Add required investigations

	Required Investigations:
🗑 ✎	73% X-ray, chest
	AND
🗑 ✎	57% ECG
	AND
🗑 ✎	56% CBC

S5b

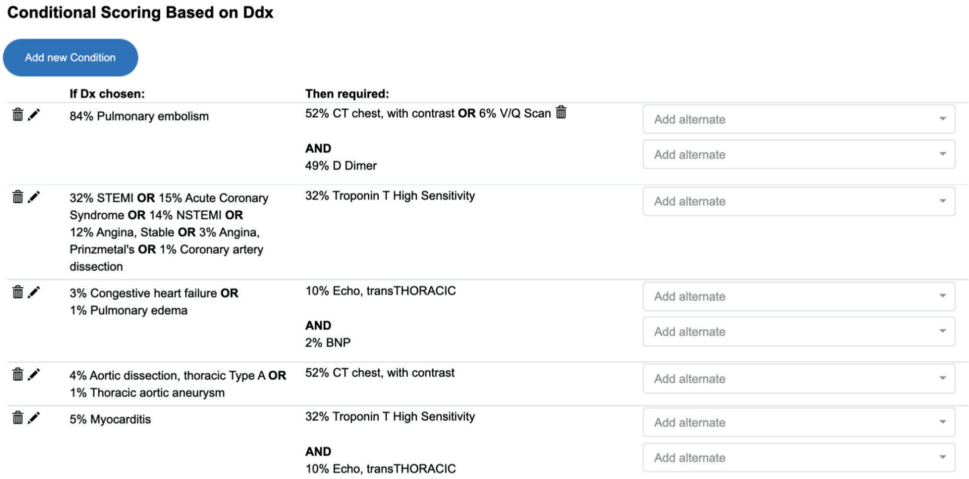


Inappropriate Investigations

Nothing selected

Inappropriate : 3% Treadmill stress test 🗑 **OR** 2% Blood culture 🗑 **OR** 1% CT head, angiogram 🗑 **OR** 1% PET scan, chest 🗑 **OR** 0% ADAMTS13 activity 🗑 **OR** 0% CSF cell count and differential 🗑 **OR** 0% Doppler, arterial leg 🗑 **OR** 0% Ketones, blood 🗑 **OR** 0% Peripheral blood smear 🗑 **OR** 0% CT head, no contrast 🗑 **OR** 0% ERCP 🗑

S5c



Conditional Scoring Based on Ddx

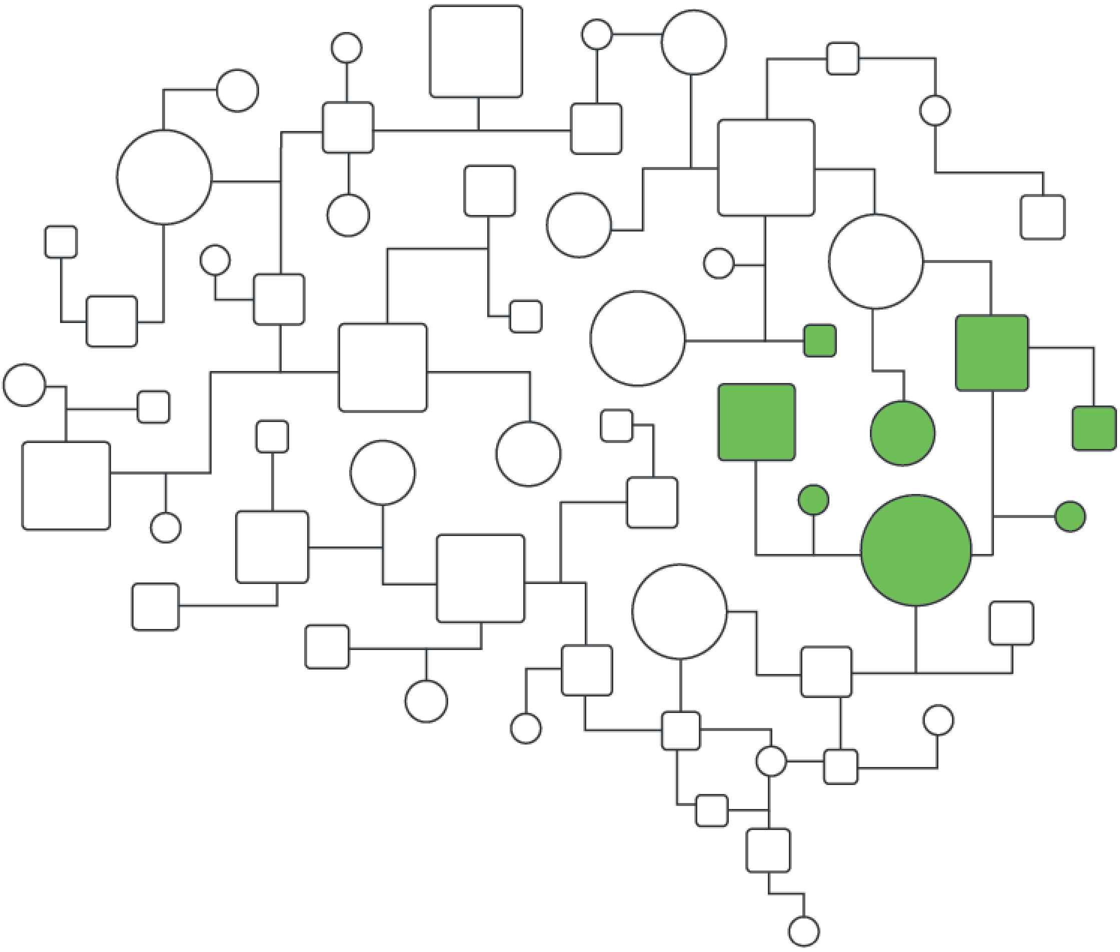
Add new Condition

	If Dx chosen:	**Then required:**	
🗑 ✎	84% Pulmonary embolism	52% CT chest, with contrast **OR** 6% V/Q Scan 🗑 **AND** 49% D Dimer	Add alternate Add alternate
🗑 ✎	32% STEMI **OR** 15% Acute Coronary Syndrome **OR** 14% NSTEMI **OR** 12% Angina, Stable **OR** 3% Angina, Prinzmetal's **OR** 1% Coronary artery dissection	32% Troponin T High Sensitivity	Add alternate
🗑 ✎	3% Congestive heart failure **OR** 1% Pulmonary edema	10% Echo, transTHORACIC **AND** 2% BNP	Add alternate Add alternate
🗑 ✎	4% Aortic dissection, thoracic Type A **OR** 1% Thoracic aortic aneurysm	52% CT chest, with contrast	Add alternate
🗑 ✎	5% Myocarditis	32% Troponin T High Sensitivity **AND** 10% Echo, transTHORACIC	Add alternate Add alternate

Figure S5. Scorecard for Investigations, consisting of a: Required Investigations; b: Inappropriate Investigations, and c: Conditional scoring based on Ddx.

INTERVENTIONS TO PREVENT DIAGNOSTIC ERRORS

III



CHAPTER 6

Impact of diagnostic checklists on the interpretation of normal and abnormal electrocardiograms

Staal, J., Zegers, R., Caljouw-Vos, J., Mamede, S., Zwaan, L.

Published in *Diagnosis* 2022.

DOI: 10.1515/dx-2022-0092.

Abstract

Background: Checklists that aim to support clinicians' diagnostic reasoning processes are often recommended to prevent diagnostic errors. Evidence on checklist effectiveness is mixed and seems to depend on checklist type, case difficulty, and participants' expertise. Existing studies primarily use abnormal cases, leaving it unclear how the diagnosis of normal cases is affected by checklist use. We investigated how content-specific and debiasing checklists impacted performance for normal and abnormal cases in electrocardiogram (ECG) diagnosis.

Methods: In this randomized experiment, 42 first year general practice residents interpreted normal, simple abnormal, and complex abnormal ECGs without a checklist. One week later, they were randomly assigned to diagnose the ECGs again with either a debiasing or content-specific checklist. We measured residents' diagnostic accuracy, confidence, patient management, and time taken to diagnose. Additionally, confidence-accuracy calibration was assessed.

Results: Accuracy, confidence, and patient management were not significantly affected by checklist use. Time to diagnose decreased with a checklist (M = 147s (77)) compared to without a checklist (M= 189s (80), *Z* = -3.10, p = 0.002). Additionally, residents' calibration improved when using a checklist (phase 1: R^2 = 0.14, phase 2: R^2 = 0.40).

Conclusions: In both normal and abnormal cases, checklist use improved confidence-accuracy calibration, though accuracy and confidence were not significantly affected. Time to diagnose was reduced. Future research should evaluate this effect in more experienced GPs. Checklists appear promising for reducing overconfidence without negatively impacting normal or simple ECGs. Reducing overconfidence has the potential to improve diagnostic performance in the long term.

Introduction

In recent years, checklists have received increasing attention as a promising tool to reduce medical errors.(1-3) This started with the successful implementation of checklists in reducing hospital-acquired infections (4) and preventing errors during surgeries.(5) These checklists aimed to reduce clinician's cognitive load and reliance on memory (6) by documenting the steps of a specific task (e.g., a surgical procedure). Following these successes, the use of checklists has also been advocated as a tool to reduce diagnostic errors (7-11), a long understudied type of medical errors (12) that occur when diagnoses are wrong, missed, or delayed.(13) Diagnostic errors are a large burden on patient safety and it is estimated that a majority of people will experience a diagnostic error during their lifetime.(13, 14) Therefore, developing successful interventions to reduce diagnostic errors is crucial.(15)

Flaws in the cognitive processes underlying reasoning are seen as a primary cause of diagnostic errors (16-21) and consequently, diagnostic checklists aim to reduce errors by supporting clinicians' reasoning processes. These checklists can generally be divided into two types.(22, 23) The first type aim to have clinicians examine and improve their reasoning processes. These process checklists generally give broad instructions to carefully reconsider your diagnosis or to check your reasoning for cognitive biases (i.e. predispositions to think in a way that leads to systematic failures in judgement (24)).(22, 25, 26) The second type includes content-specific checklists, which aim to compensate for possible knowledge deficits or mistakes (21, 27) by having clinicians examine the content of their reasoning. Content checklists can give possible diagnoses for certain symptoms (23, 26, 28) or ensure the clinician considers all relevant information for a diagnosis, as even those who were trained to follow the steps of a specific protocol will not always adhere to this protocol.(29-37) Furthermore, content checklists might have the potential to reduce clinicians' cognitive load by facilitating information integration.(7, 38)

Empirical evidence that checklists reduce diagnostic errors is scarce and inconsistent.(10, 22, 23) Reviews on error interventions generally report small to medium improvements in diagnostic accuracy (6, 22, 39), but the practical significance of this improvement is unclear. Overall, existing studies hint that checklist effectiveness might depend on the type of checklist, the relative difficulty of the clinical cases that have to be diagnosed, and the participants' level of expertise. For example, process checklists (22) were shown to be ineffective in increasing diagnostic accuracy (28, 33), with exception of one study by Sibbald et al. (31) that showed an improvement. Content checklists often led to small reductions in diagnostic errors (29-31, 40) – except in one study where no benefit was seen.(33) Furthermore, checklists were more effective in improving diagnostic accuracy for novices than for experts in two studies examining ECG interpretation and dermatological images, respectively.(30, 36) Finally, in some studies checklists only benefited the diagnosis of complex clinical cases.(28, 31, 40) Unfortunately, the factors impacting checklist effectiveness are still poorly understood and more research is necessary to determine if, and when, checklists are effective.(10, 22, 23).

Our understanding of checklist effectiveness is especially limited for settings such as general practice. For general practitioners (GPs), it is more important to recognize normal cases and to exclude certain diagnoses than it is to arrive at the precise correct diagnosis. Existing studies, however, mostly test checklists on abnormal cases that were designed to be complex. This approach is intended to create a situation where the potential for making and subsequently correcting mistakes is high, so that benefits from an intervention can be observed.(22) Furthermore, GPs are also expected to correctly manage patients, even before knowing the exact diagnosis. Existing studies primarily measure diagnostic accuracy, leaving out other such aspects of diagnostic performance. A task for which these issues are relevant is the interpretation of ECGs. At least one-third of ECGs seen in Dutch general practice are normal (41) and the most important decision GPs make is on whether or not to refer the patient to a specialist. In the Netherlands, ECG interpretation has recently shifted more and more from secondary to primary care, even though most GPs are not specialized in this task and have had limited training.(41) GP education now often implements checklists to teach this skill.(34) It is therefore crucial to understand how checklist use will impact ECG interpretation, as checklists could lead to overtesting and overdiagnosis, or unnecessary consumption of resources such as time and personnel.(22)

In this randomized experiment, we examined the impact of checklist use on the performance of GP residents when interpreting normal, simple abnormal, and complex abnormal ECGs. Performance was measured as residents' diagnostic accuracy, confidence, patient management, and time to diagnose. Additionally, residents' confidence-accuracy calibration was assessed. We studied two types of checklists – a debiasing checklist focused on detecting and correcting cognitive biases (28, 33, 38) and a content checklist focused on ensuring all ECG elements important for interpretation are checked.(34) We expected that neither checklist would benefit performance for normal cases. For simple and complex abnormal cases, we expected that only the content checklist would be beneficial. Furthermore, we expected that residents' confidence-accuracy calibration would increase for the content checklist, but decrease for the debiasing checklist.

Methods

Design

The study was a computer-based experiment with a mixed design. All methods were carried out in accordance with the relevant guidelines and regulations. In the first phase, residents interpreted ECGs in a randomized order, without a checklist. In the second phase one week later, residents were randomly allocated to using either a debiasing or a content-specific checklist to interpret the ECGs from phase 1 in a randomized order. Participants were not informed the same ECGs were shown. We chose to present the same ECGs twice to ensure a direct comparison between the two phases was possible.

Participants

First year GP residents in training at the Erasmus Medical Center Rotterdam were recruited. The study was scheduled between educational sessions. Sample size was estimated a priori in G*power for a repeated measures ANOVA (multiple analysis of variance) with between-subject factors, for a medium effect size (0.5), a power of 0.8, and an α of 0.05.(42) The estimated total sample size was 30 participants.

Materials

Checklists

The used checklist materials were taken from recent studies which showed improvements in diagnostic accuracy when using the checklists (Table 1). The debiasing checklist and instructions for use were obtained from Sibbald et al. (33) and were translated to Dutch by a native speaker (JS). The content-specific condition was the ECG10+ as it is used in Dutch GP education.(34)

Table 1. *Overview of checklist materials*.

Debiasing checklist (38)
Please check your ECG diagnosis carefully considering each of the following: Was I comprehensive? Did I consider the inherent flaws of heuristic thinking? Was my judgment affected by any cognitive bias? Were any of the following biases present (anchoring, availability, confirmation, search satisficing, framing)? What is the worst-case scenario?
Content-specific checklist (34)
Please check your ECG diagnosis carefully considering each of the following: Frequency and rate Axis P-wave PQ-interval Q-wave QRS-complex ST-interval T-wave QT-interval Rhythm After the 10 points in this checklist, a '+' is added, where participants are asked to combine all their previous findings into one interpretation of the ECG.

ECGs

Two experienced GPs with cardiology specializations selected nine anonymized ECGs from real patients with a confirmed diagnosis from an educational database targeted at GP residents. One GP (JCV) independently selected the ECGs and the second GP (RZ) judged them. Disagreements were solved via discussion. Three normal ECGs (with a sinus rhythm and no abnormalities), three simple abnormal ECGs (indicating atrial fibrillation, an easily recognizable condition with a high incidence), and three complex abnormal ECGs (indicating ischemia, a difficult to recognize condition) were selected. The ratio of normal (one third) to abnormal (two thirds) ECGs was based on a study that examined the incidence of ECG presentations in general practice in the Netherlands.(41) The ECGs were selected from a database with educational materials for GP residents. The selected cases were labeled appropriate for use in the education of first year residents in the database and were therefore deemed of appropriate difficulty for our participants. An overview of all ECGs is shown in Table 2 and the ECGs are shown in Supplemental Material 1.

Table 2. *Overview of patient information of the selected ECGs.*

ECG type	Diagnosis	Patient information	Reason for ordering ECG
Practice	Left ventricular hypertrophy	70 year old woman	Shortness of breath, chest pain
Normal	Sinus rhythm	37 year old woman	Ordered for regular check-up
Normal	Sinus rhythm	67 year old man	Dizziness, heart palpitations
Normal	Sinus rhythm	81 year old woman	Chest pain
Abnormal	Atrial fibrillation	89 year old woman	Slower heart rate than usual combined with being tired and out of breath when exercising
Abnormal	Atrial fibrillation	76 year old woman	Swollen legs, out of breath when exercising
Abnormal	Atrial fibrillation	81 year old woman	Tires quickly, dizziness
Abnormal	Ischemia	59 year old woman	Pain in the abdomen, a feeling of pressure on the elbows
Abnormal	Ischemia	68 year old woman	Cardiologist detected atypical chest pain before, patient asked for follow-up
Abnormal	Ischemia	85 year old man	Swollen legs, tires quickly

Procedure

The study was prepared in Qualtrics (an online survey tool) and residents filled out the survey at home. They had to complete both phases during the allocated time slots in their schedule. Before starting a phase, residents received an information letter and were asked to sign informed consent. Residents were informed of the study's purpose and were aware that there were two checklist conditions, although they were not informed they would see the same ECGs twice.

In phase 1, residents were asked to provide demographic information and then to interpret 9 ECGs without specific instructions. We asked them to indicate the most likely diagnosis (or indicate "normal" if there were no abnormalities). Each ECG was accompanied by the sex and age of the patient, the patients' chief complaint, and the patients' physical examination and test results. Residents had 60 minutes to complete this task. Residents were also asked for their confidence in the interpretation and if they would refer the patient based on the ECG.

A week later in phase 2, residents were randomly allocated to a checklist condition and received instructions on how to use their respective checklist (as in Sibbald et al. (33), Table 1). They had the opportunity to practice using the checklist on one ECG. Next, they had 60 minutes to interpret all 9 ECGs using either the debiasing checklist or the content-specific checklist. They were again asked for their interpretation, their confidence, and their patient management decisions. After phase 2, an experienced GP (JCV) led a 30-minute feedback session to discuss the study's ECGs and answer any questions.

Outcome measures

The between subject independent variable was checklist type: debiasing or content-specific checklist. The within subjects independent variables were ECG type (normal, simple abnormal, and complex abnormal) and phase (phase 1: interpretation without instructions and phase 2: interpretation with checklist). We further measured four dependent variables, which together characterized residents' performance: diagnostic accuracy, confidence in diagnosis, patient management, and time to diagnose.

Diagnostic accuracy was independently scored by two experienced GPs. One GP (JCV) assessed all diagnoses and the second GP (RZ) scored half of the diagnoses. Their judgements showed substantial interrater reliability (κ = 0.72, 95% CI: 0.62-0.82). Discrepancies in scoring were resolved through discussion. Diagnostic accuracy was scored as 0 if the incorrect diagnosis was given; as 0.5 if a partially correct diagnosis was given (e.g., the participant answered AF with aberration in case of an AF diagnosis), and as 1 if the correct diagnosis was given. Second, participants were asked to rate their confidence in their interpretation on a scale from 1 to 10. For each participant, overall accuracy and the confidence corresponding to that accuracy were combined to measure "calibration". Third, participants were asked where they would refer the patient based on the ECG in a multiple-choice format to measure patient management. Based on consultation with an experienced GP (RZ) and existing guidelines, patient management was rated as follows: for normal ECGs, the patient should be reassured; for atrial fibrillation ECGs, residents were expected to start their own treatment, and for ischemia ECGs, residents were expected to refer the patient to the cardiologist.(43, 44) The management decision was scored as 0 if incorrect and as 1 if correct. Fourth, Qualtrics recorded time to diagnose in seconds for each ECG. Finally, participants were asked for their age, sex, months as a resident, and level of expertise, which were measured as covariates (Table 3).

Table 3. *Participant demographics.*

	Content checklist (N = 21)	**Debiasing checklist (N = 21)**	**Total (N = 42)**
Demographics			
Sex (n female) (%)	17 (81%)	10 (45%)	28 (67%)
Age (years)			
Mean (SD)	29 (3)	31 (3)	30 (3)
Range	25 – 36	27 – 39	25 – 39
Time in residency (months)			
Mean (SD)	9 (2)	9 (1)	9 (1)
Range	7 – 15	7 – 9	7 – 15

Statistical analysis

For each dependent variable, the average was calculated for the normal, simple abnormal, and complex abnormal ECGs. Mean scores were calculated for residents who interpreted all 9 ECGs. A Shapiro-Wilk test showed that these data were not normally distributed and therefore, non-parametric tests were performed for all comparisons in IBM SPSS Statistics for Windows (Version 25.0). All tests were considered significant at the α = .05 level. A Wilcoxon test examined differences for each dependent variable between phase 1 and phase 2. Additionally, a Mann Whitney-U test compared each dependent variable between both checklists and a Friedman test compared performance for each ECG type. Finally, the calibration between residents' confidence and accuracy was averaged per participant over all cases and examined in a scatterplot to investigate whether there was a linear association between these variables. Calibration was then quantified using Spearman's rho and expressed as a goodness-of-fit measure (R^2). It was further explored by calculating absolute accuracy (the absolute difference between accuracy and confidence, where 0 is perfect and 1 is inaccurate) and bias (the signed difference between accuracy and confidence, where -1 is underconfident and +1 is overconfident), which were then compared using a Wilcoxon signed rank test. Absolute accuracy and bias were calculated as in Kuhn et al. (45).

Results

In total, 55 first year GP residents participated in at least one phase. Five residents did not give permission to use their data for research and an additional eight only completed one phase or did not interpret all ECGs. 42 residents completed both phases. 21 residents were allocated to the debiasing checklist and 21 to the content-specific checklist. Participant demographics are shown in Table 3 and Supplemental Material 2.

Residents' prior experience (specifically, the number of ECGs diagnosed) was used to test whether experience moderated the dependent variables (Table 4). Only confidence systematically varied with experience and post-hoc tests indicated that only residents who diagnosed fewer than 10 ECGs differed from the other experience groups. Therefore, these three residents were excluded to correct for experience as a covariate, leaving 18 participants in the content-specific condition. Age, sex, and time in residency did not moderate diagnostic performance.

Table 4. *Mean and standard deviation for accuracy, confidence in diagnosis, patient management, and time spent to diagnose in phase 1 and phase 2 per ECG type.*

	Phase 1 (n=42)	**Phase 2: Content (n=18)**	**Phase 2: Debiasing (n=21)**	**Moderation by number of ECGs**[b]
	M (SD)	**M (SD)**	**M (SD)**	**χ^2, *p***
Accuracy[a]				Phase 1: $\chi^2(3) = 6.93$, p = 0.074 Phase 2: $\chi^2(3) = 7.18$, p = 0.067
Normal	0.64 (0.3)	0.73 (0.3)	0.68 (0.3)	
Simple abnormal	0.61 (0.3)	0.69 (0.3)	0.65 (0.3)	
Complex abnormal	0.37 (0.2)	0.42 (0.3)	0.40 (0.3)	
Total	0.55 (0.2)	0.63 (0.2)	0.61 (0.2)	
Confidence[a]				Phase 1: $\chi^2(3) = 8.18$, p = 0.042 Phase 2: $\chi^2(3) = 10.18$, p = 0.017
Normal	5.2 (2.0)	5.1 (2.5)	5.8 (1.8)	
Simple abnormal	5.6 (2.1)	5.1 (2.3)	5.9 (1.7)	
Complex abnormal	5.2 (2.0)	4.5 (2.2)	5.4 (1.8)	
Total	5.5 (1.7)	5.1 (2.1)	5.8 (1.6)	
Management[a]				Phase 1: $\chi^2(3) = 2.36$, p = 0.501 Phase 2: $\chi^2(3) = 2.84$, p = 0.416
Normal	0.56 (0.3)	0.61 (0.3)	0.59 (0.2)	
Simple abnormal	0.40 (0.3)	0.52 (0.3)	0.38 (0.3)	
Complex abnormal	0.85 (0.2)	0.83 (0.2)	0.81 (0.3)	
Total	0.61 (0.2)	0.66 (0.2)	0.60 (0.2)	
Time[a] (in seconds)				Phase 1: $\chi^2(3) = 2.08$, p = 0.556 Phase 2: $\chi^2(3) = 0.413$, p = 0.938
Normal	185 (87)	149 (82)	116 (70)	
Simple abnormal	191 (109)	142 (93)	181 (110)	
Complex abnormal	200 (100)	172 (89)	157 (140)	
Total	189 (80)	143 (62)	144 (90)	

[a] Averages were computed without participants who diagnosed fewer than 10 ECGs during their training.

[b] Kruskal-Wallis tests tested whether the outcome measures were moderated by experience (based on the number of ECGs residents diagnosed during their studies).

Diagnostic performance

Phase 1 versus phase 2

When interpreting ECGs in phase 2 (M = 0.63, SD = 0.2) compared to phase 1 (M = 0.55, SD = 0.2), there was a trend for overall accuracy to improve (*Z* = -1.81, p = 0.070, *g* = 0.25). Checklist use did not affect residents' confidence (phase 1: M = 5.5, SD = 1.7), phase 2: M = 5.5, SD = 1.9, *Z* = -0.23, p = 0.817) and patient management (phase 1: M = 0.61, SD = 0.2, phase 2: M = 0.63, SD = 0.2, *Z* = -0.92, p = 0.358) in phase 1 compared to phase 2. Lastly, residents took less time to interpret all ECGs in phase 2 (phase 1: M = 189, SD = 80, phase 2: M = 144, SD = 76, *Z* = -3.10, p = 0.002, *g* = 0.54). Resident's performance on each outcome measure is summarized in Table 4.

Checklist type

Using either the debiasing or content-specific checklist did not differentially affect accuracy (*U* = 158, p = 0.707), confidence (*U* = 134, p = 0.270), patient management (*U* = 137, p = 0.311), or time spent to diagnose (*U* = 162, p = 0.821).

ECG type

ECG type did not affect checklist use for accuracy ($\chi^2(2) = 2.54$, p = 0.281), confidence ($\chi^2(2) = 2.74$, p = 0.254), patient management ($\chi^2(2) = 2.10$, p = 0.350) and time to diagnose ($\chi^2(2) = 1.16$, p = 0.559). For patient management, residents descriptively scored the best for complex abnormal cases, where the patient should be referred to the cardiologist. For normal and simple abnormal cases, more than 90% of the incorrect answers constituted referral to the cardiologist.

Confidence-accuracy calibration

In both phases, confidence increased when accuracy increased (phase 1: $r_s = 0.42$, p = 0.004; phase 2: $r_s = 0.67$, $p < 0.001$). Moreover, residents' confidence-accuracy calibration was lower when they interpreted ECGs without specific instructions ($R^2 = 0.14$, Figure 1) compared to when they used a checklist ($R^2 = 0.40$, Figure 2), although calibration remained moderate. Further analysis showed that their absolute accuracy did not differ between phases (*Z* = -0.59, p = 0.554). Bias showed a trend to decrease, indicating that residents became less overconfident when using a checklist (*Z* = -3.10, p = 0.055). Residents improved using either checklist compared to interpretation without a checklist, but seemed to benefit more from using the debiasing checklist ($R^2 = 0.59$) than from the content-specific checklist ($R^2 = 0.32$).

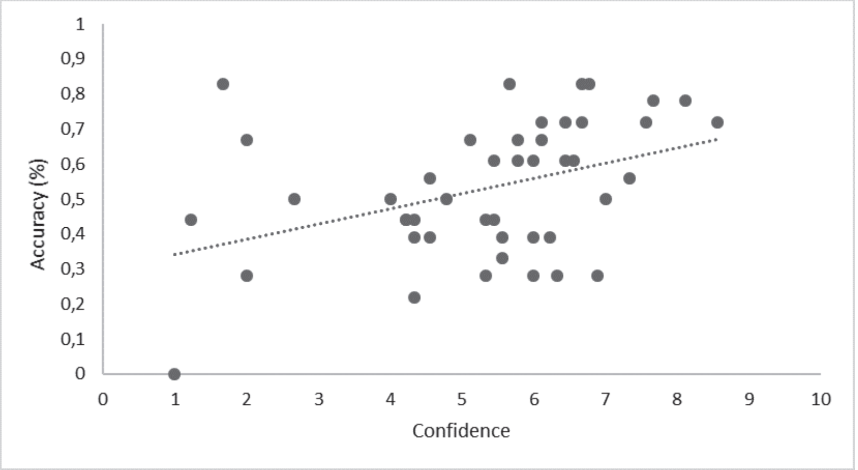


Figure 1. Scatterplot of residents' confidence-accuracy calibration in phase 1, $R^2 = 0.14$.

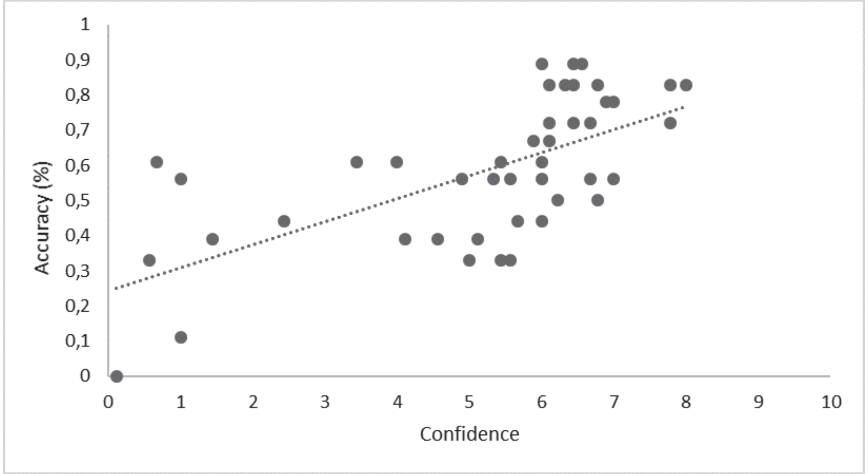


Figure 2. Scatterplot of residents' confidence-accuracy calibration in phase 2, $R^2 = 0.40$.

Discussion

This study examined the impact of checklist use on the interpretation of normal, simple abnormal, and complex abnormal ECGs. There was a trend for improvement in residents' accuracy when they used a checklist, whereas their confidence did not change. This resulted in an overall improved confidence-accuracy calibration: participants were less overconfident after using a checklist compared to when they first interpreted the ECGs. Furthermore, residents' patient management was very conservative as they consistently referred patients to the cardiologist. This was not affected by checklist use. Finally, residents took less time to interpret ECGs in phase 2. Contrary to our expectations, these findings were similar for all ECG types and both the debiasing checklist and the content-specific checklist, although the debiasing checklist seemed to improve to residents' calibration the most.

For our interpretations regarding diagnostic accuracy, we should consider the possibility of a learning effect on diagnostic performance. There was no independent control group and therefore, residents saw the ECGs twice. Furthermore, the effects were similar across all ECGs and both checklist types, which indicates that the trend for improvement in accuracy and the decrease in time to diagnose are likely due to a small learning effect and do not fully reflect the effects of checklist use. These findings contradict previous studies that found increases in accuracy when a checklist was used (23, 28-32), specifically for the content-specific checklist. In most of these studies, participants also examined cases once before verifying their diagnosis using a checklist, although this verification took place immediately after the initial diagnosis.(28-30, 32) Similar studies often did not find an improvement when using a debiasing checklist, in line with our current findings.(23, 28, 33) Despite this limitation, our study design is reflective of how checklists would be used in practice: to verify a working diagnosis or to check someone's reasoning process. Alternatively, the current lack of improvement in diagnostic accuracy could be explained by the use of singular reasoning approaches. Our participants were asked only to reason analytically, following either a feature list (given by the content-specific checklist) or a debiasing approach. This contrasts work by Eva et al. (46) and Ark et al. (47, 48), who showed in several experiments with naïve students that combining analytical and non-analytical approaches is more effective than applying singular reasoning approaches. Future studies might benefit from not only comparing singular methods but also combining reasoning strategies in error intervention studies.

Interestingly, the trend for improved accuracy did not coincide with an increase in confidence. One would expect that if a previously incorrect diagnosis was changed or if a previously correct diagnosis was confirmed with the help of a checklist, this would boost confidence, especially if residents were simply re-examining a case. Our data showed that in each phase, if residents were more accurate, they were also more confident. However, this did not translate to an increase in confidence between phases, indicating that residents became less overconfident and potentially that their insight in their own skills improved. Despite the increased confidence, the majority of residents still chose to refer the patient to a cardiologist. This is likely related to the relative inexperience of our participants. Future research should also measure participants' referral behavior, as overreferral leads to large economic costs.

The increase in residents' confidence might have been extra pronounced for the debiasing checklist because GP residents in the Netherlands are already taught to interpret ECGs using the content-specific checklist, whereas the debiasing checklist was completely new to them. The fact that the GP residents were already familiar with the content-specific checklist, and because novices often use a more analytical step-by-step approach than more experienced clinicians (49), might also have diluted potential effects of the content-specific checklist. Although our study showed no immediate benefits of improved calibration, there could be value in using checklists to reduce overconfidence. Overconfidence has previously been indicated as a cause of diagnostic errors (50) and fostering proper calibration could improve residents' diagnostic process and potentially improve their diagnostic performance in the long term. Future research should confirm whether checklists can be used to reduce overconfidence and what the long-term effects of checklist use are.

This study had several strengths and limitations. Strengths include that this was a randomized experiment that used ECGs of an appropriate level for first year GP residents. Furthermore, the ECGs were verified teaching materials from real patients with a confirmed diagnosis. A final strength was that participants performed the experiment online, from their home, and participated in multiple experiments and lectures. This greatly reduced the chances of participants discussing the study's ECGs amongst themselves.

The study is limited because of the design without an independent control group in which participants interpreted each ECG twice, which left the possibility for a learning effect to influence our results. This primarily influenced the interpretation of diagnostic accuracy and time on task, but even with a possible learning effect participants did not improve on immediate accuracy. Furthermore, we chose to have participants diagnose the same ECGs twice so we could directly compare changes in confidence, calibration, and patient management. This allowed for reliable assessment of residents' confidence, as there was no room for between-case variability. The remaining variables could be inflated by a possible learning effect and should be interpreted with caution. A second limitation is the relative inexperience of our participants. Considering that most residents had interpreted few ECGs during their studies, suddenly seeing 9 ECGs in one day was a significant increase in practice. This might have contributed to the trend for improved accuracy. Lastly, a limitation is that the overall sample size and the sample size for the separate checklist analyses were relatively small and might be underpowered, meaning these results should be interpreted with caution. The study might, additionally, have benefited from including more than 9 EKGs. The a priori power calculation was performed assuming 9 measurements but the true effect might have been smaller than the medium effect size we estimated. Future research should examine the impact of checklist use on accuracy and calibration in more experienced

GP residents, as the issue remains crucial to GPs, with a control group and a larger sample of EKGs.

In summary, checklist use did not differentially affect GP residents' diagnostic process for normal cases compared to simple abnormal or complex abnormal cases. Surprising was that residents' confidence did not increase over repeated viewing of the ECGs and that checklists improved residents' confidence-accuracy calibration, which translated in reduced overconfidence. Although more research is needed to evaluate how checklists impact residents' confidence in the long term, checklists could be promising. Reducing overconfidence, an important cause of diagnostic errors, could improve residents' insight into their own skill level, and in the long term has the potential to improve their diagnostic performance.

Acknowledgements

We would like to thank the residents who participated in the study. We would also like to thank Dr. Pieter van den Berg for his feedback on the study design.

References

1. Zia SMR, Zahid R, Ashraf H. The WHO Surgical Safety Checklist: A Systematic Literature Review. Archives of Surgical Research. 2021.
2. Thomassen Ø, Storesund A, Søfteland E, Brattebø G. The effects of safety checklists in medicine: a systematic review. Acta Anaesthesiologica Scandinavica. 2014;58(1):5-18. doi: http://dx.doi.org/10.1111/aas.12207.
3. Woodward HI, Mytton OT, Lemer C, Yardley IE, Ellis BM, Rutter PD, et al. What have we learned about interventions to reduce medical errors? Annual review of public health. 2010;31:479-97. doi: http://dx.doi.org/10.1146/annurev.publhealth.012809.103544.
4. Pronovost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. New England journal of medicine. 2006;355(26):2725-32.
5. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat A-HS, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. New England journal of medicine. 2009;360(5):491-9.
6. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: Still no easy answers. Western Journal of Emergency Medicine. 2020;21(6):125.
7. Gawande A. The checklist manifesto: How to get things right. Journal of Nursing Regulation. 2011;1(4):64. doi: http://dx.doi.org/10.1016/S2155-8256(15)30310-0.
8. Gupta A, Graber ML. Annals for hospitalists inpatient notes-just what the doctor ordered—checklists to improve diagnosis. Annals of Internal Medicine. 2019;170(8):HO2-HO3.
9. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ quality & safety. 2016;25(10):808-20.
10. Graber ML, Kissam S, Payne VL, Meyer AND, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ quality & safety. 2012;21(7):535-57.
11. Clinician Checklists [Internet]: Society to Improve Diagnosis in Medicine; 2020 [updated 2020 May 20]; cited 2021 Jul 1]. Available from: https://www.improvediagnosis.org/clinician-checklists/.
12. Wachter RM. Why diagnostic errors don't get any respect—and what can be done about them. Health Affairs. 2010;29(9):1605-10.
13. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
14. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
15. Zwaan L, El-Kareh R, Meyer AND, Hooftman J, Singh H. Advancing Diagnostic Safety Research: Results of a Systematic Research Priority Setting Exercise. Journal of General Internal Medicine. 2021:1-9. doi: http://dx.doi.org/10.1007/s11606-020-06428-3.
16. Phua DH, Tan NC. Cognitive aspect of diagnostic errors. Ann Acad Med Singapore. 2013;42(1):33-41.
17. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Academic medicine. 2003;78(8):775-80.
18. Croskerry P. From mindless to mindful practice—cognitive bias and clinical decision making. N Engl J Med. 2013;368(26):2445-8.

19. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education. 2009;14(1):27-35.
20. Elia F, Apra F, Verhovez A, Crupi V. "First, know thyself": cognition and error in medicine. Acta Diabetologica. 2016;53(2):169-75.
21. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. Journal of Evaluation in Clinical Practice. 2018;24(3):666-73.
22. Zwaan L, Staal J. Evidence on Use of Clinical Reasoning Checklists for Diagnostic Error Reduction. AHRQ Papers on Diagnostic Safety Topics [Internet]. 2020; (3).
23. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: a randomized experiment. Medical Education. 2021.
24. Kahneman D, Egan P. Thinking, fast and slow: Farrar, Straus and Giroux New York; 2011.
25. Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Annals of emergency medicine. 2003;41(1):110-20.
26. Ely JW, Graber MA. Checklists to prevent diagnostic errors: a pilot randomized controlled trial. Diagnosis. 2015;2(3):163-9. doi: http://dx.doi.org/10.1515/dx-2015-0008.
27. Norman, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine. 2017;92(1):23-30. doi: http://dx.doi.org/10.1097/ACM.0000000000001421.
28. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. Medical teacher. 2013;35(6):e1218-e29.
29. Sibbald M, de Bruin ABH, Cavalcanti RB, van Merrienboer JJG. Do you have to re-examine to reconsider your diagnosis? Checklists and cardiac exam. BMJ quality & safety. 2013;22(4):333-8.
30. Sibbald M, De Bruin ABH, van Merrienboer JJG. Finding and fixing mistakes: do checklists work for clinicians with different levels of experience? Advances in Health Sciences Education. 2014;19(1):43-51.
31. Sibbald M, de Bruin ABH, van Merrienboer JJG. Checklists improve experts' diagnostic decisions. Medical education. 2013;47(3):301-8.
32. Sibbald M, de Bruin ABH, Yu E, van Merrienboer JJG. Why verifying diagnostic decisions with a checklist can help: insights from eye tracking. Advances in Health Sciences Education. 2015;20(4):1053-60. doi: http://dx.doi.org/10.1007/s10459-015-9585-1.
33. Sibbald M, Sherbino J, Ilgen JS, Zwaan L, Blissett S, Monteiro S, et al. Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. Advances in Health Sciences Education. 2019;24(3):427-40.
34. Konings K, Willemsen R. ECG 10+: Systematisch ECG's beoordelen. Huisarts en wetenschap. 2016;59(4):166-70.
35. Berbaum K, Franken Jr EA, Caldwell RT, Schartz KM. Can a checklist reduce SOS errors in chest radiography? Academic radiology. 2006;13(3):296-304. doi: http://dx.doi.org/10.1016/j.acra.2005.11.032.
36. Kok EM, Abed A, Robben SGF. Does the use of a checklist help medical students in the detection of abnormalities on a chest radiograph? Journal of digital imaging. 2017;30(6):726-31. doi: http://dx.doi.org/10.1007/s10278-017-9979-0.

37. Krage R, Tjon Soei Len L, Schober P, Kolenbrander M, van Groeningen D, Loer SA, et al. Does individual experience affect performance during cardiopulmonary resuscitation with additional external distractors? Anaesthesia. 2014;69(9):983-9.

38. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. Academic Medicine. 2011;86(3):307-13.

39. Abimanyi-Ochom J, Mudiyanselage SB, Catchpool M, Firipis M, Dona SWA, Watts JJ. Strategies to reduce diagnostic errors: a systematic review. BMC medical informatics and decision making. 2019;19(1):1-14. doi: http://dx.doi.org/10.1186/s12911-019-0901-1.

40. Nedorost S. A diagnostic checklist for generalized dermatitis. Clinical, Cosmetic and Investigational Dermatology. 2018;11:545.

41. Rutten FH, Kessels AGH, Willems FF, Hoes AW. Is elektrocardiografie in de huisartspraktijk nuttig? Huisarts en wetenschap. 2001;44(11):179-83.

42. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91.

43. Boode BSP, Frijling BD, Heeringa J, Rutten FH, Van den Berg PJ, Zwietering PJ, et al. NHG-standaard Atriumfibrilleren. NHG-Standaarden 2009: Springer; 2009. p. 67-86.

44. Rutten FH, Grundmeijer H, Grijseels EWM, Van Bentum STB, Hendrick JMA, Bouma M, et al. NHG-Standaard Acuut coronair syndroom. NHG-Standaarden 2009: Springer; 2009. p. 3-24.

45. Kuhn J, van den Berg P, Mamede S, Zwaan L, Bindels P, van Gog T. Improving medical residents' self-assessment of their diagnostic accuracy: does feedback help? Advances in Health Sciences Education. 2021:1-12.

46. Eva KW, Hatala RM, LeBlanc VR, Brooks LR. Teaching from the clinical reasoning literature: combined reasoning strategies help novice diagnosticians overcome misleading information. Medical education. 2007;41(12):1152-8.

47. Ark TK, Brooks LR, Eva KW. The benefits of flexibility: the pedagogical value of instructions to adopt multifaceted diagnostic reasoning strategies. Medical education. 2007;41(3):281-7.

48. Ark TK, Brooks LR, Eva KW. Giving learners the best of both worlds: do clinical teachers need to guard against teaching pattern recognition to novices? Academic Medicine. 2006;81(4):405-9.

49. Coderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. Medical education. 2003;37(8):695-703.

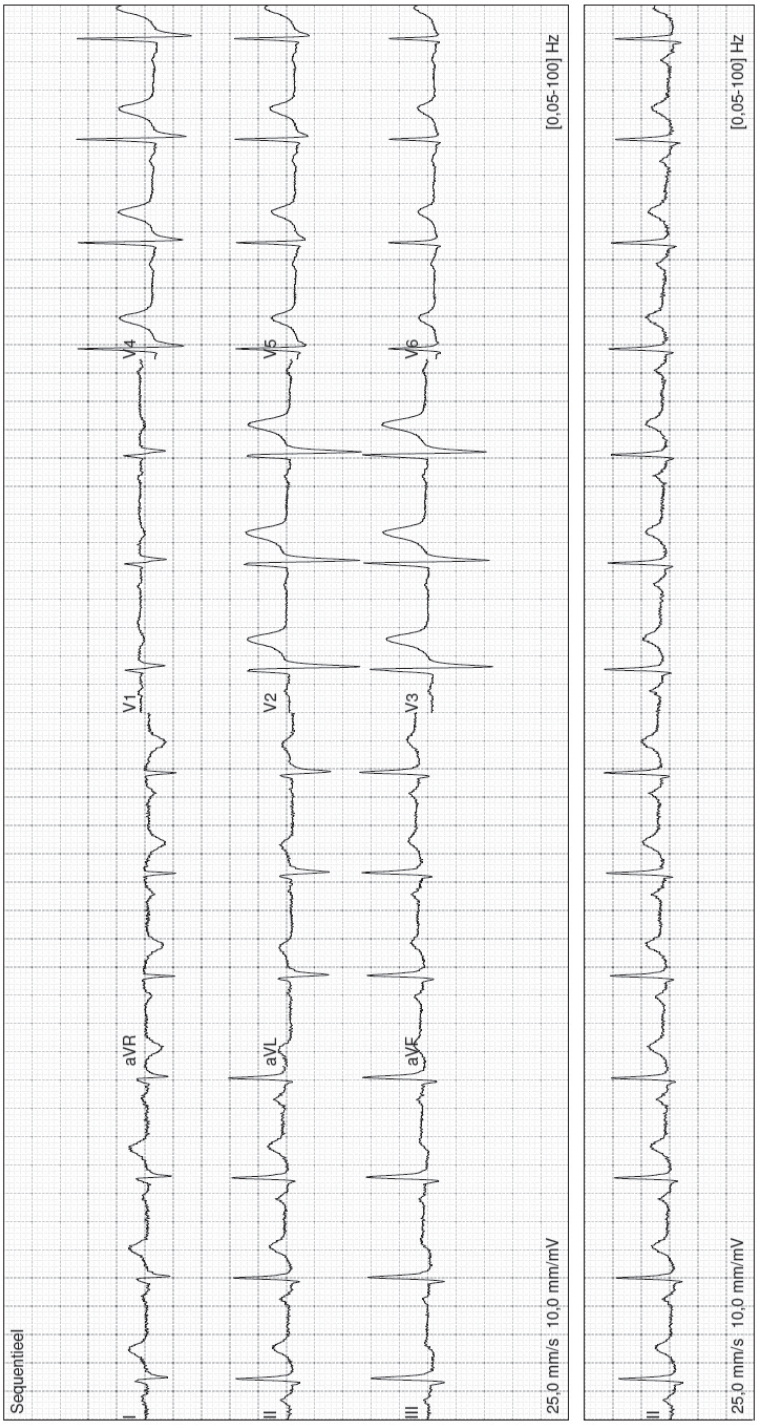
50. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. The American journal of medicine. 2008;121(5):S2-S23.

Supplemental Material 1 – Overview of included ECGs

Normal ECGs

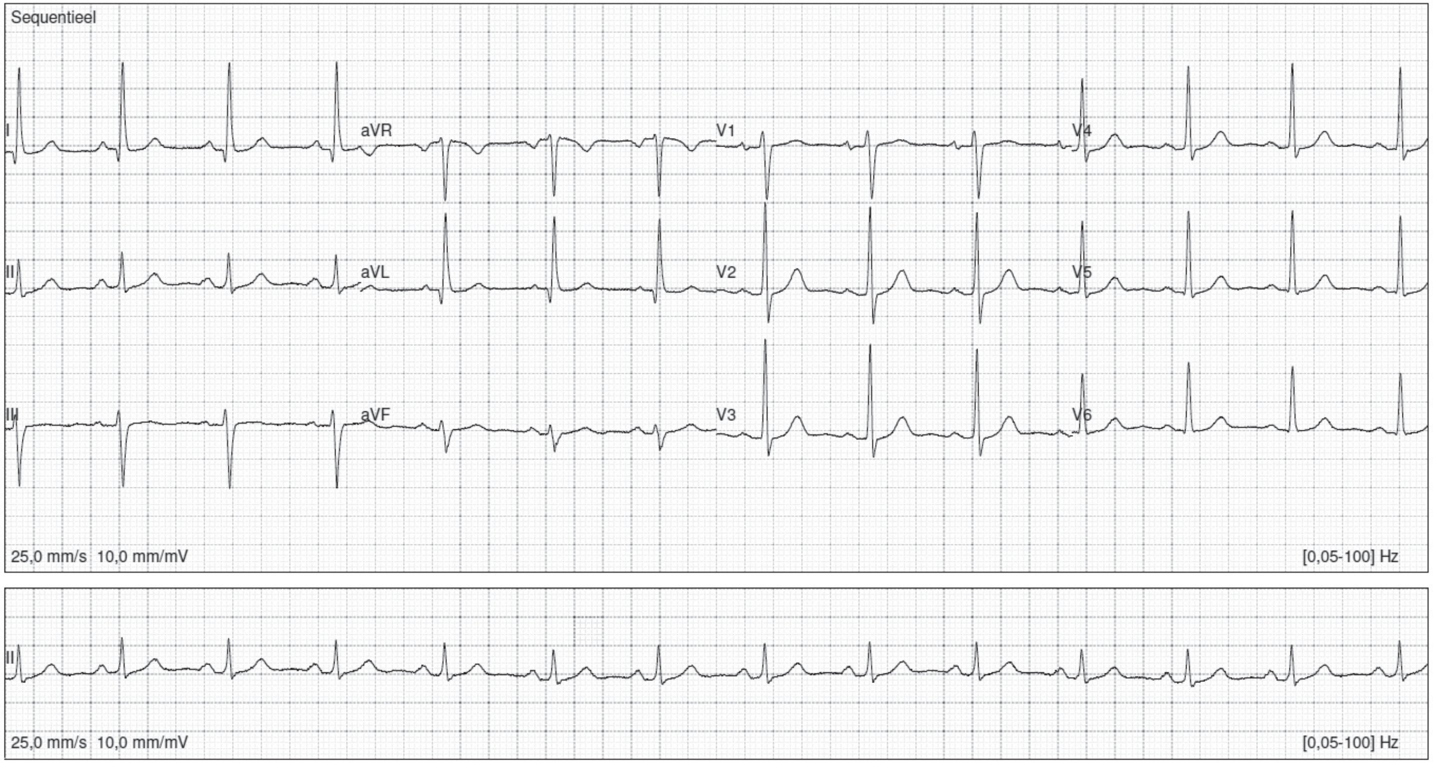
<u>Normal ECG 1</u>

37 year old woman, ECG ordered for regular check-up.



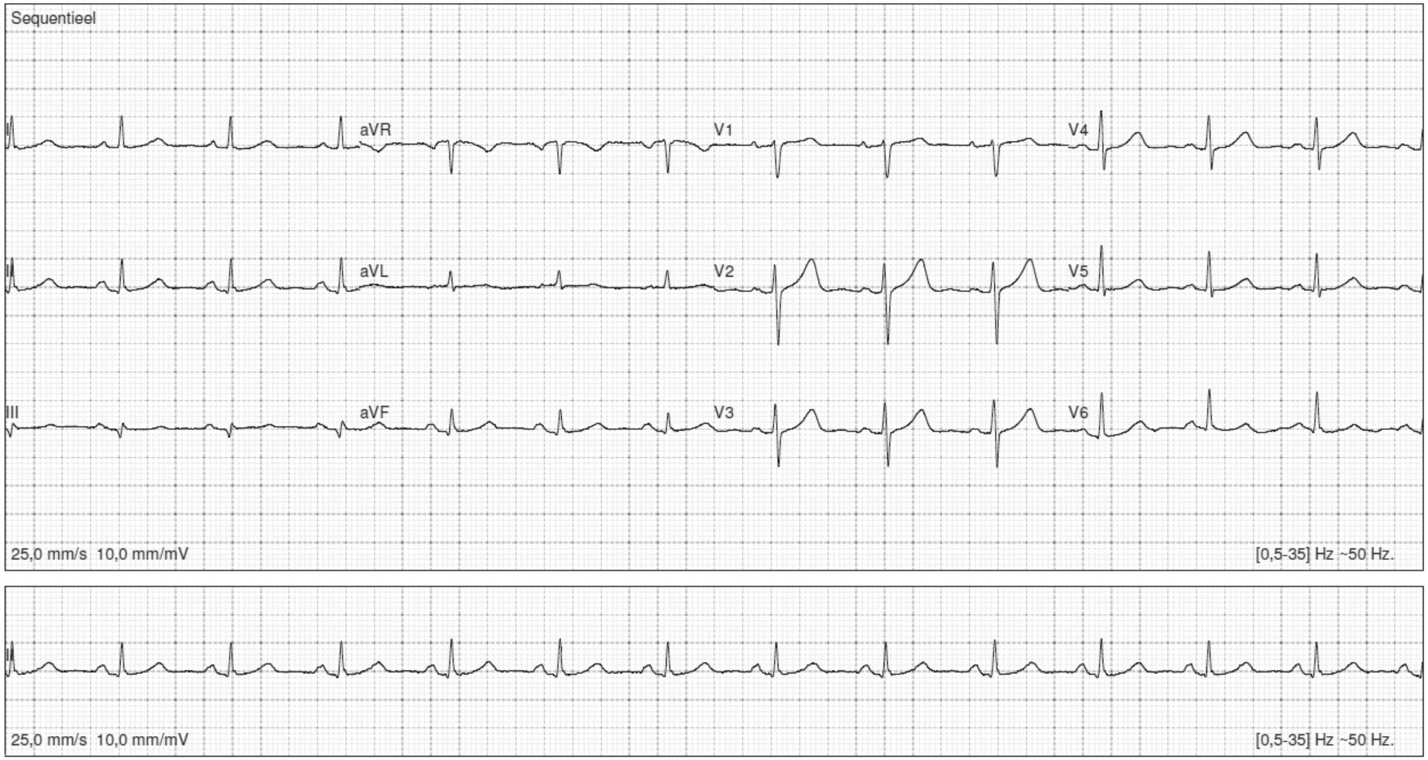
Normal ECG 2

67 year old man, suffering dizziness and heart palpitations.



Normal ECG 3

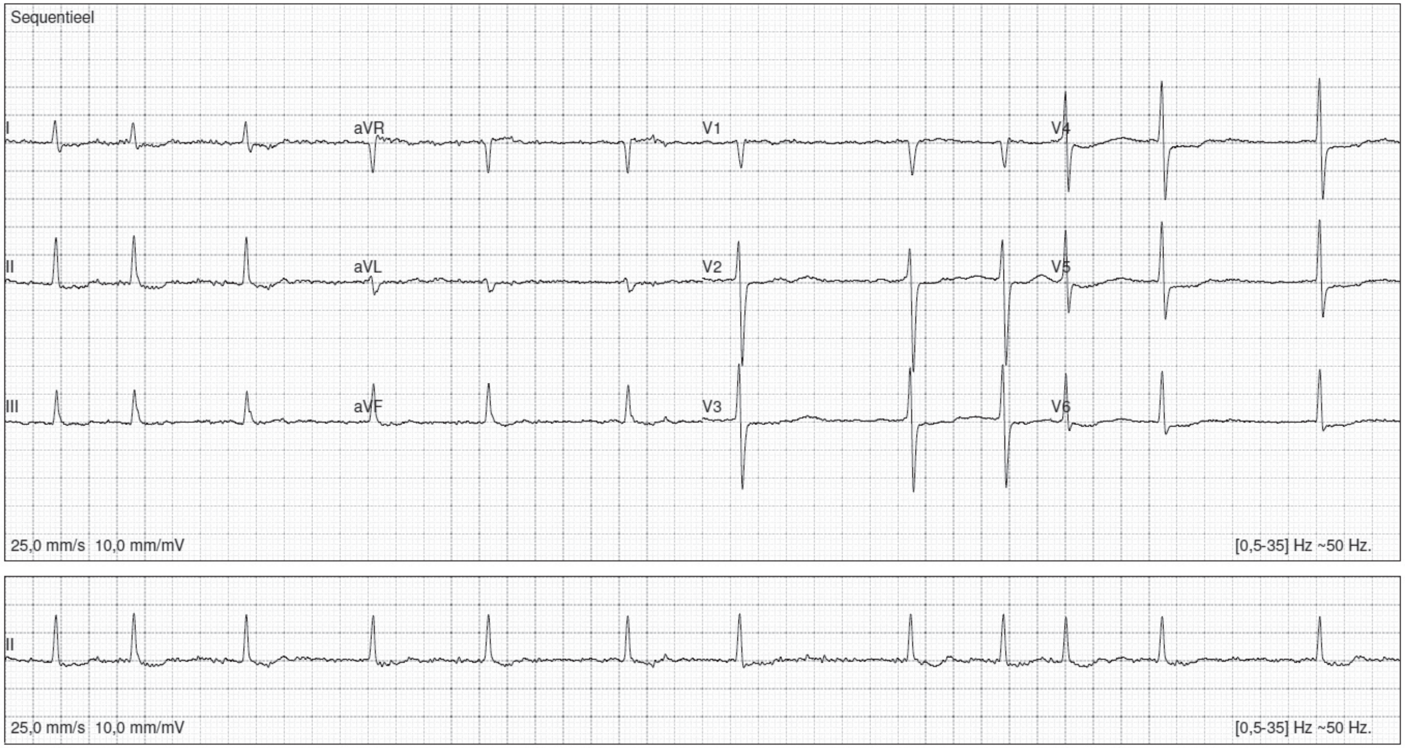
81 year old woman, sent in with chest pain.



Atrial fibrillation

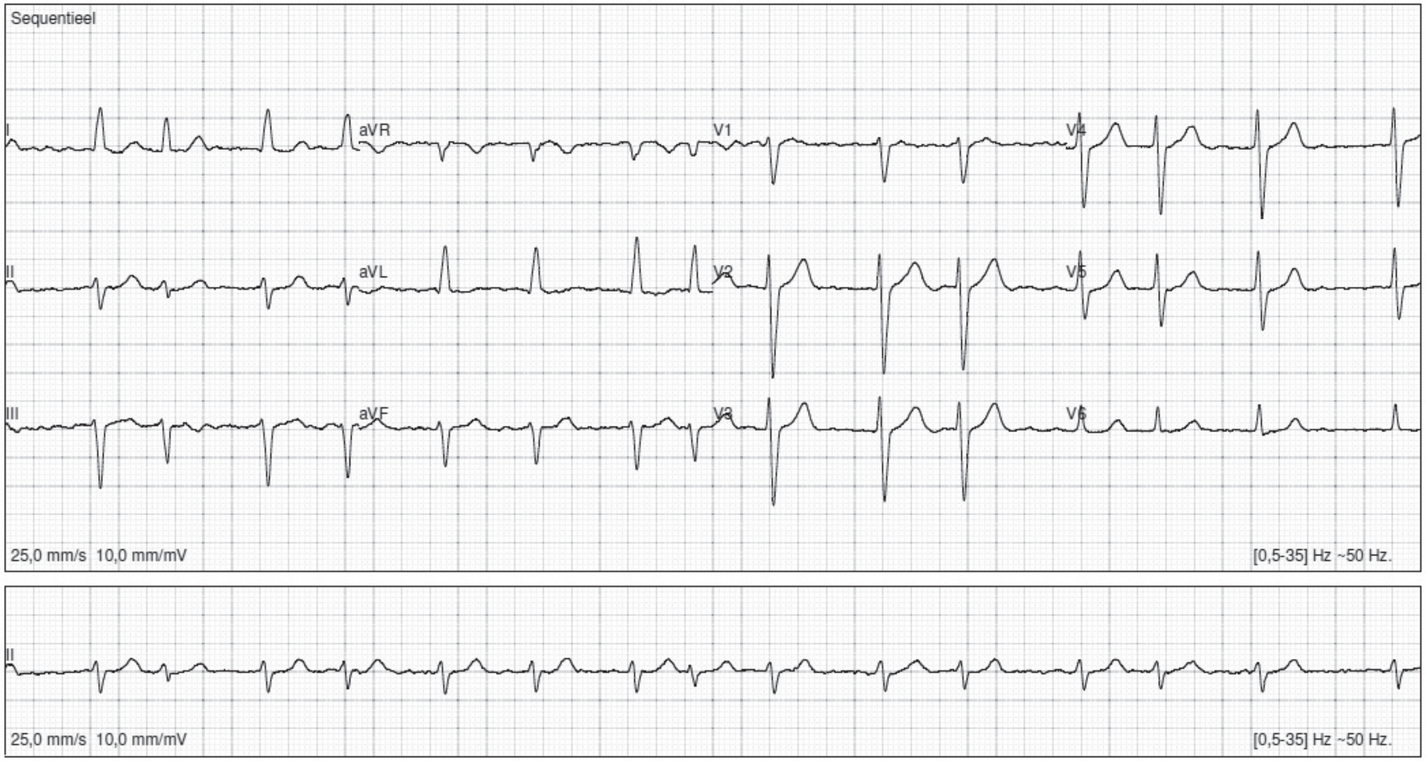
Atrial fibrillation ECG 1

89 year old woman, sent in because of a slower heart rate than usual combined with being tired and out of breath when exercising.



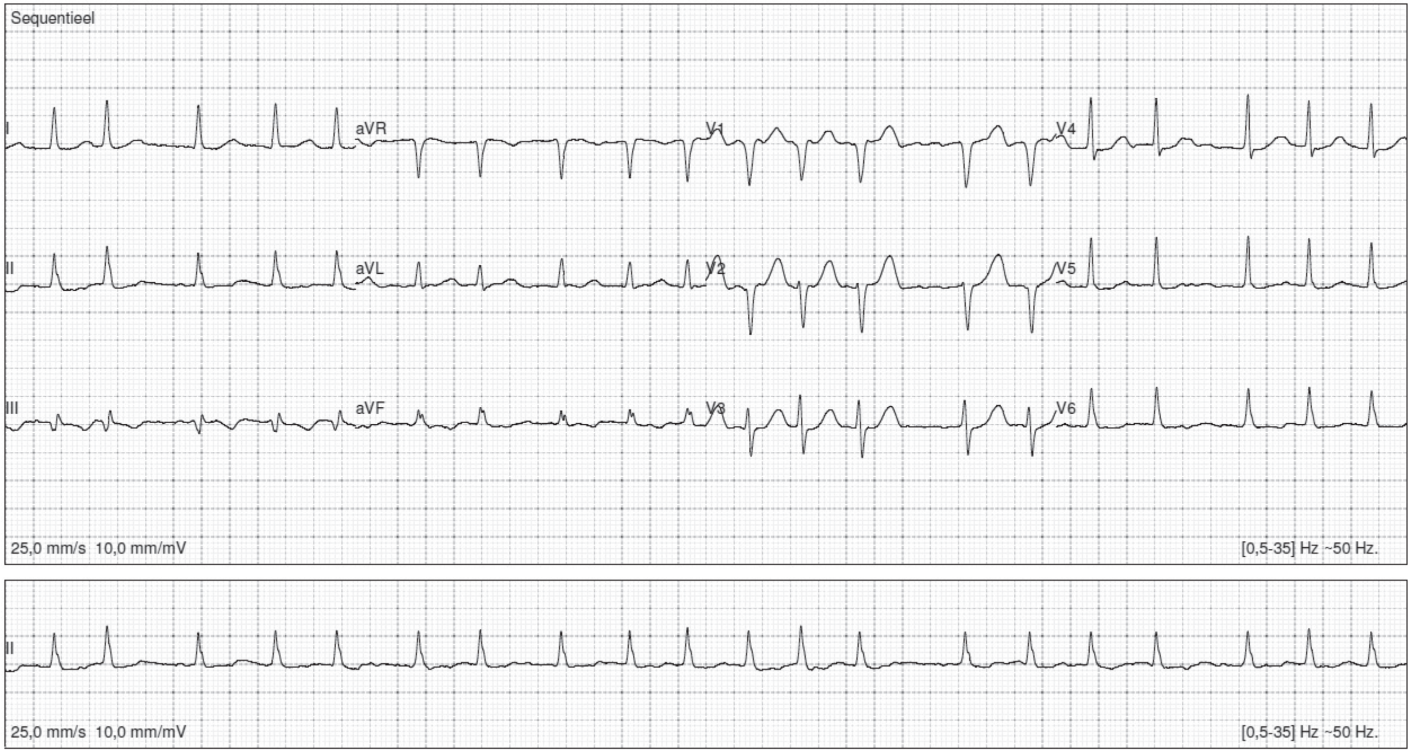
Atrial fibrillation ECG 2

76 year old woman, with swollen legs and because she is out of breath when exercising.



Atrial fibrillation ECG 3

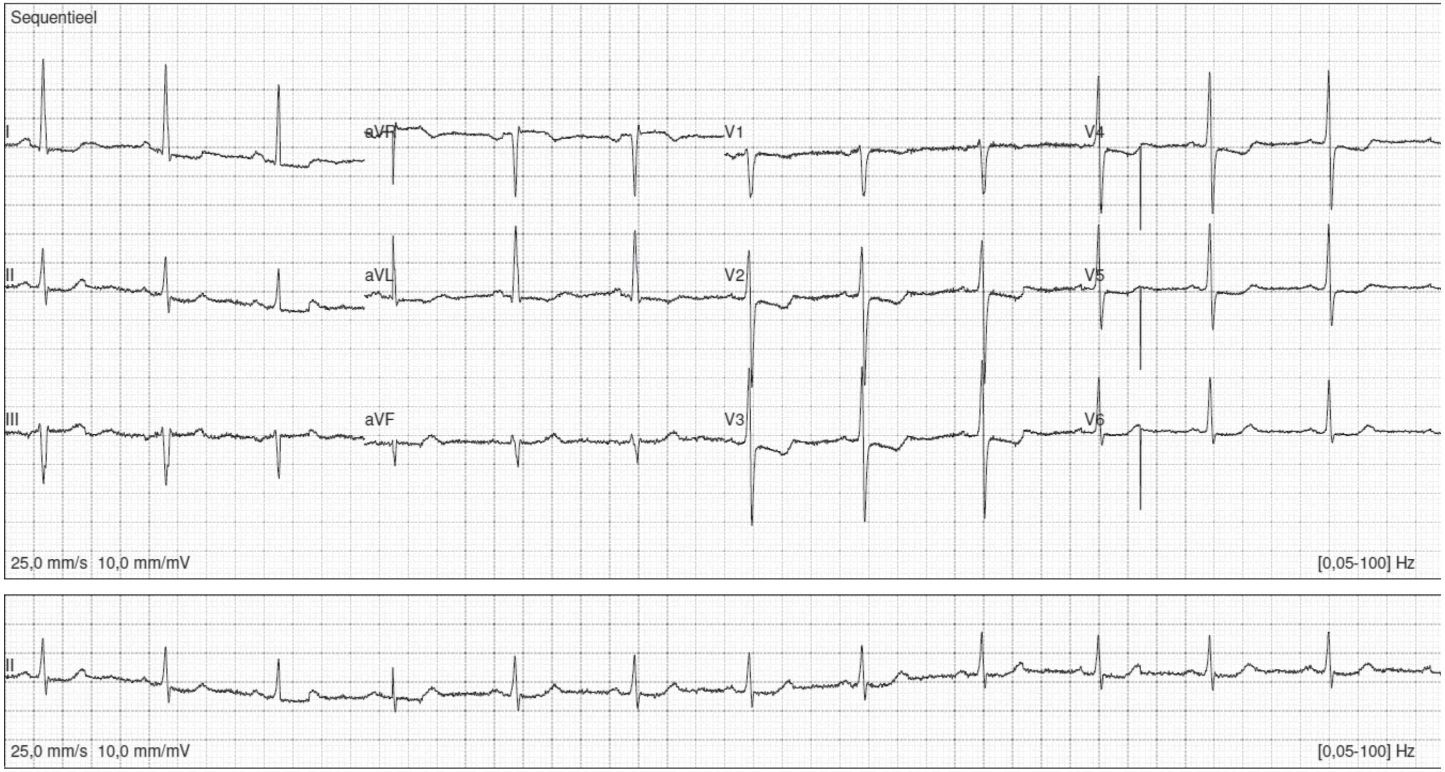
81 year old woman, she tires quickly and is dizzy.



Ischemia (based on T-abnormalities)

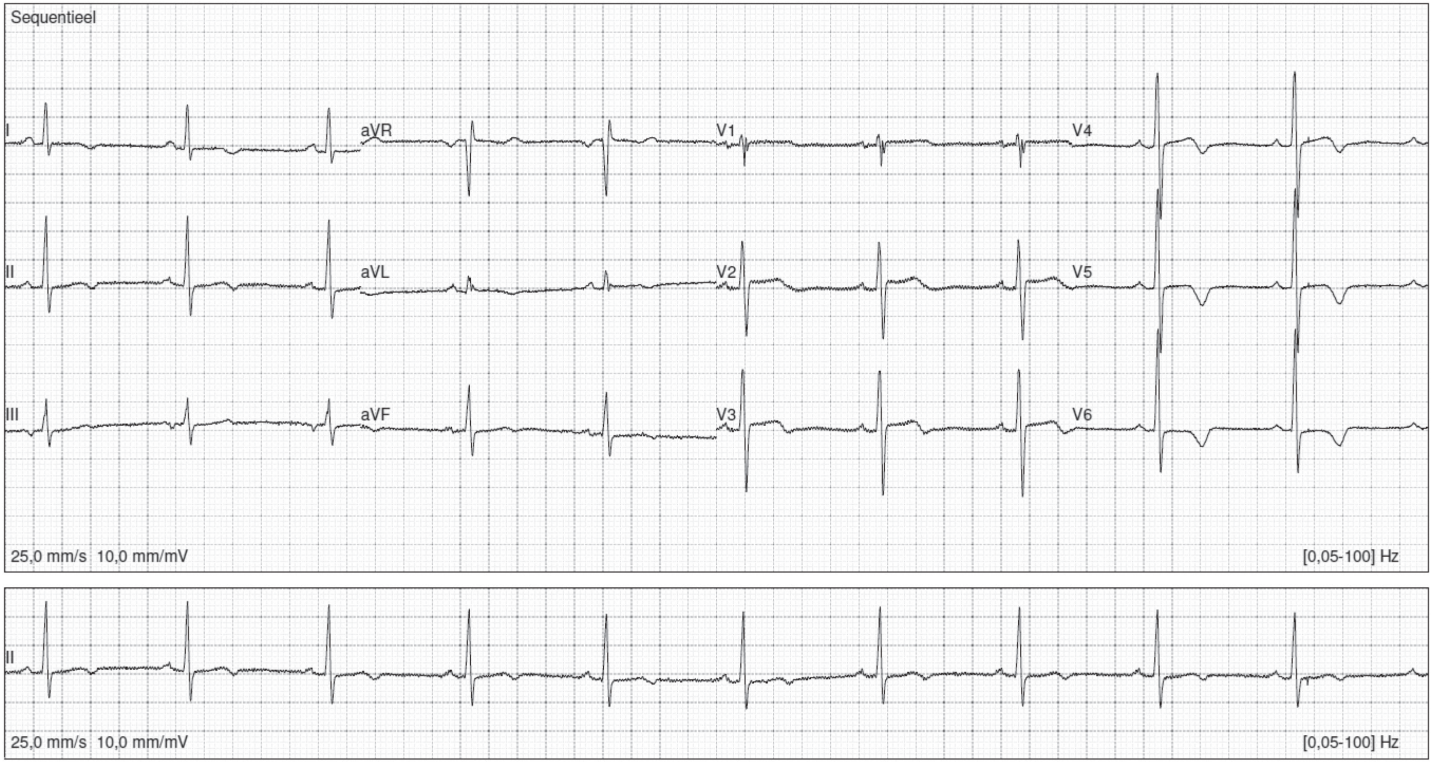
Ischemia ECG 1

59 year old woman with pain in the abdomen, a feeling of pressure on the elbows.



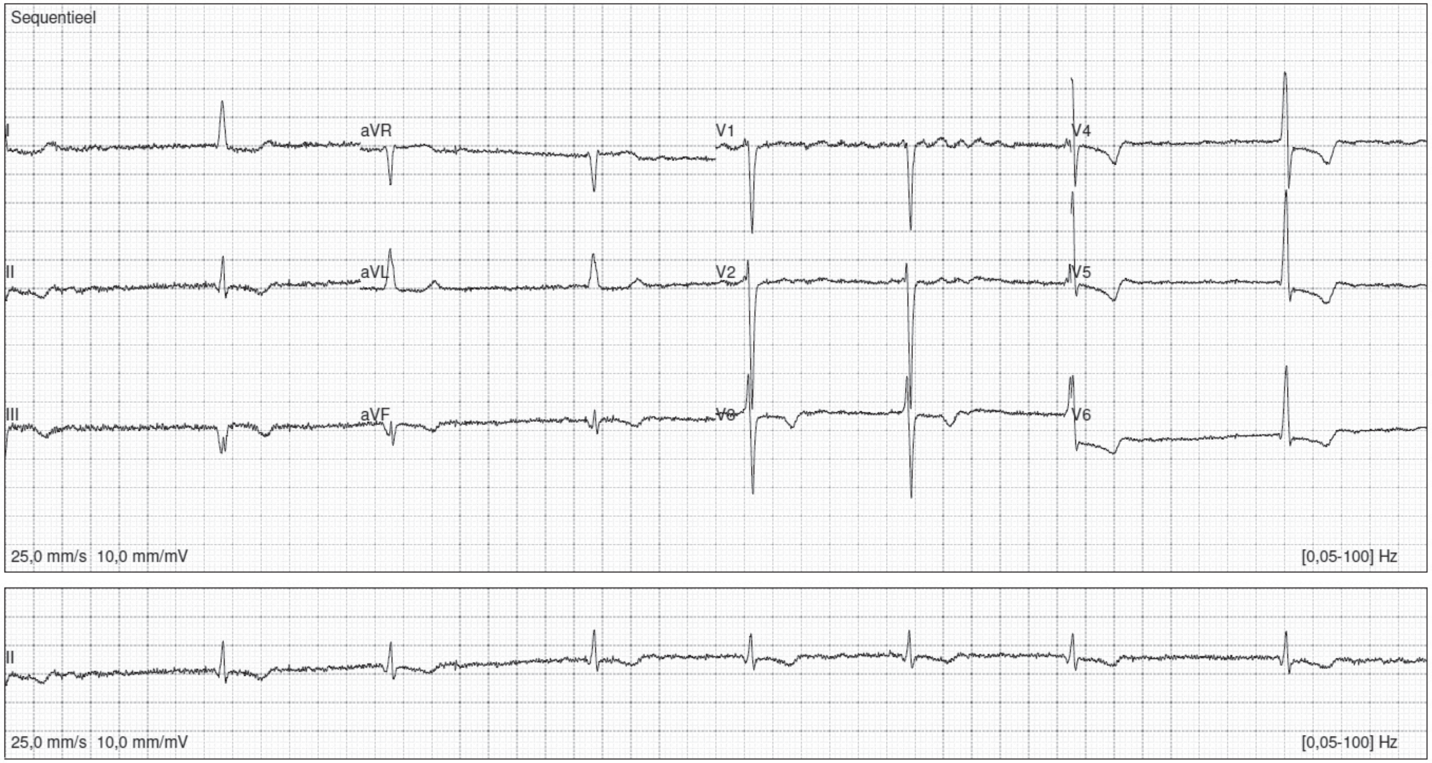
Ischemia ECG 2

68 year old woman, the cardiologist detected atypical chest pain before and the patient asked for a follow-up.



Ischemia ECG 3

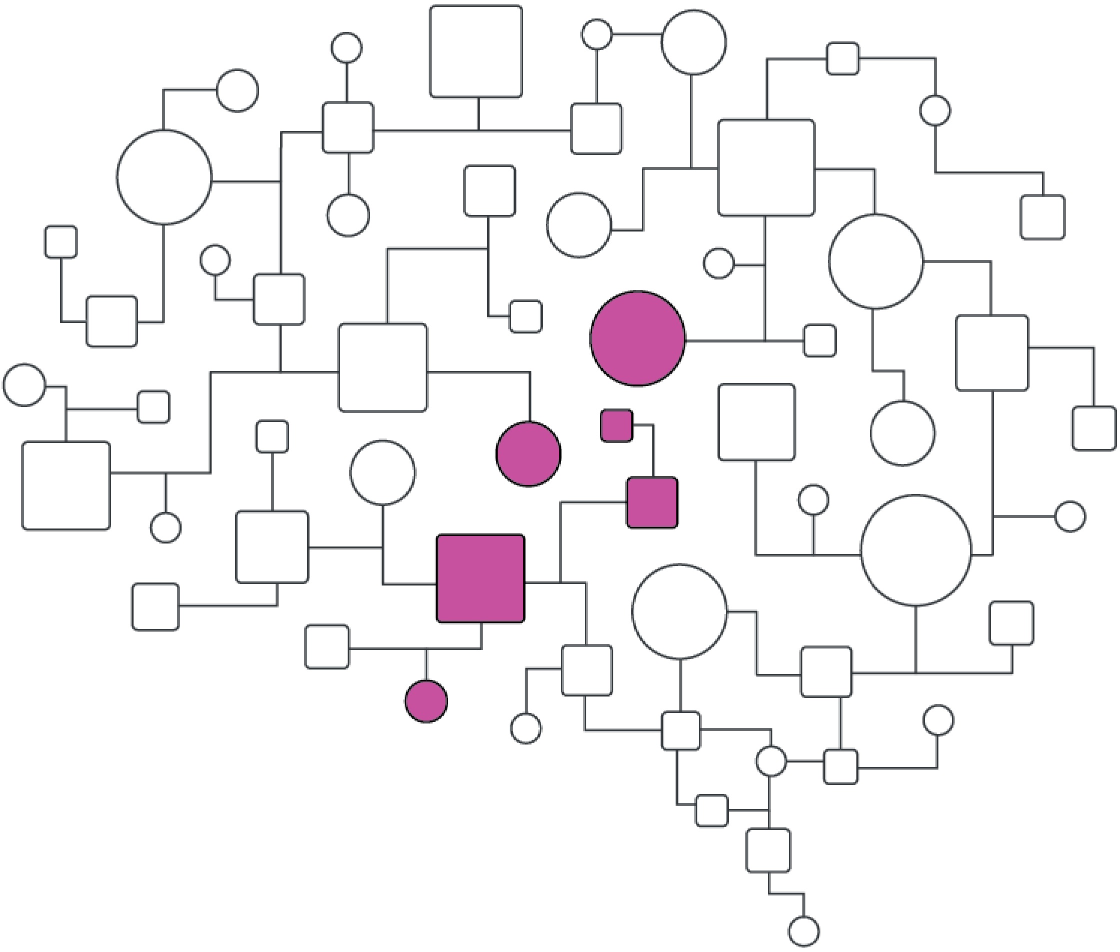
85 year old man, sent in with swollen legs and because he tires quickly.



Supplemental Material 2. Participant experience levels.

Table 1. *Response frequency on each category of all experience items.*

Experience	n	n	n
I am experienced at ECG interpretation			
Completely disagree	3	1	4
Disagree	4	3	7
Neutral	7	9	17
Agree	6	8	14
Completely agree	1	0	1
I am a capable ECG interpreter			
Completely disagree	2	1	3
Disagree	4	5	10
Neutral	11	7	18
Agree	4	8	12
Completely agree	0	0	0
I interpret ECGs in clinical practice			
Completely disagree	3	3	6
Disagree	9	4	14
Neutral	5	4	9
Agree	3	8	11
Completely agree	1	2	3
Number of ECGs interpreted			
< 10	3	1	4
< 25	7	3	10
< 50	2	7	10
> 50	9	10	19



CHAPTER 7

Impact of performance and information feedback on medical interns' confidence-accuracy calibration

Staal, J.*, Katarya, K.*, Speelman, M., Brand, R., Alsma, J., Sloane, J., Van den Broek, W.W., Zwaan, L.

*J. Staal and K. Katarya are shared first authors; they contributed equally to the work.

Published in *Advances in Health Sciences Education* 2023.

DOI: 10.1007/s10459-023-10252-9.

Abstract

Background: Diagnostic errors are a major, largely preventable, patient safety concern. Error interventions cannot feasibly be implemented for every patient that is seen. To identify cases at high risk of error, clinicians should have a good calibration between their perceived and actual accuracy. This experiment studied the impact of feedback on medical interns' calibration and diagnostic performance.

Methods: In a two-phase experiment, 125 medical interns from Dutch university medical centers were randomized to receive no feedback (control), feedback on their accuracy (performance feedback), or feedback with additional information on why a certain diagnosis was correct (information feedback) on 20 chest X-rays they diagnosed in a feedback phase. A test phase immediately followed this phase and had all interns diagnose an additional 10 X-rays without feedback. Outcome measures were confidence-accuracy calibration, diagnostic accuracy, confidence, and time to diagnose.

Results: Both feedback types improved overall confidence-accuracy calibration ($R^2_{\text{No Feedback}} = 0.05$, $R^2_{\text{Performance Feedback}} = 0.12$, $R^2_{\text{Information Feedback}} = 0.19$), in line with the individual improvements in diagnostic accuracy and confidence. We also report secondary analyses to examine how case complexity affected calibration. Time to diagnose did not differ between conditions.

Conclusion: Feedback improved interns' calibration. However, it is unclear whether this improvement reflects better confidence estimates or an improvement in accuracy. Future research should examine more experienced participants and non-visual specialties. Our results suggest that feedback is an effective intervention that could be beneficial as a tool to improve calibration, especially in cases that are of not too complex for learners.

Keywords: calibration, clinical reasoning, diagnostic error, feedback, medical education

Introduction

Diagnostic errors are defined as unintentionally missed, delayed, or wrong diagnoses and form a threat to achieving high quality care.(1) It is estimated that in the United States alone, 12 million adults are affected by diagnostic errors yearly(2), even though 80% are estimated to be preventable.(3) Moreover, diagnostic errors resulted in higher mortality rates when compared with other adverse events (i.e., errors that resulted in unintended harm).(3) Given the major implications for patient safety, it is crucial to develop strategies to prevent diagnostic errors.

Research shows that diagnostic errors are primarily caused by flaws in clinician's cognitive processes, often in combination with technical and organizational factors.(4) Ideally, error interventions would be reserved for cases at a high risk of error, as taking extra time for every patient that is seen is not feasible.(5) To identify high risk cases, clinicians should be able to accurately predict when they need help. This concept is measured as calibration, or the alignment of a clinician's confidence in their accuracy and their actual accuracy. Unfortunately , the confidence-accuracy calibration of clinicians is poor(6) and they are often overconfident(7), which impedes them in obtaining help. On top of that, calibration is found to get even worse as cases get more difficult.(6) Improving calibration could help clinicians in recognizing when they are at risk of making an error and this may prevent diagnostic errors.(8-11)

One strategy to improve calibration is feedback.(8-13) Feedback is generally defined as information concerning one's performance or understanding of a task.(14) It is suggested that this information will raise awareness of the mismatch between estimated performance and actual performance, which will help to close that gap.(12, 14) A comprehensive review by Wisniewski et al.(15) has shown that feedback is beneficial overall, but that specific forms of feedback are more effective. For example, feedback on correct responses or concrete feedback focused on specific goals were shown to result in larger improvements.(16) High information feedback is also seen as very valuable: this type of feedback not only helps students understand what mistake they made, but also why they made it and how they can avoid it in the future.(15) Another, less effective, form of feedback is performance feedback, which only informs the recipient of whether their response was correct or not.(14, 17) Despite this evidence, appropriate forms of feedback are rarely provided in clinical practice(8, 11, 18) and even less is known on how this feedback impacts clinicians' calibration. Previous studies have shown that performance feedback could improve calibration on easy clinical cases(19) but not on difficult cases.(20) It has been suggested that information feedback is needed to improve diagnostic performance,(17, 21) though evidence for its effects remains scarce.(22) While this is an important first step in understanding how feedback affects calibration, more research is needed to compare the different types of feedback, especially given that prior research suggests performance feedback may also be effective.

This study examined the effect of performance feedback and information feedback on calibration and other aspects of diagnostic performance, compared to a control condition that did not receive feedback. Performance was measured as the diagnostic accuracy, confidence, calibration, and time to diagnose of medical interns diagnosing chest X-rays. We hypothesized that information feedback would be more beneficial for diagnostic accuracy than performance feedback, as it has the potential to fill knowledge gaps by addressing mistakes in interpretations.(21) Further, we hypothesized that both feedback types would improve calibration because both give an opportunity to update the recipient's estimate of their performance. Last, we expected that receiving information feedback would reduce time to diagnose in the test phase compared to the no feedback condition, because interns could learn how to correctly recognize the X-ray diagnoses in the feedback phase. Conversely, we expected performance feedback would increase time to diagnose as we expected it would make students more aware of their limitations but would not provide them with ways to better diagnose the X-rays.

We further explored confidence and calibration in additional analyses. First, previous research suggests better performing participants can better estimate their performance regardless of the intervention.(23) To better understand the relation between accuracy and confidence, we compared the confidence of the 25% lowest and 25% highest scoring (on accuracy) interns. Second, prior research has shown poorer calibration for more difficult cases.(6) With a wider gap between accuracy and confidence, we were interested in exploring whether feedback would have an even larger impact on difficult cases relative to easier cases.

Methods

Ethics approval

The study was approved by the medical ethical committee of the Erasmus University Medical Center (Erasmus MC) (MEC-2021-0808). All participants gave informed consent. All methods were carried out in accordance with the relevant guidelines and regulations.

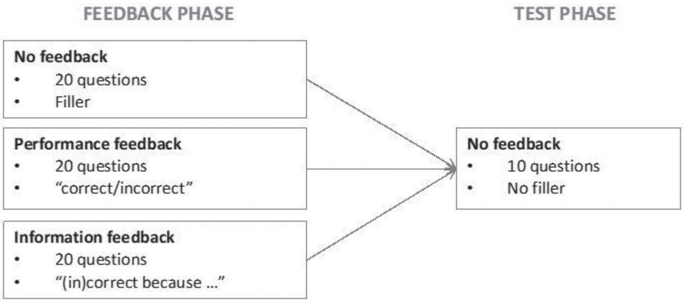


Figure 1. Study design.

Design

We conducted a computer-based experiment with a 2 (phase) x 3 (feedback condition) mixed design. In the first phase, the feedback phase, participants were randomly divided into one of three conditions (no feedback, performance feedback, or information feedback) and diagnosed 20 chest X-rays (Figure 1). Each condition provided an additional layer of feedback. After participants entered a diagnosis, those in the no feedback condition were shown the X-ray a second time with no extra information on their diagnosis or the X-ray itself. Those in the performance feedback condition saw the X-ray again as well, but were additionally told whether or not their diagnosis was correct and what the actual correct diagnosis was. Finally, participants in the information feedback condition also saw the X-ray again and received the correct diagnosis, with the addition of an explanation on how the correct diagnosis could be identified (Appendix A). In the second phase, the test phase that immediately followed the feedback phase, participants diagnosed 10 new X-rays without receiving feedback.

Participants

Interns in at least their fourth year of Dutch medical school, who were about to start clinical internships, were recruited during class, through online student portals, and via social media. The estimated sample size was calculated using G-power 3.1.9.7 (24) for one-way analysis of variance (ANOVA) with a power of 0.80, α of 0.05, and a medium effect size of 0.3 based on Nederhand et al.(19) This resulted in an estimated sample size of 111 participants.

Materials

Thirty chest X-rays representing five diagnoses (i.e., atelectasis, pleural effusion, pneumothorax, tumor, or no abnormality) were selected from the Erasmus MC database and external open access databases. The diagnoses were confirmed by CT scans. Per diagnosis, four X-rays were selected for the feedback phase and two for the test phase. Cases were matched across phases on diagnosis and difficulty level, ensuring that the cases were comparable. The difficulty level was judged for the level of medical interns with little experience and confirmed by an internist (JA), a medical doctor (RB), and a final year medical student (MS). The cases were classified as easy if all three experts could diagnose the X-ray correctly and as difficult if only two of the three experts could diagnose the X-ray correctly.

Procedure

The experiment was conducted using an online questionnaire prepared in Qualtrics (an online survey tool). Upon starting the experiment, participants received an information letter and were asked to sign informed consent. They were fully informed about the goal of the study. Participants then filled out general demographics. During the feedback phase, participants were randomized into one of the three feedback conditions. For each case, they had to select the most likely diagnosis out of five possible diagnoses from a drop-down menu and then were asked to indicate how confident they were in this diagnosis. Then, in the test phase, participants diagnosed ten new chest X-rays without feedback and marked their confidence per case. After completing the experiment, all participants received information feedback on the test phase X-rays and in addition, the no feedback condition received information feedback on the feedback phase X-rays (Appendix A).

Outcome measures

The independent variable was the type of feedback participants received in the feedback phase. This was no feedback (control condition), performance feedback, or information feedback. The dependent variables were diagnostic accuracy, confidence, confidence-accuracy calibration, and time to diagnose. For diagnostic accuracy, selection of the correct diagnosis was scored as 1, any other answer was scored as 0, based on pre-established diagnoses. We further measured confidence on a scale from 0-10. Confidence-accuracy calibration was derived from the diagnostic accuracy and confidence measures. Finally, time to diagnose was measured in seconds.

Statistical analysis

We performed a Kolmogorov-Smirnov to test for normal distribution. If the data were normally distributed, we performed a one-way ANOVA to compare the outcome measures between the three conditions in the test phase. If this test was significant, we performed a Kruskal-Wallis test (non-parametric ANOVA) instead. We focused on the results from the test phase because the intervention needed to be finished before its effects could be measured. If the comparison of the three conditions was significant, we performed post-hoc Bonferroni tests. We assumed significance if $p < 0.05$. All tests were performed in IBM SPSS Statistics (Version 28, Armonk, NY: IBM Corp).

Confidence-accuracy calibration was derived by plotting the mean diagnostic accuracy and mean confidence for each condition. For this, the mean accuracy was converted into a percentage and the mean confidence was multiplied by ten to make it comparable to accuracy. Calibration was additionally quantified using the R^2 as a measure of goodness-to-fit to a scatterplot of the mean confidence and mean accuracy per condition. This was done according to the method described by Staal et al.(25) in which a higher R^2-value indicated a better calibration.

Furthermore, we performed exploratory analyses to further investigate confidence and calibration. We compared average confidence over all test phase cases for the 25% worst and 25% best performing students and compared the outcomes using a between subjects t-test. Secondly, we compared the effects of feedback on diagnostic accuracy, confidence, calibration, and time to diagnose separately for easy and difficult cases using a paired t-test.

Results

Demographics

A total of 125 medical interns participated. 45 participants were randomized into the no feedback condition, 38 into the performance feedback condition, and 42 into the information feedback condition. Both phases were completed by 116 participants and only these participants were included for analysis. Participant demographics are displayed in Table 1. Means of all outcome measures for the three feedback conditions are listed in Table 2.

Table 1. *Participant demographics. A total of 125 interns participated.*

Age (mean (SD))	**Sex (N (%) female)**	**University (N (%) Erasmus MC)**	**Time studying medicine (Mean (SD))**	**Attended clinical clerkships (N (%))**		
				None	**Internal medicine**	**Multiple**
23 (2) years	93 (74.4%)	118 (94.4%)	53 (21) months	51 (40.8%)	53 (42.4%)	21 (16.8%)

Table 2. *Overview of means and 95% confidence interval for performance in the test phase, per feedback condition.*

		Condition					
		No feedback		**Performance feedback**		**Information feedback**	
Outcome measure		*Mean (SD)*	*95% CI*	*Mean (SD)*	*95% CI*	*Mean (SD)*	*95% CI*
	Diagnostic accuracy (0-1)	0.49 (0.2)	[0.43- 0.55]	0.65 (0.2)	[0.58- 0.72]	0.68 (0.2)	[0.60-0.75]
	Confidence (0-10)	5.74 (1.1)	[5.40-6.09]	6.38 (1.6)	[5.82-6.93]	6.39 (1.2)	[6.02-6.75]
	Time to diagnose (in seconds)	16.98 (7.2)	[14.70 -19.27]	15.02 (5.0)	[13.27-16.78]	19.13 (16.1)	[14.06-24.20]

Main analyses

Data for diagnostic accuracy and time taken to diagnose were not normally distributed, so we performed a Kruskal-Wallis test.

Diagnostic accuracy. Diagnostic accuracy between feedback conditions differed significantly overall (F(2) = 18.06, $p < 0.001$). Post-hoc analysis showed that the no feedback condition scored lower than the performance feedback condition (F(2) = -25.25, p = 0.003, d = 0.79) and the information feedback condition (F(2) = -29.02, $p < 0.001$, d = 0.86). The feedback conditions did not differ significantly (F(2) = -3.78, p = 1.000).

Confidence

Overall, confidence differed significantly between all feedback conditions (F(2) = 3.29, p = 0.041); however, no significant differences were found in the pairwise post-hoc comparisons between the conditions ($p > 0.050$ for all).

Confidence-accuracy calibration

We now present the main variable of interest, which is derived from the preceding data on accuracy and confidence. Mean diagnostic accuracy was overall well-aligned with mean confidence (Figure 2). The confidence-accuracy calibration was lowest in the no feedback condition ($R^2 = 0.05$). Both feedback conditions achieved better calibration, with information feedback showing the highest calibration (performance feedback: $R^2 = 0.12$; information feedback: $R^2 = 0.19$) (Appendix B).

Time to diagnose

Between the three conditions, there were no significant differences in time spent on diagnosing the cases (F (2) = 3.24, p = 0.197).

Exploratory analyses

As mentioned in the introduction, exploratory analyses were performed to further understand our results and the impact of feedback.

First, we selected the 25% lowest scoring interns (N = 32, average test phase accuracy ≤ 0.4) and the 25% highest scoring interns (N = 39, average test phase accuracy ≥ 0.8). Confidence for the lowest scoring interns was not normally distributed (p = 0.042), though it was normally distributed for the highest scoring interns (p = 0.200). Given that a non-parametric test gave the same results as the t-test, we reported the t-test. The 25% best performing interns were more confident (M = 6.8, SD = 1.26) than the 25% worst performing interns (M = 5.4, SD = 1.34; $p < 0.001$). The best performing interns were underconfident whereas the worst performing interns were overconfident about their performance.

Second, we plotted the results separately for easy and difficult cases (See Figures 3 and 4). Overall, mean diagnostic accuracy was significantly lower (t(115) = 7.37, $p < 0.001$) for difficult cases (M = 0.40, SD = 0.37) compared to easy cases (M = 0.65, SD = 0.24). The same was true for mean confidence (t(115) = 8.17, $p < 0.001$) for difficult (M = 5.41, SD = 1.57) compared to easy cases (M = 6.34, SD = 1.35). Confidence-accuracy calibration was better for easy cases ($R^2 = 0.18$) (Figure 3A), compared to difficult cases ($R^2 = 0.02$). The calibration for easy cases was worst in the no feedback condition ($R^2 = 0.06$) and improved in the feedback conditions, with information feedback showing the highest calibration (performance feedback: $R^2 = 0.11$, information feedback: $R^2 = 0.22$). Feedback did not improve calibration in difficult cases (no feedback: $R^2 = 0.01$, performance feedback: $R^2 = 0.02$, information feedback: $R^2 = 0.01$).

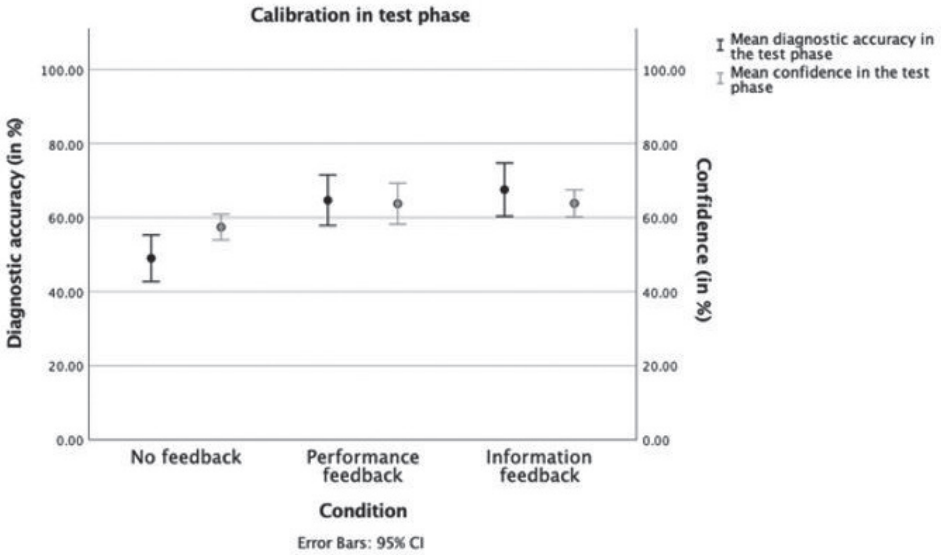


Figure 2. Mean accuracy and confidence results of the test phase per feedback condition. Error bars represent the 95% confidence interval.

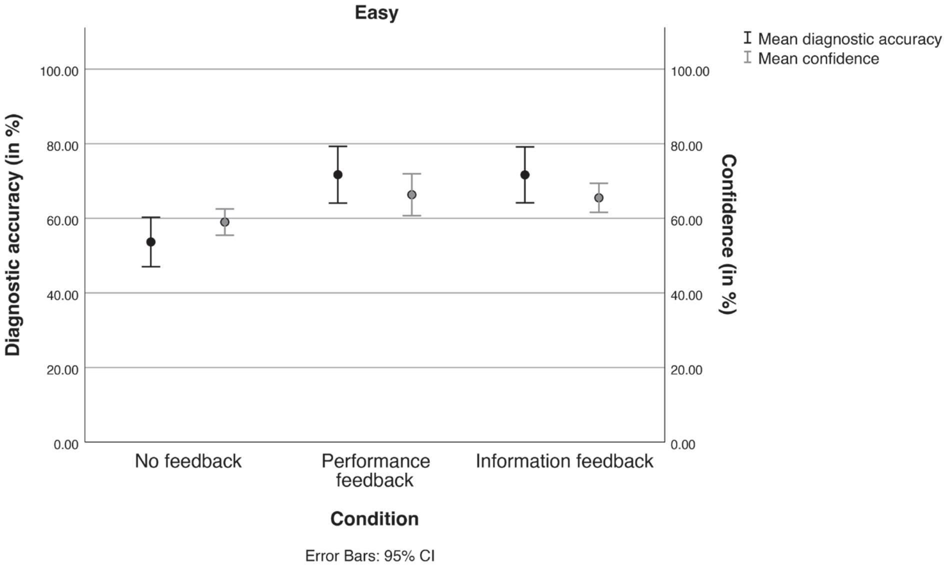


Figure 3. Interns' mean diagnostic accuracy and confidence scores per feedback condition for easy cases. Error bars represent the 95% confidence interval.

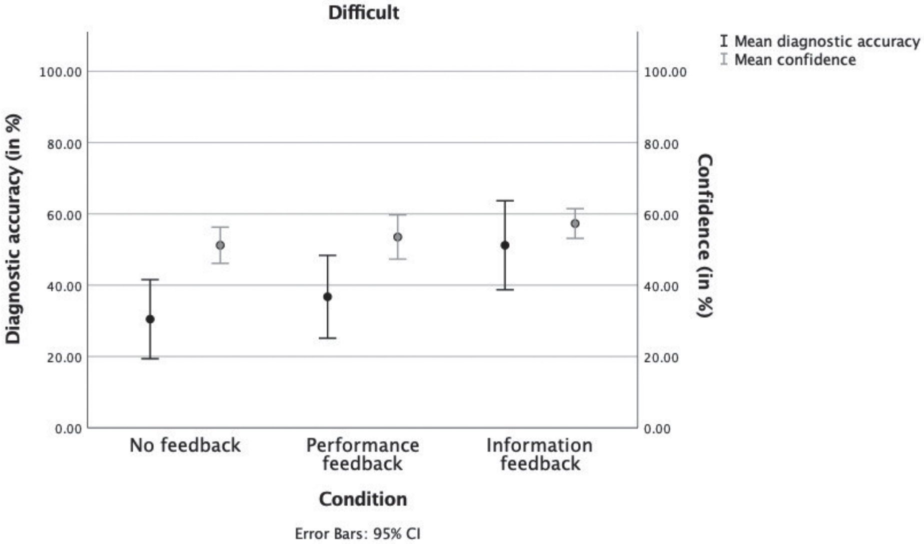


Figure 4. Interns' mean diagnostic accuracy and confidence scores per feedback condition for difficult cases. Error bars represent the 95% confidence interval.

Discussion

The current study examined the impact of performance feedback and information feedback, compared to a control condition who did not receive feedback, on the confidence-accuracy calibration and diagnostic performance of medical interns who diagnosed chest X-rays. Contrary to our hypothesis, both types of feedback improved diagnostic accuracy. Confidence increased in both feedback conditions; this increase especially stands out compared to the small confidence intervals around interns' average reported confidence. Although the difference was no longer significant in the post-hoc tests, it indicates that confidence was influenced by feedback. In line with our hypothesis, overall calibration improved in both feedback conditions as compared to the no feedback condition. Contrary to our hypothesis, time to diagnose did not differ between the conditions.

Further exploratory analyses indicated that interns' confidence seemed at least somewhat sensitive to their performance, as the 25% worst performing interns reported lower confidence than the 25% best performing interns and confidence was lower for more difficult cases. However, we cannot be sure of the underlying mechanisms and should keep in mind that people often show a tendency to score more towards the middle of a scale (to 50% confidence in this case), which would also result in the pattern we observe. For easy cases, interns were overall well-calibrated and calibration increased in the feedback conditions; for difficult cases calibration was poor and was not affected by feedback condition, though future research should replicate these results in a larger sample of cases as the difficult case sample only consisted of two cases.

Our results regarding the positive impact of performance feedback on diagnostic accuracy and overall calibration are in line with previous studies.(13, 19, 26) We found good calibration in easy cases, similarly to Nederhand et al. (19), along with an increase in calibration in the feedback conditions. In line with Kuhn et al. (20), we also observed poorer calibration in difficult cases, but we did not replicate their observation that participants became underconfident. If anything, participants in our study appeared to be more overconfident as opposed to underconfident. The positive effects of information feedback on diagnostic performance we observed are in line with previous work, though this work was not specifically aimed at medical education.(14, 15) Lastly, we observed that performance feedback and information feedback were equally effective, contrary to Ryan et al. (21), who proposed that information feedback was superior as it has the potential to fill knowledge gaps.

Although our study indicated that feedback was overall beneficial to calibration, it remains difficult to determine what processes underlie this improvement. One possible explanation is that calibration improved as a result of interns' improved accuracy rather than a change in their confidence. We observed a similar pattern as Meyer et al. (6) who showed that clinician's confidence was less sensitive to changes in their accuracy, as confidence was relatively stable across easy and difficult cases despite larger fluctuations in accuracy. On the other hand, our exploratory analyses suggested that interns' were at least somewhat sensitive to case difficulty, as confidence was significantly lower for the 25% worst performing interns compared to the 25% best performing interns, and confidence was lower for difficult cases relative to easy cases. Further research is necessary to understand what exactly we are measuring when we ask clinicians for their subjective confidence: perhaps confidence also reflects clinicians' decision threshold, or how certain they want to be before they decide on a diagnosis. In that case, the measure would be expected to remain stable. It will be crucial to understand clinician's confidence and how we measure it before we can improve calibration. In summary, the current study shows that clinicians' calibration can be improved by feedback. However, this improvement was mostly limited to easier cases, suggesting that another approach will likely be needed to improve calibration in difficult cases. Feedback relies on the ability of the learner to recognize and improve on their mistakes, which is difficult to achieve in tasks that have a high complexity for the learner.(16) If implemented over the course of an entire curriculum, however, learners might gain more insight in their general performance and might become more effective learners over time. After all, as they are taught more, less material will be too complex and more material will become easier, which would also increase the impact of feedback. Overall, feedback remains a valuable intervention, given its effectiveness in improving diagnostic accuracy without significantly increasing time spent to diagnose. The latter might be attributed to our use of chest X-rays, as visual cases are usually diagnosed quicker. Furthermore, suggestions to give feedback on the diagnostic process of clinicians are becoming more frequent and our findings support this endeavor.(11) There are ideas to standardize communicating the final diagnosis of a patient to the clinician who had seen the patient.(27-29) Future research should replicate the current findings in more experienced clinicians and test the implementation of both feedback types in practice.

This study has several strengths and limitations. Strengths include the experimental design with control condition, ensuring that effects seen in the between subjects analyses could be distinguished from learning effects between the two phases. Furthermore, all included chest X-rays had confirmed diagnoses and we could include a large number of cases because we used visual cases. This is important because sufficient practice is necessary to see effects of feedback. Limitations include that we only tested medical interns on visual images, meaning that the results are not generalizable to other levels of expertise, other types of cases, or to practice. Further, the test phase occurred immediately after the feedback phase. A time gap would have allowed participants more time to incorporate the intervention in their learning and thus have a larger effect in the test phase.(30) Another limitation was the multiple choice format for diagnosis: participants could have selected the correct diagnosis per exclusionem. However, providing too many options (i.e., via free text response) could have overwhelmed our relatively inexperienced participants. Future research should investigate if the effects of feedback remain when these factors are accounted for.

In conclusion, clinicians' confidence-accuracy calibration could be improved with both performance and information feedback, though exploratory results indicate this was limited to easier cases. More research will be needed to understand the relationship between feedback and calibration, however, for example by replicating these results in other, non-visual specialties, and in more experienced participants. Overall, feedback is a promising intervention that has the potential to improve both clinicians' actual diagnostic accuracy and their estimation of their own accuracy in cases that are not too complex for the learner, as well as the potential to reduce diagnostic errors.

Acknowledgements

We would like to thank drs. Attrach from the radiology department of the Erasmus MC for his help in recruiting participants, and the residents who participated in the study.

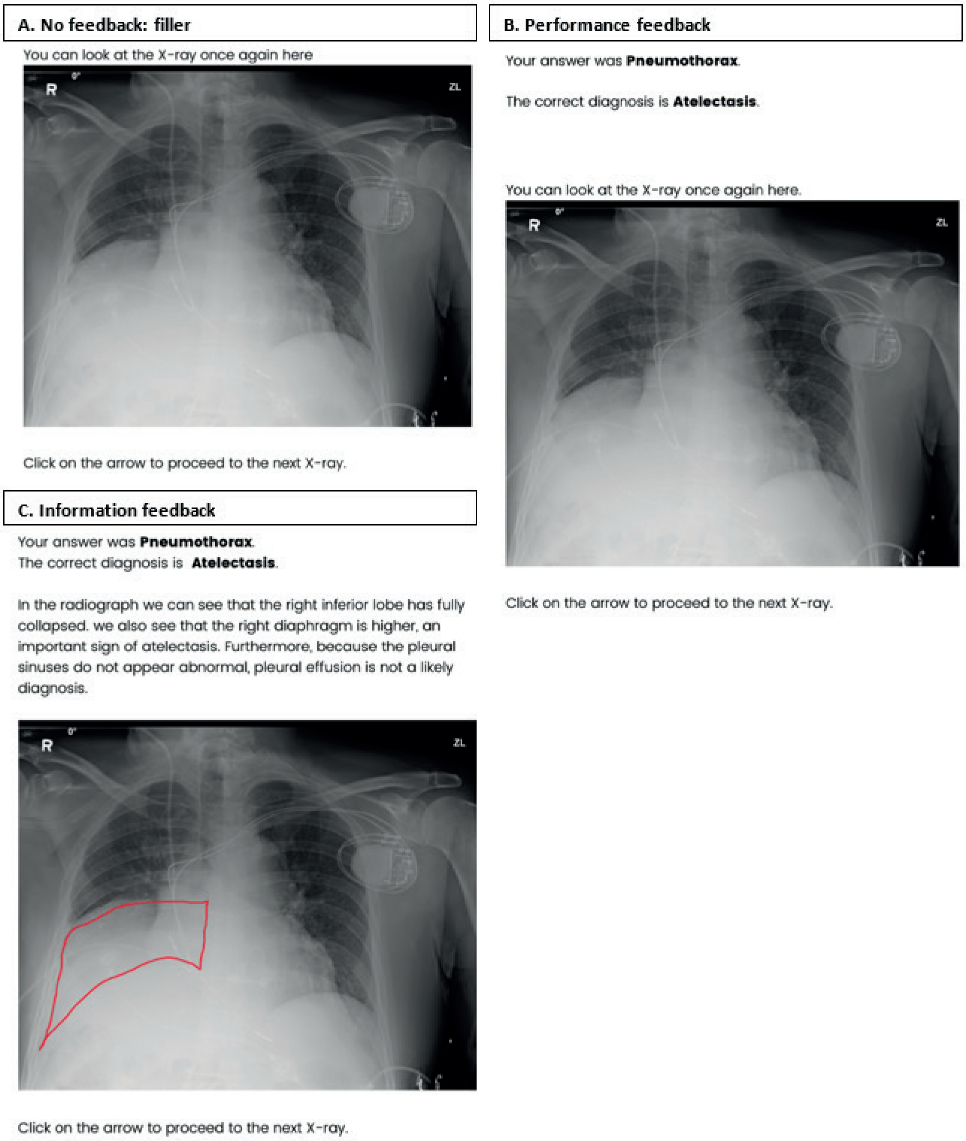
References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ quality & safety. 2014;23(9):727-31.
3. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
4. Singh H, Zwaan L. Annals for hospitalists inpatient notes-reducing diagnostic error—a new horizon of opportunities for hospital medicine. Annals of Internal Medicine. 2016;165(8):HO2-HO4.
5. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis. 2015;2(2):97-103.
6. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine. 2013;173(21):1952-8.
7. Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al. Do physicians know when their diagnoses are correct? Journal of General Internal Medicine. 2005;20(4):334-9.
8. Zwaan L, Hautz WE. Bridging the gap between uncertainty, confidence and diagnostic accuracy: calibration is key. BMJ Publishing Group Ltd; 2019. p. 352-5.
9. Meyer AND, Singh H. The path to diagnostic excellence includes feedback to calibrate how clinicians think. Jama. 2019;321(8):737-8.
10. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. The American journal of medicine. 2008;121(5):S2-S23.
11. Schiff GD. Minimizing diagnostic error: the importance of follow-up and feedback. The American journal of medicine. 2008;121(5):S38-S42.
12. Rawson KA, Dunlosky J. Improving students' self-evaluation of learning for key concepts in textbook materials. European Journal of Cognitive Psychology. 2007;19(4-5):559-79.
13. Dunlosky J, Rawson KA. Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. Learning and Instruction. 2012;22(4):271-80.
14. Hattie J, Timperley H. The power of feedback. Review of educational research. 2007;77(1):81-112.
15. Wisniewski B, Zierer K, Hattie J. The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology. 2020;10:3087.
16. Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological bulletin. 1996;119(2):254.
17. Archer JC. State of the science in health professional education: effective feedback. Medical education. 2010;44(1):101-8.
18. Burgess A, van Diggele C, Roberts C, Mellis C. Feedback in the clinical setting. BMC medical education. 2020;20(2):1-5.
19. Nederhand ML, Tabbers HK, Splinter TAW, Rikers RMJP. The effect of performance standards and medical experience on diagnostic calibration accuracy. Health Professions Education. 2018;4(4):300-7.

20. Kuhn J, van den Berg P, Mamede S, Zwaan L, Bindels P, van Gog T. Improving medical residents' self-assessment of their diagnostic accuracy: does feedback help? Advances in Health Sciences Education. 2022;27(1):189-200.
21. Ryan A, Judd T, Swanson D, Larsen DP, Elliott S, Tzanetos K, et al. Beyond right or wrong: More effective feedback for formative multiple-choice tests. Perspectives on Medical Education. 2020;9(5):307-13.
22. Kornegay JG, Kraut A, Manthey D, Omron R, Caretta-Weyer H, Kuhn G, et al. Feedback in medical education: a critical appraisal. AEM education and training. 2017;1(2):98-109.
23. Nederhand ML, Tabbers HK, Rikers RMJP. Learning to calibrate: Providing standards to improve calibration accuracy for different performance levels. Applied Cognitive Psychology. 2019;33(6):1068-79.
24. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91.
25. Staal J, Alsma J, Mamede S, Olson APJ, Prins-van Gilst G, Geerlings SE, et al. The relationship between time to diagnose and diagnostic accuracy among internal medicine residents: a randomized experiment. BMC medical education. 2021;21(1):1-9.
26. Lichtenstein S, Fischhoff B. Training for calibration. Organizational behavior and human performance. 1980;26(2):149-71.
27. Shenvi EC, Feupe SF, Yang H, El-Kareh R. "Closing the loop": a mixed-methods study about resident learning from outcome feedback after patient handoffs. Diagnosis. 2018;5(4):235-42.
28. Lavoie CF, Schachter H, Stewart AT, McGowan J. Does outcome feedback make you a better emergency physician? A systematic review and research framework proposal. Canadian Journal of Emergency Medicine. 2009;11(6):545-52.
29. Branson CF, Williams M, Chan TM, Graber ML, Lane KP, Grieser S, et al. Improving diagnostic performance through feedback: the Diagnosis Learning Cycle. BMJ quality & safety. 2021;30(12):1002-9.
30. Mamede S, van Gog T, Moura AS, de Faria RMD, Peixoto JM, Rikers RMJP, et al. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. Medical education. 2012;46(5):464-72.

Appendix A. Feedback conditions

Figure 1 shows an example of the feedback and fillers participants received in each condition



A. No feedback: filler

You can look at the X-ray once again here

Click on the arrow to proceed to the next X-ray.

B. Performance feedback

Your answer was **Pneumothorax**.

The correct diagnosis is **Atelectasis**.

You can look at the X-ray once again here.

Click on the arrow to proceed to the next X-ray.

C. Information feedback

Your answer was **Pneumothorax**.
The correct diagnosis is **Atelectasis**.

In the radiograph we can see that the right inferior lobe has fully collapsed. we also see that the right diaphragm is higher, an important sign of atelectasis. Furthermore, because the pleural sinuses do not appear abnormal, pleural effusion is not a likely diagnosis.

Click on the arrow to proceed to the next X-ray.

Figure 1A. Filler task in the control condition, *B*. Performance feedback, *C*. Information feedback. After the experiment, all participants received information feedback (C) on cases they had not previously received feedback for.

Appendix B - Calibration

Scatterplots of the relationship between mean accuracy and mean confidence over all cases (no feedback group: Figure 1; performance feedback group: Figure 2; information feedback group: Figure 3). The R^2 is a measure for calibration, which is expresses how well the data fit a linear model.

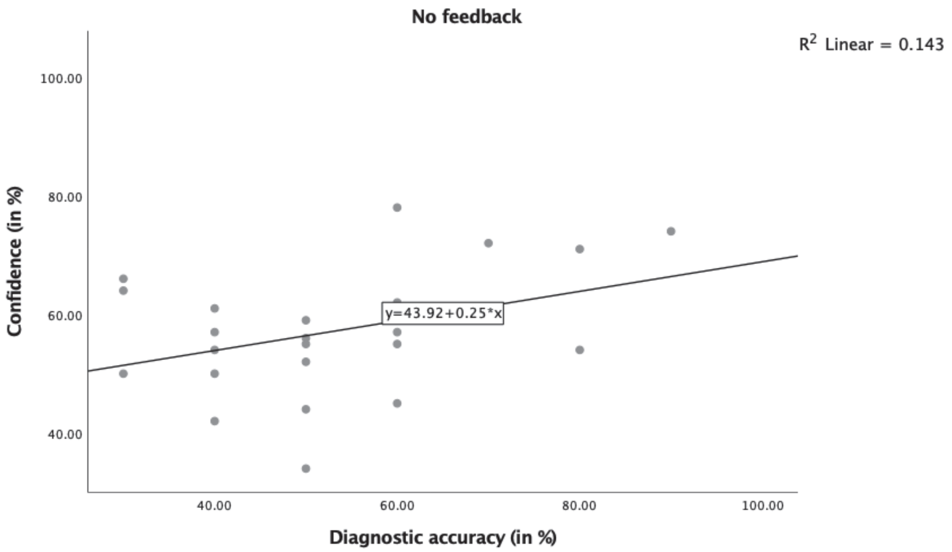


Figure 1. Scatterplot of the mean confidence and mean diagnostic accuracy of each participant in the no feedback condition.

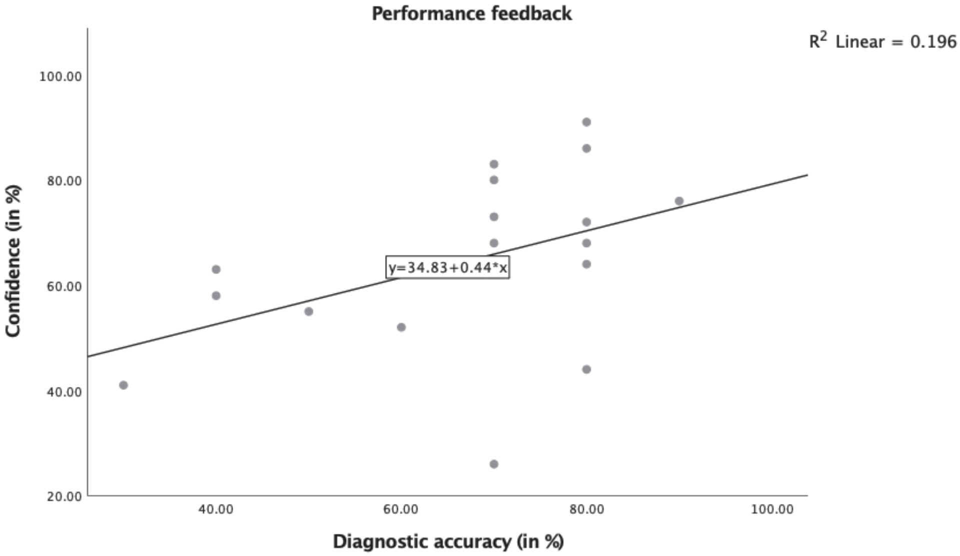


Figure 2. Scatterplot of the mean confidence and mean diagnostic accuracy of each participant in the performance feedback condition.

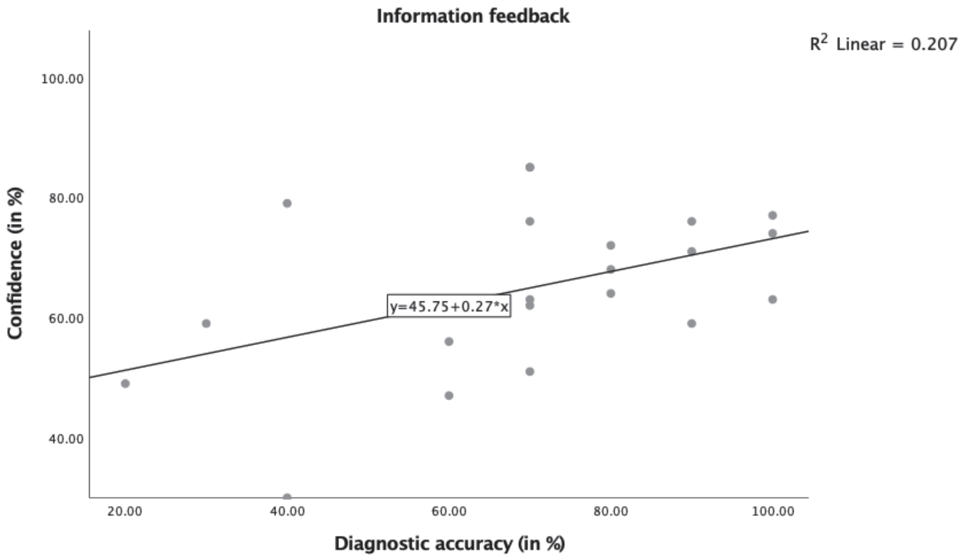


Figure 3. Scatterplot of the mean confidence and mean diagnostic accuracy of each participant in the information feedback condition.

CHAPTER 8

The effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis

Staal, J., Hooftman, J., Gunput, S.T.G., Mamede, S., Frens, M.A., Van den Broek, W.W., Alsma, J., Zwaan, L.

Published in *BMJ Quality & Safety* 2022; 31(12): 899-910.

DOI: 10.1136/bmjqs-2022-014865.

Abstract

Background: Preventable diagnostic errors are a large burden on healthcare. Cognitive reasoning tools, i.e., tools that aim to improve clinical reasoning, are commonly suggested interventions. However, quantitative estimates of tool effectiveness have been aggregated over both workplace-oriented and educational-oriented tools, leaving the impact of workplace-oriented cognitive reasoning tools alone unclear. This systematic review and meta-analysis aims to estimate the effect of cognitive reasoning tools on improving diagnostic performance among medical professionals and students, and to identify factors associated with larger improvements.

Methods: Controlled experimental studies that assessed whether cognitive reasoning tools improved the diagnostic accuracy of individual medical students or professionals in a workplace setting were included. Embase.com, Medline ALL via Ovid, Web of Science Core Collection, Cochrane Central Register of Controlled Trials, and Google Scholar were searched from inception to October 15, 2021, supplemented with hand searching. Meta-analysis was performed using a random-effects model.

Results: The literature search resulted in 4546 articles of which 29 studies with data from 2732 participants were included for meta-analysis. The pooled estimate showed considerable heterogeneity (I^2 = 70%). This was reduced to I^2 = 38% by removing three studies that offered training with the tool before the intervention effect was measured. After removing these studies, the pooled estimate indicated that cognitive reasoning tools led to a small improvement in diagnostic accuracy (Hedges' *g* = 0.20, 95% CI: 0.10-0.29, $p <$ 0.001). There were no significant subgroup differences.

Conclusions: Cognitive reasoning tools resulted in small but clinically important improvements in diagnostic accuracy in medical students and professionals, although no factors could be distinguished that resulted in larger improvements. Cognitive reasoning tools could be routinely implemented to improve diagnosis in practice, but going forward, more large-scale studies and evaluations of these tools in practice are needed to determine how these tools can be effectively implemented.

PROSPERO registration number: CRD42020186994

Keywords: checklist; clinical reasoning; cognitive bias; diagnostic error; meta-analysis

What is already known on this: Cognitive reasoning tools i.e., tools that aim to improve clinical reasoning, are often recommended to reduce diagnostic errors. Quantitative effect estimates have been aggregated over workplace-oriented and education-oriented tools. It is unknown what the impact of workplace-oriented cognitive reasoning tools is and what factors are associated with greater effectiveness.

What this study adds: Workplace-oriented cognitive reasoning tools lead to small improvements in diagnostic accuracy, but based on the current evidence no factors could be isolated that lead to greater improvements.

How this study might affect research, practice, or policy: This meta-analysis suggests that cognitive reasoning tools could improve diagnostic accuracy in practice, but that more large-scale studies are necessary to evaluate the effects of cognitive reasoning tools in practice and under which circumstances cognitive reasoning tools are most effective.

Introduction

Diagnostic errors, defined as missed, delayed, and wrong diagnoses, are a large burden on healthcare and a threat to patient safety. The National Academies of Sciences, Engineering, and Medicine (NASEM), the collective national academy of the United States, estimated that most people will experience a diagnostic error in their lifetime, sometimes with devastating consequences.(1) A significant portion of diagnostic errors is considered preventable and effective interventions are crucial to reduce these errors.(2-4)

The use of interventions focused on cognitive factors is often recommended (3, 5-8): these factors are thought to be a primary cause of errors which have been identified in more than 75% of error cases.(4, 9-11) Such interventions, referred to as cognitive reasoning tools in this study, are aimed at improving clinical reasoning and decision-making skills by improving clinicians' intuitive and rational processing during diagnosis.(3) Examples include checklists (12), reflective practices (2, 7, 12-15), cognitive forcing strategies (12), and clinical decision support systems.(12, 16) Experiments testing the effectiveness of cognitive reasoning tools are relatively scarce (3, 17), but overall the current literature indicates these tools could improve diagnostic accuracy. Previous studies seem to suggest that this effect differs between subgroups: for example, tool effectiveness between studies differed depending on the participants' level of expertise and the difficulty level of the cases.(18)

Previous quantitative estimates of the impact of these tools on diagnostic accuracy were made by Prakash et al. (2) and Kwan et al. (16), who examined the impact of reflective practices and decision support systems, respectively. Crucially, these meta-analyses and other reviews (3, 7, 19, 20) have aggregated studies which focused on cognitive reasoning tools settings where the tools are used to improve learning and competence (education-oriented settings) with settings where the tools are used to improve performance (workplace-oriented settings), a distinction commonly made in the literature.(7, 21) The exact impact of cognitive reasoning tools on performance in workplace-oriented settings remains unknown. This study therefore aimed to separate both settings and provide insight in the effectiveness of cognitive reasoning tools aimed at workplace-oriented settings. Additionally, there is no consensus on what factors make an effective reasoning tool. In this systematic review and meta-analysis, we aimed to extend on the estimate of the effect of cognitive reasoning tools on improving diagnostic accuracy among medical students and professionals. Secondly, we aimed to identify factors in study or intervention design that were associated with higher overall effectiveness.

Methods

The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions (22) was followed in this study. The review's objectives and methods were specified in advance in the Prospero Database (registration number: CRD42020186994).

Data sources and searches

All searches were conducted with the assistance of biomedical information specialists of the medical library. The complete search strategy is documented in Appendix A. The following electronic databases were searched: Embase.com (1971-Present), Medline ALL (1946-Present) via Ovid, Web of Science Core Collection (1975-Present), Cochrane Central Register of Controlled Trials (1992-Present). Additionally, a search was performed in Google Scholar from which the 200 most relevant references were downloaded. All searches included unpublished "grey" literature. After the original search was performed in April 2020, the search was last updated on October 15, 2021. Further studies were identified by reviewing reference lists of included studies and conference proceedings (Diagnostic Error in Medicine conferences in Diagnosis) and asking colleagues about unpublished work. Authors were contacted for missing information if necessary.

Study selection

Three reviewers independently performed the title and abstract screening. An article was included for full text review if one reviewer included it. For articles that were not available in English a translation was generated via Google Translate and checked by an author who understood the language (i.e., Dutch, French, German, Swedish, Russian). No other languages were encountered. Two reviewers subsequently screened all selected full-text studies. Disagreements were solved via consensus, and if no consensus was reached, via consultation of the third reviewer. Interrater reliability was assessed using Cohen's kappa statistic.(23)

We included all studies that evaluated cognitive reasoning tools focused on medical specialists (including students and those in training) with the aim to improve diagnosis. Although we excluded educational interventions, studies that included medical students could still be considered if they measured performance using workplace-oriented tools. We defined cognitive reasoning tools as structured tools that focus on improving clinical reasoning and decision making skills.(3) There were no restrictions for publication status or publication year. Searching was limited to controlled studies (quasi-experimental or experimental studies, controlled and cross-over trials, or before-after designs) that measured diagnostic performance (either as diagnostic error or diagnostic accuracy).

We excluded tools that focused on specific diseases (e.g., diagnostic guidelines) because these present a set of decision rules that predict whether or not the patient should be diagnosed with a certain disease, instead of improving the diagnostic process in general. We further excluded studies in which the tool was not explicitly available while diagnosing cases (e.g., studies that focused on using the tool for learning and education and not on implementing it into practice). Lastly, we excluded studies focused on psychiatric diseases, because psychiatric diagnosis is largely based on identifying a certain number of behaviors in a patient that match to a disorder in the Diagnostic and Statistical Manual of Mental Disorders (DSM) (24), which is similar to using a checklist-like tool. We expected that the effectiveness of cognitive reasoning tools in psychiatric settings would not be comparable to other clinical settings.

Data extraction and quality assessment

Two reviewers independently performed data extraction and quality assessment for 30% of the studies. Disagreements were resolved via discussion and the task proceeded with a single evaluator. Data were extracted using the Cochrane Data Collection Form for intervention reviews on RCTs and non-RCTs (version 12-08-2013).(25) This form was adapted by removing questions specific for medication trials and questions specific to cognitive reasoning tools were added. Information extracted from each study included year of publication, country, participant characteristics (years of experience, level of expertise, area of expertise); type of intervention (type of tool, phase of the diagnostic process where the tool is used, diagnostic tasks the tool applies to, whether the tool's items have to be acknowledged or reported); outcome measure (measure of cases diagnosed correctly or incorrectly); setting; and research design (control group, randomization). The adapted form was pilot-tested on five randomly-selected included studies.

The methodological quality of included studies was assessed using the Cochrane Collaboration Risk of Bias (version 2.0) template.(26) This form assessed study randomization, deviations from the intended intervention, allocation concealment and blinding, outcome measures, and selective outcome reporting. On each domain, a study could be rated as at high, medium, or low risk of bias. If insufficient information was available, the domain was rated as 'no information (NI)' and the study authors were contacted. The final bias assessment was equivalent to the highest received sub-assessment.

The overall strength of the evidence was assessed using the Grading of Recommendations Assessment, Development, and Evaluation group's tool (GRADE).(27) This tool assesses the quality of evidence along the domains of risk of bias, consistency, directness, precision, and publication bias. The tool rates the confidence in the evidence as high, moderate, low, or very low.

Two studies reported diagnostic error rates (28, 29); these percentages were inversed to be comparable to diagnostic accuracy rates.

Data synthesis

The primary outcome was the difference in diagnostic performance between the control group or baseline measurement and the intervention group. For continuous data, the mean and standard deviation of diagnostic performance were used to compute the standardized mean difference (Hedges' g) and the 95% confidence interval of g; for dichotomous data, the reported effect size (i.e., odds ratio) was transformed to Hedges' g. These results were pooled using a random-effects model meta-analysis with the Hartung-Knapp adjustment (30), using the restricted maximum likelihood (REML) method to estimate variation between studies. One trial was included per study in the main analysis. If a study directly compared a control group or baseline measurement with the intervention group, this comparison was included; if there were multiple comparisons in one study, comparisons that satisfied our inclusion criteria were aggregated. Between-study heterogeneity was estimated using the I2-statistic, which was categorized as: might not be important (0-40%), moderate (30-60%), substantial (50-90%), and considerable (75-100%).(31) It was considered feasible to combine the included studies for meta-analysis if heterogeneity did not exceed 40%, which indicated consistency in the study outcomes. Further study differences could then be explored using subgroup analyses. Heterogeneity was further explored via influence and sensitivity analyses based on the risk of bias assessment. Influence was measured using leave-one-out estimates of heterogeneity and covariance ratios, where a study was considered influential if the covariance ratio was below 1. Publication bias was assessed using a funnel plot and Egger's regression.(32)

Subgroup analyses were performed for participant expertise, several intervention characteristics (i.e., intervention type, moment of intervention, intervention items), and study characteristics (i.e., diagnostic task, case difficulty, same cases used with and without intervention, study intention). Variable definitions are given in Table 1. The subgroup analyses for level of expertise and intervention characteristics were pre-specified; the analyses for study characteristics were based on observations made during study characteristic extraction. Analyses were performed with the metafor package (33) in R (version 1.4.1106) (34), with significance levels set at $p < 0.05$.

Table 1. *Definitions of the characteristics used in subgroup analyses*.

Characteristics	**Definition**
Expertise	The level of participant expertise was classified as novice (i.e., medical students), intermediate (i.e., residents, fellows), or expert (i.e., specialists, faculty).
Intervention characteristics	
Intervention type	Interventions were placed in one of four categories: checklists, computerized decision support systems, instructions at test, or guided reflection. Instructions at test were defined as interventions that instruct participants to use a certain reasoning strategy, where the instructions are provided together with the cases on which performance is measured, i.e. the test cases. Additionally, interventions were classified based on the focus of their items. Interventions could be process-focused (i.e., the intervention was applicable to any task and supported participants' general diagnostic process, such as a debiasing checklist) or content-specific (i.e., the intervention focused on the steps taken in a specific diagnostic task, such as electrocardiogram diagnosis).
Intervention moment	Interventions were classified on whether the tool was used during initial diagnosis, or to verify the initial diagnosis afterwards.
Intervention items	Interventions were classified as either tools that only required the participant to read, but not respond to the items (i.e., acknowledge), or tools that required the participant to respond to the items (i.e., report).
Study characteristics	
Case difficulty	Case difficulty was classified as either simple or complex. This characteristic was only recorded if the authors specifically reported the intended difficulty of their case sample.
Diagnostic task	The diagnostic task in a study was classified as patient diagnosis (either standardized or real patients), visual diagnosis (e.g. electrocardiogram, radiograph, dermatology diagnosis), or written case diagnosis.
Same cases used with and without intervention	This variable was recorded for a study if participants had the opportunity to diagnose the same cases before and after the intervention was implemented, and the control group did not get this opportunity. Seeing the same cases could give participants more opportunities for considering the case and could therefore make it difficult to ascribe improvements in accuracy to the intervention tool, instead of to the intervention of simply revisiting a case.
Study intention	Study intention was recorded as stated in the study aim. Studies were classified as either having the goal to evaluate the performance of a cognitive reasoning tool, or to induce errors in participants and evaluate whether the tool could fix these induced errors.

Results

Our database search yielded 4546 studies and an additional 24 studies were identified through other search activities (Figure 1). After removing duplicates, 2963 studies remained for initial screening. Of these, 2822 studies were excluded because their title and abstract did not meet the inclusion criteria, leaving 141 studies for full-text screening. Interrater reliability was moderate to substantial for title and abstract screening and substantial for full text screening, although the overall rate of agreement was almost perfect (Appendix B). 112 studies did not meet our inclusion criteria. Examples of excluded studies were studies where the intervention under study was not focused on supporting cognitive processes (35, 36), studies that did not measure diagnostic accuracy or diagnostic errors (37-40), or studies that did not describe an experiment.(41, 42) The remaining 29 studies were included for review and meta-analysis. All studies were available in English. All studies were published except for unpublished data from Staal et al. (2021). This unpublished experiment compared diagnostic accuracy on ECGs diagnosis using a debiasing checklist and a ECG-specific checklist. The data were obtained from the authors. Three studies (43, 44) (Staal et al., 2021) contained two trials (two separate interventions were tested and compared with, in these cases, the same control group). These trials were aggregated for calculation of the main effect to prevent double counting of the control group. The different interventions were evaluated separately in a subgroup analysis. The characteristics of the included studies are detailed in Table 2. The findings of the individual included studies are reported in Appendix C.

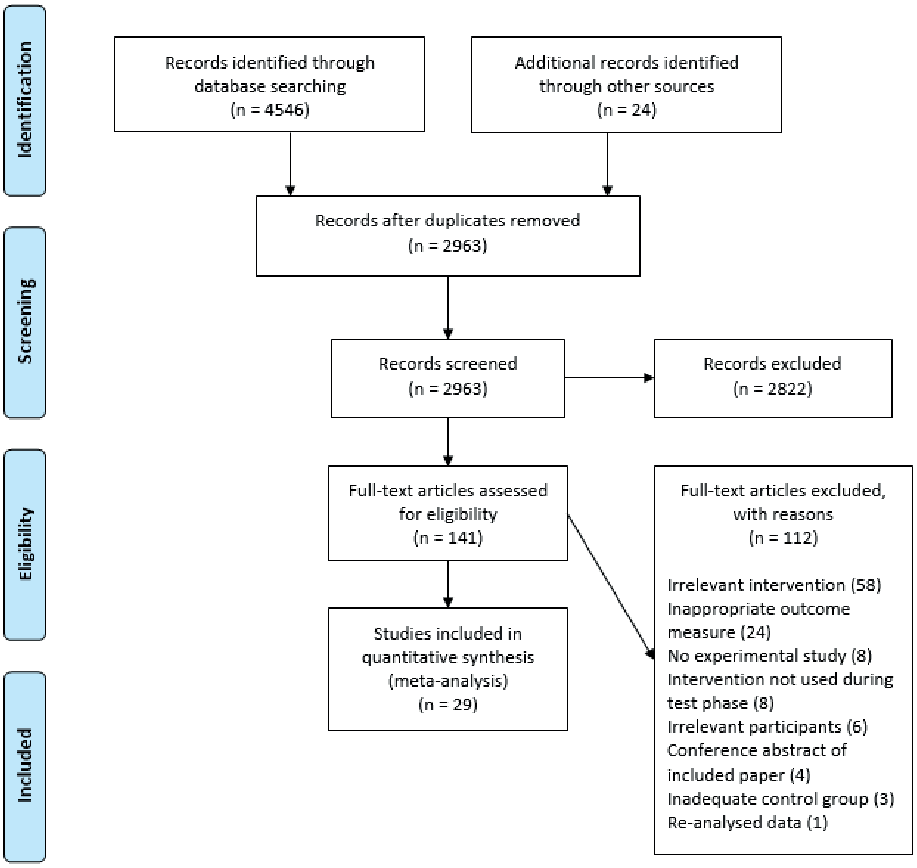

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 4546)
Additional records identified through other sources (n = 24)
Records after duplicates removed (n = 2963)
Records screened (n = 2963)
Records excluded (n = 2822)
Full-text articles assessed for eligibility (n = 141)
Full-text articles excluded, with reasons (n = 112)
Irrelevant intervention (58)
Inappropriate outcome measure (24)
No experimental study (8)
Intervention not used during test phase (8)
Irrelevant participants (6)
Conference abstract of included paper (4)
Inadequate control group (3)
Re-analysed data (1)
Studies included in quantitative synthesis (meta-analysis) (n = 29)

Figure 1. Study inclusion flowchart (PRISMA).

Table 2. *Characteristics of included studies.*

Category	**Characteristics**	**Overall, N = 29* (%)**
Region	Asia	4 (17%) (43, 45-47)
	Europe	11 (38%) (15, 29, 44, 48-54) (Staal et al., 2021)
	North America	12 (41%) (28, 55-65)
	South America	2 (7%) (66, 67)
Setting	Clinical practice	2 (7%) (28, 47)
	Experiment	27 (93%) (15, 29, 43-46, 48-67) (Staal et al., 2021)
Specialism**	Emergency medicine	5 (24%) (47, 59-61, 63)
	Family medicine, general practice	4 (19%) (48, 52, 67) (Staal et al., 2021)
	Internal medicine	9 (42%) (15, 45, 50, 51, 54, 57, 59, 60, 64)
	Osteopathy	1 (5%) (65)
	Paediatry	1 (5%) (56)
	Radiology	1 (5%) (55)
Intervention	Checklist	13 (45%) (28, 29, 43, 44, 47-49, 55, 61, 63-65) (Staal et al., 2021)
	Computerized decision support	2 (7%) (58, 67)
	Instructions at test (e.g., red flag)	2 (7%) (15, 56)
	Guided reflection	12 (41%) (45, 46, 50-54, 57, 59, 60, 62, 66)
Comparator	Non-analytical instructions	8 (28%) (43, 50-54, 59, 60)
	Diagnosis without the tool	20 (69%) (15, 29, 44-49, 55-58, 61-67) (Staal et al., 2021)
	Usual care	1 (3%) (28)
Expertise	Novice	13 (35%) (29, 43, 44, 46, 48, 49, 53, 58-60, 62, 65, 66)
	Intermediate	17 (46%) (29, 45, 47, 50-54, 56, 57, 59-61, 63, 64, 67) (Staal et al., 2021)
	Experts	7 (19%) (15, 28, 29, 48, 55, 59, 60)
Intervention type	Content-focus	29 (85%) (15, 28, 29, 43-67) (Staal et al., 2021)
	Process-focus	5 (15%) (29, 43, 44, 49) (Staal et al., 2021)
Intervention moment	During initial diagnosis	12 (41%) (28, 44, 48, 49, 55, 58-60, 62, 63, 65, 67)
	Verification after initial diagnosis	17 (59%) (15, 29, 43, 45-47, 50-54, 56, 57, 61, 64, 66) (Staal et al., 2021)
Intervention items	Acknowledge	12 (41%) (28, 29, 43, 44, 49, 56, 59-61, 63, 65) (Staal et al., 2021)
	Report	17 (59%) (15, 45-48, 50-55, 57, 58, 62, 64, 66, 67)
Case difficulty	Simple	7 (41%) (50, 52, 53, 57, 59, 60) (Staal et al., 2021)
	Complex	10 (59%) (15, 50, 52, 53, 57-61) (Staal et al., 2021)
Diagnostic task	Patient diagnosis	2 (7%) (28, 46)
	Visual diagnosis	9 (31%) (47-49, 55, 63-66) (Staal et al., 2021)
	Written case diagnosis	18 (62%) (15, 29, 43-45, 50-54, 56-62, 67)

Table 2. *Continued*

Category	Characteristics	Overall, N = 29* (%)
Same cases used with and without intervention	Yes	13 (45%) (43, 45, 50-54, 57, 61, 62, 64, 65) (Staal et al., 2021)
	No	16 (55%) (15, 28, 29, 44, 46-49, 55, 56, 58-60, 63, 66, 67)
Study intention	Evaluate	25 (86%) (15, 28, 29, 43, 44, 46-50, 53, 55-67) (Staal et al., 2021)
	Fix	4 (14%) (45, 51, 52, 54)

*Totals in columns do not equal "overall" as some studies did not report a variable or included multiple categories.

**Studies with medical students were excluded as students are generally not specialized.

Interventions

A variety of interventions was included for analysis, which were divided into 4 categories based on Lambe et al. (7): checklists, computerized decision support systems, instructions at test (i.e., interventions that instruct participants to use a certain reasoning approach), and guided reflection (Table 2). First, checklists were paper-based or online lists that guided participants through all important factors that need to be considered before coming to a final diagnosis. Second, computerized decision support tools were electronic algorithms that guided participants by suggesting differential diagnoses for certain symptoms. Third, interventions providing instructions at test aimed to guide participants' thinking in a certain way which was hypothesized to reduce errors. Finally, reflective reasoning tools were based on the deliberate reflection procedure designed by Mamede et al.(50). In some cases, similar procedures were named differently, e.g., Ilgen et al. (59, 60) used an abbreviated deliberate reflection which they called "directed search instructions". Reflective reasoning tools ask participants to consider a diagnosis for a case, then consider all information in the case that confirms or contradicts that diagnosis and information that would have been expected if the diagnosis were correct, but is not presented. Participants are then asked to repeat this process for all differential diagnosis they come up with and finally, all diagnoses are ranked in order of likelihood. Details on the interventions of each individual study and how these have been classified are listed in Table 3.

Despite variations in the format of these interventions, most shared the common focus on prompting participants to consider certain information in a specific manner (content-specific) or to consider one's reasoning processes during diagnosis (process-focused) (Table 1).

Table 3. *Descriptions of the interventions in each study and the category the intervention was assigned to.*

Study	Category	Intervention
Berbaum et al. (2006)(55)	Checklists	Checklist listing separate anatomical regions to assist visual search of chest radiographs
Cairns et al. (2016)(48)	Checklists	Interactive system (IPI) to guide ECG interpretation via a series of systematic subtasks (e.g., rhythm interpretation)
Chartan et al. (2019)(56)	Instructions at test	Prompt to identify “red flags” in clinical cases (I-RED strategy)
Costa Filho et al. (2019)(66)	Guided reflection	Reflective reasoning tool (based on Mamede et al. (2008)(50))
Dinardo et al. (2018)(57)	Guided reflection	Reflective reasoning tool (adapted from Mamede et al. (2008)(50))
Ely et al. (2015)(28)	Checklist	Checklist providing differential diagnoses depending on the patient’s complaint
Graber et al. (2009)(58)	Computerized decision support system	ISABEL, a computerized system that produces a ranked list of diagnoses based on a set of clinical findings
Ilgen et al. (2011)(60)	Guided reflection	Directed search instructions (based on Mamede et al. (2008)(50))
Ilgen et al. (2013)(59)	Guided reflection	Directed search instructions (based on Mamede et al. (2008)(50))
Kämmer et al. (2021)(44)	Checklists	General debiasing checklist, which provided prompts to carefully consider the diagnosis, or a checklist which presented differential diagnoses based on the patient’s complaint
Kilian et al. (2019)(61)	Checklists	Checklist based on reflective reasoning tools (ACT tool: seeking Alternative explanations, exploring the Consequences of missing the alternative diagnosis, identifying Traits that may contradict the provisional diagnosis)
Kok et al. (2017)(49)	Checklists	Checklist for diagnosing chest radiographs focusing on anatomy, potential pitfalls, and frequently missed diagnoses
Lambe et al. (2018)(62)	Guided reflection	Reflective reasoning tools (based on Mamede et al. (2008)(50))
Li et al. (2020)(45)	Guided reflection	Reflective reasoning tool (based on Mamede et al. (2008)(50))
Mamede et al. (2008)(50)	Guided reflection	Reflective reasoning tool, which asks participants to think of several differential diagnoses and provide information in the case that confirms or contradicts these diagnoses, or information that would have been expected but is not present. Finally, the diagnoses are ranked from most to least likely
Mamede et al. (2010a)(53)	Guided reflection	Reflective reasoning tool (as used in Mamede et al. (2008)(50))

Table 3. *Continued*

Study	**Category**	**Intervention**
Mamede et al. (2010b)(54)	Guided reflection	Reflective reasoning tool (as used in Mamede et al. (2008)(50))
Mamede et al. (2020)(15)	Instructions to test	Instructions to write down findings that favored the initial diagnosis (confirmatory), contradicted the initial diagnosis (contradictory), or both
Martinez-Franco et al. (2018)(67)	Computerized decision support system	DXplain, a computerized system that produces a ranked list of diagnoses based on a set of clinical findings
Myung et al. (2013)(46)	Guided reflection	Reflective reasoning tool (similar to Mamede et al. (2008)(50))
Nickerson et al. (2020)(63)	Checklists	Checklist of conditions that can cause syncope (for ECG diagnosis)
O'Sullivan et al. (2019)(29)	Checklists	Mnemonic checklist that asks clinicians to slow down and review their decisions (SLOW tool: Sure about that? Why?; Look at the data, What is Lacking, does it all Link together?; Opposite – What if the opposite is true?; Worst case scenario, What else could this be?)
Schmidt et al. (2014)(51)	Guided reflection	Reflective reasoning tool (similar to Mamede et al. (2008)(50))
Schmidt et al. (2017)(52)	Guided reflection	Reflective reasoning tool (similar to Mamede et al. (2008)(50))
Shimizu et al. (2013)(43)	Guided reflection	General debiasing checklist, which provided prompts to carefully consider the diagnosis, or a checklist which presented differential diagnoses based on the patient's complaint
Sibbald et al. (2013)(64)	Checklists	Checklist that breaks up the process of performing a cardiac physical exam in systematic subtasks
Staal et al. (2021)	Checklists	General debiasing checklist, which provided prompts to carefully consider the diagnosis, or a checklist which breaks ECG diagnosis into systematic subtasks
Talebian et al. (2014)(47)	Checklists	ECG checklist (DECKlist) for interpreting ECGs based on 12 items (such as rhythm and heart axis)
Thompson et al. (2017)(65)	Checklists	Mnemonic checklist listing separate anatomical regions to assist visual search of lateral chest radiographs

Risk of bias assessment

For 25 studies, risk of bias was low in all categories except in "Selection of reported results", because these studies had no preregistered analysis plans available to verify whether selection bias was present (Staal et al., 2021). Only one study was preregistered. Three studies were assessed as high risk of bias. First, O'Sullivan et al. (29) had an medium risk of bias due to a large drop-out rate during the study. Second, Shimizu et al. (43) was scored at high risk because of their quasi-random participant allocation. Third, Cairns et al. (48) was scored at high risk because of missing outcome data: participants were asked to diagnose at least one ECG, with a maximum of 10, but only 6 participants completed 2 or more ECGs. Interrater reliability for the total risk of bias score could not be calculated using Cohen's kappa, but overall agreement was high (Appendix B). See Appendix D for the overall risk of bias assessment score.

Main analysis

Data on diagnostic accuracy were available for 29 studies. This resulted in analyzable data for 2732 participants. A random-effect meta-analysis showed that the use of cognitive reasoning tools led to a small improvement in diagnostic accuracy (0.28, 95% CI: 0.14-0.43, $p < 0.001$). There was evidence of considerable heterogeneity in this estimate ($I^2 = 70\%$, $x^2(28) = 93.82$, $p < 0.001$), although this was not unexpected given the broad inclusion of cognitive reasoning tools. Retrospective exploration of influential studies indicated that Martinez-Franco et al. (67), Talebian et al. (47), and Thompson et al. (65) seemed to differ from the other studies: their participants had received training with the intervention directly before measuring diagnostic accuracy in the intervention group. Excluding these studies reduced heterogeneity ($I^2 = 38\%$, $x^2(25) = 40.22$, $p = 0.028$) sufficiently to interpret the meta-analysis. The effect estimate was slightly reduced (0.20, 95% CI: 0.10-0.29, $p < 0.001$), although the effect magnitude and direction remained unchanged (Fig. 2). A more elaborate exploration of the heterogeneity is presented in Appendix E.

Publication bias

A funnel plot was drawn to check for small study effects due to publication bias and to further explore heterogeneity (Appendix F). The funnel plot did not show significant asymmetry based on Egger's regression test ($t(27) = 1.84$, $p = 0.077$). This indicated there was no reason to suspect an influence of small study effects, nor did the funnel plot offer an explanation for the heterogeneity.

Subgroup analyses

Several subgroup analyses were performed to explore study heterogeneity and possible moderators of the effectiveness of clinical reasoning tools. The results for each subgroup are detailed in Appendix G. Only the type of diagnostic task seemed to moderate the effect of clinical reasoning tools: studies using real or standardized patients had a higher effect estimate than studies using visual tasks or written cases ($Q(2) = 22.10$, $p < 0.001$). However, only two studies had participants diagnose real or virtual patients (28, 46), reducing the reliability of the comparison. There was no difference in performance between visual or written diagnostic tasks (Q(1) = 0.63, *p* = 0.426). No significant differences were found for the other subgroup comparisons.

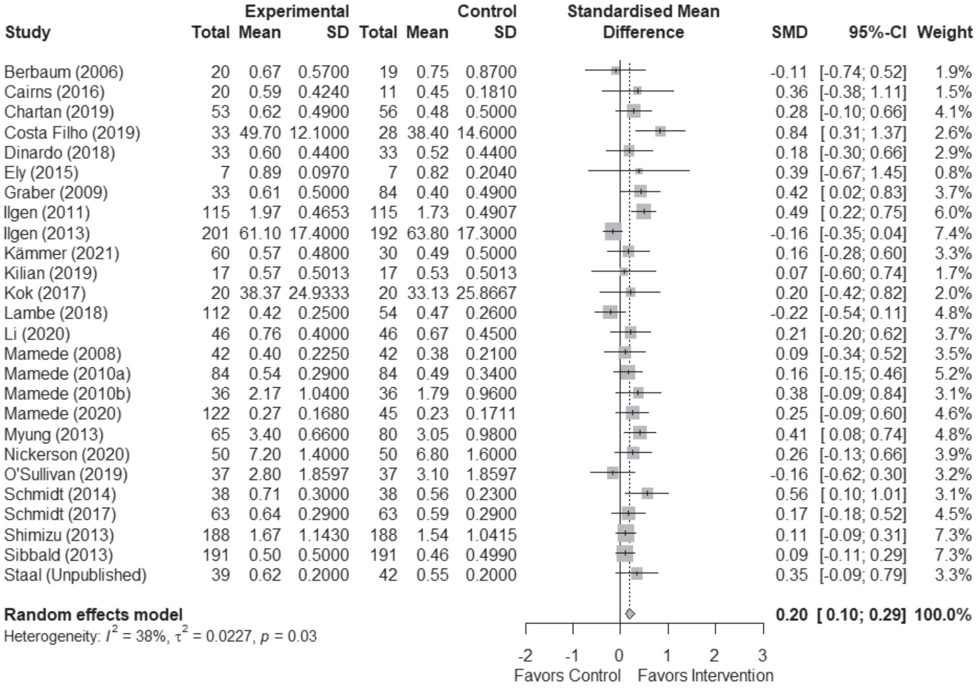


Figure 2. Forest plot of the overall pooled estimate.

Descriptively, participants of an intermediate level (i.e., residents and fellows) seemed to benefit more from using cognitive reasoning tools than novices (i.e., medical students). Experts seemed to benefit somewhat more than novices, but less than intermediates. Furthermore, content interventions seemed more effective than process interventions. Finally, studies where errors were induced and then remedied with the tool were more successful than studies that simply evaluated their tool, although it should be noted that only 4 studies induced and then remedied errors.

GRADE assessment

Finally, overall evidence was qualified for the meta-analysis excluding studies with extensive training (47, 65) (Table 4). The GRADE assessment indicated moderate quality of evidence, which shows that cognitive reasoning tools may benefit diagnostic performance as opposed to diagnosis without such a tool. The level of evidence was downgraded because of the moderate risk of bias on the selection of reported results, since pre-specified analysis plans were available for only one study (Staal et al., 2021).

Table 4. *GRADE certainty of evidence assessment.*

Certainty assessment						
No. of studies (participants)	**Risk of bias***	**Inconsistency**	**Indirectness**	**Imprecision**	**Publication bias**	**Overall quality of evidence**
26** (2539)	Serious	Not serious	Not serious	Not serious	Not detected	Moderate

*Risk of bias was rated as serious because most studies did not include a preregistered analysis plan. Despite 23 studies scoring "not serious" on all other dimensions of the risk of bias assessment, this resulted in a moderate risk assessment.

**Meta-analysis without Martinez-Franco (67), Talebian et al. (47), and Thompson et al. (65).

Discussion

This systematic review and meta-analysis of 29 studies involving 2732 medical students and physicians showed that workplace-oriented cognitive reasoning tools modestly improved diagnostic accuracy (0.28, 95% CI: 0.14-0.43, $p < 0.001$). This estimate exhibited substantial heterogeneity ($I^2 = 70\%$), which was largely attributable to three studies that offered training with their tool before measuring performance.(47, 65, 67) Removing these studies resulted in a lower, but more precise effect size (0.20, 95% CI: 0.10-0.29, $p < 0.001$) and reduced heterogeneity ($I^2 = 38\%$). Further subgroup analyses indicated that participant expertise, intervention characteristics (type of intervention, moment of intervention, and intervention items), and design characteristics (study design, case difficulty, same cases used with and without intervention, and study intention) could not explain the remaining between-study heterogeneity (Table 1). Only type of diagnostic task influenced tool effectiveness: the diagnosis of real or simulated patients seemed more effective (0.41, 95% CI: 0.33-0.49) than for written (0.16, 95% CI: 0.05-0.28) or visual cases (0.16, 95% CI: 0.05-0.28). However, because only 2 studies included patient encounters this result should be interpreted cautiously and verified in future research.

The modest improvement in diagnostic accuracy when using cognitive reasoning tools is largely in line with existing narrative and systematic reviews. Many of these reviews examined a broad range of interventions and outcomes, among which several interventions that were defined as cognitive reasoning tools in the current review. Recommended interventions primarily included reflection strategies (2, 3, 7, 12, 13, 68), clinical decision support systems (12, 19, 20, 69), cognitive forcing strategies (7, 12), and checklists (12, 20, 68). However, these recommendations were given with a cautionary note as evidence was often mixed and study designs were too divergent to draw strong conclusions.(12, 15, 68) A more direct comparison can be made with Graber et al. (3) and Lambe et al. (7), who specifically examined cognitive interventions. They concluded the interventions seemed promising but also cautioned that empirical evidence was scarce and preliminary. Lastly, the current estimate is in line with the meta-analysis by Prakash et al. (2), who reported a modest improvement of diagnostic decision making when using reflection strategies (0.38, 95% CI: 0.23-0.52, I^2 = 31%). The discrepancy in effect size with our estimate might be explained by differences in the included studies. Prakash et al. only quantified the effect of reflection strategies and did not consider other tools, whereas we included a range of tools. Additionally, Prakash et al. included both education-oriented studies (i.e., studies that tested interventions with the aim to teach someone how to solve cases in the future) and workplace-oriented studies (i.e., studies that tested interventions with the aim to measure performance when the tool is used for diagnosis). We quantified the effect of workplace-oriented studies alone, so Prakash et al.'s larger effect size could reflect differences in how effective cognitive reasoning tools are for teaching versus practical use. Taken together, cognitive reasoning tools are often recommended in the literature as promising interventions and this is corroborated by the improvement in accuracy we found. Caution should, however, be taken when interpreting this improvement due to the limited underlying evidence base.

The factors determining the effectiveness of cognitive reasoning tools remain unclear. Although several individual studies suggested that cognitive reasoning tools are more effective in specific subgroups (15, 18, 38, 43, 47), the current review found little indication of this. Of note might be the subset of three studies we excluded due to their contribution to the heterogeneity.(47, 65, 67) These studies were methodologically different because participants trained with the diagnostic task and intervention before performance was measured, which seemed to result in better performance than the other included studies. When considering all subset analyses, it would be premature to take our findings as evidence that cognitive reasoning tools are equally effective under most circumstances. This is due to the many different factors that might theoretically impact tool effectiveness and the combinations of these factors across studies. For example, several studies showed that process-focused interventions (i.e., aimed at preventing flaws in reasoning processes) were often less effective than content-focused interventions (i.e., aimed at providing or triggering relevant knowledge).(18) However, this distinction was difficult to make in the current review, as most interventions included both process and content elements to a certain extent. It was furthermore difficult to account for interactions between process or content interventions and other factors: for example, content interventions might be more beneficial for one subgroup whereas process interventions might be more useful for another subgroup. There are many potential influences on tool effectiveness and not enough studies with the same combination of factors. The current evidence base is simply not extensive enough to reliably assess such interactions and as a result we were unable to isolate the effect of individual factors or determine under which circumstances the tools are most effective.

In summary, cognitive reasoning tools modestly improved diagnostic accuracy. This effect should, however, be considered within the context of clinical practice. Diagnostic errors occur in about 10% of diagnoses, meaning the majority of diagnoses is correct.(1) The small improvement in overall diagnostic accuracy would therefore translate to a larger and clinically important improvement in the small subset of diagnostic errors, indicating that cognitive reasoning tools are a promising type of intervention. Whether this effect can be maximized to increase its potential use in practice will depend on our understanding of the factors that influence tool effectiveness.

Future research should focus on performing more large scale studies, as the small sample sizes contribute to mixed conclusions in the literature. Additional studies should be performed that examine factors that might influence tool effectiveness in order to determine the effects in different subgroups. Indications for potentially interesting factors may be taken from descriptive differences in our subgroup comparisons (Appendix G), which suggest diagnostic task and intervention type (content or process-focused intervention) as factors of interest. Furthermore, the excluded subset of studies(47, 65, 67) seemed to indicate the effect of the interventions was larger when participants were first given time to practice. This effect could translate well to medical education and especially cognitive reasoning tools that offer structured guidance (such as deliberate reflection(50) or checklists(64)) might provide benefits to learners. Finally, this effect could give an indication of what the effect of cognitive reasoning tools in practice could be: after all, clinicians will first be trained to use any tool before it will be used on real diagnoses. Future research should investigate the implementation of cognitive reasoning tools in practice to determine whether the improvement of accuracy can be replicated.

Limitations

Our review has three important limitations based on the studies included in the review and the review process. The first limitation is the high heterogeneity in the initial study sample which likely reflected the methodological and statistical differences between the interventions included based on our broad inclusion criteria. We explored this heterogeneity by examining the influence each individual study had on the estimate and excluded 3 studies that allowed participants to train with the tool before using it.(47, 65, 67) This reduced heterogeneity sufficiently to allow interpretation of the meta-analysis. Because we expected some heterogeneity, we used a random effects meta-analysis model which extra variance in underlying population distributions into account. As a result, our pooled estimate is an accurate estimate of the effectiveness of cognitive reasoning tools based on the available literature. Additionally, the broad inclusion criteria we applied are also a strength of the review: it allowed us to give a generalizable overview of the effectiveness of similar tools in different settings.

A second limitation is that only studies measuring diagnostic accuracy or diagnostic errors in percentages could be compared in this meta-analysis. Several studies measured diagnostic performance in other ways that were not comparable to the predominant measure of accuracy in the literature, such as the number of errors made (37-39, 70), whether the correct diagnosis was included in the differential (71, 72), or whether a new diagnostic plan was made for a patient based on the leading diagnosis (73). There were too few studies with these measures to perform an additional meta-analysis. However, given that these studies mostly show small, positive improvements, we would expect a summary of these diagnostic performance measures to be in line with the current estimate.

The third limitation concerns the available literature: studies that tested their intervention in practice are lacking, which is a result of the trade-off between performing well-designed and methodologically strong experimental studies and evaluating a tool in a less controlled, but more relevant environment. The current estimate of workplace-oriented tools is generalizable to different diagnostic tasks and specialisms in artificial settings, but the effectiveness of cognitive reasoning tools in practice remains unclear. Although there have been calls to reconfirm current findings in practice for the last decade (3, 7, 12, 69), for this review only two studies could be identified that were performed outside of an artificial setting (28, 47). Additionally, the long-term effects of cognitive reasoning tools are also unknown, as the included studies use single session designs. Future research should replicate the findings of existing studies and measure tool effectiveness in practice.

Conclusion

In conclusion, cognitive reasoning tools led to a small but clinically important improvement in diagnostic accuracy. Going forward, more studies should aim to identify the factors that influence tool effectiveness and under which conditions these tools are the most beneficial.

Cognitive reasoning tools could be routinely implemented in practice to improve diagnosis. However, a larger evidence base, consisting of more large-scale studies and evaluations of cognitive reasoning tools in practice, is needed to guide the implementation of cognitive reasoning tools in such a way that their effectiveness is optimized.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. Medical teacher. 2019;41(5):517-24.
3. Graber ML, Kissam S, Payne VL, Meyer AND, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ quality & safety. 2012;21(7):535-57.
4. Zwaan L, de Bruijne M, Wagner C, Thijs A, Smits M, van der Wal G, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Archives of internal medicine. 2010;170(12):1015-21.
5. Clinician Checklists [Internet]: Society to Improve Diagnosis in Medicine; 2020 [updated 2020 May 20]. Available from: https://www.improvediagnosis.org/clinician-checklists/.
6. Gawande A. The checklist manifesto: How to get things right. Journal of Nursing Regulation. 2011;1(4):64.
7. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ quality & safety. 2016;25(10):808-20.
8. Gupta A, Graber ML. Annals for hospitalists inpatient notes-just what the doctor ordered—checklists to improve diagnosis. Annals of internal medicine. 2019;170(8):HO2-HO3.
9. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.
10. Schiff GD, Hasan O, Kim S, Abrams R, Cosby K, Lambert BL, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. Archives of internal medicine. 2009;169(20):1881-7.
11. Newman-Toker DE, Schaffer AC, Yu-Moe CW, Nassery N, Tehrani ASS, Clemens GD, et al. Serious misdiagnosis-related harms in malpractice claims: the "Big Three"–vascular events, infections, and cancers. Diagnosis. 2019;6(3):227-40.
12. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the Basics of Cognitive Error in Emergency Medicine: Still No Easy Answers. Western Journal of Emergency Medicine. 2020;21(6):125.
13. Mamede S, Schmidt HG. Reflection in medical diagnosis: a literature review. Health Professions Education. 2017;3(1):15-25.
14. Astik GJ, Olson APJ. Learning from Missed Opportunities Through Reflective Practice. Critical Care Clinics. 2022;38(1):103-12.
15. Mamede S, Hautz WE, Berendonk C, Hautz SC, Sauter TC, Rotgans J, et al. Think twice: effects on diagnostic accuracy of returning to the case to reflect upon the initial diagnosis. Academic medicine. 2020;95(8):1223-9.
16. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. bmj. 2020;370.
17. Winters BD, Aswani MS, Pronovost PJ. Commentary: reducing diagnostic errors: another role for checklists? Academic Medicine. 2011;86(3):279-81.
18. Zwaan L, Staal J. Evidence on Use of Clinical Reasoning Checklists for Diagnostic Error Reduction. AHRQ Papers on Diagnostic Safety Topics [Internet]. 2020; (3).

19. McDonald KM, Matesic B, Contopoulos-Ioannidis DG, Lonhart J, Schmidt E, Pineda N, et al. Patient safety strategies targeted at diagnostic errors: a systematic review. Annals of internal medicine. 2013;158(5_Part_2):381-9.

20. Dave N, Bui S, Morgan C, Hickey S, Paul CL. Interventions targeted at reducing diagnostic error: systematic review. BMJ quality & safety. 2021.

21. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 2: impediments to and strategies for change. BMJ Qual Saf. 2013;22 Suppl 2(Suppl 2):ii65-ii72.

22. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009;62(10):e1-e34.

23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.

24. American Psychiatric Association AP, American Psychiatric A. Diagnostic and statistical manual of mental disorders: DSM-5: Washington, DC: American psychiatric association; 2013.

25. (EPOC) CEPaOoC. EPOC Resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services 2013 [updated Aug 24 2017]. Available from: https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/screening_data_extraction_and_management.pdf.

26. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration�s tool for assessing risk of bias in randomised trials. Bmj. 2011;343.

27. Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC health services research. 2004;4(1):1-7.

28. Ely JW, Graber MA. Checklists to prevent diagnostic errors: a pilot randomized controlled trial. Diagnosis. 2015;2(3):163-9.

29. O'Sullivan ED, Schofield SJ. A cognitive forcing tool to mitigate cognitive bias–a randomised control trial. BMC medical education. 2019;19(1):1-8.

30. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC medical research methodology. 2015;15(1):1-7.

31. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019.

32. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997;315(7109):629-34.

33. Viechtbauer W, Viechtbauer MW. Package metafor. The Comprehensive R Archive Network. Package 'metafor'. 2017.

34. Team R. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020. 2020.

35. Walter F, Prevost T, Vasconcelos J, Hall P, Burrows N, Morris H, et al. THE DIAGNOSTIC ACCURACY OF THE 7-POINT CHECKLIST TO ASSESS PIGMENTED SKIN LESIONS IN PRIMARY CARE: 734. Asia-pacific Journal of Clinical Oncology. 2012;8.

36. Letourneau KM, Horne D, Soni RN, McDonald KR, Karlicki FC, Fransoo RR. Advancing prenatal detection of congenital heart disease: a novel screening protocol improves early diagnosis of complex congenital heart disease. Journal of Ultrasound in Medicine. 2018;37(5):1073-9.

37. Sibbald M, de Bruin ABH, van Merrienboer JJG. Checklists improve experts' diagnostic decisions. Medical education. 2013;47(3):301-8.

38. Sibbald M, De Bruin ABH, van Merrienboer JJG. Finding and fixing mistakes: do checklists work for clinicians with different levels of experience? Advances in Health Sciences Education. 2014;19(1):43-51.

39. Sibbald M, de Bruin ABH, Yu E, van Merrienboer JJG. Why verifying diagnostic decisions with a checklist can help: insights from eye tracking. Advances in Health Sciences Education. 2015;20(4):1053-60.

40. Segal MM, Athreya B, Son MBF, Tirosh I, Hausmann JS, Ang EYN, et al. Evidence-based decision support for pediatric rheumatology reduces diagnostic errors. Pediatric Rheumatology. 2016;14(1):67.

41. Billingsley S. Evaluating chest X-rays. Use mnemonics to develop a systematic approach. Advance for nurse practitioners. 2009;17(2):24-5.

42. Dryver E, Johannsson G, Mokhtari A, Larsson D, Khoshnood A, Ekelund U. Checklists and" crowdsourcing" for increased patient safety in the emergency department. Lakartidningen. 2014;111(11):493-4.

43. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. Medical teacher. 2013;35(6):e1218-e29.

44. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: a randomized experiment. Medical Education. 2021.

45. Li P, yan Cheng Z, lin Liu G. Availability Bias Causes Misdiagnoses by Physicians: Direct Evidence from a Randomized Controlled Trial. Internal Medicine. 2020;59(24):3141-6.

46. Myung SJ, Kang SH, Phyo SR, Shin JS, Park WB. Effect of enhanced analytic reasoning on diagnostic accuracy: a randomized controlled study. Medical teacher. 2013;35(3):248-50.

47. Talebian MT, Zamani MM, Toliat A, Ghasemzadeh R, Saeedi M, Momeni M, et al. Evaluation of emergency medicine residents competencies in electrocardiogram interpretation. Acta Medica Iranica. 2014:848-54.

48. Cairns AW, Bond RR, Finlay DD, Breen C, Guldenring D, Gaffney R, et al. A computer-human interaction model to improve the diagnostic accuracy and clinical decision-making during 12-lead electrocardiogram interpretation. Journal of biomedical informatics. 2016;64:93-107.

49. Kok EM, Abed A, Robben SGF. Does the use of a checklist help medical students in the detection of abnormalities on a chest radiograph? Journal of digital imaging. 2017;30(6):726-31.

50. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Medical education. 2008;42(5):468-75.

51. Schmidt HG, Mamede S, Van Den Berge K, Van Gog T, Van Saase JLCM, Rikers RMJP. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. Academic Medicine. 2014;89(2):285-91.

52. Schmidt HG, Van Gog T, Schuit SCE, Van den Berge K, Van Daele PLA, Bueving H, et al. Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. BMJ quality & safety. 2017;26(1):19-23.

53. Mamede S, Schmidt HG, Rikers RMJP, Custers EJFM, Splinter TAW, van Saase JLCM. Conscious thought beats deliberation without attention in diagnostic decision-making: at least when you are an expert. Psychological research. 2010;74(6):586-92.

54. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203.

55. Berbaum K, Franken Jr EA, Caldwell RT, Schartz KM. Can a checklist reduce SOS errors in chest radiography? Academic radiology. 2006;13(3):296-304.

56. Chartan C, Singh H, Krishnamurthy P, Sur M, Meyer A, Lutfi R, et al. Isolating red flags to enhance diagnosis (I-RED): An experimental vignette study. International Journal for Quality in Health Care. 2019;31(8):G97-G102.

57. DiNardo D, Tilstra S, McNeil M, Follansbee W, Zimmer S, Farris C, et al. Identification of facilitators and barriers to residents' use of a clinical reasoning tool. Diagnosis. 2018;5(1):21-8.

58. Graber ML, Tompkins D, Holland JJ. Resources medical students use to derive a differential diagnosis. Medical teacher. 2009;31(6):522-7.

59. Ilgen JS, Bowen JL, McIntyre LA, Banh KV, Barnes D, Coates WC, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. Academic Medicine. 2013;88(10):1545-51.

60. Ilgen JS, Bowen JL, Yarris LM, Fu R, Lowe RA, Eva K. Adjusting our lens: can developmental differences in diagnostic reasoning be harnessed to improve health professional and trainee assessment? Academic Emergency Medicine. 2011;18:S79-S86.

61. Kilian M, Sherbino J, Hicks C, Monteiro SD. Understanding diagnosis through ACTion: evaluation of a point-of-care checklist for junior emergency medical residents. Diagnosis. 2019;6(2):151-6.

62. Lambe KA, Hevey D, Kelly BD. Guided reflection interventions show no effect on diagnostic accuracy in medical students. Frontiers in psychology. 2018;9:2297.

63. Nickerson J, Taub ES, Shah K. A checklist manifesto: Can a checklist of common diagnoses improve accuracy in ECG interpretation? The American journal of emergency medicine. 2020;38(1):18-22.

64. Sibbald M, de Bruin ABH, Cavalcanti RB, van Merrienboer JJG. Do you have to re-examine to reconsider your diagnosis? Checklists and cardiac exam. BMJ quality & safety. 2013;22(4):333-8.

65. Thompson M, Johansen D, Stoner R, Jarstad A, Sorrells R, McCarroll ML, et al. Comparative effectiveness of a mnemonic-use approach vs. self-study to interpret a lateral chest X-ray. Advances in physiology education. 2017;41(4):518-21.

66. Costa Filho GB, Moura AS, Brandão PR, Schmidt HG, Mamede S. Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. Perspectives on medical education. 2019;8(4):230-6.

67. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, Hernandez-Torres I, Rivero-Lopez C, Spicer T, et al. Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. Diagnosis. 2018;5(2):71-6.

68. Griffith PB, Doherty C, Smeltzer SC, Mariani B. Education initiatives in cognitive debiasing to improve diagnostic accuracy in student providers: A scoping review. J Am Assoc Nurse Pract. 2020.

69. Abimanyi-Ochom J, Mudiyanselage SB, Catchpool M, Firipis M, Dona SWA, Watts JJ. Strategies to reduce diagnostic errors: a systematic review. BMC medical informatics and decision making. 2019;19(1):1-14.

70. Sibbald M, Sherbino J, Ilgen JS, Zwaan L, Blissett S, Monteiro S, et al. Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. Advances in Health Sciences Education. 2019;24(3):427-40.

71. Chew KS, Durning SJ, Van Merriënboer JJ. Teaching metacognition in clinical decision-making using a novel mnemonic checklist: an exploratory study. Singapore medical journal. 2016;57(12):694.

72. Chew KS, van Merrienboer JJG, Durning SJ. Investing in the use of a checklist during differential diagnoses consideration: what's the trade-off? BMC medical education. 2017;17(1):234.

73. Walayat S, Chaucer B, Kim M, Pflederer BR. Diagnostic Reboot: A Proposal to Improve Diagnostic Reasoning: ncbi.nlm.nih.gov; 2021.

Appendix A - Search strategies

Embase.com 1997

(‘checklist’/de OR ‘mnemonics’/de OR ‘metacognition’/de OR ‘diagnostic reasoning’/de OR ‘clinical reasoning’/de OR (checklist* OR check-list* OR mnemonic* OR metacognit* OR instruction* OR (cognit* NEAR/3 intervent*) OR ((analytic* OR tool* OR conscious) NEAR/3 (reason* OR thought)) OR (cognit* NEAR/3 forc* NEAR/3 tool*) OR ((reflectiv* OR diagnostic* OR clinical) NEAR/2 (reasoning*)) OR ((deliberate*) NEAR/3 (reflection))):ab,ti,kw) **AND** (‘diagnostic error’/de OR ‘diagnostic accuracy’/de OR ‘delayed diagnosis’/de OR ‘missed diagnosis’/de OR ‘differential diagnosis’/de OR ‘early diagnosis’/de OR ‘medical error’/’prevention’ OR (((diagnos* OR radiograph* OR radiolog*) NEAR/3 (error* OR miss* OR delay* OR wrong* OR accur* OR differential* OR earl*)) OR misdiagnos*):ab,ti,kw) NOT [review]/lim NOT ((animal/exp OR animal*:de OR nonhuman/de) NOT (‘human’/exp)) AND (‘physician’/exp OR ‘medical staff’/de OR ‘resident’/de OR ‘medical personnel’/de OR (physician* OR doctor* OR practitioner* OR ((medical OR hospital*) NEAR/3 (staff* OR student* OR expert*)) OR hospitalist* OR resident* OR specialist* OR surgeon* OR internist* OR radiologist* OR clinician*):ab,ti,kw) NOT (psychiat* OR schizophren* OR autism* OR autistic* OR depression* OR psychosis OR psychoses OR bipolar* OR laborator*):ti

Medline Ovid

(Checklist / OR Metacognition/ OR Clinical Reasoning/ OR (checklist* OR check-list* OR mnemonic* OR metacognit* OR instruction* OR (cognit* ADJ3 intervent*) OR ((analytic* OR tool* OR conscious) ADJ3 (reason* OR thought*)) OR (cognit* ADJ3 forc* ADJ3 tool*) OR ((reflectiv* OR diagnostic* OR clinical) ADJ2 (reasoning*)) OR ((deliberate*) ADJ3 (reflection))).ab,ti,kf.) AND (exp Diagnostic Errors/ OR Delayed Diagnosis/ OR Diagnosis, Differential/ OR Early Diagnosis/ OR ((diagnos* OR radiograph* OR radiolog*) ADJ3 (error* OR miss* OR delay* OR wrong* OR accur* OR differential* OR earl*)).ab,ti,kf.) AND (exp Physicians/ OR exp Medical Staff/ OR exp Health Personnel/ OR (physician* OR doctor* OR practitioner* OR ((medical OR hospital*) ADJ3 (staff* OR student* OR expert*)) OR hospitalist* OR resident* OR specialist* OR surgeon* OR internist* OR radiologist* OR clinician*).ab,ti,kf) NOT (psychiat* OR schizophren* OR autism* OR autistic* OR depression* OR psychosis OR psychoses OR bipolar* OR laborator*).ti. *NOT (review).pt.*

Web of Science

TS=(((checklist* OR check-list* OR mnemonic* OR metacognit* OR instruction* OR (cognit* NEAR/2 intervent*) OR ((analytic* OR tool* OR conscious) NEAR/2 (reason* OR thought*)) OR (cognit* NEAR/2 forc* NEAR/2 tool*)) OR ((reflectiv* OR diagnostic* OR clinical) NEAR/1 (reasoning*)) OR ((deliberate*) NEAR/2 (reflection))) AND (((diagnos* OR radiograph* OR radiolog*) NEAR/2 (error* OR miss* OR delay* OR wrong* OR accur* OR differential* OR earl*))) AND (physician* OR doctor* OR practitioner* OR ((medical OR hospital*) NEAR/2 (staff* OR student* OR expert*)) OR hospitalist* OR resident* OR specialist* OR surgeon* OR internist* OR radiologist* OR clinician*)) AND DT=(Article OR Letter OR Early Access) NOT TI=(psychiat* OR schizophren* OR autism* OR autistic* OR depression* OR psychosis OR psychoses OR bipolar* OR laborator*)

Cochrane Central

((checklist* OR check-list* OR mnemonic* OR metacognit* OR instruction* OR (cognit* NEAR/3 intervent*) OR ((analytic* OR tool* OR conscious) NEAR/3 (reason* OR thought)) OR (cognit* NEAR/3 forc* NEAR/3 tool*) OR ((reflectiv* OR diagnostic* OR clinical) NEAR/2 (reasoning*)) OR ((deliberate*) NEAR/3 (reflection))):ab,ti,kw) **AND** ((((diagnos* OR radiograph* OR radiolog*) NEAR/3 (error* OR miss* OR delay* OR wrong* OR accur* OR differential* OR earl*)) OR misdiagnos*):ab,ti,kw) AND ((physician* OR doctor* OR practitioner* OR ((medical OR hospital*) NEAR/3 (staff* OR student* OR expert*)) OR hospitalist* OR resident* OR specialist* OR surgeon* OR internist* OR radiologist* OR clinician*):ab,ti,kw) NOT (psychiat* OR schizophren* OR autism* OR autistic* OR depression* OR psychosis OR psychoses OR bipolar* OR laborator*):ti

Google Scholar

checklist|“check list” “diagnostic|diagnosis error|missed|delayed|wrong|accurate|differential|early” physician|doctor|practitioner|resident|specialist|surgeon|internist|radiologist|clinician| ”medical|hospital staff|student|expert”

checklist|’check list’ ‘diagnostic|diagnosis error|missed|delayed|wrong|accurate|differential|early’ physician|doctor|practitioner|resident|specialist|surgeon|internist|radiologist|clinician| ’medical|hospital staff|student|expert’

Appendix B – Interrater reliability

Table 1. *Interrater reliability measured as overall agreement and Cohen's kappa*.

Review phase	Reviewers	Agreement	Cohen's kappa (95% CI)
Title-abstract screening	JH, JS	97%	0.65 (0.50 - 0.80)
Title-abstract screening	JS, LZ	97%	0.44 (0.22 - 0.66)
Title-abstract screening	JH, LZ	99%	0.66 (0.41 - 0.91)
Full-text screening	JH, JS	86%	0.65 (0.47 - 0.83)
Risk of bias assessment*	JH, JS	88%	-

*The estimate of kappa could not be calculated because one rater only scored in one category.

Appendix C – Study outcomes per included study

Table 1. *Study outcomes for included study in the main analysis, where intervention groups were aggregated if there was more than one intervention condition. The means and standard deviations are displayed as reported in each study. These outcomes represent the average diagnostic accuracy in both the comparator group and the intervention group over all cases. Additionally, the interpretation of the results given by the authors of the original study and the source from which the data were obtained are listed.*

Study	Comparator			Intervention			Effect estimate Hedges' g (95% CI)	Interpretation	Source
	N	Mean	SD	N	Mean	SD			
Berbaum et al. (2006)(1)	19	0.75 ROC^a	0.87	20	0.67 ROC^a	0.57	-0.11 (-0.74; 0.52)	"[…] use of a self-prompting checklist to counteract satisfaction of search is not warranted…"	Journal article; SD and effect size were computed from available information
Cairns et al. (2016)(2)	11	45%	18%	20	59%	42%	0.36 (-0.38; 1.11)	"[…] this approach improves diagnostic accuracy at the expense of time…"	Journal article; effect size was calculated from available information
Chartan et al. (2019)(3)	56	48%	50%	53	62%	49%	0.28 (-0.10; 0.66)	"A cognitive strategy of prompting red flag isolation prior to differential diagnosis did not improve diagnostic accuracy of 'do-not-miss diagnoses'."	Journal article; effect size was calculated from available information
Costa Filho et al. (2019)(4)	28	38.4%	14.6%	33	49.7%	12.1%	0.84 (0.31-1.37)	"Deliberate reflection increased diagnostic accuracy in dermatology but did not affect confidence and calibration."	Journal article; effect size was calculated from available information.
Dinardo et al. (2018)(5)	33	52%	44%	33	60%	44%	0.18 (-0.30; 0.66)	"A clinical reasoning tool improved residents' diagnostic accuracy on written cases."	Journal article; SD obtained via correspondence and effect size was calculated from available information
Ely et al. (2015)(6)	7	82%	20%	7	89%	10%	0.39 (-0.67; 1.45)	"Checklists did not improve the diagnostic error rate in this study."	Journal article; effect size was calculated from available information

Table 1. *Continued*

Study	Comparator			Intervention			Effect estimate Hedges' g (95% CI)	Interpretation	Source
Graber et al. (2009)(7)	84	40%	49%	33	61%	50%	0.42 (0.02; 0.83)	"In summary, our study indicates that medical trainees use and misuse electronic decision support products in deriving a differential diagnosis for a challenging cases."	Journal article; effect size was calculated from available information
Ilgen et al. (2011)(8)	115	1.73[b]	0.49	115	1.97[b]	0.47	0.49 (0.22; 0.75)	"Overall, mean diagnostic accuracy and the extent to which the test consistently discriminated between subjects was higher when participants were given directed search instructions than when they were given first impression instructions."	Journal article; effect size was calculated from available information
Ilgen et al. (2013)(9)	192	63.8%	17.3%	201	61.1%	17.4%	-0.16 (-0.35; 0.04)	"Instructions to trust one's first impressions result in similar performance when compared with instructions to consider clinical information in a systematic fashion, but have greater utility when used for the purposes of assessment."	Journal article; effect size was calculated from available information.
Kämmer et al. (2021)(10)	30	0.49[c]	0.50	60	0.57[c]	0.48	0.16 (-0.28; 0.60)	"The use of DDXC but not of GDBC may enhance diagnostic accuracy, without influencing the diagnostic reasoning process, but only if it contains the correct diagnosis."	Journal article, effect size was calculated from available information
Kilian et al. (2019)(11)	17	53%	50%	17	57%	50%	0.07 (-0.60; 0.74)	"Decisions to revise diagnoses may be cued by the detection of contradictory evidence."	Journal article; effect size was calculated from available information
Kok et al. (2017)(12)	20	33.13%	25.87%	20	38.37%	24.93%	0.20 (-0.42; 0.82)	"In conclusion, a checklist can help medical students to detect abnormalities in chest radiographs."	Journal article; effect size was calculated from available information

Table 1. *Continued*

Study	Comparator			Intervention			Effect estimate Hedges' g (95% CI)	Interpretation	Source
Lambe et al. (2018)(13)	54	0.47 [c]	0.26	112	0.42 [c]	0.25	-0.22 (-0.54; 0.11)	"This study finds no evidence to support the use of the guided reflection method as a diagnostic aid for novice diagnosticians…"	Journal article; effect size was calculated from available information
Li et al. (2020)(14)	46	0.67 [c]	0.45	46	0.76 [c]	0.40	0.21 (-0.20; 0.62)	"Availability bias led to diagnostic errors. Misdiagnoses cannot always be repaired by solely adopting a reflective approach."	Journal article; effect size was calculated from available information
Mamede et al. (2008)(15)	42	0.38 [c]	0.21	42	0.40 [c]	0.23	0.09 (-0.34; 0.52)	"Reflective practice had a positive effect on diagnosis of complex, unusual cases. Non-analytical reasoning was shown to be as effective as reflective reasoning for diagnosing routine clinical cases."	Journal article; effect size was calculated from available information
Mamede et al. (2010a)(16)	84	0.49 [c]	0.34	84	0.54 [c]	0.29	0.16 (-0.15; 0.46)	"Experts benefit from consciously thinking about complex problems; for novices thinking does not help in those cases."	Journal article; raw data were obtained via correspondence and effect sizes were calculated from available information
Mamede et al. (2010b)(17)	36	0.23 [b]	0.17	36	2.17 [b]	1.04	0.38 (-0.09; 0.84)	"When faced with cases similar to previous ones and using nonanalytic reasoning, second-year residents made errors consistent with the availability bias. Subsequent application of diagnostic reflection tended to counter this bias; it improved diagnostic accuracy in both first- and second-year residents."	Journal article; effect size was calculated from available information
Mamede et al. (2020)(18)	45	0.23 [c]	0.17	122	0.27 [c]	0.17	0.25 (-0.09; 0.60)	"Physicians' diagnostic accuracy improved after reflecting upon initial diagnoses provided for difficult cases, independently of the evidence searched for while reflecting."	Journal article; effect size was calculated from available information

Table 1. *Continued*

Study	Comparator			Intervention			Effect estimate Hedges' g (95% CI)	Interpretation	Source
Martinez-Franco et al. (2018)(19)	44	74.1%	9.4%	43	82.4%	8.5%	0.92 (0.47; 1.36)	"Family Medicine residents have appropriate diagnostic accuracy that can improve with the use of DXplain. This could help decrease diagnostic errors, improve patient safety and the quality of medical practice."	Journal article; effect size was calculated from available information
Myung et al. (2013)(20)	80	3.05[b]	0.98	65	3.40 [b]	0.66	0.41 (0.08; 0.74)	"Enhancement of analytic reasoning may improve diagnostic accuracy in novice doctors."	Journal article; effect size was calculated from available information
Nickerson et al. (2020)(21)	50	6.80[b]	1.60	50	7.20[b]	1.40	0.26 (-0.13; 0.66)	"Using a checklist with common syncope-related pathology when interpreting an ECG for a patient with clinical scenario of syncope may improve residents' ability to recognize some clinically important pathologies; however it could lead to increased interpretation and suspicion of pathology that is not present."	Journal article; effect size was calculated from available information
O'Sullivan et al. (2019)(22)	37	3.10[b]	1.86	37	2.80[b]	1.86	-0.16 (-0.62; 0.30)	"There is insufficient evidence to recommend this tool in clinical practice, however the qualitative data suggests such an approach has some merit and face validity to users."	Journal article; SD and effect size were calculated from available information
Schmidt et al. (2014)(23)	38	0.56[c]	0.23	38	0.71[c]	0.3	0.56 (0.10; 1.01)	"Availability bias may arise simply from exposure to media-provided information about a disease, causing diagnostic errors. The bias's effect can be substantial. It is apparently associated with nonanalytical reasoning and can be counteracted by reflection."	Journal article; effect size was calculated from available information

Table 1. *Continued*

Study	Comparator			Intervention			Effect estimate Hedges' g (95% CI)	Interpretation	Source
Schmidt et al. (2017)(24)	63	0.59[c]	0.29	63	0.64[c]	0.29	0.17 (-0.18; 0.52)	"Disruptive behaviours displayed by patients seem to induce doctors to make diagnostic errors."	Journal article; effect size was calculated from available information
Shimizu et al. (2013)(25)	188	1.55[d]	1.04	188	1.67[d]	1.14	0.11 (-0.09; 0.31)	"The use of DDXC, not GDBC, may improve the diagnostic performance in difficult cases, while intuitive process may still be better for simpler cases."	Journal article; raw data were obtained via correspondence; SD and effect size were calculated from available information
Sibbald et al. (2013)(26)	191	46%	50%	191	51%	50%	0.09 (-0.11; 0.29)	"Verifying diagnostic decisions with a checklist improved diagnostic accuracy. This benefit was only seen when more information could be collected."	Journal article; effect size was calculated from available data
Staal et al. (Unpublished)	42	55%	20%	39	62%	20%	0.35 (-0.09; 0.31)	"Checklist use improved confidence-accuracy calibration in normal as well as abnormal cases."	Raw data obtained from authors
Talebian et al. (2014)(27)	31	42%	49%	31	76%	43%	0.74 (0.50; 0.97)	"The increased accuracy of ECG interpretation and final diagnosis can be attributed to the utilization of a checklist by residents, especially in the first year and second year residents."	Journal article; raw data were obtained via correspondence; SD and effect size were calculated from available data
Thompson et al. (2017)(28)	14	77%	5%	14	92%	5%	2.96 (1.84; 4.08)	"This study demonstrates students can quickly and effectively learn to read a lateral chest film using this novel mnemonic."	Journal article; effect size was calculated from available data

[a] ROC: receiving operator characteristic.

[b] Average number of cases correct (range: 0-4 ([16 20]), 0-10 ([21 22]), 0-12 ([8])).

[c] Each case was scored as incorrect (0), partially correct (0.5), or correct (1). These scores were averaged over all cases.

[d] A score of 3 was given if the correct diagnosis was the first most likely diagnosis, a score of 2 if the correct diagnosis was the second most likely diagnosis, a score of 1 if the correct diagnosis was the third most likely diagnosis, and 0 if the diagnosis was not included. These scores were averaged over all cases.

References

1. Berbaum K, Franken Jr EA, Caldwell RT, et al. Can a checklist reduce SOS errors in chest radiography? *Academic radiology* 2006;13(3):296-304. doi: http://dx.doi.org/10.1016/j.acra.2005.11.032
2. Cairns AW, Bond RR, Finlay DD, et al. A computer-human interaction model to improve the diagnostic accuracy and clinical decision-making during 12-lead electrocardiogram interpretation. *Journal of biomedical informatics* 2016;64:93-107. doi: http://dx.doi.org/10.1016/j.jbi.2016.09.016
3. Chartan C, Singh H, Krishnamurthy P, et al. Isolating red flags to enhance diagnosis (I-RED): An experimental vignette study. *International Journal for Quality in Health Care* 2019;31(8):G97-G102. doi: http://dx.doi.org/10.1093/intqhc/mzz082
4. Costa Filho GB, Moura AS, Brandão PR, et al. Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. *Perspectives on medical education* 2019;8(4):230-36. doi: http://dx.doi.org/10.1007/s40037-019-0522-5
5. DiNardo D, Tilstra S, McNeil M, et al. Identification of facilitators and barriers to residents' use of a clinical reasoning tool. *Diagnosis* 2018;5(1):21-28. doi: http://dx.doi.org/10.1515/dx-2017-0037
6. Ely JW, Graber MA. Checklists to prevent diagnostic errors: a pilot randomized controlled trial. *Diagnosis* 2015;2(3):163-69. doi: http://dx.doi.org/10.1515/dx-2015-0008
7. Graber ML, Tompkins D, Holland JJ. Resources medical students use to derive a differential diagnosis. *Medical teacher* 2009;31(6):522-27. doi: http://dx.doi.org/10.1080/01421590802167436
8. Ilgen JS, Bowen JL, Yarris LM, et al. Adjusting our lens: can developmental differences in diagnostic reasoning be harnessed to improve health professional and trainee assessment? *Academic Emergency Medicine* 2011;18:S79-S86. doi: http://dx.doi.org/10.1111/j.1553-2712.2011.01182.x
9. Ilgen JS, Bowen JL, McIntyre LA, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. *Academic Medicine* 2013;88(10):1545-51. doi: http://dx.doi.org/10.1097/ACM.0b013e3182a31c1e
10. Kämmer JE, Schauber SK, Hautz SC, et al. Differential diagnosis checklists reduce diagnostic error differentially: a randomized experiment. *Medical Education* 2021 doi: http://dx.doi.org/10.1111/medu.14596
11. Kilian M, Sherbino J, Hicks C, et al. Understanding diagnosis through ACTion: evaluation of a point-of-care checklist for junior emergency medical residents. *Diagnosis* 2019;6(2):151-56. doi: http://dx.doi.org/10.1515/dx-2018-0073
12. Kok EM, Abed A, Robben SGF. Does the use of a checklist help medical students in the detection of abnormalities on a chest radiograph? *Journal of digital imaging* 2017;30(6):726-31. doi: http://dx.doi.org/10.1007/s10278-017-9979-0
13. Lambe KA, Hevey D, Kelly BD. Guided reflection interventions show no effect on diagnostic accuracy in medical students. *Frontiers in psychology* 2018;9:2297. doi: http://dx.doi.org/10.3389/fpsyg.2018.02297
14. Li P, yan Cheng Z, lin Liu G. Availability Bias Causes Misdiagnoses by Physicians: Direct Evidence from a Randomized Controlled Trial. *Internal Medicine* 2020;59(24):3141-46. doi: http://dx.doi.org/10.2169/internalmedicine.4664-20
15. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Medical education* 2008;42(5):468-75. doi: http://dx.doi.org/10.1111/j.1365-2923.2008.03030.x

16. Mamede S, Schmidt HG, Rikers RMJP, et al. Conscious thought beats deliberation without attention in diagnostic decision-making: at least when you are an expert. *Psychological research* 2010;74(6):586-92. doi: http://dx.doi.org/10.1007/s00426-010-0281-8

17. 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(11):1198-203. doi: http://dx.doi.org/10.1001/jama.2010.1276

18. Mamede S, Hautz WE, Berendonk C, et al. Think twice: effects on diagnostic accuracy of returning to the case to reflect upon the initial diagnosis. *Academic medicine* 2020;95(8):1223-29. doi: http://dx.doi.org/10.1097/ACM.0000000000003153

19. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, et al. Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. *Diagnosis* 2018;5(2):71-76. doi: http://dx.doi.org/10.1515/dx-2017-0045

20. Myung SJ, Kang SH, Phyo SR, et al. Effect of enhanced analytic reasoning on diagnostic accuracy: a randomized controlled study. *Medical teacher* 2013;35(3):248-50. doi: http://dx.doi.org/10.3109/0142159X.2013.759643

21. Nickerson J, Taub ES, Shah K. A checklist manifesto: Can a checklist of common diagnoses improve accuracy in ECG interpretation? *The American journal of emergency medicine* 2020;38(1):18-22. doi: http://dx.doi.org/10.1016/j.ajem.2019.03.048

22. O'Sullivan ED, Schofield SJ. A cognitive forcing tool to mitigate cognitive bias–a randomised control trial. *BMC medical education* 2019;19(1):1-8. doi: http://dx.doi.org/10.1186/s12909-018-1444-3

23. Schmidt HG, Mamede S, Van Den Berge K, et al. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Academic Medicine* 2014;89(2):285-91. doi: http://dx.doi.org/10.1097/ACM.0000000000000107

24. Schmidt HG, Van Gog T, Schuit SCE, et al. Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ quality & safety* 2017;26(1):19-23. doi: http://dx.doi.org/10.1136/bmjqs-2015-004109

25. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. *Medical teacher* 2013;35(6):e1218-e29. doi: http://dx.doi.org/10.3109/0142159X.2012.742493

26. Sibbald M, de Bruin ABH, Cavalcanti RB, et al. Do you have to re-examine to reconsider your diagnosis? Checklists and cardiac exam. *BMJ quality & safety* 2013;22(4):333-38. doi: http://dx.doi.org/10.1136/bmjqs-2012-001537

27. Talebian MT, Zamani MM, Toliat A, et al. Evaluation of emergency medicine residents competencies in electrocardiogram interpretation. *Acta Medica Iranica* 2014:848-54.

28. Thompson M, Johansen D, Stoner R, et al. Comparative effectiveness of a mnemonic-use approach vs. self-study to interpret a lateral chest X-ray. *Advances in physiology education* 2017;41(4):518-21. doi: http://dx.doi.org/10.1152/advan.00034.2017

Appendix D – Overall risk of bias assessment

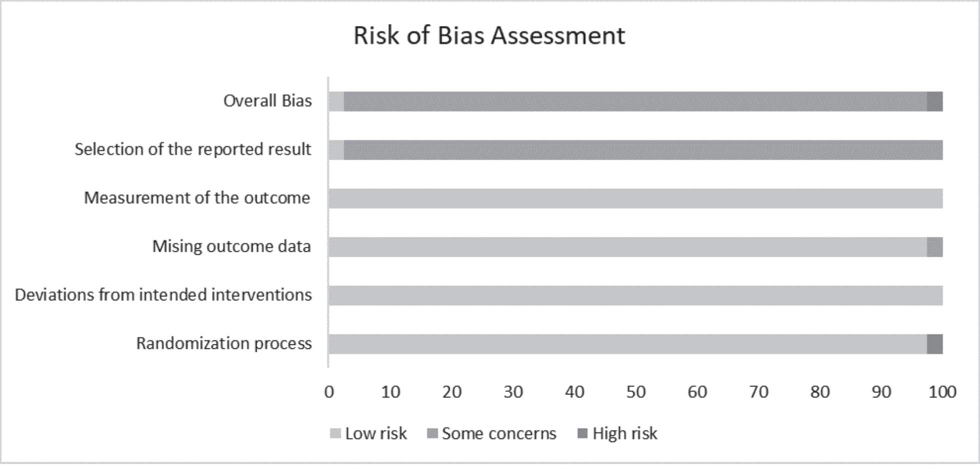


Figure 1. Risk of bias assessment.

Appendix E – Exploration of heterogeneity between the studies included for meta-analysis

Sensitivity analyses were performed to explore the robustness of the effect estimate and its heterogeneity. The meta-analysis was repeated excluding studies with a high risk of bias (1, 2) or a medium risk of bias assessment in any domain other than "Selection of reported results"(3). The effect estimate increased (0.31, 95% CI: 0.15-0.47, $p < 0.001$), but heterogeneity remained practically unchanged ($I^2 = 72\%$, $x^2(25) = 88.87$, $p < 0.001$). Next, influence analyses were performed by repeating the meta-analysis while leaving one study out of each iteration and examining the influence of each separate study. The effect estimate varied between 0.26-0.31 in all iterations. Omitting either Ilgen (4), Talebian et al. (5), or Thompson et al. (6) decreased heterogeneity the most, to around $I^2 = 64\text{-}66\%$. The influence analysis additionally identified Costa Filho et al. (7), Lambe et al. (8), and Martinez-Franco et al. (9) as influential studies. Removing all influential studies reduced the effect estimate (0.22, 95% CI = 0.15-0.29, $p < 0.001$) and removed heterogeneity ($I^2 = 0\%$, $x^2(22) = 16.64$, $p = 0.783$).

Martinez-Franco et al. (9), Talebian et al. (5), and Thompson et al. (6) seemed to differ from the other studies: their participants had received training with the intervention directly before measuring diagnostic accuracy in the intervention group. Exclusion of these studies is discussed in the Main analysis section of the manuscript. The heterogeneity in the remaining studies by Costa Filho et al. (7), Ilgen et al. (4), and Lambe et al. (8) can likely be partially attributed to statistical heterogeneity: for example, Ilgen et al. and Lambe et al. have negative effect estimates with a relatively large study weight. These characteristics set the studies apart from the average included study, but do not warrant exclusion.

References

1. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. Medical teacher. 2013;35(6):e1218-e29.
2. Cairns AW, Bond RR, Finlay DD, Breen C, Guldenring D, Gaffney R, et al. A computer-human interaction model to improve the diagnostic accuracy and clinical decision-making during 12-lead electrocardiogram interpretation. Journal of biomedical informatics. 2016;64:93-107.
3. Abimanyi-Ochom J, Mudiyanselage SB, Catchpool M, Firipis M, Dona SWA, Watts JJ. Strategies to reduce diagnostic errors: a systematic review. BMC medical informatics and decision making. 2019;19(1):1-14.
4. Ilgen JS, Bowen JL, McIntyre LA, Banh KV, Barnes D, Coates WC, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. Academic Medicine. 2013;88(10):1545-51.
5. Talebian MT, Zamani MM, Toliat A, Ghasemzadeh R, Saeedi M, Momeni M, et al. Evaluation of emergency medicine residents competencies in electrocardiogram interpretation. Acta Medica Iranica. 2014:848-54.
6. Thompson M, Johansen D, Stoner R, Jarstad A, Sorrells R, McCarroll ML, et al. Comparative effectiveness of a mnemonic-use approach vs. self-study to interpret a lateral chest X-ray. Advances in physiology education. 2017;41(4):518-21.
7. Costa Filho GB, Moura AS, Brandão PR, Schmidt HG, Mamede S. Effects of deliberate reflection on diagnostic accuracy, confidence and diagnostic calibration in dermatology. Perspectives on medical education. 2019;8(4):230-6.
8. DiNardo D, Tilstra S, McNeil M, Follansbee W, Zimmer S, Farris C, et al. Identification of facilitators and barriers to residents' use of a clinical reasoning tool. Diagnosis. 2018;5(1):21-8.
9. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, Hernandez-Torres I, Rivero-Lopez C, Spicer T, et al. Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. Diagnosis. 2018;5(2):71-6.

Appendix F – Funnel plot

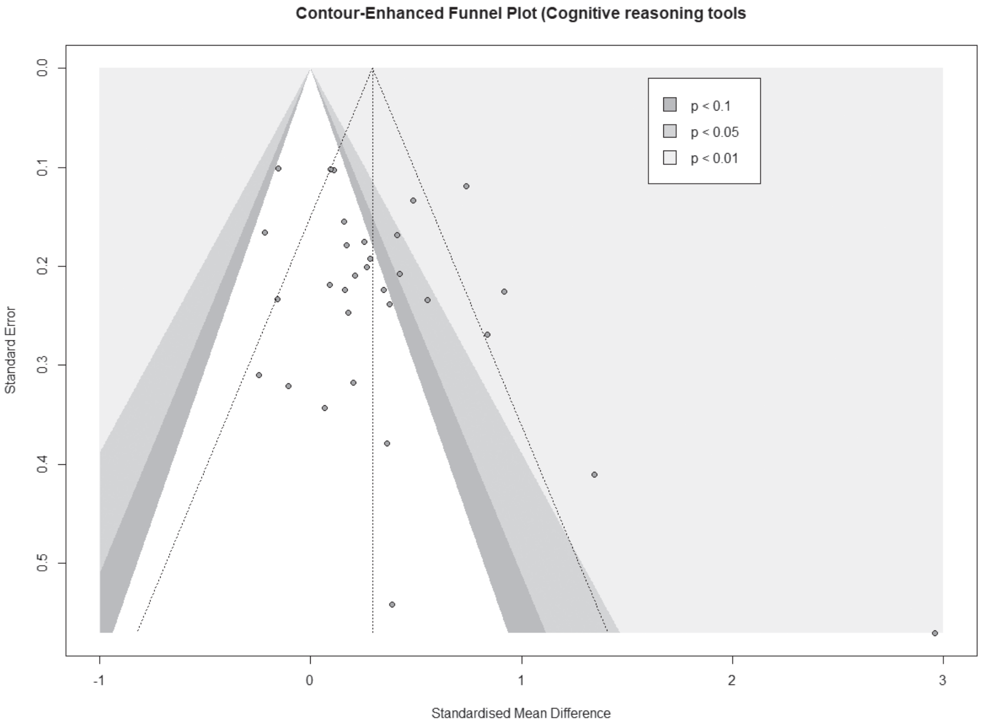


Figure 1. Contour-enhanced funnel plot. Each dot represents the effect size and the corresponding standard error for each study included in the meta-analysis. The dashed lines indicate the expected distribution of study sample sizes if there were no small study bias. The contoured area represents the actual sample size distribution and the significance level of this effect.

Appendix G – Subgroup analyses

Table 1. Overview of subgroup analysis results.

Analysis	Number of trials	Effect estimate (95% CI)	Heterogeneity (I^2)	Subgroup differences (Q)
Case difficulty: simple	6	0.23 (-0.02-0.48)	I^2 = 41%, $x^2(5)$ = 8.52, p = 0.130	Q(1) = 0.12, p = 0.730
Case difficulty: complex	9	0.27 (0.14-0.41)	I^2 = 0%, $x^2(8)$ = 6.91, p = 0.550	
Diagnostic task: patient (actors)	2	0.41 (0.33-0.49)	I^2 = 0%, $x^2(1)$ = 0, p = 0.970	Q(2) = 22.10, p < 0.001
Diagnostic task: visual diagnosis	7	0.25 (0.01-0.50)	I^2 = 28%, $x^2(6)$ = 8.35, p = 0.210	
Diagnostic task: written clinical vignettes	17	0.16 (0.05-0.28)	I^2 = 45%, $x^2(16)$ = 28.81, p = 0.030	
Expertise: novice (medical students)	10	0.22 (-0.01-0.44)	I^2 = 58%, $x^2(9)$ = 21.58, p = 0.010	Q(2) = 0.15, p = 0.929
Expertise: intermediate (residents, fellows)	13	0.23 (0.14-0.32)	I^2 = 0%, $x^2(12)$ = 6.69, p = 0.880	
Expertise: expert (faculty, specialists)	3	0.19 (-0.29-0.66)	I^2 = 0%, $x^2(2)$ = 1.12, p = 0.570	
Intention: evaluate tool performance	22	0.18 (0.07-0.28)	I^2 = 42%, $x^2(21)$ = 36.41, p = 0.020	Q(1) = 1.45, p = 0.229
Intention: fix induced errors with tool	4	0.30 (0.03-0.57)	I^2 = 0%, $x^2(3)$ = 2.00, p = 0.570	
Intervention (1): checklists	11	0.12 (0.01-0.21)	I^2 = 0%, $x^2(10)$ = 4.33, p = 0.930	Q(1) = 0.55, p = 0.460
Intervention (2): guided reflection	11	0.19 (0.02-0.36)	I^2 = 63%, $x^2(10)$ = 26.72, p < 0.001	
Intervention (2): content	24	0.21 (0.11-0.31)	I^2 = 40%, $x^2(23)$ = 38.16, p = 0.020	Q(1) = 1.94, p = 0.163
Intervention (2): process	3	0.12 (-0.07-0.35)	I^2 = 0%, $x^2(2)$ = 0.49, p = 0.780	
Intervention moment: during initial diagnosis	10	0.15 (-0.05-0.35)	I^2 = 61%, $x^2(9)$ = 22.80, p = 0.010	Q(1) = 0.29, p = 0.593
Intervention moment: verification after initial diagnosis	16	0.20 (0.11-0.30)	I^2 = 3%, $x^2(15)$ = 15.47, p = 0.420	
Intervention items: acknowledge	10	0.22 (0.09-0.36)	I^2 = 0%, $x^2(9)$ = 8.60, p = 0.480	Q(1) = 0.16, p = 0.691
Intervention items: report	16	0.19 (0.05-0.33)	I^2 = 50%, $x^2(15)$ = 30.21, p = 0.010	

Table 1. Continued

Analysis	Number of trials	Effect estimate (95% CI)	Heterogeneity (I^2)	Subgroup differences (Q)
Same cases used with and without intervention: yes	12	0.14 (0.03-0.24)	I^2 = 0%, $x^2(11)$ = 10.14, p = 0.520	Q(1) = 1.59, p = 0.207
Same cases used with and without intervention: yes	14	0.25 (0.09-0.40)	I^2 = 56%, $x^2(12)$ = 29.30, p < 0.010	

DISCUSSION AND SUMMARY

IV

CHAPTER 9

General discussion
Summary
Samenvatting
Dankwoord
Curriculum Vitae
List of publications
Portfolio

General discussion

Reducing diagnostic errors is an important step towards improving patient safety. A majority of people will encounter a diagnostic error in medicine during their lifetime and for many, this will result in serious harm.(1) Flaws in the cognitive processes underlying diagnostic reasoning are a major cause of diagnostic errors.(2) Given that these cognitive flaws are thought to be highly preventable (3), it is vital to capitalize on this room for improvement to reduce errors in clinicians' reasoning processes. The challenge set is to elucidate the processes that contribute to cognitive diagnostic errors and to find strategies that prevent these errors, with the goal of reducing the burden on patient safety. However, efforts to train and support clinicians are still limited by a scarcity of prospective and experimental studies that clarify how diagnostic errors occur and consequently, how they can be prevented.

This thesis aimed to contribute to the understanding and prevention of diagnostic errors. Chapter 2 to 5 reported experimental and observational studies where participants' diagnostic processes were compared between cases in which an error was made and cases where the correct diagnosis was reached. This comparison allowed insight in which processes were compromised when an error occurred and thus, could hint at where the error originated. Chapter 6 to 8 addressed experiments and a systematic review and meta-analysis that assessed the effectiveness of several error interventions in reducing diagnostic errors. In Chapter 9, the main findings of the preceding chapters are summarized and related to the literature. Implications as well as future directions for research and practice are discussed.

Summary of main findings

Two main questions were formulated to address the thesis aims. For the first aim, we investigated how clinicians' cognitive processes differed in cases where an error occurred versus where it did not. In the literature, it is primarily hypothesized that in error cases either a cognitive bias or a knowledge deficit has occurred.(4, 5) Neither biases nor knowledge deficits can be observed directly; instead we attempted to induce bias via the study design and examined participants' knowledge through their use of available case information, on the assumption that appropriate existing knowledge is prerequisite to selecting the right information to come to a diagnosis. Furthermore, the chapters in this thesis include a wide range of diagnostic tasks (i.e., written case vignettes of varying difficulty, radiographs, ECGs), medical specialties (i.e., internal medicine, emergency medicine, general practice, radiology, cardiology), and clinicians with widely varying ranges of expertise (i.e., from medical interns to experienced residents) to give more generalizable insights when taken together.

The first section discusses the main findings relating to the first thesis aim (Chapter 2 to 5). To study cognitive biases, we prospectively induced availability bias (Chapter 2) and confirmation bias (Chapter 3) in an experimental setting and examined participants' diagnostic processes to gain more insight in the relationship between biases and the diagnostic process. We then expanded on these studies by combining the induction of confirmation bias with the measurement of what case information participants used via eye-tracking methods (Chapter 4). Furthermore, we observed how proficient participants were in using case information to justify their diagnoses in an educational setting, without inducing bias (Chapter 5). These studies aimed to give insight in the causes of cognitive diagnostic errors. The second aim (Chapter 6 to 8) is discussed after summarizing the findings of the first aim. We evaluated how effective several promising cognitive error interventions were, with the aim to contribute to the existing evidence base. The investigated interventions in this thesis were: checklists for ECG diagnosis (Chapter 6), feedback on X-ray diagnosis (Chapter 7), and cognitive reasoning tools (i.e., any tool focused on supporting or improving clinicians' cognitive reasoning processes during diagnosis), which were examined in a more general systematic review and meta-analysis (Chapter 8). In Chapter 2 to 7, the diagnostic process was primarily operationalized as diagnostic accuracy, confidence in the diagnosis, time taken to diagnose, and, if appropriate, the type of information used to reach a diagnosis. In the systematic review and meta-analysis in Chapter 8 we only considered diagnostic accuracy.

Cognitive processes underlying diagnostic error

Chapter 2

The multi-center laboratory experiment in Chapter 2 investigated the assumption that diagnostic errors are primarily caused by faster diagnostic reasoning. As mentioned in the introduction, diagnostic reasoning is understood according to dual process theory (6). It is theorized that cognitive biases (i.e., predispositions to think in a way that leads to systematic failures in judgement (6) in reasoning occur because System 1 uses quick reasoning shortcuts, which do not consider all relevant information. To counteract these errors, then, it is often recommended to slow down diagnostic reasoning and deliberately engage System 2 reasoning. This view has been challenged by previous literature, such as studies by Norman et al. (7) and Sherbino et al. (8) which showed that correct diagnoses were reached just as fast, or even faster, than incorrect diagnoses. However, the interpretation of these results is somewhat limited because of their between-subjects designs: the time participants took to diagnose was compared between groups of different physicians. The findings could therefore also be explained by the hypothesis that better diagnosticians were simply also faster diagnosticians. We aimed to replicate these results in a within-subjects design. Internal medicine residents diagnosed written clinical case vignettes meant to induce availability bias (using Mamede et al.'s methodology (9)). Next, they diagnosed new cases with diagnoses that resembled, but were different from, the availability bias-induced diagnoses they encountered before.

The results showed that, on average, correct diagnoses were reached faster than incorrect diagnoses. This replicated and expanded upon the previous work by Norman et al. (7) and Sherbino et al. (8), by showing that the same effect could be found within individual clinicians. Further, residents were more confident in their correct diagnoses, although their overall confidence-accuracy calibration was poor. The latter finding speaks against the alternative explanation that these results could be due to differences in case difficulty: for example, if residents found a case easier they might spend less time, whereas they might take longer when in doubt. However, residents were poor judges of their own performance and it would be unlikely that they, on average, would consistently speed up or slow down for specific cases. Lastly, an exploratory analysis showed a trend for bias-induced errors to be faster than other types of errors but not faster than correct diagnoses, indicating there might be similar cognitive processes underlying both correct and flawed diagnostic reasoning.

These results imply that faster reasoning is not necessarily wrong and that slower reasoning is not necessarily right. Although this finding is not novel in itself, it provides stronger evidence for the assertion that faster reasoning is a valuable part of the diagnostic process and further implies that other causes for diagnostic errors rather than just speed of diagnosis should be considered. Additionally, simply slowing down reasoning will likely not be an effective error reduction strategy: it is not practically possible to slow down for every patient and the strategy to slow down only when necessary is hampered by residents' poor estimation of their own performance. Further research will be necessary to understand what factors cause cognitive diagnostic errors. In conclusion, fast reasoning seems to underlie both correct and erroneous diagnoses, and the pervasive notion that fast reasoning is only related to errors should be revisited.

Chapter 3

Chapter 3 presents a laboratory experiment that examined the effect of diagnostic suggestions on the diagnostic reasoning process. In the emergency department, clinical information (e.g., symptoms or test results) from the patients' referral letter is often used in diagnostic decisions. The referral letter can also provide a diagnostic hypothesis. Diagnostic suggestions in patient referral could potentially lead to incorrect interpretations of clinical information or cognitive biases, which could in turn result in diagnostic errors. As explained in the previous study, cognitive biases are seen as an important cause of errors.(6) We investigated the effect of a general practitioner's (GP) referral letter to the emergency department on the diagnostic process of medical interns to further elucidate the processes underlying such errors.

Medical interns diagnosed written clinical case vignettes with a GP referral, aimed at emergency medicine diagnoses, under three conditions. One third of the cases contained a correct diagnostic suggestion, one third contained an incorrect diagnostic suggestion, and one third did not contain any diagnostic suggestion. We measured interns' diagnostic accuracy, confidence, number of differential diagnoses, and time to diagnose for each case.

Interns who received either a correct or an incorrect diagnostic suggestion included fewer differential diagnoses than interns who did not receive any suggestion. However, interns seemed to be able to overcome the bias of the diagnostic suggestion when formulating their final diagnosis, as their diagnostic accuracy, confidence, and time to diagnose were not affected. This contradicted literature showing that diagnostic suggestions can bias clinicians in the direction of the suggestion.(10, 11) The relative inexperience of our participants might explain the difference with existing literature, as less experienced clinicians are thought to rely less on pattern recognition and more on analytical reasoning.(12) Such an approach might make interns less susceptible to the diagnostic suggestion, as deliberate reflection has been shown to reduce diagnostic errors due to bias.(9) This is in line with the positive association we found between diagnostic accuracy and confidence: students seemed to be at least partially aware of when they were correct. Interestingly, the effect of diagnostic suggestions seemed case specific, as the correct diagnostic suggestion could improve diagnostic accuracy in some cases, but reduced it in others. The effect of the diagnostic suggestion might be dependent on the specific case, prior knowledge, or the mental flexibility necessary to consider a suggestion.

Failure to consider the correct diagnosis is an important cause of diagnostic errors.(13) As such, the quality of a diagnostic suggestion will determine whether it could lead to an error, especially in practice. For example, narrowing the differential diagnosis list correctly could provide guidance and increase the efficiency of follow-up investigations – but only if the diagnostic suggestion was correct. In an educational setting, of which our study and participants are more representative, it could be valuable for educators to be aware of the potential influence of diagnostic suggestions on their students' differential diagnosis. Perhaps students could practice with cases that do and do not contain a suggestion, or with constructing a broad differential diagnosis. To consider implications in clinical practice, future research should evaluate whether this effect also occurs in more experienced clinicians.

Chapter 4

The laboratory eye-tracking experiment in Chapter 4 compared residents' diagnostic information processing between correct and incorrect diagnoses. This expanded on the previous studies (Chapter 2, Chapter 3) by not only examining outcomes of the diagnostic process (i.e., accuracy, confidence, time to diagnose) but also what information was used during diagnosis. The diagnostic process is naturally guided by the clinical information available in a case. However, clinicians can arrive at the wrong diagnosis even if all necessary information is available. Selectivity in information processing might influence the clinical information that is considered and consequently, could lead to diagnostic errors if the wrong information is considered. For example, if a clinician attempts to confirm a diagnostic hypothesis, they might focus primarily on information relevant to that hypothesis and overlook other information – which might point at another diagnosis. We aimed to investigate how diagnostic errors and cognitive biases related to selectivity in information processing.

Internal medicine and emergency medicine residents diagnosed written case vignettes with a suggested working diagnosis. Half of these suggestions was correct and the other half was a likely, but incorrect, alternative. Residents were asked to indicate whether or not they agreed with the suggestion and if not, what alternative diagnosis they would suggest. We used eye-tracking technology to measure residents' eye movements during diagnosis. Eye movements are a relatively objective measure of information processing as they occur without conscious awareness.(14) It is assumed that, the longer someone looks at certain information, the more they are processing that information. We created two categories of information in each case: information that was necessary to arrive at the correct diagnosis or information necessary for the incorrect alternative diagnosis. We then measured how long (fixation time) and how often residents (number of fixations) looked at either type of information and compared these information processing characteristics for correct and incorrect diagnoses, and cases with a correct or incorrect diagnostic suggestion. Additionally, we measured residents' confidence and time to diagnose.

Overall, residents' selectivity in information processing did not differ between error cases and correct cases, or between cases with an incorrect or a correct diagnostic suggestion. However, an interaction between diagnostic accuracy and the diagnostic suggestion was found. Regions of interest relevant for the correct diagnosis were fixated more often if residents received an incorrect suggestion but arrived at the correct final diagnosis regardless. This was in comparison with conditions where residents did not arrive at the correct diagnosis at all, independent of the diagnostic suggestion, or where they made the correct diagnosis after receiving a correct suggestion. The interaction between diagnostic accuracy and the diagnostic suggestion seemed similar for both the number of fixations and the relative fixation time but was only significant for the number of fixations. Residents' confidence and time to diagnose did not differ under any condition. These findings are partially in line with the hypothesis that selective information processing could lead to diagnostic errors.(15) The interaction shows that residents who focused more on relevant information were able to overcome the bias of the suggestion, which is in line with Mamede et al.'s (16) conclusion that clinicians with higher knowledge were less susceptible to bias. On the other hand, no such selectivity was observed in error cases specifically. One possible explanation might be that residents who overcame the bias in a certain clinical case might have been more "expert" on that specific diagnosis than their peers. Eye-tracking research in diagnostic reasoning has shown that experts spend more time focusing on relevant information than novices (17) and similarly, our participants might have shown a more "expert" search pattern when they were able to detect the suggestion was wrong and corrected it. Taken together, differences in selectivity in information processing might be an indication of changes in cognitive processes related to the diagnostic process, rather than a cause for errors themselves.

In summary, selectivity in information processing does not exclusively occur in error cases. Rather, appropriate selectivity occurred when correcting a diagnostic error. It is likely that selectivity plays a complex role in the diagnostic process, rather than being a direct cause of errors. However, this study does suggest that being able to select the relevant information from a case is important to arrive at the correct diagnosis and can assist in overcoming confirmation bias. Future research should explore the influence of many more factors, such as different types of information or types of error or biases, on clinicians' information processing.

Chapter 5

Chapter 5 described an observational study that aimed to further understand the use of clinical information in diagnostic reasoning. Diagnostic justification, or the skill to determine whether certain clinical information increases (pertinent positive information) or decreases (pertinent negative information) the probability of a certain diagnosis, is a crucial part of the diagnostic process. Previous studies have shown that clinicians' ability to accurately assign pertinent information differentiates experts from novices: novices had limited knowledge on pertinent information and underreported pertinent negative information.(18-20) Lacking skill in diagnostic justification might cause students to incorrectly assign pertinent information to diagnoses, which could lead to diagnostic errors. Whether pre-clerkship medical students show similar patterns has, to our knowledge, not been studied before.

We aimed to investigate what clinical information first and second year students used when diagnosing clinical cases and how this was associated with their diagnostic accuracy. We additionally examined whether practice with these aspects of the diagnostic process would increase students' performance.

First and second year medical students diagnosed written case vignettes in an online learning environment. This environment allowed students to select information in the patient history, physical exam, and investigations, and to assign this information as either increasing or decreasing the probability of their differential diagnoses. They also had to select a final most likely diagnosis at the end of the case. Diagnostic performance was measured by scoring students' final diagnostic accuracy, their created differential diagnosis, diagnostic justification, and ordered investigations.

The results showed that pre-clerkship students performed well on diagnostic accuracy and creating a differential diagnosis, average on investigations, and poor on diagnostic justification. Their scores were worst for the information in the physical exam compared to the history and investigations. Students consistently underreported all pertinent information, especially pertinent negative information. Interestingly, this pattern did not differ between correct and incorrect diagnoses, although this might have been distorted by the low amount of pertinent (especially negative) information assigned overall or students' high diagnostic accuracy. These findings are in line with studies investigating diagnostic justification in more experienced medical students: the deficiencies in diagnostic justification skill are apparent early on and are not remedied during the clerkship period. The good news, though, was that diagnostic justification skill did increase as students practiced with more cases, in line with previous literature (21) showing a 'dose-response' relationship between simulation attendance and diagnostic justification scores. Overall, students achieved a high final diagnostic accuracy despite their diagnostic justification scores being poor. This might be explained by the issue that students often receive cases depicting classic presentations of common diseases, mostly diseases they are concurrently learning about in the curriculum at that moment.(22) Under these circumstances, students can often correctly guess the diagnosis correctly, even without the proper underlying diagnostic skills. This can lead to errors further in their career, where such guesswork is no longer reliable. Of course, due to the observational design of the study, no such conclusions could be drawn from the current data.

The current study indicated that the ability to identify pertinent information, especially negative pertinent information, is lacking in pre-clerkship students. Properly developing students' diagnostic justification skills will likely be crucial to ensure good quality diagnostic reasoning going forward. Future research could examine students' diagnostic justification skills in more difficult cases: the cases in the current study were easy to suit the level of the students but using difficult cases could provide different insights in which diagnostic processes could be improved. It should further be determined whether practice leads to robust improvements in diagnostic justification and whether this is related to improvements in diagnostic accuracy.

Discussion of main findings: Cognitive processes underlying diagnostic error

Chapter 2 through 5 together provide insight in the origins of diagnostic errors. In summary, most diagnostic reasoning processes occurred both when reasoning was correct or when it failed to produce the right answer, and no process differences were identified that exclusively occurred in error cases (Chapter 2-5). We focused on time taken to diagnose (i.e., how long it takes the clinician to submit a final diagnosis) and confidence in the final diagnosis, as faster diagnostic reasoning (23) and overconfidence are often proposed as causes of cognitive diagnostic errors.(24) Neither faster time to diagnose (Chapter 2-4) nor overconfidence (Chapter 2 and 4) were associated with diagnostic errors. In fact, correct diagnoses were reached faster than erroneous diagnoses in general (Chapter 2). The current results seem to indicate that diagnostic errors do not originate from changes in clinicians' time to diagnose or confidence directly. However, reasoning faster or being overconfident might cause a clinician to miss an existent error (from another source) and prevent them from detecting or correcting it.

We also assessed differences in the use of clinical information during the diagnostic process. Diagnostic suggestions influenced the flexibility of medical interns' reasoning process, as they generated a less extensive differential diagnosis under influence of a suggestion. Differential diagnoses did not differ depending on the accuracy of the diagnostic suggestion or of interns' final diagnosis (Chapter 3), so this was again similar between correct diagnoses and error cases. However, further exploration of the results showed that the effects of the diagnostic suggestion were case-specific, meaning that the suggestion improved accuracy in some cases but reduced accuracy in others. This case-specificity could indicate that the content of the case, and by extension, perhaps interns' knowledge of that specific disease, were important to arriving at the correct diagnosis and might be lacking if an error is made. Further, the type of clinical information used during diagnosis did not differ between correct diagnoses and error cases. Medical interns were generally lacking in their diagnostic justification ability, specifically in recognizing and using pertinent information (Chapter 5). Knowledge on pertinent information has previously been indicated as an explanation of why experts are better diagnosticians than novices.(25) Pertinent information was underreported regardless of diagnostic accuracy. However, these results have to be interpreted with caution: we suspected that students were able to arrive at the correct diagnosis often, even if their information use was suboptimal, because they might be able to guess which diagnosis is relevant based on the current focus of their curriculum and the limited number of diagnoses that they have learned.(22) It would, therefore, be possible that the ceiling effect of accuracy obscured differences between information use in correct diagnoses and error cases.

Interestingly, we observed an interaction between bias induction and clinicians' diagnostic accuracy. Clinicians' diagnostic reasoning was only affected when a bias was induced, e.g., by presenting an incorrect diagnostic suggestion, and subsequently overcome to arrive at the correct diagnosis. In Chapter 4, residents who overcame the bias allocated a larger percentage of their time to processing clinical information relevant to the correct diagnosis compared to the residents who did not overcome the bias. Similarly, the medical interns in Chapter 3 generally overcame the bias of an incorrect diagnostic suggestion: their diagnostic accuracy was not reduced by the diagnostic suggestion. These interns also showed a higher confidence in their correct diagnoses than in their incorrect diagnoses. In Chapter 2 and 4, where the clinicians did not overcome the induced bias, their confidence did not distinguish between correct and incorrect diagnoses.

When taking these findings together, we might infer that the processes underlying diagnostic reasoning and the information used during diagnosis do not explain the origins of diagnostic errors per se – after all, they occur similarly in correct diagnoses and errors due to bias. However, the interaction observed when clinicians overcame a bias suggests a more nuanced view. Detecting a wrong diagnostic suggestion was associated with changes in the relative time spent on relevant information - though not on absolute time to diagnose - confidence, and type of clinical information used. We only observed these changes when a clinician arrived at the correct diagnosis in spite of the bias and not when a clinician "corrected" a bias suggestion that was already correct. In both scenarios, a conflict exists between the suggested diagnosis and the answers the clinician believed to be likely, but only in one scenario did we observe changes in clinician's behaviors. This could indicate that having the knowledge necessary to arrive at the correct diagnosis is crucial, as it will allow the clinician to detect the error in the diagnostic suggestion. Our case vignettes were designed to have two likely diagnoses but only one diagnosis was ultimately correct, based on several crucial pieces of pertinent clinical information which supported the correct diagnosis but not the incorrect diagnosis. The observed behavioral changes that followed the correction could indicate the conflict detection and subsequent attempt to mitigate the error. When an incorrect diagnosis was given in spite of a correct diagnostic suggestion, this did not result in changes in information processing, which might indicate that the pertinent information necessary to distinguish between the correct diagnosis and the suggestion was not recognized as such. We cannot know this for certain as we can only measure clinicians' diagnostic reasoning process indirectly; however, these findings could hint at the origin of diagnostic errors.

To summarize the prior discussion, we infer from our studies that the process factors and the clinical information used in diagnostic reasoning do not differ between correct diagnoses and error cases. Neither therefore seem to be a cause of diagnostic errors. Rather, when we observed changes in clinicians' diagnostic reasoning processes, this seemed to reflect the additional activity required to detect and mitigate a bias-induced error instead of the cause of an error. In these cases, clinicians spent relatively more time processing the information that was necessary to make the correct diagnoses (Chapter 4), hinting that the ability to recognize this pertinent information was vital. The next section will discuss the findings related to our second aim, regarding diagnostic error interventions, before synthesizing all findings in one framework to answer both aims.

Interventions to prevent diagnostic error

Chapter 6

The mixed design laboratory experiment in Chapter 6 compared the effect of checklist interventions aimed at reducing diagnostic errors between normal and abnormal electrocardiograms (ECGs). Previously, checklists have been successful in reducing hospital-acquired infections (26) and preventing errors during surgery.(27) Therefore, checklists are also proposed as interventions to reduce one of the most prevalent types of diagnostic errors: cognitive errors.(28-30) Such checklists can generally be divided into two types: process checklists (or debiasing checklists), and content checklists.(31, 32) Additionally, existing experiments are primarily conducted using complex and abnormal cases, to create a situation where the chance of diagnostic errors is high, so that possible benefits from the checklist can be observed.(31) This limits our understanding of checklist effectiveness for normal and simple abnormal cases, which limits insights of checklist research for settings such as general practice. In the Netherlands, general practitioners (GPs) often use checklists to interpret ECGs and to decide whether the patient should be referred to the hospital or not.(33) Approximately one third of the ECGs they encounter are normal, and another third is a simple rather than a complex abnormal ECG.(34) Because checklists could potentially lead to the unnecessary use of resources, or overtesting and overdiagnosis (31), we studied the impact of a debiasing checklist (13) and a content checklist (33) on the diagnostic process of GPs when interpreting normal, simple abnormal, and complex abnormal ECGs.

GP residents were recruited during an educational day focused on research. They were asked to diagnose ECGs in two sessions: one during the educational day, and one a week later. In the first session, they diagnosed ECGs with a short patient description as they normally would, without a checklist. In the second session, they received the same ECGs and diagnosed these using a checklist. Half of the residents received instructions to use the debiasing checklist; the other half received instructions for the content checklist. ECGs were randomized in both sessions. In both sessions, residents reported their most likely diagnosis, their confidence in this diagnosis, and how they would manage this patient based on the ECG (i.e., wait and see, start treatment, refer patient to cardiologist). Additionally, their confidence-accuracy calibration was calculated from the accuracy and confidence measurements.

Residents' performance did not differ for normal, simple abnormal, and complex abnormal ECGs when they were using a checklist versus when they were not. Performance also did not differ between the debiasing checklist and the content checklist. Previous literature on debiasing checklist generally does not show an improvement in performance (35, 36), as we observed, but content checklists are generally found to have some positive effect (37-40), which was not replicated. Overall, residents showed a small learning effect between sessions, likely explained because they saw the same ECGs twice, as time to diagnose was decreased in the second session. A trend for improved accuracy was observed but this was not significant. Interestingly, residents' confidence did not change, despite the trend in increased accuracy. This resulted in an overall improved confidence-accuracy calibration which indicated residents seemed to become less overconfident in the second session. After all, their confidence did not increase with their accuracy and as a result, their previous overconfidence was tempered.

In conclusion, checklist use did not differentially affect the diagnosis of normal and abnormal ECGs in GP residents. Checklists appear promising in reducing overconfidence without negatively affecting the diagnosis of normal ECGs, however, this effect should be replicated in more experienced GPs. In the long term, reduced overconfidence might improve residents' insights in their own performance and enable them to improve their diagnostic performance.

Chapter 7

The effectiveness of a feedback intervention in improving diagnosis was assessed in the laboratory experiment in Chapter 7. Oftentimes, intervention tools cannot be used for every patient the clinician encounters, or the clinician is not aware that they have encountered a case at high risk for diagnostic error.(41) In an ideal situation, clinicians would be aware when they need help but unfortunately, clinician's calibration (i.e., the alignment between their estimated performance and their actual performance) is poor.(42) Feedback is regarded as a promising intervention to improve calibration, as it is theorized to raise awareness of any discrepancies between estimated and actual performance.(43) There are two main types of feedback: performance feedback, where the recipient is informed whether their answer was correct or incorrect; and information feedback, where the recipient additionally receives an explanation of how the correct answer was reached.(44) Previous studies on performance feedback show that it can improve diagnostic accuracy, especially in easy cases (45), although evidence is mixed.(46) It has been suggested that information feedback might be necessary to improve diagnostic performance (44, 47) but despite this, evidence remains scarce. We studied the effect of performance feedback and information feedback on the diagnostic performance of medical interns when diagnosing chest X-rays.

Medical interns were asked to diagnose chest X-rays in two phases. In the first phase, the feedback phase, they were randomized to diagnose X-rays in one of three conditions: either they received performance feedback, information feedback, or no feedback on their diagnosis. They could select the X-ray diagnosis from a pre-specified list of five diagnoses (i.e., atelectasis, pleural effusion, pneumothorax, tumor, or no abnormality). Immediately after this phase, they diagnosed new X-rays without feedback in the test phase. Interns' performance was measured in terms of diagnostic accuracy, confidence, and time to diagnose. Confidence-accuracy calibration was calculated based on diagnostic accuracy and confidence. Performance was compared between the feedback conditions and between easy and difficult X-rays.

Both types of feedback improved interns' diagnostic accuracy and confidence compared to the control group, although the latter difference was no longer significant in post hoc tests. As a result, their overall calibration improved as well. Time to diagnose was not affected by feedback condition. Calibration improved the most in the information feedback condition. Easy cases were overall well-calibrated and performance improved with feedback, whereas calibration and performance remained poor for difficult cases regardless of feedback condition. These findings are in line with previous findings regarding the effectiveness of performance feedback (45, 48, 49) and the theoretical work on information feedback in medical diagnosis (50), although we did not replicate the finding that interns became underconfident after receiving feedback.(46) If anything, our interns remained more overconfident than underconfident.

Overall, both performance feedback and information feedback improved interns' diagnostic accuracy, confidence, and calibration, without taking extra time. It remains unclear, however, which mechanisms underlie the improvements in performance. One possibility is that interns' awareness of their own performance improved, but conversely it is also possible that interns' accuracy improved while their confidence remained stable, which would lead to an improved calibration without an actual change in estimated performance.

Chapter 8

The systematic review and meta-analysis in Chapter 8 aimed to examine the effectiveness of error interventions focused on cognitive diagnostic errors. Cognitive reasoning tools, i.e., tools meant to improve clinical reasoning and decision making skills by improving clinicians' intuitive and rational processing during diagnosis (51), are often recommended to reduce diagnostic errors.(28-30, 51) Experimental evidence of the effectiveness of such tools is, however, relatively scarce.(51, 52) Moreover, existing reviews have aggregated interventions aimed at improving performance in an educational setting with interventions aimed at improving performance in the workplace (29, 53-56), leaving the effectiveness of the latter unknown. Therefore, we aimed to estimate how effective cognitive reasoning tools are in improving diagnostic accuracy, and whether any study or intervention characteristics could be identified that were associated with a higher effectiveness.

We conducted a systematic review and meta-analysis according to PRISMA guidelines.(57) We included experimental studies with a control group or baseline measurement, which investigated the effectiveness of cognitive reasoning tools on improving diagnostic accuracy in medical students and healthcare professionals. Subgroup comparisons were made based on participant expertise, intervention characteristics (type of intervention, moment of intervention, intervention items), or study characteristics (case difficulty, diagnostic task, whether the same cases were seen in the intervention and the control condition, and the intention of the study).

Three studies were removed from the meta-analysis because participants received extensive training with the tool before diagnostic accuracy was measured. The remaining studies showed a small improvement of diagnostic accuracy with cognitive reasoning tools. This was in line with reviews on similar interventions.(29, 51, 53) No significant subgroup differences were found.

Cognitive reasoning tools resulted in a small, but clinically relevant improvement in diagnostic accuracy. Given the high prevalence of diagnostic errors, even a small improvement can make a valuable difference. However, caution should be taken due to the relatively small underlying evidence base. The factors underlying the effectiveness of cognitive reasoning tools remain unclear: it was difficult to isolate the effects of specific study or intervention characteristics due to the many potential influences on tool effectiveness. Notably, though, we observed a larger improvement in diagnostic accuracy in the studies that were excluded because their participants received extensive training with the tool. This might transfer well to medical education, as practice with cognitive reasoning tools seems to be beneficial. Future research should elucidate under which circumstances cognitive reasoning tools are most effective and the positive effect of the tools should be replicated in practice.

Discussion of main findings: Interventions to prevent diagnostic error

Chapter 6 through 8 investigated the effectiveness of several diagnostic error interventions by testing interventions that are regarded as promising in the literature, and by aggregating currently available evidence via meta-analysis. The data from Chapter 6 was also included in the meta-analysis in Chapter 8. Taken together, these studies suggest that cognitive reasoning tools are effective in improving overall diagnostic accuracy (Chapter 7, 8). The checklist intervention in Chapter 6 was not effective, but when pooled with other studies in the meta-analysis (Chapter 8), a small positive effect still emerged. This was in line with previous reviews and meta-analyses.(29, 51, 53-56) We further attempted to identify factors associated with greater tool effectiveness. Previous literature shows indications that factors such as the type of intervention (i.e., focused on cognitive processes or on task content) or the difficulty of the case to be diagnosed (31) might mediate the effect on diagnostic accuracy. In Chapter 7, we indeed found that feedback only improved performance in easy cases. However, we found no differences in several methodological and participant factors in the meta-analysis, likely because extensive heterogeneity in the methods of the included studies ((e.g., in the used tools, settings, instructions, case complexity) made it difficult to reliably identify such factors. The exception were studies that allowed extensive practice with the intervention tool reported greater effect sizes (Chapter 8). This difference might be explained by the type of intervention: the meta-analysis focused on tools that were meant to be used in a workplace setting, whereas the feedback intervention was aimed at an educational setting. Perhaps differences in the use and application of tools between these settings produce differences in tool effectiveness.

In addition to accuracy, we also examined the impact of interventions on other aspects of the diagnostic process. First on the relation between confidence and diagnostic accuracy, termed confidence-accuracy calibration. Generally, clinicians were overconfident regardless of the use of tools. However, when tools were used, confidence remained relatively stable and did not increase as much as their accuracy, which resulted in an increase in overall confidence-accuracy calibration (Chapter 6, 7). Calibration especially improved when information feedback was provided, which gave an explanation of how the correct answer was reached and calibration improved mostly in easy cases but not in difficult cases (Chapter 7). However, overall calibration remained relatively poor and clinicians were generally not good at estimating their own performance. Our findings were in line with previous studies on feedback, though our participants did not become underconfident after receiving feedback.(45, 46, 48, 49) The lack of sensitivity of clinicians' confidence to their accuracy raises some questions about how these findings ought to be interpreted. On the one hand, we could hypothesize that the diagnostic error interventions improve clinicians' insight in their accuracy: the intervention leads to increases in diagnostic accuracy and because they were overconfident from the start, confidence would not increase along with accuracy and result in a better calibration. It could, on the other hand, also be the case that confidence is not affected by the interventions, and that only accuracy changes while confidence remains stable. The current studies cannot distinguish between these two scenarios. However, it would be worthwhile to consider what we are measuring when we ask clinicians for their confidence. We assume that we measure how certain they are that a diagnosis is correct, but seeing as medical diagnosis always carries a degree of uncertainty, it might also be possible that we are instead measuring a form of decision threshold, or how certain someone wants to be before they commit to an answer. In that case, a stable level of confidence would be expected. There might also be individual differences in how confidence is interpreted. Because confidence, and as an extension, calibration, are seen as possible markers for when a clinician would require support to make a diagnosis and when an intervention should be implemented, it is crucial that we understand what we are measuring.

Lastly, the use of interventions did not increase clinicians' time to diagnose, indicating that the interventions could be used without a significant cost in time spent on each case (Chapter 6, 7). This was partially in line with a study by Ely et al. (13) conducted in practice, where consulting a checklist added minimal extra time. However, considering how short patient consults in practice can be, it could be argued that even a minimal addition could be costly.

Surprisingly, the studies in this thesis showed no differences in effectiveness between process interventions, such as debiasing strategies, or content interventions, such as deliberate reflection (Chapter 6-8). Previous studies comparing similar interventions generally concluded that content interventions, focused on a specific task, were more effective than general reasoning or processing instructions.(31) Given that several of the individual trials showing a difference between process and content interventions were included, it would seem that the difference was not large enough to be detected in the pooled result. Perhaps scenarios might exist where either of the interventions is more successful due to specific circumstances, but as was discussed in Chapter 8, the current evidence base for error interventions is not extensive enough to draw conclusions about tool effectiveness in specific subgroups. Based on Chapter 8, we conclude that both types of interventions result in a small improvement in diagnostic accuracy: however, we do not preclude that future research might further nuance our findings.

In summary, cognitive reasoning tools as a category are successful in improving diagnostic reasoning, without detracting from other aspects of the diagnostic process. Although the overall effect is small, even such a small improvement can have relevant impact, given the high prevalence of cognitive diagnostic errors. We do, however, encourage future research to further examine factors that might be implicated in greater effectiveness and to more frequently expand the study of interventions beyond the laboratory, into practice. In the following section, the results of this thesis will be presented in light of a theoretical framework, the mindware framework by Stanovich.(58)

Theoretical framework

Our current findings fit well with the cognitive processing framework developed by Stanovich.(58) This framework posits that in addition to processing skills, available knowledge structures are vital to task performance. Stanovich applies the framework to literature concerning heuristics and biases in reasoning and dual process theory, which directly connects the framework to our current understanding of the diagnostic process. Generally, the framework states that task performance is dependent on the availability and integration of relevant knowledge (Figure 1). This knowledge is termed *mindware*, as taken from earlier work by Perkins.(59) The integration of knowledge, or how well-learned it is, is referred to as *mindware instantiation*. Any task involves mindware to some degree, so having the proper mindware available is a prerequisite to successfully completing a task. After all, errors can only be avoided when a conflict is detected between a solution to the task and the relevant mindware. After detecting an error, the incorrect solution needs to be overridden by the solution taken from the relevant mindware. If mindware for the task is not available, one cannot be expected to arrive at a correct solution. Existing literature on heuristics and biases in reasoning has given much insight in human reasoning processes (6, 60) but has focused mainly on process skills, while neglecting the interaction between these process skills and the available mindware for a task.(58) As a result, heuristics and biases are often solely seen as causes of errors, without considering the quality of the underlying mindware. This has led to hypotheses such as that faster reasoning must be flawed and slower reasoning must be correct.(61-63)

From the perspective of the mindware framework (58), both the automatic and fast System 1 and the conscious and slow System 2 can generate correct and incorrect responses. When posed with a problem, System 1 generates a normative response based on implicitly learned rules and associations, such as heuristics, and a response based on well-instantiated mindware that can be accessed immediately. If the relevant mindware is extremely well-instantiated, it might even become the normative response. The availability of relevant mindware to System 1 will determine whether the quick response is correct or not. System 2 also has access to less well-instantiated mindware, and will engage in a slower, more time consuming and effortful process to generate a response. This parallel activation of System 1 (i.e., generating multiple answers, such as the normative response and the response based on mindware) and the serial activation of first System 1 (as its activation is automatic) and then System 2, is in line with the hybrid dual process model proposed by de Neys.(64) Based on the quality and relevance of the available mindware, either system can arrive at the correct solution. Consequently, a faster response would not necessarily be wrong and a slower response would not necessarily be correct, which is in line with the results of this thesis.

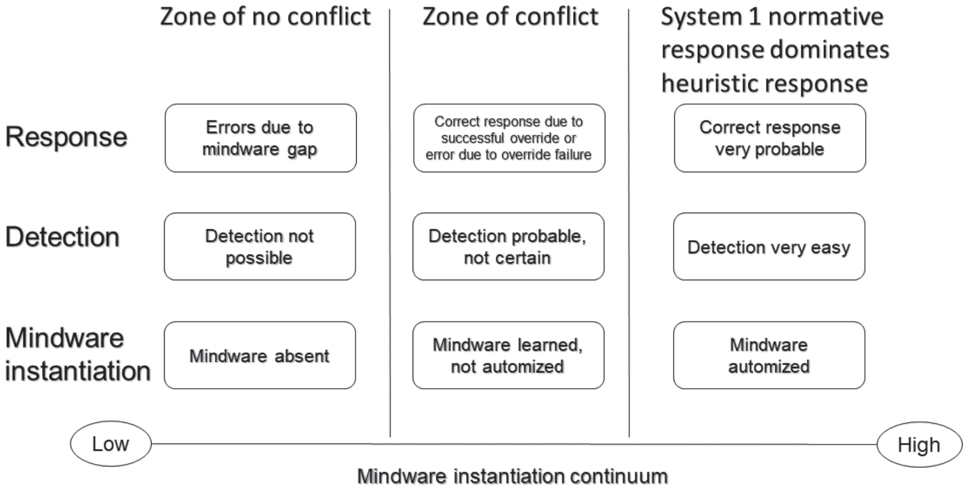

Zone of no conflict
Zone of conflict
System 1 normative response dominates heuristic response
Response
Errors due to mindware gap
Correct response due to successful override or error due to override failure
Correct response very probable
Detection
Detection not possible
Detection probable, not certain
Detection very easy
Mindware instantiation
Mindware absent
Mindware learned, not automized
Mindware automized
Low
High
Mindware instantiation continuum

Figure 1. Schematic representation of the states of information processing on the mindware continuum as proposed in the mindware framework. Adapted from “Miserliness in human cognition: The interaction of detection, override and mindware.” by Stanovich KE (2018). *Thinking & Reasoning, 24*(4), 423–444. Copyright (2018) by Taylor & Francis.(58)

Conflict detection and override

The difference between a correct and an incorrect response thus lies in the available mindware (Figure 1). Only if the appropriate mindware can be accessed, can an error be detected and then overridden. In the scenario where someone is checking an existing solution, a conflict can be detected between the existing solution and the generated solution(s), and again either can be overridden. Whether an override occurs, however, depends on the instantiation of the relevant mindware. Stanovich's (58) framework proposes that a conflict cannot be detected under two circumstances: either someone is so expert and has such well-instantiated mindware that the correct response has become the normative response, hence leaving no room for errors to occur; or someone is not equipped to handle a task at all and has no relevant mindware, meaning an error will almost always occur. In situations where the mindware is learned but not automated, it has to be retrieved consciously, which takes more time and effort. This process is referred to as a sustained override. If the mindware is not instantiated sufficiently, no response can be generated that is strong enough to compete with the normative response or the override cannot be completed (or, sustained). If the mindware is instantiated sufficiently, the normative response can be overridden.

In this grey area where a sustained override can succeed or fail, two causes of errors are differentiated. Either, a failure of override occurred, which is a process failure, or an absence of mindware occurred, which is a knowledge deficit. Again, depending on the level of mindware instantiation, either type of error becomes more or less likely. When the mindware is well-instantiated, it is more likely a process failure occurred as the appropriate mindware is available, and vice versa. It is virtually impossible to distinguish between knowledge deficits and process failures in practice.(65) However, mindware is a crucial component for task performance in either scenario. The importance of mindware is supported by previous studies: for example, Šrol and de Neys (66) concluded that mindware instantiation was the best predictor of participants' ability to detect conflicts and their susceptibility to making errors due to bias. Burič and Konradova (67) showed that mindware instantiation and conflict detection efficiency explained 10% of the variance in the accuracy of individual responses. These studies were performed using well-defined logic tasks, such as the ball-and-bat problem.(68) In such tasks, a wrong normative answer is easily triggered and there is only one correct answer based on logical principles like probability. Furthermore, Janssen et al. (69) determined that conflict detection also occurred when evaluating solutions provided by others, instead of only detecting internally generated solutions, showing that mindware is also important in such circumstances.

Mindware and diagnostic errors

Stanovich's (58) framework can directly be applied to diagnostic reasoning. Appropriate mindware in medicine constitutes the illness scripts (70), exemplars, and prototypes (71) that encode medical knowledge. How successful performance is on the task of diagnostic reasoning subsequently depends on the instantiation of these knowledge structures. In the event that an incorrect diagnosis is generated, this can either be detected and overridden, or not detected and missed. The latter scenario would result in a diagnostic error. If the clinician's illness scripts were incomplete or not readily available, this diagnostic error would most likely be a result of a knowledge deficit. On the other hand, if the clinician's illness scripts were well-learned, the diagnostic error would probably occur due to a failure to override – a process failure. Within this framework, cognitive biases still have a place as causes of diagnostic errors: accepting a normative response over a competing response, where the normative response is generated by implicitly learned rules, would be considered an error due to a cognitive bias resultant from a heuristic. Only now, cognitive biases would also be explained in light of someone's mindware and not from the perspective of process skills alone.

It should be kept in mind that the studies on which the framework is based primarily presented participants with clearly defined logic tasks, which is not the case in medicine. Well-defined tasks might be difficult or formulated in a misleading fashion to induce incorrect normative responses, but they still have one accepted correct answer that could be known from the outset. In medicine, however, our knowledge is not quite so complete. Just as we have no idea of the contents of about 80% of Earth's oceans (72), there are diagnostic errors, referred to as 'unavoidable errors', that can occur simply because medical knowledge on diseases or disease presentations is not yet advanced enough.(73) For example, diseases can present atypically or relevant symptoms can be occluded by comorbidities. In some scenarios clinicians are not able to gather all possible information, as diagnostic testing also carries risks, such as harming the patient or resulting in unnecessary treatment due to overdiagnosis. Therefore, unlike the ball-and-bat problem, a patient case does not have a definitive correct answer. A patient's true diagnosis can often not be known until after an autopsy has been performed. Because medicine always carries a degree of uncertainty, it is classified as a task requiring *fuzzy logic*. Fuzzy logic applies to situations where decisions are made based on vague or imprecise information, and where the answer can be anywhere in between completely true and completely false.(74) Stanovich's (58) framework can be applied both to well-defined and fuzzy logic tasks, only with the caveat that conflict detection becomes more difficult and it becomes harder to override incorrect responses due to the inherent uncertainty regarding the correct diagnosis.

Main conclusions

In summary, the first aim of this thesis was to provide insight in the cognitive causes of diagnostic errors. Our results indicated that neither aspects of cognitive processing (i.e., time taken to diagnose and confidence in diagnosis) nor the clinical information used to arrive at a diagnosis (i.e., pertinent information, necessary to distinguish between the correct diagnosis and the diagnostic suggestion) differed between correct diagnoses and error cases. However, changes were observed when clinicians overcame a bias suggestion and arrived at the correct diagnosis compared to when they made a mistake or arrived at the correct diagnosis under a correct suggestion. From these results we hypothesized that the ability to detect an incorrect suggestion, in conjunction with having the appropriate knowledge to arrive at the correct diagnosis, seemed crucial to overcoming an error and could therefore hint at what processes are compromised when an error does occur. These observations are in line with Stanovich's (58) cognitive reasoning framework, which posits that task performance depends on the degree of how available and well-learned one's knowledge relevant to the task is. From there, when an error occurs, it can either be because the error was not detected or overridden (a process failure) or because the appropriate mindware was not instantiated (a knowledge deficit). Therefore, we hypothesize that a lack of availability and instantiation of relevant mindware is a crucial cause of diagnostic errors, from which process failures and knowledge deficits can both spring forth.

The second aim of this thesis was to investigate possible diagnostic error interventions. Overall, cognitive reasoning tools aimed at improving diagnostic reasoning performance for workplace settings showed a small, but clinically relevant, increase in diagnostic accuracy. In perspective of the mindware framework, it could be hypothesized that these tools improved clinicians' mindware either by providing support specifically for a certain task (i.e., by suggesting diagnoses for certain symptoms or by providing a list of steps necessary to complete for specific tasks) or by supporting general reasoning processes (i.e., by guiding clinicians through a structured format of reasoning). Several of our more surprising observations could be explained by hypothesizing that a lack of appropriate mindware causes errors. First, feedback might be more effective for easy than for difficult cases because medical interns can make better use of the feedback when the appropriate mindware is available, which is more likely in easier cases. Furthermore, extensive practice with a tool increased its effectiveness: practice with a tool could increase its instantiation and therefore lead to more effective use. Lastly, when reasoned from the mindware framework, process failures and knowledge deficits both originate from differing degrees of mindware instantiation. Perhaps no differences between process and content interventions are observed in the pooled effect because error interventions might remedy both types of errors, which would occur differently depending on the sample and other methodological characteristics of individual studies. This is merely theoretical, however, as research on cognitive reasoning tools is relatively scarce and data regarding the tools' effectiveness in practice and under which circumstances the tools could be most useful is lacking. Taken together, we conclude that interventions aimed at reducing diagnostic errors due to cognitive deficits could be effective, as even a small improvement in accuracy could mitigate a substantial amount of errors in light of their high prevalence. Caution should, however, be taken and more research will be necessary to properly implement the tools in practice.

Strengths and limitations

Several overarching strengths and limitations of this thesis should be addressed when interpreting the main findings. This section will discuss several issues relevant to most or all studies in this thesis, as specific strengths and limitations are discussed in their respective chapters. First, a limitation is that our studies cover a wide range of medical specialties and participants of differing levels of expertise. By aggregating over studies with varying methods, we might have missed nuances in different subgroups: perhaps underlying reasoning processes differ for students when compared to residents, or for a visual diagnostic specialty, such as radiology, compared to internal medicine. However, being able to compare and aggregate the results over studies performed in varying settings also presents a strength. Even across this wide range of settings, we found generally consistent patterns in our results. Therefore, we can present a more robust estimation of how cognitive reasoning processes occur, rather than being confined to one setting. Perhaps it would be best to draw both overarching conclusions, across a wide range of settings, and specific conclusions, by aggregating results within specific subgroups, to properly evaluate the rapidly expanding evidence base in this field.

A second limitation is that all our chapters concern laboratory studies, which favor a controlled environment in exchange for ecological validity. We cannot account for variations in cognitive processes triggered by additional factors, such as time pressure, stress, or input from peers. Additionally, the way in which we present cases to clinicians differs from how they approach diagnosis in the clinic. In our studies, with the exception of Chapter 5, all necessary information is presented immediately. There is no need to further collect information, or to gradually build a diagnosis based on ordering multiple rounds of investigations. Therefore, our ability to generalize our findings to practice is limited. This problem is present in most studies in the field. Especially in the case of error interventions, it will be absolutely necessary replicate these findings in the clinic. We deliberately chose to focus on laboratory studies in order to examine clinicians' cognitive processes in detail and to eliminate several confounders which might otherwise explain our results. Therefore, we were able to isolate certain effects and processes pertaining to diagnostic reasoning using available information in our studies. These results should still, however, be supplemented with replications in practice, as we will only then know whether cognitive diagnostic errors can truly be reduced for real patients.

Third, the clinical cases used in our studies were often designed to result on average in an accuracy of 50 to 60% for our participants. This is a problem not only present in the current thesis but also in many studies in the field as a whole. Cases are generally designed to be more difficult than in clinical practice because researchers ideally need a sufficient numbers of correct and incorrect responses to apply statistical analyses to their data. If there are too few correct or incorrect answers, no comparison can be made. The cognitive processes underlying diagnosis might differ for easier and more difficult cases, and which cases are perceived as easy or difficult by a clinician will probably be case-specific due to differences in individual knowledge bases. Future research should investigate whether easier or more difficult case have additional or different effects on the diagnostic reasoning process.

Further strengths of the studies in this thesis include that most studies included relatively large samples of participants and were able to reach the minimum requirements for 80% power. This indicates that if no significant effect was found, this was likely not because the studies were underpowered but because there was no difference to be detected. Additionally, the studies were designed and conducted by a team of both clinical and content experts, to ensure both the methodological quality and clinical relevance of the studies. And finally, a strength of this thesis was that all studies were preregistered in the Open Science Framework (OSF) repository and the review was preregistered on Prospero, a systematic review registration repository from the University of York. We also endeavored to publish our articles with Open Access rights. Open Science aims to be transparent about each step of the research process to enhance reproducibility of results and to improve accessibility.(75) Preregistration is one of the responses against issues in academic publishing, which are termed questionable research practices.(76) These practices include things such as changes hypotheses after the results have been seen to make it appear like the hypotheses were verified (hypothesizing after the results are known, or HARKing) or cherry-picking significant results while not reporting non-significant findings. By preregistering, such practices can be quickly identified, as a time-stamped protocol from before data collection or analysis is available. Preregistration and other Open Science practices allow for a more accurate assessment of the scientific evidence base in general and contribute to improving our ability to draw accurate conclusions from scientific findings.

Implications for research and practice

The findings in this thesis provide insights and guidance for practical implications and future research directions. Our findings support the hypothesis that the availability and accessibility of one's mindware underlies cognitive diagnostic errors, rather than differences in the reasoning process itself. Furthermore, we demonstrated that interventions aimed specifically at improving and supporting clinicians' cognition can improve diagnostic accuracy. These findings suggest that clinicians' mindware and their ability to use that mindware should be a focus in our aim to reduce diagnostic errors. By applying the mindware framework to diagnosis, we hypothesize that a gap between the existing and the necessary mindware for a task is a main cause for diagnostic errors. Naturally then, the question arises: how do we address this mindware gap, in order to reduce diagnostic errors? The first option would be to improve clinician's available mindware. Although it is impossible to possess every bit of medical information available in the 21st century, which will only expand in the future (77), it should not be ignored that deficits in current education likely exist and could be improved upon. Identifying and improving such problem areas in students' diagnostic reasoning could be a step forward to improving students' mindware. Practicing with a wide variety of diseases and disease presentations can be a first step to building more elaborate and better instantiated knowledge structures. Additionally, feedback might be a valuable strategy as it can both add to clinicians' knowledge by providing the correct answer, and improve their calibration, making them better equipped to detect errors in their diagnostic reasoning in the future.(78)

Furthermore, instead of focusing on the content of mindware, interventions could also be aimed at improving diagnostic reasoning in general. A paper by Croskerry (79) also addresses this issue. He discusses the existence of a mindware gap, defining it as a lack of "the cognitive resources needed to think rationally".(80) Croskerry posits there is a gap between routine expertise, or what clinicians are taught, and adaptive expertise (81), the skills and knowledge that are actually needed in practice. To narrow this gap, thinking strategies could be taught to augment clinician's diagnostic reasoning skills. Mentioned are strategies that would aim to increase clinician's rational thinking (i.e., via debiasing strategies), metacognition (thinking about thinking, i.e., reflection on one's reasoning process), critical thinking, training in the medical humanities (i.e., developing diagnostic reasoning skills through observing artwork), lateral thinking (i.e., flexible and creative thinking that ignores traditional stepwise reasoning), and distributed thinking (i.e., distributing cognitive load and approaching diagnosis from a team-view rather than from the view of an individual clinician). The effectiveness of such general strategies might be limited, as it is unclear to what extent clinicians' routine expertise truly differs from the necessary adaptive expertise. Many strategies overlap with what clinicians have been taught and might therefore not result in a noticeable benefit. Additionally, the effectiveness of many of these strategies in reducing diagnostic errors has not been tested, although deliberate reflection (a metacognitive approach) has overall been found to lead to small improvements in accuracy.(53)

For future research, it will be valuable to identify possible areas of improvement in clinicians' diagnostic reasoning process. For example, Chapter 5 showed that diagnostic justification skill in medical students is a weakness in the diagnostic process, both pre- and post-clerkship.(22) Identifying and improving specific reasoning skills will likely allow students to better apply their mindware and to improve their diagnostic ability. Future research should compare the diagnostic reasoning process in subgroups, such as participants with differing levels of expertise or from different specialties, as differences might exist across settings. Additionally, future studies should ensure that our current results, both regarding the origins of cognitive diagnostic errors and error interventions, are replicable in practice.

It should be kept in mind that not all improvements can be realized internally, within the cognition of the clinician. It is simply not possible for one individual to know all available information, let alone to use it appropriately. For this reason, it is crucial to realize that the concept of "available knowledge" extends beyond just the information that someone has learned during their studies. In our current era of technology, clinicians are also able to rely on external sources of information and reasoning support, such as patients' electronic health records, electronic triggers and reminders in such records, or artificial intelligence such as computerized decision support systems or more general programs like ChatGTP.(82-84) Artificial intelligence has the potential to drastically change – and hopefully improve – medical diagnosis. However, artificial intelligence is not ready yet to be fully implemented in the clinical workflow. This will require further research into how clinicians work together such systems, as the question remains how the diagnostic process can most optimally be supported by artificial intelligence. Previous literature shows that artificial intelligence is already quite good at making diagnoses but often fails to be used by clinicians due to various barriers and limitations.(85) Important here is to not see artificial intelligence as a replacement of clinicians or experts but as a supplement that will further support their reasoning processes, for example by recommending diagnoses that might not have been included in the differential diagnosis yet.

The relevance of this thesis should also be considered in the broader perspective of diagnostic errors in general. We specifically focused on gaining insight in cognitive errors and interventions that directly targeted cognition, as cognitive errors are the most frequently identified cause of errors and are crucially involved in the foundation of diagnostic reasoning. However, cognitive errors are ultimately only one part of the puzzle. Diagnostic errors are often the result of multifactorial causes (86, 87), and interactions between cognitive, organizational, and system breakdowns can lead to errors in many ways. Therefore, although it is promising that cognitive errors could be reduced by cognitive reasoning tools, it should be emphasized that this will only improve part of the problem.

Other changes or interventions will be necessary to reduce diagnostic errors as they occur in the clinic. Furthermore, improving cognitive causes of diagnostic errors is quite challenging and cognitive errors will likely never fully be eradicated. In light of that, it will also be valuable to develop interventions that target other causes of errors and that might be able to compensate for clinicians' cognitive flaws. For example, it will likely be much easier to develop and implement system interventions, rather than targeting the cognitive processes of individual healthcare professionals. This will also be extremely valuable: diagnostic errors often occur due to the interaction of multiple breakdowns that aggravate each other and will only lead to a breakdown in the diagnostic process when the combination of factors is severe enough. Preventing or reducing several points of breakdown, whether that be focused on cognition or the system, will make the diagnostic process less likely to be compromised. Therefore, although cognitive interventions seem promising, interventions that target other or multiple processes will also be valuable going forward. Such interventions could, for example, be improvements in information accessibility in the electronic health record or in properly relaying test results et cetera between clinicians. Such system changes prevent system breakdowns and ameliorate clinicians' ability to access the available information and from there, their ability to arrive at the correct diagnosis. Improving the information and knowledge that clinicians can easily access will also improve their mindware. In short, it should always be remembered that cognitive diagnostic errors are an important part of diagnostic errors but not the only part nor the only point for intervention. Prevention of diagnostic errors will have to target, and be tailored to, the diagnostic environment as a whole.

As research into diagnostic errors and ways to prevent them progresses, attention should also be directed to better understanding when interventions should be used. Although very well-implemented interventions might have the potential to be used for each patient, currently, successful implementation of interventions is limited by feasibility issues. Even if an intervention requires five extra minutes to complete, that is five multiplied by the large volume of patients that have to be seen by a clinician. Therefore, the issue remains that we would ideally not only have a good intervention to implement, but also an idea of when to implement it. A straightforward case, for example, would present a lower chance for diagnostic errors than a complex case, and the intervention would only be necessary for the complex case. This knowledge, of course, is only available retrospectively and therefore, it remains of interest to further research possible markers of when an intervention would be necessary. Clinicians' confidence and calibration might be promising markers, although they are still poorly understood. Despite several findings suggesting that confidence and calibration might be improved through feedback or practice, and could then perhaps provide an indication of when intervention might be necessary, there are also studies indicating that confidence is largely insensitive to accuracy and that clinicians' calibration remains poor. Better understanding confidence and calibration might provide opportunities for signaling when interventions could be implemented.

Conclusions

This thesis provides insight in the causes of cognitive diagnostic errors and interventions aimed at reducing those errors. The studies presented show that neither differences in cognitive processes nor the type of information used during diagnosis are related to diagnostic errors. Instead, the results support the hypothesis that clinicians' available mindware (i.e., the relevant knowledge they have access to on a specific subject) and the degree of mindware instantiation (i.e., how automated and easily accessible this knowledge is) determine whether a correct diagnosis can be made. From there, the mindware framework suggests that either problems in cognitive processes when overriding a wrong diagnosis or knowledge deficits result in cognitive diagnostic errors. We further demonstrated that cognitive reasoning tools were effective in improving diagnostic accuracy, via interventions that improve clinicians' mindware, either by targeting the knowledge individual clinicians' have stored or by providing external mindware through systems such as the electronic health record. Further research will be necessary to examine possible differences in cognitive processes and intervention effectiveness in different subgroups, such as clinicians of different levels of expertise or different clinical specialties. Additionally, the current results should be replicated in practice. In conclusion, cognitive causes of diagnostic errors are an important part of the problem and can be reduced, but the full context of clinical practice needs to be considered to effectively reduce diagnostic errors as a whole.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Archives of internal medicine. 2005;165(13):1493-9.
3. Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Annals of emergency medicine. 2003;41(1):110-20.
4. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. Journal of Evaluation in Clinical Practice. 2018;24(3):666-73.
5. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine. 2017;92(1):23-30.
6. Kahneman D. Thinking, fast and slow: Macmillan; 2011.
7. Norman G, Sherbino J, Dore K, Wood T, Young M, Gaissmaier W, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. Academic Medicine. 2014;89(2):277-84.
8. Sherbino J, Dore KL, Wood TJ, Young ME, Gaissmaier W, Kreuger S, et al. The relationship between response time and diagnostic accuracy. Academic Medicine. 2012;87(6):785-91.
9. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203.
10. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. Jama. 2004;292(13):1602-9.
11. LeBlanc VR, Brooks LR, Norman GR. Believing is seeing: the influence of a diagnostic hypothesis on the interpretation of clinical features. Academic Medicine. 2002;77(10):S67-S9.
12. Coderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. Medical education. 2003;37(8):695-703.
13. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. Academic Medicine. 2011;86(3):307-13.
14. Findlay JM, Gilchrist ID. Active vision: The psychology of looking and seeing: Oxford University Press; 2003.
15. Zwaan L, Thijs A, Wagner C, Timmermans DRM. Does inappropriate selectivity in information use relate to diagnostic errors and patient harm? The diagnosis of patients with dyspnea. Social science & medicine. 2013;91:32-8.
16. Mamede S, Goeijenbier M, Schuit SCE, de Carvalho Filho MA, Staal J, Zwaan L, et al. Specific disease knowledge as predictor of susceptibility to availability bias in diagnostic reasoning: a randomized controlled experiment. Journal of general internal medicine. 2021;36(3):640-6.
17. Al-Moteri MO, Symmons M, Plummer V, Cooper S. Eye tracking to investigate cue processing in medical decision-making: A scoping review. Computers in Human Behavior. 2017;66:52-66.
18. Gilliland S. Clinical reasoning in first-and third-year physical therapist students. Journal of Physical Therapy Education. 2014;28(3):64-80.
19. Walling A, Moser SE, Dickson G, Zackula RE. Are students less likely to report pertinent negatives in post-encounter notes? Family Medicine-Kansas City. 2012;44(1):22.

20. Corbett A, Sandholdt C, Bakerjian D, editors. Learning analytics with virtual patient data reveals subgroup of students who miss pertinent findings. Innovate Learning Summit; 2020: Association for the Advancement of Computing in Education (AACE).

21. Hayden EM, Petrusa E, Sherman A, Feinstein DM, Khoury K, Krupat E, et al. Association of Simulation Participation With Diagnostic Reasoning Scores in Preclinical Students. Simulation in Healthcare. 2022;17(1):35-41.

22. Williams RG, Klamen DL. Examining the diagnostic justification abilities of fourth-year medical students. Academic Medicine. 2012;87(8):1008-14.

23. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Advances in health sciences education. 2009;14(1):27-35.

24. Berner ES. Diagnostic error in medicine: introduction. Springer; 2009. p. 1-5.

25. Hobu PPM, Schmidt HG, Boshuizen HPA, Patel VL. Contextual factors in the activation of first diagnostic hypotheses: expert-novice differences. Medical education. 1987;21(6):471-6.

26. Pronovost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. New England journal of medicine. 2006;355(26):2725-32.

27. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat A-HS, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. New England journal of medicine. 2009;360(5):491-9.

28. Gawande A. The checklist manifesto: How to get things right. Journal of Nursing Regulation. 2011;1(4):64.

29. Lambe KA, O’Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ quality & safety. 2016;25(10):808-20.

30. Gupta A, Graber ML. Annals for hospitalists inpatient notes-just what the doctor ordered—checklists to improve diagnosis. Annals of Internal Medicine. 2019;170(8):HO2-HO3.

31. Zwaan L, Staal J. Evidence on Use of Clinical Reasoning Checklists for Diagnostic Error Reduction. AHRQ Papers on Diagnostic Safety Topics [Internet]. 2020; (3).

32. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: A randomised experiment. Medical education. 2021;55(10):1172-82.

33. Konings K, Willemsen R. ECG 10+: Systematisch ECG’s beoordelen. Huisarts en wetenschap. 2016;59(4):166-70.

34. Rutten FH, Kessels AGH, Willems FF, Hoes AW. Is elektrocardiografie in de huisartspraktijk nuttig? Huisarts en wetenschap. 2001;44(11):179-83.

35. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing checklist on diagnostic performance in comparison to intuitive diagnosis. Medical teacher. 2013;35(6):e1218-e29.

36. Sibbald M, Sherbino J, Ilgen JS, Zwaan L, Blissett S, Monteiro S, et al. Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. Advances in Health Sciences Education. 2019;24(3):427-40.

37. Sibbald M, de Bruin ABH, van Merrienboer JJG. Checklists improve experts’ diagnostic decisions. Medical education. 2013;47(3):301-8.

38. Sibbald M, de Bruin ABH, Cavalcanti RB, van Merrienboer JJG. Do you have to re-examine to reconsider your diagnosis? Checklists and cardiac exam. BMJ quality & safety. 2013;22(4):333-8.

39. Nedorost S. A diagnostic checklist for generalized dermatitis. Clinical, Cosmetic and Investigational Dermatology. 2018;11:545.

40. Sibbald M, De Bruin ABH, van Merrienboer JJG. Finding and fixing mistakes: do checklists work for clinicians with different levels of experience? Advances in Health Sciences Education. 2014;19(1):43-51.

41. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis. 2015;2(2):97-103.

42. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine. 2013;173(21):1952-8.

43. Rawson KA, Dunlosky J. Improving students' self-evaluation of learning for key concepts in textbook materials. European Journal of Cognitive Psychology. 2007;19(4-5):559-79.

44. Archer JC. State of the science in health professional education: effective feedback. Medical education. 2010;44(1):101-8.

45. Nederhand ML, Tabbers HK, Splinter TAW, Rikers RMJP. The effect of performance standards and medical experience on diagnostic calibration accuracy. Health Professions Education. 2018;4(4):300-7.

46. Kuhn J, van den Berg P, Mamede S, Zwaan L, Bindels P, van Gog T. Improving medical residents' self-assessment of their diagnostic accuracy: does feedback help? Advances in Health Sciences Education. 2021:1-12.

47. Ryan A, Judd T, Swanson D, Larsen DP, Elliott S, Tzanetos K, et al. Beyond right or wrong: More effective feedback for formative multiple-choice tests. Perspectives on Medical Education. 2020;9(5):307-13.

48. Dunlosky J, Rawson KA. Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. Learning and Instruction. 2012;22(4):271-80.

49. Lichtenstein S, Fischhoff B. Training for calibration. Organizational behavior and human performance. 1980;26(2):149-71.

50. Hattie J, Timperley H. The power of feedback. Review of educational research. 2007;77(1):81-112.

51. Graber ML, Kissam S, Payne VL, Meyer AND, Sorensen A, Lenfestey N, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ quality & safety. 2012;21(7):535-57.

52. Winters BD, Aswani MS, Pronovost PJ. Commentary: reducing diagnostic errors: another role for checklists? Academic Medicine. 2011;86(3):279-81.

53. Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. Medical Teacher. 2019;41(5):517-24.

54. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: Still no easy answers. Western Journal of Emergency Medicine. 2020;21(6):125.

55. Mamede S, Schmidt HG. Reflection in medical diagnosis: a literature review. Health Professions Education. 2017;3(1):15-25.

56. Astik GJ, Olson APJ. Learning from Missed Opportunities Through Reflective Practice. Critical Care Clinics. 2022;38(1):103-12.

57. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009;62(10):e1-e34.

58. Stanovich KE. Miserliness in human cognition: The interaction of detection, override and mindware. Thinking & Reasoning. 2018;24(4):423-44.

59. Perkins D. Outsmarting IQ: The emerging science of learnable intelligence. 1995.

60. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annual review of psychology. 2011;62(1):451-82.

61. Achtziger A, Alós-Ferrer C. Fast or rational? A response-times study of Bayesian updating. Management Science. 2014;60(4):923-38.

62. Jimenez N, Rodriguez-Lara I, Tyran J-R, Wengström E. Thinking fast, thinking badly. Economics Letters. 2018;162:41-4.

63. Moore R. Fast or slow: Sociological implications of measuring dual-process cognition. Sociological Science. 2017;4:196-223.

64. De Neys W. Bias, conflict, and fast logic: Towards a hybrid dual process future? Dual process theory 20: Routledge; 2017. p. 47-65.

65. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Medical education. 2010;44(1):94-100.

66. Šrol J, De Neys W. Predicting individual differences in conflict detection and bias susceptibility during reasoning. Thinking & Reasoning. 2021;27(1):38-68.

67. Burič R, Konrádová Ľ. Mindware instantiation as a predictor of logical intuitions in the Cognitive Reflection Test. Studia Psychologica. 2021;63(2):114-28.

68. Frederick S. Cognitive reflection and decision making. Journal of Economic perspectives. 2005;19(4):25-42.

69. Janssen EM, Velinga SB, de Neys W, Van Gog T. Recognizing biased reasoning: Conflict detection during decision-making and decision-evaluation. Acta Psychologica. 2021;217:103322.

70. Schmidt HG, Boshuizen H. On acquiring expertise in medicine. Educational psychology review. 1993;5(3):205-21.

71. Nosofsky RM. Exemplars, prototypes, and similarity rules. Essays in honor of William K Estes. 1992;1:149-67.

72. Society NG. Ocean: National Geographic Society; 2022 [updated 15-07-2022. Available from: https://education.nationalgeographic.org/resource/ocean.

73. Newman-Toker DE. A unified conceptual model for diagnostic errors: underdiagnosis, overdiagnosis, and misdiagnosis. Diagnosis. 2014;1(1):43-8.

74. Novák V, Perfilieva I, Mockor J. Mathematical principles of fuzzy logic: Springer Science & Business Media; 2012.

75. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348(6242):1422-5.

76. Andrade C. HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. The Journal of Clinical Psychiatry. 2021;82(1):25941.

77. Densen P. Challenges and opportunities facing medical education. Transactions of the American Clinical and Climatological Association. 2011;122:48.

78. Schiff GD. Minimizing diagnostic error: the importance of follow-up and feedback. The American journal of medicine. 2008;121(5):S38-S42.

79. Croskerry P. Narrowing the mindware gap in medicine. Diagnosis. 2022;9(2):176-83.

80. Stanovich KE. Rational and irrational thought: The thinking that IQ tests miss. Scientific American. 2015;23:12-7.

81. Croskerry P. Adaptive expertise in medical decision making. Medical teacher. 2018;40(8):803-8.

82. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. bmj. 2020;370.

83. Murphy DR, Meyer AND, Sittig DF, Meeks DW, Thomas EJ, Singh H. Application of electronic trigger tools to identify targets for improving diagnostic safety. BMJ Quality & Safety. 2019;28(2):151-9.

84. Graber ML, Byrne C, Johnston D. The impact of electronic health records on diagnosis. Diagnosis. 2017;4(4):211-23.

85. Mirbabaie M, Stieglitz S, Frick NRJ. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health and Technology. 2021;11(4):693-731.

86. Schiff G, Volodarskaya M, Nieva HR, Singh H, Wright A, editors. Diagnostic Pitfalls: A New Paradigm to Understand and Prevent Diagnostic Error. Journal of general internal medicine; 2016: SPRINGER ONE NEW YORK PLAZA, SUITE 4600, NEW YORK, NY, UNITED STATES.

87. Schiff GD, Hasan O, Kim S, Abrams R, Cosby K, Lambert BL, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. Archives of internal medicine. 2009;169(20):1881-7.

Summary

In 2015, the National Academies of Sciences, Engineering, and Medicine (NASEM) raised awareness for diagnostic errors, which had been a neglected part of medical errors, through their report *Improving diagnosis in healthcare*.(1) They estimated that "most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences." Diagnostic errors are generally defined as "the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient".(1) Cognitive causes, such as faulty information gathering or interpretation, are seen as major contributors to diagnostic errors.(2) Given the complexity of the diagnostic process, detecting and measuring diagnostic errors is a challenging task, and much is still unclear about the mechanisms underlying the cognitive causes of diagnostic errors and the interventions that could prevent them. This thesis consists of four studies that sought to increase our understanding of the cognitive mechanisms underlying diagnostic errors, and three studies with the aim to determine the effectiveness of interventions meant to prevent cognitive errors.

Chapter 1 describes the current theoretical framework of diagnostic errors and the debate on whether cognitive biases or knowledge deficits should be considered the main cause of cognitive diagnostic errors. Additionally, different error prevention strategies in line with either side of the debate are discussed, along with the available, though relatively scarce, empirical evidence for these strategies. Subsequently, **Chapter 2 through 5** focus on investigating how several cognitive factors related to the occurrence of diagnostic errors, and whether these cognitive factors could be seen as mechanisms of cognitive errors. **Chapter 6 through 8** examines diagnostic error interventions and their efficacy in improving diagnostic accuracy. Finally, **Chapter 9** synthesizes the results of the thesis with the current literature and supports an update to the existing theoretical framework to understand cognitive diagnostic errors. Implications for future research and practice are also discussed.

The first aim of the thesis concerned cognitive factors that might function as mechanisms for cognitive diagnostic errors. In several experimental and observational studies, we investigated how clinicians' cognitive processes differed in cases where an error occurred versus where it did not. Since cognitive processes cannot be observed directly, we attempted to observe bias by inducing cognitive bias via the study design and examined potential knowledge deficits through clinicians' use of available clinical information.

Chapter 2 concerned a multi-center laboratory experiment that investigated the hypothesis that diagnostic errors are primarily caused by cognitive biases that result from faster diagnostic reasoning.(3) Previous studies by Norman et al. (4) and Sherbino et al. (5) already showed that correct diagnoses were reached just as fast, or even faster, than incorrect diagnoses. However, their results were still open to the interpretation that faster participants are simply better diagnosticians. We therefore replicated these experiments in a within-subjects design. Internal medicine residents diagnosed clinical cases that aimed to induce availability bias (using Mamede et al.'s methodology (6)). Availability bias occurs when someone bases themselves on information that easily comes to mind, for example, when a clinician's diagnosis of a patient is influenced by a similar patient they saw before. The residents in our experiment first diagnosed cases with a diagnostic suggestion and then diagnosed new cases with diagnoses that resembled, but were different from, the suggested diagnoses they encountered before. The results indicate that correct diagnoses were, on average, reached faster than incorrect diagnoses. Residents were also more confident in their correct diagnoses, although their overall confidence-accuracy calibration was poor. Exploratory results showed a trend for bias-induced errors to be faster than other types of errors, but not faster than correct diagnoses. Perhaps this indicates that similar cognitive processes underlie both correct and flawed diagnostic reasoning. In line with the previous literature, this experiment suggests faster reasoning seems to underlie both correct and erroneous diagnoses. Faster reasoning is likely a valuable part of the diagnostic process and future research should consider causes for diagnostic errors other than the speed of diagnosis.

Other than the intrinsic quality of a clinicians' speed of diagnosis, available clinical information also has the potential to bias a clinicians' reasoning process. **Chapter 3** examined the effect of externally provided diagnostic suggestions, specifically of the suggested diagnosis found in a patients' referral letter from the general practitioner to the emergency department. In theory, incorrect diagnostic suggestions could lead to incorrect interpretations of clinical information or cognitive biases, which would finally result in diagnostic errors. Medical interns diagnosed written case vignettes that were randomly displayed with a correct diagnostic suggestion, an incorrect diagnostic suggestion, or no suggestion at all. The case vignettes were formatted to resemble genuine anonymized referral letters. Interns who received a diagnostic suggestion included fewer differential diagnoses than interns who did not receive any suggestion. However, overall diagnostic accuracy, confidence, and time to diagnose were not affected. The interns in this study were not misled by incorrect diagnostic suggestions. A case-by-case examination of the results revealed that the effect of diagnostic suggestions could be dependent on the specific clinical case, as the correct diagnostic suggestion improved diagnostic accuracy in some cases, but not in others. Interns' prior knowledge or the mental flexibility necessary to consider a suggestion could be moderators of this effect. In conclusion, diagnostic suggestions could serve both to correctly guide the clinical reasoning process or to unnecessarily restrict it. It is valuable for educators to be aware of the potential influences of diagnostic suggestions on their students' differential diagnoses. Perhaps practicing with both cases that do or do not contain a suggestion could help with learning to construct a broad differential diagnosis.

Chapter 4 zoomed in on residents' diagnostic information processing between correct and incorrect diagnoses by measuring specifically what information was used during diagnosis. Even if all necessary information to make a diagnosis is available, clinicians can still arrive at the incorrect diagnosis. It is hypothesized that incorrect selectivity in information processing could influence the clinical information that is or is not considered. Consequently, if the wrong information is considered, this could result in a diagnostic error. Internal medicine and emergency medicine residents diagnosed written case vignettes with either a correct or an incorrect suggested working diagnosis. Internal medicine residents' eye movements were recorded while they diagnosed written case vignettes. We measured how long and how often residents looked at either information that supported the correct working diagnosis or the incorrect working diagnosis, and how this related to their final diagnostic accuracy. Overall, information processing did not differ between correct and incorrect diagnoses. However, an interaction was observed between diagnostic accuracy and the diagnostic suggestion where information that supported the correct working diagnosis was looked at more often if residents saw a case with an incorrect working diagnosis but arrived at the correct final diagnosis regardless. This interaction seemed similar for both how often and how long residents looked at the information, but was only significant for how often they looked. Their confidence and time to diagnose, the other aspects of the residents' diagnostic process, did not differ under any condition. Although this interaction suggests that residents who focused more on relevant information were able to overcome the bias of the suggestion, no specific selectivity in information processing was observed in error cases alone. Therefore, selectivity might be more of an indication of changes in cognitive processes related to diagnosis, rather than a direct cause for errors. This experiment suggests that the ability to recognize the relevant information from a case is important to arrive at the correct diagnosis and can assist in overcoming confirmation bias.

To further supplement our understanding of our information processing occurs in diagnosis, **Chapter 5** reports an observational study that measured medical students' diagnostic justification, or the skill to determine whether certain information increases or decreases the probability of a certain diagnosis. This goes beyond examining what information is used and shows how students apply this information to the construction of their differential diagnosis and the eventual selection of a final diagnosis. Incorrectly assigning certain clinical information to diagnoses might be an explanation of how diagnostic errors occur. First and second year medical students diagnosed written case vignettes in an online learning environment where they could select specific pieces of information and indicate whether they increased or decreased the likelihood of a diagnosis in their differential diagnosis. They had to do the same thing for investigations that they ordered. At the end of the case, they provided a final diagnosis from their differential. Despite that students performed well in their final diagnostic accuracy and in creating an initial differential diagnosis, they performed poorly on ordering investigations and diagnostic justification. Their high diagnostic accuracy might partially be explained by the issue that students often have to diagnose cases about diseases they are currently learning about in the curriculum, making it easier to guess the diagnosis of the case even without being a good diagnostician.(7) Students consistently underreported clinical information, especially if it decreased the likelihood of a certain diagnosis. This pattern did not differ between correct and incorrect diagnoses, though this relationship might have been distorted by students' high final diagnostic accuracy or the low amounts of clinical information assigned to diagnoses overall. This study shows that deficiencies in diagnostic justification are present early on, and research with students who are further along in their studies shows that these deficiencies are barely remedied during their clerkship period.(7) Students' diagnostic justification skills were improved as they practiced more with the written case vignettes, however, possibly due to the specific feedback they were provided after each vignette. Properly developing students' diagnostic justification skills is likely crucial to fostering good diagnostic reasoning skills.

Taken together, these chapters suggest that diagnostic errors do not originate from changes in clinicians' time to diagnose, confidence, or information processing directly, seeing as these processes occur similarly in correct diagnoses and errors due to bias. However, the knowledge necessary to arrive at the correct diagnosis seems crucial, as it likely allows clinicians to detect the error in an incorrect diagnostic suggestion – or in their own reasoning.

The second aim of this thesis was to evaluate the effectiveness of several promising cognitive error interventions. We performed experimental studies on the use of diagnostic checklists and feedback and compiled available experimental evidence on the effectiveness of cognitive reasoning tools (i.e., any tool focused on supporting or improving clinicians' cognitive reasoning processes during diagnosis) in a systematic review and meta-analysis.

Chapter 6 compared the effect of checklist interventions at reducing diagnostic errors in normal and abnormal electrocardiograms (ECGs). Because the effectiveness of error intervention strategies is often tested only with abnormal cases, it is unknown how normal cases are affected. In the Netherlands, general practitioners often use checklists to interpret ECGs and to decide whether the patient should be referred to the hospital or not.(8) Approximately one third of the ECGs they encounter are normal, and another third is simple rather than a complex abnormal ECG.(9) Possibly, checklists could lead to the unnecessary consumption of resources, overtesting, or overdiagnosis, as checklists might lead to clinicians overexamining cases even when there is nothing to find.(10) To test this, we asked general practitioner residents to diagnose ECGs with and without a checklist. We examined both process (or debiasing) checklists and content checklists.(10, 11) Process checklists focus on not missing any relevant steps of the diagnostic process, or explicitly checking for biases and errors in reasoning, whereas content checklists focus on activating prior knowledge and ensuring all relevant information is collected. Half of the residents used the debiasing checklist, half used the content checklist. Checklists did not differentially affect residents' performance for normal, simple, and complex abnormal ECGs. Performance also did not differ between the debiasing checklist and the content checklist. A trend for improved accuracy was observed, but this was not significant. Interestingly, residents' confidence did not change, which together resulted in an overall improved confidence-accuracy calibration. Residents became less overconfident when using the checklist. Taken together, checklists could offer value by reducing overconfidence without differentially affecting normal and abnormal ECGs.

Another promising intervention for diagnostic errors is feedback. Often, error interventions cannot be implemented as using them for each patient that is seen would simply take too much time. The solution would be to only use interventions for difficult cases that need extra support, but clinicians' confidence-accuracy calibration is notably poor.(12) **Chapter 7** examined whether feedback could improve clinicians' calibration. Medical interns diagnosed chest X-rays in a learning phase, followed by a test phase. They received either no feedback, feedback on whether their chosen diagnosis was correct (performance feedback), or feedback on whether their chosen diagnosis was correct in addition to an explanation of how the correct diagnosis could be recognized (information feedback). Both types of feedback improved diagnostic accuracy and confidence, compared to the control group. As a result, interns' overall calibration improved as well. There was no difference in time spent during diagnosis. An exploratory subanalysis showed that calibration improved especially for easy X-rays, but not for difficult X-rays. Overall, if interns had the appropriate knowledge or level to diagnose an X-ray, feedback was valuable, whereas cases that were too difficult for their level did not benefit. The improvement in interns' calibration might be explained by an improvement in their awareness of their performance, but conversely it is also possible that interns' accuracy improved due to the feedback whereas their confidence levels remained stable. However, this suggests that at least as performance increased, confidence did not, and previous overconfidence was corrected.

The systematic review and meta-analysis in **Chapter 8** aimed to investigate the effectiveness of cognitive reasoning tools on improving diagnostic accuracy. Existing reviews have aggregated interventions aimed at improving performance in an educational setting with interventions aimed at improving performance in the workplace.(13-17) This review aimed to estimate the effectiveness of cognitive reasoning tools in the workplace and in addition aimed to identify study or intervention characteristics that were associated with higher effectiveness. Three studies were removed from the meta-analysis because participants received extensive training with the tool before diagnostic accuracy was measured, which resulted in a large improvement in diagnostic accuracy. The remaining studies showed a small improvement of diagnostic accuracy. No specific characteristics related to higher diagnostic accuracy were discovered, though it was difficult to isolate the effects of specific characteristics due to the many potential influences on tool effectiveness. It should be kept in mind that the evidence base underlying this meta-analysis was relatively limited and that more research will be necessary to understand under which circumstances cognitive reasoning tools are most effective. Additionally, most underlying studies were laboratory experiments, so replication in practice will also be necessary.

Taken together, cognitive reasoning tools overall seem successful in improving diagnostic accuracy, without detracting from other aspects of the diagnostic process. Although the overall effect is small, even a small improvement can have relevant impact on patient safety given the high prevalence of diagnostic errors.

Chapter 9 synthesized the previously discussed findings and linked them to a cognitive processing framework developed by Stanovich.(18) This thesis proposes that neither aspects of cognitive processes (i.e., time taken to diagnose and confidence in diagnosis) nor the clinical information used to arrive at a diagnosis differed between correct and error cases. We only observed changes in information processing when a biased diagnostic suggestion was corrected. The ability to detect an incorrect suggestion, in conjunction with having the appropriate knowledge to arrive at the correct diagnosis, is likely crucial to overcoming diagnostic error and could hint at the cognitive processes that are compromised when an error occurs.

These findings can be placed within the mindware framework by Stanovich (18), which proposes that in addition to processing skills, available knowledge structures are vital to task performance. The framework allows for both cognitive biases and knowledge deficits to explain diagnostic errors. Mindware is defined as the availability and integration of information relevant to the task, which are proposed to determine task performance, and deficiencies in mindware can both lead to cognitive biases and knowledge deficits. In medicine, mindware consists of illness scripts, exemplars, and prototypes (19, 20) in which medical knowledge is encoded. If this knowledge is well-learned, diagnostic reasoning is likely successful. If an incorrect diagnosis is generated, but the mindware is well-learned, the error might still be detected and corrected. If the mindware is poorly learned, the error will likely evade detection and persist. The mindware framework is primarily based on research that used clearly defined tasks, whereas medical diagnosis if a task with inherent uncertainty. The framework can be applied to medical diagnosis but it should be taken into account that both detecting and correcting errors is more difficult due to this uncertainty.

Furthermore, this thesis showed that cognitive reasoning tools aimed at improving diagnostic performance in the workplace showed a small, but clinically relevant increase in diagnostic accuracy. In perspective of the mindware framework, it could be hypothesized that these tools improved clinicians' mindware either by providing support for a certain task (i.e., by suggesting diagnoses for certain symptoms a patient can present with) or by supporting general reasoning processes (i.e., by providing a structured format for reasoning). However, more research will be necessary to determine under which circumstances error interventions can most benefit clinicians.

The findings in this thesis, when synthesized with the literature, suggest that clinicians' mindware and their ability to use that mindware should be a focus in our aim to reduce diagnostic errors. The gap between the existing and necessary mindware for a task could be a main cause for diagnostic errors. To reduce this gap, it will be necessary to improve clinicians' available mindware. For example, practicing with a wide variety of diseases and disease presentations can be a first step to building more elaborate and well learned knowledge structures. Furthermore, clinicians' general reasoning processes might also be improved via practice or critical thinking. It should, however, be kept in mind that not all mindware improvements can be realized within the cognition of the clinician, as medical knowledge is too extensive to be learned and used by one person. It is crucial to realize that the concept of "available knowledge" extends beyond what is learned during medical school. Technology provides external sources of information and reasoning support, such as patients' electronic health records, electronic triggers, or artificial intelligence. Successfully implementing such interventions in the workflow will likely provide clinicians with the opportunity to effectively make use of the available knowledge. Importantly, electronic supports are not intended as a replacement of clinicians but as a supplement to their reasoning processes.

Overall, this thesis provides insight in the causes of cognitive diagnostic errors and interventions that target these errors. Still, cognitive errors are only a part of the puzzle. Causes of errors are often multifactorial and though reducing cognitive errors is promising, it will only impact part of the problem. Other changes and interventions will be necessary to reduce diagnostic errors as they occur in the clinic. Furthermore, cognitive errors can likely never be fully eradicated and instead, interventions for other causes of diagnostic errors or ways to compensate for clinicians' cognitive flaws should be considered. It is likely easier to implement system changes, for example, than to target cognitive flaws in individual clinicians. Prevention of diagnostic errors will have to target, and be tailored to, the diagnostic environment as a whole.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. European journal of internal medicine. 2013;24(6):525-9.
3. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. BMJ quality & safety. 2013;22(Suppl 2):ii58-ii64.
4. Norman G, Sherbino J, Dore K, Wood T, Young M, Gaissmaier W, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. Academic Medicine. 2014;89(2):277-84.
5. Sherbino J, Dore KL, Wood TJ, Young ME, Gaissmaier W, Kreuger S, et al. The relationship between response time and diagnostic accuracy. Academic Medicine. 2012;87(6):785-91.
6. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203.
7. Williams RG, Klamen DL. Examining the diagnostic justification abilities of fourth-year medical students. Academic Medicine. 2012;87(8):1008-14.
8. Konings K, Willemsen R. ECG 10+: Systematisch ECG's beoordelen. Huisarts en wetenschap. 2016;59(4):166-70.
9. Rutten FH, Kessels AGH, Willems FF, Hoes AW. Is elektrocardiografie in de huisartspraktijk nuttig? Huisarts en wetenschap. 2001;44(11):179-83.
10. Zwaan L, Staal J. Evidence on Use of Clinical Reasoning Checklists for Diagnostic Error Reduction. AHRQ Papers on Diagnostic Safety Topics [Internet]. 2020; (3).
11. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: A randomised experiment. Medical education. 2021;55(10):1172-82.
12. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine. 2013;173(21):1952-8.
13. Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. Medical Teacher. 2019;41(5):517-24.
14. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ quality & safety. 2016;25(10):808-20.
15. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: Still no easy answers. Western Journal of Emergency Medicine. 2020;21(6):125.
16. Mamede S, Schmidt HG. Reflection in medical diagnosis: a literature review. Health Professions Education. 2017;3(1):15-25.
17. Astik GJ, Olson APJ. Learning from Missed Opportunities Through Reflective Practice. Critical Care Clinics. 2022;38(1):103-12.
18. Stanovich KE. Miserliness in human cognition: The interaction of detection, override and mindware. Thinking & Reasoning. 2018;24(4):423-44.

19. Schmidt HG, Boshuizen H. On acquiring expertise in medicine. Educational psychology review. 1993;5(3):205-21.

20. Nosofsky RM. Exemplars, prototypes, and similarity rules. Essays in honor of William K Estes. 1992;1:149-67.

Samenvatting

Summary in Dutch

In 2015 wezen de National Academies of Sciences, Engineering, and Medicine (NASEM) door middel van hun rapport *Improving diagnosis in healthcare* op het belang van diagnosefouten, die tot dan toe een verwaarloosd onderdeel van medische fouten waren.(1) Ze schatten dat de meeste mensen minstens één diagnosefout zullen ervaren in hun leven, soms met verstrekkende gevolgen. Diagnosefouten worden over het algemeen gedefinieerd als "het niet (a) vaststellen van een nauwkeurige en tijdige verklaring van het gezondheidsprobleem of de gezondheidsproblemen van de patiënt of (b) het niet tijdig communiceren van die verklaring aan de patiënt".(1)

Cognitieve oorzaken, zoals foutieve verzameling of interpretatie van klinische informatie, worden gezien als belangrijke oorzaken van diagnosefouten.(2) Gezien de complexiteit van het diagnostische proces is het detecteren en meten van diagnosefouten een uitdagende taak. Er is nog veel onduidelijk over de mechanismen die ten grondslag liggen aan de cognitieve oorzaken van diagnosefouten en de interventies die deze fouten kunnen voorkomen. Dit proefschrift bestaat uit vier studies die trachten meer inzicht te verschaffen in deze cognitieve mechanismen en drie studies die als doel hadden de effectiviteit van interventies, bedoeld om cognitieve diagnosefouten tegen te gaan, te onderzoeken.

Hoofdstuk 1 beschrijft het huidige theoretische kader van diagnosefouten en het debat over cognitieve *biases* of kennisgebreken als belangrijkste oorzaak van cognitieve diagnosefouten moeten worden beschouwd. Daarnaast worden verschillende interventies in lijn met beide zijdes van het debat besproken, samen met het relatief schaarse beschikbare empirische bewijs voor deze interventies. Vervolgens richten **hoofdstuk 2 tot en met 5** zich op het onderzoeken van verschillende cognitieve factoren die verband houden met diagnosefouten, en of deze cognitieve factoren kunnen worden beschouwd als mechanismen van cognitieve fouten. **Hoofdstuk 6 tot en met 8** onderzoeken interventies voor diagnosefouten en hun effectiviteit bij het verbeteren van de diagnostische nauwkeurigheid. Tot slot brengt **hoofdstuk 9** de resultaten van het proefschrift in verband met de huidige literatuur en ondersteunt het een update van het bestaande theoretische kader om cognitieve diagnosefouten te begrijpen. Ook worden implicaties voor toekomstig onderzoek en de praktijk besproken.

Het eerste doel van het proefschrift betrof het onderzoeken van cognitieve factoren die mechanismen voor cognitieve diagnosefouten kunnen zijn. In verschillende experimentele en observationele studies onderzochten we hoe de cognitieve processen van clinici verschilden in diagnoses waarin een fout optrad versus diagnoses waarin dit niet het geval was. Aangezien cognitieve processen niet direct waarneembaar zijn, probeerden we *bias* waar te nemen door cognitieve *bias* op te wekken via de onderzoeksopzet en onderzochten we mogelijke kennisgebreken door te meten hoe clinici gebruik maken van beschikbare klinische informatie.

Hoofdstuk 2 betrof een multicenter laboratoriumexperiment dat de hypothese onderzocht dat diagnosefouten voornamelijk worden veroorzaakt door cognitieve *bias* die voortkomt uit sneller diagnostisch redeneren.(4) Eerdere studies van Norman et al. (5) en Sherbino et al. (6) toonden al aan dat correcte diagnoses net zo snel, of zelfs sneller, werden gemaakt als incorrecte diagnoses. Hun resultaten lieten echter nog steeds de interpretatie toe dat snellere deelnemers gewoon betere diagnostici waren. We hebben deze experimenten daarom gerepliceerd in een studie waarin vergelijkingen tussen proefpersonen werden gemaakt. Artsen in opleiding tot specialist (AIOS) bij de interne geneeskunde diagnosticeerden klinische casussen die bedoeld waren om beschikbaarheidsbias op te wekken (met behulp van de methodologie van Mamede et al. (7)). Beschikbaarheidsbias treedt op wanneer iemand zich baseert op informatie die gemakkelijk herinnerd wordt, bijvoorbeeld wanneer de diagnose van een patiënt wordt beïnvloed door een soortgelijke patiënt die de clinicus eerder heeft gezien. De AIOS in ons experiment diagnosticeerden eerst casussen met diagnostische suggesties en diagnosticeerden vervolgens nieuwe casussen met diagnoses die leken op, maar verschillend waren van, de diagnostische suggesties die ze eerder tegenkwamen. De resultaten laten zien dat correcte diagnoses gemiddeld sneller werden gesteld dan incorrecte diagnoses. AIOS hadden ook meer vertrouwen in hun correcte diagnoses, hoewel hun algehele kalibratie van vertrouwen en nauwkeurigheid slecht was. Exploratieve resultaten toonden een trend aan waarin diagnoses sneller werden gesteld in casussen waar *bias*fouten voorkwamen vergeleken met andere soorten fouten, maar dit was niet sneller dan in casussen met de correcte diagnose. Dit kan wellicht aantonen dat vergelijkbare cognitieve processen ten grondslag liggen aan zowel correcte als aan foutieve diagnoses. In lijn met de eerdere literatuur suggereert dit experiment dat sneller redeneren plaatsvindt in zowel correcte als incorrecte diagnoses. Sneller redeneren is waarschijnlijk een waardevol onderdeel van het diagnostische proces en in vervolgonderzoek moeten oorzaken van diagnosefouten onafhankelijk van de snelheid van diagnose overwogen worden.

Afgezien van de intrinsieke factor van hoe snel een clinicus een diagnose kan stellen, heeft beschikbare klinische informatie ook de potentie om het redeneerproces van een clinici te beïnvloeden. **Hoofdstuk 3** onderzocht het effect van extern verstrekte diagnostische suggesties, specifiek van de suggestie die te vinden is in de verwijsbrief van de huisarts naar de spoedeisende hulp. In theorie kunnen incorrecte diagnostische suggesties leiden tot incorrecte interpretaties van klinische informatie of cognitieve *biases*, wat uiteindelijk zou resulteren in diagnosefouten. Coassistenten diagnosticeerden casussen die willekeurig een correcte diagnostische suggestie, een incorrecte diagnostische suggestie, of helemaal geen suggestie bevatten. De casussen leken op echte geanonimiseerde verwijsbrieven. Coassistenten die een diagnostische suggestie kregen, includeerden minder diagnoses in hun differentiaaldiagnose dan coassistenten die geen suggestie kregen. Echter, algehele diagnostische nauwkeurigheid, vertrouwen en tijd om te diagnosticeren werden niet beïnvloed. De coassistenten in dit onderzoek werden niet op het verkeerde spoor gezet door incorrecte diagnostische suggesties. Een analyse van de resultaten per casus onthulde dat het effect van diagnostische suggesties afhankelijk kan zijn van de specifieke klinische casus, aangezien de correcte diagnostische suggestie in sommige gevallen de diagnostische nauwkeurigheid verbeterde, maar in andere niet. De voorkennis van de coassistenten, of de mentale flexibiliteit die nodig is om een suggestie te overwegen, kunnen moderators zijn van dit effect. Samengevat, diagnostische suggesties kunnen het klinische redeneerproces in de correcte richting leiden of onnodig beperken. Het is waardevol voor onderwijzers om zich bewust te zijn van de mogelijke invloeden van diagnostische suggesties op de differentiaaldiagnose van hun studenten. Eventueel kan oefenen met casussen die zowel een als geen suggestie bevatten, helpen bij het leren opstellen van een brede differentiaaldiagnose.

Hoofdstuk 4 zoomde in op de informatieverwerking van klinische informatie van AIOS, door specifiek te meten welke informatie tijdens het stellen van een diagnose werd gebruikt en hoe dat verschilde tussen correcte en incorrecte diagnoses. Zelfs als alle benodigde informatie om een diagnose te stellen beschikbaar is, kunnen clinici nog steeds bij de verkeerde diagnose uitkomen. De hypothese hierbij is dat selectiviteit tijdens het verwerken van klinische informatie, bijvoorbeeld een incorrecte focus op een bepaalde diagnose, ertoe kan leiden dat bepaalde klinische informatie wel of niet overwogen wordt. Als gevolg hiervan kan er een diagnosefout gemaakt worden. AIOS interne geneeskunde en AIOS spoedeisende geneeskunde diagnosticeerden casussen met een correcte of een incorrecte werkdiagnose. Hun oogbewegingen werden geregistreerd terwijl ze de casus diagnosticeerden. We maten hoe lang en hoe vaak AIOS keken naar informatie die de correcte of incorrecte werkdiagnose ondersteunde en hoe dit verband hield met hun uiteindelijke diagnostische nauwkeurigheid. Over het algemeen verschilde hun informatieverwerking niet tussen correcte en incorrecte diagnoses. Er werd echter een interactie gevonden tussen de diagnostische nauwkeurigheid en de diagnostische suggestie, waarbij informatie die de correcte werkdiagnose ondersteunde vaker werd bekeken als assistenten een casus met een incorrecte werkdiagnose zagen, maar desondanks toch de correcte uiteindelijke diagnose vonden. Deze interactie leek vergelijkbaar voor zowel hoe lang als hoe vaak AIOS naar bepaalde informatie keken, maar was alleen significant voor hoe vaak AIOS keken. Hun vertrouwen en tijd om te diagnosticeren, de andere aspecten van het diagnostische proces, verschilden niet. Hoewel deze interactie suggereert dat AIOS die zich meer op relevante informatie richtten in staat waren om de *bias* van de suggestie te overwinnen, vonden we geen specifieke selectiviteit in de informatieverwerking bij incorrecte diagnoses. Daarom is selectiviteit wellicht meer een indicatie van veranderingen in de cognitieve processen die verband houden met het diagnostisch proces, dan een directe oorzaak van fouten. Dit experiment suggereert dat het vermogen om de relevante informatie uit een casus te herkennen belangrijk is om tot de correcte diagnose te komen en kan helpen om *bias* te overwinnen.

Om ons begrip van hoe informatie verwerkt wordt bij diagnose verder aan te vullen, rapporteert **Hoofdstuk 5** over een observationele studie die de diagnostische justificatie van medische studenten heeft gemeten, oftewel hun vaardigheid in het bepalen of klinische informatie de waarschijnlijkheid van een bepaalde diagnose verhoogt of verlaagt. Dit gaat verder dan enkel het onderzoeken welke informatie wordt gebruikt en laat zien hoe studenten deze informatie toepassen bij het opstellen van hun differentiaaldiagnose en de uiteindelijke selectie van een definitieve diagnose. Het verkeerd toewijzen van bepaalde klinische informatie aan diagnoses kan een oorzaak zijn van diagnosefouten. Eerste- en tweedejaars medische studenten diagnosticeerden casussen in een online leeromgeving waar ze specifieke informatie konden selecteren om aan te geven of dit de waarschijnlijkheid van een diagnose in hun differentiaaldiagnose verhoogde of verlaagde. Ze moesten hetzelfde doen voor onderzoeken die ze aanvroegen. Aan het einde van de casus gaven ze een definitieve diagnose vanuit hun differentiaaldiagnose. Ondanks dat de studenten goed presteerden wat betreft hun uiteindelijke diagnostische nauwkeurigheid en het creëren van een initiële differentiaaldiagnose, presteerden ze slecht op het aanvragen van onderzoeken en diagnostische justificatie. Hun hoge diagnostische nauwkeurigheid kan gedeeltelijk worden verklaard doordat studenten vaak casussen moeten diagnosticeren over ziektes waar ze op dat moment over leren in het curriculum, waardoor het gemakkelijker wordt om de diagnose van de casus te raden zonder ook een goede diagnosticus te zijn. (8) Studenten rapporteerden consistent weinig klinische informatie, vooral informatie die de waarschijnlijkheid van een bepaalde diagnose verlaagde. Dit patroon verschilde niet tussen correcte en incorrecte diagnoses, hoewel deze relatie mogelijk werd vertekend door de hoge uiteindelijke diagnostische nauwkeurigheid van studenten, of door de lage hoeveelheid klinische informatie die überhaupt aan diagnoses werd toegewezen. Deze studie toont aan dat tekortkomingen in diagnostische justificatie al vroeg bij studenten aanwezig zijn, en onderzoek met studenten die verder zijn in hun studie laat zien dat deze tekortkomingen nauwelijks worden verholpen tijdens hun coschappen.(8) De diagnostische justificatie van studenten verbeterde naarmate ze meer oefenden met de casussen, mogelijk door de specifieke feedback die ze kregen na elk vignet. Het goed ontwikkelen van de diagnostische justificatie van studenten is waarschijnlijk cruciaal voor het bevorderen van goede diagnostische redeneervaardigheden.

Samen suggereren deze hoofdstukken dat diagnosefouten niet ontstaan als een direct gevolg van hoe snel of langzaam clinici een diagnose stellen, hun zelfvertrouwen of informatieverwerking, aangezien deze processen vergelijkbaar zijn in zowel correcte diagnoses als fouten door *bias*. Echter lijkt de kennis die nodig is om tot de correcte diagnose te komen cruciaal te zijn, omdat dit waarschijnlijk clinici in staat stelt om fouten te detecteren in een incorrecte diagnostische suggestie- of in hun eigen diagnostisch proces.

Het tweede doel van dit proefschrift was om de effectiviteit van verschillende veelbelovende interventies, met als doel om cognitieve fouten te verminderen, te evalueren. We hebben experimentele studies uitgevoerd naar de effectiviteit van diagnostische checklists en feedback en hebben het beschikbare experimentele bewijsmateriaal over de effectiviteit van cognitieve redenatie *tools* (dat wil zeggen, elke *tool* gericht op het ondersteunen of verbeteren van het cognitieve redenatieproces van clinici tijdens diagnose) samengesteld in een systematische review en meta-analyse.

Hoofdstuk 6 vergeleek het effect van checklist-interventies op het verminderen van diagnosefouten in normale en abnormale elektrocardiogrammen (ECG's). Omdat de effectiviteit van interventies voor het verminderen van diagnosefouten vaak alleen wordt getest met abnormale casussen, is het onbekend hoe normale casussen worden beïnvloed. In Nederland gebruiken huisartsen vaak checklists om ECG's te interpreteren en te beslissen of de patiënt al dan niet moet worden doorverwezen naar het ziekenhuis.(9) Ongeveer een derde van de ECG's die ze tegenkomen is normaal, een derde is simpel abnormaal, en een derde is complex abnormaal.(10) Mogelijk kunnen checklists leiden tot onnodig verbruik van middelen, overmatig testen of overdiagnostiek, omdat checklists ervoor kunnen zorgen dat clinici casussen overmatig gaan onderzoeken, zelfs als er niets te vinden is.(11) Om dit te testen, lieten we AIOS bij de huisartsgeneeskunde om ECG's te diagnosticeren met en zonder een checklist. We onderzochten zowel proces (of *debiasing*) checklists als inhoudschecklists.(11, 12) Processchecklists richten zich op het niet missen van relevante stappen in het diagnostische proces of op het expliciet controleren op *bias* en fouten in redenering, terwijl inhoudschecklists zich richten op het activeren van eerdere kennis en het verzamelen van alle relevante informatie. De helft van de AIOS gebruikte de processchecklist, de andere helft gebruikte de inhoudschecklist. Checklists hadden geen verschillend effect op de het diagnostisch proces van AIOS voor normale, simpele en complexe abnormale ECG's. Het diagnostisch proces verschilde ook niet tussen de processchecklist en de inhoudschecklist. Er was een trend richting verbeterde nauwkeurigheid waargenomen, maar dit was niet significant. Interessant genoeg veranderde het vertrouwen van de AIOS in hun diagnose niet, wat resulteerde in een algemene verbetering in kalibratie van vertrouwen en nauwkeurigheid. AIOS werden minder overmoedig bij het gebruik van de checklist. Samengevat kunnen checklists waarde bieden door kalibratie te verbeteren zonder te leiden tot overdiagnostiek op zowel normale als abnormale ECG's.

Een andere veelbelovende interventie voor diagnosefouten is feedback. Vaak kunnen interventies niet worden geïmplementeerd, omdat het te veel tijd zou kosten om ze voor elke patiënt te gebruiken. De oplossing zou zijn om interventies alleen te gebruiken voor moeilijke casussen die extra ondersteuning nodig hebben, maar de kalibratie van het vertrouwen in hun diagnose en de echte nauwkeurigheid van clinici is opmerkelijk slecht.(13) **Hoofdstuk 7** onderzocht of feedback de kalibratie van clinici kon verbeteren. Coassistenten diagnosticeerden röntgenfoto's van de thorax in een leerfase, gevolgd door een testfase. Ze kregen ofwel geen feedback, ofwel feedback of hun gekozen diagnose correct was (prestatiefeedback), ofwel feedback over of hun gekozen diagnose correct was plus een uitleg over hoe de correcte diagnose kon worden herkend (informatiefeedback). Beide soorten feedback verbeterden de diagnostische nauwkeurigheid en het vertrouwen in de diagnose, vergeleken met de controlegroep. Als gevolg hiervan verbeterde de algemene kalibratie van de coassistenten ook. Er was geen verschil in tijd besteed aan de diagnose. Een exploratieve analyse toonde aan dat kalibratie vooral verbeterde voor simpele röntgenfoto's, maar niet voor complexe röntgenfoto's. Over het algemeen was feedback waardevol als coassistenten de juiste kennis of vaardigheidsniveau hadden om een röntgenfoto te diagnosticeren, terwijl feedback geen voordeel had voor casussen die te complex waren voor hun niveau. De verbetering in kalibratie van de coassistenten kan worden verklaard door een verbetering in hun inzicht in hun prestaties, maar anderzijds is het ook mogelijk dat de nauwkeurigheid van de coassistenten verbeterde als gevolg van de feedback, terwijl hun niveau van vertrouwen stabiel bleef. Dit suggereert dat, naarmate de diagnostische nauwkeurigheid verbeterde, het vertrouwen niet in gelijke mate toenam en dat eerdere overmoed werd gecorrigeerd.

De systematische review en meta-analyse in **Hoofdstuk 8** had als doel de effectiviteit van cognitieve redenatie *tools* bij het verbeteren van diagnostische nauwkeurigheid te onderzoeken. Bestaande reviews hebben interventies die gericht zijn op het verbeteren van het diagnostisch proces in een educatieve setting samengevoegd met interventies gericht op het verbeteren van het diagnostisch proces op de werkplek. (14-18) Deze review had als doel de effectiviteit van cognitieve redenatie *tools* op de werkplek in te schatten en te onderzoeken welke studie- of interventiekenmerken geassocieerd waren met hogere effectiviteit. Drie studies werden uit de meta-analyse verwijderd omdat de deelnemers uitgebreide training met de tool kregen voordat de diagnostische nauwkeurigheid werd gemeten, wat resulteerde in een grote verbetering van de diagnostische nauwkeurigheid. De overgebleven studies lieten een kleine verbetering van de diagnostische nauwkeurigheid zien. Er werden geen specifieke kenmerken ontdekt die verband hielden met een hogere diagnostische nauwkeurigheid, hoewel het moeilijk was om de effecten van specifieke kenmerken te isoleren vanwege de vele potentiële invloeden op de effectiviteit van de *tool*. Het moet in gedachten worden gehouden dat de bewijsbasis onder deze meta-analyse relatief beperkt was en dat er meer onderzoek nodig zal zijn om te begrijpen onder welke omstandigheden cognitieve redenatie *tools* het meest effectief zijn. Bovendien waren de meeste onderliggende studies laboratoriumexperimenten, dus replicatie in de praktijk zal ook nodig zijn.

Alles bij elkaar genomen lijken cognitieve redenatie *tools* over het algemeen succesvol te zijn bij het verbeteren van diagnostische nauwkeurigheid, zonder afbreuk te doen aan andere aspecten van het diagnostische proces. Hoewel het algehele effect klein is, kan zelfs een kleine verbetering een klinisch relevant effect hebben op patiëntveiligheid gezien de hoge prevalentie van diagnosefouten.

Hoofdstuk 9 heeft de eerder besproken bevindingen samengevat en verbonden aan een theorie ontwikkeld door Stanovich.(19) Dit proefschrift stelt dat verschillende aspecten van cognitieve processen (bijvoorbeeld de tijd die nodig is om een patiënt te diagnosticeren en het vertrouwen in de diagnose) of de klinische informatie die wordt gebruikt om tot een diagnose te komen niet verschillen tussen correcte en incorrecte casussen. We observeerden alleen veranderingen in informatieverwerking wanneer een incorrecte diagnostische suggestie werd gecorrigeerd. Het vermogen om een incorrecte suggestie te detecteren, in combinatie met de juiste kennis om tot de correcte diagnose te komen, is waarschijnlijk cruciaal om diagnosefouten te verminderen en zou kunnen aanduiden welke cognitieve processen beïnvloed worden wanneer een diagnosefout wordt gemaakt.

Deze bevindingen kunnen geïnterpreteerd worden in het kader van de *mindware*-theorie van Stanovich (19), welke stelt dat naast cognitieve redenatieprocessen ook beschikbare kennisstructuren van vitaal belang zijn voor uiteindelijke prestatie. De theorie kan diagnosefouten zowel aan de hand van cognitieve *biases* als kennistekorten verklaren. *Mindware* wordt gedefinieerd als de beschikbaarheid en integratie van informatie die relevant is voor een taak en dit wordt gezien als een bepalende factor voor taakprestatie. Tekortkomingen in *mindware* kunnen zowel leiden tot cognitieve *biases* als kennistekorten. In de geneeskunde bestaat *mindware* uit scripts, voorbeelden en prototypen van verschillende ziektebeelden (20, 21) waarin medische kennis is gecodeerd. Als deze kennis goed beschikbaar is, zal diagnostisch redeneren waarschijnlijk succesvol zijn. Als er een verkeerde diagnose wordt gesteld, maar de *mindware* goed beschikbaar is, kan de fout nog steeds worden gedetecteerd en gecorrigeerd. Als de *mindware* slecht beschikbaar is, zal de fout waarschijnlijk onopgemerkt blijven. De *mindware* theorie is voornamelijk gebaseerd op onderzoek dat gebruik maakt van taken met een duidelijk correcte oplossing, terwijl medische diagnostiek een taak met inherente onzekerheid is. De theorie kan worden toegepast op medische diagnostiek, maar er moet rekening mee gehouden worden dat zowel het detecteren als het corrigeren van fouten moeilijker is vanwege deze onzekerheid.

Dit proefschrift toonde verder aan dat cognitieve redenatietools gericht op het verbeteren van de diagnostische prestaties op de werkvloer een kleine, maar klinisch relevante toename van de diagnostische nauwkeurigheid lieten zien. In het perspectief van de *mindware* theorie kan worden verondersteld dat deze tools de *mindware* van clinici verbeteren door ondersteuning te bieden voor een bepaalde taak (bijvoorbeeld door het suggereren van diagnoses voor bepaalde symptomen die een patiënt kan hebben) of door het ondersteunen van algemene redeneringsprocessen (bijvoorbeeld door het bieden van een gestructureerd format voor redeneren). Er is echter meer onderzoek nodig om te bepalen onder welke omstandigheden clinici het meeste baat hebben bij deze tools.

De bevindingen in dit proefschrift, wanneer in verband gebracht met de literatuur, suggereren dat de *mindware* van clinici en hun vermogen om die *mindware* te gebruiken belangrijk zou moeten zijn in ons streven om diagnosefouten te verminderen. Het verschil tussen de bestaande en noodzakelijke *mindware* voor een taak kan een belangrijke oorzaak zijn van diagnosefouten. Om dit verschil te verkleinen, zal het nodig zijn om de beschikbare *mindware* van clinici te verbeteren. Zo kan het bijvoorbeeld een eerste stap zijn om te oefenen met een breed scala aan ziekten en ziektepresentaties om meer uitgebreide en beter beschikbare kennisstructuren op te bouwen. Bovendien kunnen de algemene redenatieprocessen van clinici ook worden verbeterd door middel van oefening met casuïstiek of kritisch denken. Het is echter belangrijk om in te zien dat niet alle verbeteringen van *mindware* kunnen worden gerealiseerd binnen het cognitieve vermogen van een clinicus, omdat medische kennis te uitgebreid is om door één persoon te worden geleerd en gebruikt. Het is essentieel om te beseffen dat het concept van “beschikbare kennis” verder gaat dan wat er tijdens de medische opleiding is geleerd. Technologie biedt externe bronnen van informatie en ondersteuning, zoals de elektronische gezondheidsdossiers van patiënten, elektronische triggers of kunstmatige intelligentie. Het succesvol implementeren van dergelijke interventies in de workflow zal clinici waarschijnlijk de mogelijkheid bieden om effectief gebruik te maken van de beschikbare kennis. Belangrijk is dat elektronische ondersteuning niet bedoeld is als vervanging van clinici, maar als aanvulling op hun redenatieprocessen.

Samenvattend biedt dit proefschrift inzicht in de oorzaken van cognitieve diagnosefouten en interventies die deze fouten kunnen voorkomen. Toch zijn cognitieve fouten slechts een deel van de puzzel. Oorzaken van fouten zijn vaak multifactorieel en hoewel het verminderen van cognitieve fouten veelbelovend is, zal het slechts een deel van het probleem aanpakken. Andere veranderingen en interventies zullen nodig zijn om diagnosefouten te verminderen zoals ze voorkomen in de kliniek. Bovendien kunnen cognitieve fouten waarschijnlijk nooit volledig worden uitgeroeid en zouden interventies voor andere oorzaken van diagnosefouten of manieren om voor de cognitieve gebreken van clinici te compenseren, moeten worden overwogen. Het is waarschijnlijk simpeler om veranderingen in het systeem te implementeren dan om cognitieve gebreken bij individuele clinici aan te pakken. Het voorkomen van diagnosefouten moet gericht zijn op, en worden aangepast aan, de diagnostische omgeving als geheel.

References

1. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. 2015. Washington, DC: The National Academies Press.
2. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. European journal of internal medicine. 2013;24(6):525-9.
3. Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. BMJ quality & safety. 2013;22(Suppl 2):ii58-ii64.
4. Norman G, Sherbino J, Dore K, Wood T, Young M, Gaissmaier W, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. Academic Medicine. 2014;89(2):277-84.
5. Sherbino J, Dore KL, Wood TJ, Young ME, Gaissmaier W, Kreuger S, et al. The relationship between response time and diagnostic accuracy. Academic Medicine. 2012;87(6):785-91.
6. Mamede S, van Gog T, van den Berge K, Rikers RMJP, van Saase JLCM, van Guldener C, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. Jama. 2010;304(11):1198-203.
7. Williams RG, Klamen DL. Examining the diagnostic justification abilities of fourth-year medical students. Academic Medicine. 2012;87(8):1008-14.
8. Konings K, Willemsen R. ECG 10+: Systematisch ECG's beoordelen. Huisarts en wetenschap. 2016;59(4):166-70.
9. Rutten FH, Kessels AGH, Willems FF, Hoes AW. Is elektrocardiografie in de huisartspraktijk nuttig? Huisarts en wetenschap. 2001;44(11):179-83.
10. Zwaan L, Staal J. Evidence on Use of Clinical Reasoning Checklists for Diagnostic Error Reduction. AHRQ Papers on Diagnostic Safety Topics [Internet]. 2020; (3).
11. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: A randomised experiment. Medical education. 2021;55(10):1172-82.
12. Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine. 2013;173(21):1952-8.
13. Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. Medical Teacher. 2019;41(5):517-24.
14. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ quality & safety. 2016;25(10):808-20.
15. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: Still no easy answers. Western Journal of Emergency Medicine. 2020;21(6):125.
16. Mamede S, Schmidt HG. Reflection in medical diagnosis: a literature review. Health Professions Education. 2017;3(1):15-25.
17. Astik GJ, Olson APJ. Learning from Missed Opportunities Through Reflective Practice. Critical Care Clinics. 2022;38(1):103-12.
18. Stanovich KE. Miserliness in human cognition: The interaction of detection, override and mindware. Thinking & Reasoning. 2018;24(4):423-44.
19. Schmidt HG, Boshuizen H. On acquiring expertise in medicine. Educational psychology review. 1993;5(3):205-21.
20. Nosofsky RM. Exemplars, prototypes, and similarity rules. Essays in honor of William K Estes. 1992;1:149-67.

Dankwoord

Dankwoord Checklist

(Acknowledgements in Dutch)

Voor deze quiz was geen toestemming van de Medisch Ethische Commissie nodig. Deelnemers blijven anoniem en antwoorden worden niet opgeslagen. De quiz is in lijn met de relevante voorschriften van de Declaratie van Helsinki. Meedoen is vrijblijvend. Deelnemers mogen stoppen op ieder moment van de quiz zonder verdere toelichting te geven. De quiz is samengesteld op basis van expert opinion, maar niet gevalideerd. De uitslagen kunnen mogelijk afwijken van het echte resultaat. Voor een kwalitatieve beoordeling van deze data verwijs ik u graag naar het volledige Dankwoord.

1) Heb je Justine ondersteund bij het realiseren van dit proefschrift?

a. Ik was een cheerleader, misschien wat meer aan de zijlijn, maar wel betrokken
b. Ik heb bijdrage geleverd aan specifieke onderdelen van het proefschrift
c. Ik was bij alle, of de meeste, onderdelen betrokken van begin tot eind

2) Kon Justine voor niet-proefschrift gerelateerde zaken bij jou terecht?

a. Ik was *in spirit* bij Justine
b. We hebben productieve gesprekken gehad
c. Justine mocht altijd bij mij aankloppen

3) Hoe gepassioneerd ben je over het onderwerp van dit proefschrift?

a. Waar gaat dit proefschrift ook alweer over?
b. Heel interessant, maar niet geschikt voor *binge-reading*
c. Ik had zelf dit proefschrift willen schrijven

4) Heb je aan Justine's experimenten meegewerkt?

a. Nee, ik viel buiten de inclusiecriteria
b. Ja, ik heb wel eens meegedacht
c. Ja, ik heb bijgedragen aan het bestaan van de experimenten – bijv. als deelnemer (geen zorgen, deze quiz is anoniem)

5) Vond je deze quiz leuk?

a. Ja?
b. Ja.
c. Ja!

Dit is het einde van deze vragenlijst. Sla de pagina om voor de uitslag.

Uitslag

A = 0 punten; B = 1 punt; C = 2 punten.

0-3 punten
Dank voor je bijdrage, alle hulp werd gewaardeerd!

4-6 punten
Bedankt voor je waardevolle input, dit proefschrift is alleen maar beter geworden door je bijdrage!

7-10 punten
Dit proefschrift is mede mogelijk gemaakt door jou, ontzettend bedankt!

Dankwoord

Acknowledgements in Dutch

"My journey took me somewhat further down the rabbit hole than I intended and though I dirtied my fluffy white tail I have emerged, enlightened." – Sherlock (2009)

Tot mijn (onterecht) grote verrassing was mijn PhD traject niet alleen een weg naar academische ontwikkeling, maar ook naar persoonlijke ontwikkeling. Mijn doel was om de vaardigheden van een goede onderzoeker op te doen en me te verdiepen in het veld van patiëntveiligheid. Dat is ook zeker gelukt in de afgelopen jaren, maar ik had nooit voorzien hoeveel ik persoonlijk zou leren en groeien in deze tijd. Graag wil ik nu een moment nemen om iedereen te bedanken die me in deze periode geholpen en gesteund heeft. Door jullie is de weg naar het afronden van mijn proefschrift een mooie en leerzame tocht geworden. In het bijzonder wil ik de volgende personen bedanken.

Allereerst dr. Laura Zwaan, dr. Jelmer Alsma, prof. dr. Walter van den Broek en prof. dr. Maarten Frens voor de begeleiding tijdens mijn PhD traject. Het was een eer om van jullie te mogen leren.

Laura, bedankt voor alle ondersteuning die je hebt geboden, zoals onze wekelijkse meetings en de projecten waar we samen over gebrainstormd hebben. Samen hebben we allerlei verschillende soorten studies kunnen doen op allerlei onderwerpen, waar ik veel van geleerd heb. Via jou heb ik kennis mogen maken met dit interessante veld en heb ik onderzoek kunnen uitvoeren dat bijdraagt aan beter begrijpen van diagnosefouten, een ontzettend belangrijk onderwerp. Ook hebben we het gezellig kunnen hebben op congressen en trainingsdagen en hoefde het nooit altijd over werk te gaan. Ik ben je ontzettend dankbaar voor de ruimte die je me gegeven hebt om te groeien; hierdoor heb ik veel meer uit mijn PhD traject kunnen halen dan inhoudelijke expertise. Ik ben vereerd met een onderzoeker zoals jij samengewerkt te hebben!

Jelmer, bedankt voor je inhoudelijke expertise en je nuchtere blik. Je hebt veel creativiteit aangesproken om ingewikkelde *bias* casuïstiek voor dit project te schrijven en dan toch was er altijd nog creativiteit over om interessante onderzoeksvragen te bedenken. Ik ben je dankbaar voor alle 0/0.5/1 scores en *regions of interest* die je voor dit project hebt gedefinieerd. Verder was het erg leerzaam om met je mee te mogen kijken op de spoedeisende hulp, waar het me veel duidelijker is geworden hoe het diagnostisch proces echt in de praktijk plaatsvindt – in plaats van hoe we in theorie denken dat het werkt. Je hebt me veel geleerd over de klinische en praktische kanten van dit onderzoek en ook kon ik altijd rekenen op je feedback. Wie weet kunnen we dat onderzoeksidee over of de presentatie van een patiënt, bijvoorbeeld op een bed of op een stoel, of met en zonder bloeddrukmeter, invloed heeft op het diagnostisch proces nog eens met een masterstudent uitvoeren.

Walter, bedankt voor de steun en inzichten die je geboden hebt. Het is verleidelijk om verloren te raken in details, maar met jouw hulp was het altijd weer mogelijk om uit te zoomen en het project in zijn geheel te bekijken. Ook ben ik dankbaar voor de mogelijkheid die ik heb gekregen om deel te nemen aan de *Female Talent Class*, wat een belangrijk onderdeel van mijn persoonlijke groei is geweest. En natuurlijk bedankt voor je hulp bij het begeleiden en beoordelen van de masterstudenten die bij ons hun scriptie-onderzoek hebben gedaan en bij allerlei andere administratieve zaken. Door jou heb ik altijd overzicht kunnen houden.

Maarten, bedankt voor je expertise en kritische inzichten. Een *eye-tracking* onderzoek opzetten was een uitdaging op zich en ik ben blij dat ik van jouw kennis gebruik mocht maken. En natuurlijk van het *eye-tracking* lab. Ook heb ik veel gehad aan je ondersteuning tijdens het project in zijn geheel, bijvoorbeeld doordat je hielp met op koers blijven en de hoofd- en bijzaken van mijn PhD traject te scheiden. Je kennis over de academische wereld en universitaire bureaucratie was ook onmisbaar.

De leden van mijn promotiecommissie, dr. Fop van Kooten, prof. dr. ir. Lex Burdorf en prof. dr. Daniëlle Timmermans, wil ik bedanken voor het kritisch lezen en beoordelen van dit proefschrift en het deelnemen aan de oppositie. Daarnaast wil ik de overige leden van de promotiecommissie, dr. Maarten van Aken, dr. Wolf Hautz en prof. dr. Gerda Croiset, bedanken voor hun inzichten en gedachtewisseling over dit proefschrift.

Graag wil ik het Erasmus Medisch Centrum (Erasmus MC) bedanken, waar dr. Laura Zwaan via een Erasmus MC Fellowship de mogelijkheid kreeg om dit project naar de cognitieve mechanismen achter diagnosefouten uit te voeren.

Daarnaast wil ik ook de mensen bedanken die hebben bijgedragen aan de totstandkoming van dit proefschrift. Hoewel ik vol trots “mijn” proefschrift mag zeggen, is werk in de wetenschap altijd een product waar vele anderen aan bijdragen. Ik heb de steun en hulp van veel mensen mogen genieten tijdens dit hele traject. In het bijzonder wil ik de volgende mensen bedanken.

iMERR collega’s, bedankt voor de warme en gezellige omgeving waar ik mijn PhD mocht beginnen. Hoewel we een relatief klein clubje zijn vergeleken met andere afdelingen, is er veel expertise in huis en is iedereen bereid om te helpen, te brainstormen, of mentale steun te bieden. Ook een speciaal bedankje voor mijn lieve PhD collega's, in het bijzonder Jacky, Josepha, Ligia, Wendy, Tjitske, Suzanne, Inge en Vera. Op kantoor hebben we een leuke club gecreëerd, waar veel persoonlijke aandacht en hulp was en iedereen er voor elkaar is. Ik voel me bevoorrecht dat ik met jullie projectmanagement heb mogen doen, intervisies heb kunnen houden, vooral in de pandemie elke ochtend met de dagstart door jullie gemotiveerd kon worden, en op kantoor de Feestkamer gezellig heb kunnen maken. Hopelijk kunnen we een traditie maken van het PhD kerstdiner! Graag wil ik Jacky ook apart bedanken; wij zijn al collega's sinds we allebei onderzoeksassistent waren voor Josepha en daarna ben jij ook je PhD op het onderwerp van patiëntveiligheid komen doen! En natuurlijk ben ik erg blij dat je mijn paranimf wil zijn. Soms ben ik nog wel eens in de war als ik over mijn computermonitors heen kijk en jou niet zie zitten. Bedankt voor alle gezellige middagen en gesprekken en hopelijk kunnen we meer games vinden om samen te spelen! Also, dear Ligia, here a shout-out that you will be able to read! Thank you for the great conversations and book recommendations, and your support. You did a great job and I hope to see you again soon.

Also a huge thank you to all co-authors and collaborators on the various projects I could be a part of. Your insights and help were invaluable! Hopefully we will get to collaborate again in the future.

Binnen het Erasmus MC zijn er ook veel collega's die ik graag zou willen bedanken. Jos van der Geest, bedankt voor je hulp en *troubleshooting* bij de eye-tracking studie, en je goede advies. Lizzy Boonen en de andere geweldige vrouwen van de *Female Talent Class*, wat geweldig dat jullie dit organiseren en bedankt voor alles wat ik van jullie heb mogen leren. Miranda, Marja en Rita, en de collega's van het OBA secretariaat die nu werkzaam zijn, bedankt voor al jullie ondersteuning bij het regelen van administratieve zaken, ruimtes reserveren, agenda's volplannen, of obscure regelgevingen vinden. OBA collega's, we hebben wellicht geen onderzoek samen gedaan, maar bedankt voor de gezellige tijd die we samen op kantoor hebben doorgebracht! Verder wil ik specifiek onze iMERR stagiaires Maaike, Kamya en Anne bedanken voor de leuke samenwerkingen aan jullie scriptieprojecten en de interessante studies. En natuurlijk bedankt aan alle opleiders, clinici, AIOS en medisch studenten die de onderzoeken in dit proefschrift mogelijk hebben gemaakt.

Daarnaast wil ik ook graag mijn collega's mij de Medische Informatica noemen. Peter, Nikkie, Lieke, Renske en Lana, bedankt dat jullie vertrouwen hadden dat ik mijn proefschrift af kon maken naast het lesgeven en me de kans hebben gegeven bij het team aan te sluiten. Ik kijk ernaar uit om te blijven leren bij jullie!

Mijn familie en vrienden wil ik bedanken voor hun algehele steun en het delen van successen en teleurstellingen. Jullie zijn onmisbaar!

Mijn ouders en zus, Dorien, jullie hebben onbewust van alles kunnen leren over diagnosefouten en de academische wereld, wie weet hebben jullie er ooit nog wat aan. Bedankt voor jullie begrip en steun. Dorien, bedankt dat je mijn paranimf bent! Ik weet dat jij meer gestrest gaat zijn voor mij dan ik zelf, dus dan kan ik me wat meer relaxen.

Mijn vrienden die alles hebben mogen aanhoren en in ieder geval weten dat ik, hoe dan ook, mijn best heb gedaan, bedankt voor alles! Specifiek wil ik de volgende personen bedanken. Mijn businesspartners in Appel-Kiwi aandelen, Kevin, Rosemarijn, Linda, Gwain, Amy en Dylan. Mijn universiteits- en DnD helden, Nina, Emma, Daphne, Sakshi en Maud (en Pien). De ontwerper van de voorkant van dit proefschrift (maar veel meer mijn lieve vriendin), Annemieke. Mijn persoonlijke investeerder, Jessica: De Leukste herinneringen hebben we lachend op de grond of in de auto gemaakt.

Bedankt aan iedereen, door jullie is dit een prachtig PhD traject geworden!

Justine, 2023

Curriculum Vitae

Justine Staal was born in Goes, the Netherlands, on the 31st of August in 1995. She completed her secondary education at the Pieter Zeeman Lyceum in Zierikzee in 2013 and continued on to study Psychology at the Erasmus University Rotterdam, where she obtained her bachelor's degree in 2016. From there, she started the Research Master Neuroscience at the Erasmus Medical Center in Rotterdam, where she obtained her master's degree in 2018. In May 2018, she became a PhD candidate at the Institute of Medical Education Research Rotterdam (iMERR) at the Erasmus Medical Center in Rotterdam, given her interest in cognitive psychology and reasoning. This project concerned possible causes and solutions for cognitive diagnostic errors in clinicians' reasoning processes. It was funded from an Erasmus Medical Center Fellowship obtained by dr. Laura Zwaan, who also was the daily supervisor of the project. During her PhD term, Justine collaborated with experts in the field of clinical reasoning and medical education and regularly attended various conferences and symposia associated with diagnostic errors and reasoning. She also fostered her personal interest in open science and scientific methodologies by yearly attending the conference hosted by the Society of Psychological Science (SIPS) and collaborating on open science projects with other SIPS members. She furthermore got the opportunity to develop her personal skills and career goals in the Female Talent Class hosted by the Erasmus Medical Center Rotterdam In addition, she worked as an teaching assistant for scientific and statistical courses at Leiden University, supervised medical students' master theses, organized several events such as symposia, peer-reviewed articles for journals in the medical field. Justine is currently working as a teacher in Academic Skills at the Erasmus Medical Center Rotterdam, focusing on teaching medical students to understand and utilize scientific publications to formulate appropriate management plans for their patients.

List of publications

Hoogerheide V, Staal J, Schaap L, & van Gog T. Effects of study intention and generating multiple choice questions on expository text retention. Learning and Instruction. 2019; 60: 191-198.

Zwaan L, Staal J. Evidence on use of clinical reasoning checklists for diagnostic error reduction. Agency for Healthcare Research and Quality. 2020: 1-2.

Staal J, Mattace-Raso F, Daniels HAM, van der Steen J, & Pel J. To explore the predictive power of visuomotor network dysfunctions in mild cognitive impairment and Alzheimer's disease. Front. Neurosci. 15:654003. doi: 10.3389/fnins.2021.654003.

Mamede S, Goeijenbier M, Schuit SC, de Carvalho Filho MA, Staal J, Zwaan L & Schmidt HG. Specific disease knowledge as predictor of susceptibility to availability bias in diagnostic reasoning: a randomized controlled experiment. J Gen Intern Med. 2021; 36: 640-646.

Staal J, Alsma J, Mamede S., Olson APJ, Prins-van Gilst G, Geerling SE, Plesac M, Sundberg MA, Frens MA, Schmidt HG, van den Broek WW, & Zwaan L. The relationship between time to diagnose and diagnostic accuracy among internal medicine residents: a randomized experiment. BMC Med Educ. 2021; 21: 227. https://doi.org/10.1186/s12909-021-02671-2.

Staal J, Speelman M, Brand R, Alsma J, &Zwaan L. Does a suggested diagnosis in a general practitioners' referral question impact diagnostic reasoning: an experimental study. BMC Med Educ. 2022; 22: 256. https://doi.org/10.1186/s12909-022-03325-7.

Baum AU, Hart A, Elsherif, MM, Ilchovska ZG, Moreau D, Dokovova M, LaPlume AA, Krautter K, & Staal J. Research without borders: how to identify and overcome potential pitfalls in international large-team online research projects. SAGE Research Methods: Doing Research Online (SAGE Research Methods Cases). https://doi.org/10.4135/9781529602074

Staal J, Hooftman J, Gunput STG, Mamede S, Frens, MA, van den Broek WW, Alsma J, & Zwaan L. Effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis. BMJ Qual Saf. 2022; 31(12): 899-910.

Staal J, Zegers R, Caljouw-Vos J, Mamede S, & Zwaan L. Impact of diagnostic checklists on the interpretation of normal and abnormal electrocardiograms. Diagnosis. 2022; https://doi.org/10.1515/dx-2022-0092.

Staal J, Katarya K, Speelman M, Brand R, Alsma J, Sloane J, van den Broek WW, & Zwaan L. Impact of performance and information feedback on medical interns' confidence-accuracy calibration. Adv in Health Sci Educ. 2023; https://doi.org/10.1007/s10459-023-10252-9.

Staal J, Waechter J, Allen J, Hee Lee C, & Zwaan L. Deliberate practice of diagnostic clinical reasoning reveals low performance and improvement of diagnostic justification in pre-clerkship students. Submitted.

Staal J, Alsma J, Mamede S, Jansen E, van den Broek WW, Frens MA, & Zwaan L. Selectivity in information processing in correct and incorrect diagnosis: a randomized controlled eye-tracking experiment. Submitted.

Portfolio

Summary of PhD training and activities	Date	Workload (ECTS) 1 EC = 28 uur
Courses		
Systematische reviews: van literatuur search naar literatuurreview (workshop)	December 04, 2018	0.3
Improving your statistical inferences (Coursera, online course)	February, 2019	0.7
Bayesian Statistics: From Concept to Analysis (Coursera, online course)	May, 2019	0.7
Cursus 'Wetenschappelijke Integriteit'	April 18, 2019	0.3
How to write and publish a scientific paper	May, 2019	0.4
Introduction to systematic review and meta-analysis	May-June, 2019	1.3
Improving your statistical questions (Coursera, online course)	August, 2019	0.7
Writing in the sciences	August-September, 2019	1.4
Systematic literature retrieval (in Pubmed)	September, 2019	0.3
English grammar and punctuation	July, 2020	0.4
Planning and project managing course	December, 2020	0.3
ECG beoordelen van ritmes (cursus via Eduplaza EMC)	January, 2021	0.2
Female talent class (EMC)	April-October, 2021	1.2
Open Science course	May 23-24, 2022	0.6
Summary of PhD training and activities	**Date**	**Workload (ECTS) 1 EC = 28 uur**
Symposia, Conferences, Workshops, Seminars		
Helmholtz retreat	June 27-29, 2018	0.9
Diagnostic Error in Medicine conference	August 30/31, 2018	0.6
RIME conference, Copenhagen	May 23-24, 2019	0.6
NVMO PhD Day	April 12, 2019	0.3
Symposium "Understanding Diagnostic Error"	August 15, 2019	0.3
12th Diagnostic Error in Medicine conference Washington D.C.	November 10-13, 2019	0.8
SIPS 2020 (online conference)	June 22/23, 2020	0.6
Open Research: A Vision for the Future (RIOT Science Club)	March 02, 2021	0.2
Open Research: Hidden flexibility and *p*-values (RIOT Science Club)	May 11, 2021	0.1
SIPS 2021 (online conference)	June 23-25, 2021	0.8
Diagnostic Error in Medicine conference 2021	October, 2021	0.8

ReThink Conference 2022	February 16, 2022	0.1
SIPS 2022	June 27-29, 2022	0.8
AMEE Lyon	August 28-31, 2022	1.2
Summary of PhD training and activities	**Date**	**Workload (ECTS)** **1 EC = 28 uur**
Presentations and Posters		
Helmholtz retreat	June 29, 2018	0.5
Diagnostic Error in Medicine conference	August 30/31, 2018	0.5
RIME, Copenhagen	May 23-24, 2019	0.5
Symposium “Understanding Diagnostic Error”	August 15, 2019	0.5
12th Diagnostic Error in Medicine conference Washington D.C.	November 10-13, 2019	0.5
Diagnostic Error in Medicine conference 2021	November, 2021	0.5
AMEE Lyon	August 28-31, 2022	0.5
Peer-review of articles and conference abstracts		
EuroDEM abstracts	July 2020	0.3
Diagnosis	February 2020	0.5
BMC Medical Informatics and Decision Making	January 2021	0.3
Diagnosis	March 2021	0.3
SIDM	May 2021	0.2
Diagnosis	June 2021	0.2
Medical Education	June 2021	0.2
SIDM conference abstracts	June 2021, 2022	0.5
Perspectives on Medical Education	November 2021	0.2
International Journal of General Medicine	April 2022	0.2
Summary of PhD training and activities	**Date**	**Workload (ECTS)** **1 EC = 28 uur**
Teaching		
Teaching assistant Medical Statistics LUMC	December, 2021 – December, 2022	
Design and analysis of biomedical studies		1.5
Wetenschappelijke vorming jaar 1		0.4
Methoden en technieken		0.4
Supervising Master student research project	June, 2020 – July, 2021	3.6
Supervising Master student research project	November 2021 – April 2022	3.0
Total ECTS		**32.2**