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

Justine Staal,1,2,3,4 Jacky Hooftman,1,2 Sabrina T G Gunput,3 Silvia Mamede,1,4 Maarten A Frens,5 Walter W Van den Broek,1 Jelmer Alsma,6 Laura Zwaan

ABSTRACT
Background Preventable diagnostic errors are a large burden on healthcare. Cognitive reasoning tools, that is, 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 15 October 2021, supplemented with handsearching. 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 to 0.29, $p<0.001$). There were no significant subgroup differences.

Conclusion 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.

WHAT IS ALREADY KNOWN ON THIS TOPIC
⇒ Cognitive reasoning tools that is, 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 AND/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, the collective national academy of the USA, estimated that most people will...
experience a diagnostic error in their lifetime, sometimes with devastating consequences. 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 recommended¹⁵–¹⁷: these factors are thought to be a primary cause of errors which have been identified in more than 75% of error cases.¹⁹–¹¹ 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 strategies¹² 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 al² and Kwan et al,¹⁶ who examined the impact of reflective practices and decision support systems, respectively. Crucially, these meta-analyses and other reviews¹⁷ 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. Second, we aimed to identify factors in study or intervention design that were associated with higher overall effectiveness.

**METHODS**

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions²² was followed in this study. The review’s objectives and methods were specified in advance in the PROSPERO Database.

**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 online supplemental 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) and 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 15 October 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 (ie, 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. Inter-rater 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 crossover trials or before–after designs) that measured diagnostic performance (either as diagnostic error or diagnostic accuracy).

We excluded tools that focused on specific diseases (eg, 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 (eg, 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 behaviours in a patient that match to a disorder in the Diagnostic and Statistical Manual of Mental Disorders,\textsuperscript{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 randomised controlled trials (RCTs) and non-RCTs (version 12-08-2013).\textsuperscript{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, randomisation). 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 (RoB 2) template.\textsuperscript{26} This form assessed study randomisation, 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’ and the study authors were contacted. The final bias assessment was equivalent to the highest received subassessment.

The overall strength of the evidence was assessed using the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) group’s tool.\textsuperscript{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\textsuperscript{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 SD of diagnostic performance were used to compute the standardised mean difference (Hedges’g) and the 95% CI of; for dichotomous data, the reported effect size (ie, OR) was transformed to Hedges’g. These results were pooled using a random-effects model meta-analysis with the Hartung-Knapp adjustment,\textsuperscript{30} using the restricted maximum likelihood 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 I\textsuperscript{2} statistic, which was categorised as: might not be important (0%–40%), moderate (30%–60%), substantial (50%–90%) and considerable (75%–100%).\textsuperscript{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.\textsuperscript{32}

Subgroup analyses were performed for participant expertise, several intervention characteristics (ie, intervention type, moment of intervention, intervention items) and study characteristics (ie, diagnostic task, case difficulty, same cases used with and without intervention, study intention). Variable definitions are given in table 1. The subgroup analyses for the level of expertise and intervention characteristics were prespecified; the analyses for study characteristics were based on observations made during study characteristic extraction. Analyses were performed with the metafor package\textsuperscript{33} in R (V.1.4.1106),\textsuperscript{34} with significance levels set at p<0.05.

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. Inter-rater 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 (online
supplemental appendix B). One hundred and twelve 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, studies that did not measure diagnostic accuracy or diagnostic errors or studies that did not describe an experiment. 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 one study (Staal et al, Impact of diagnostic checklists on the interpretation of normal and abnormal electrocardiograms, 2021). This unpublished experiment compared diagnostic accuracy on ECGs’ diagnosis using a debiasing checklist and an ECG-specific checklist. The data were obtained from the authors. Three studies (Staal et al, Impact of diagnostic checklists on the interpretation of normal and abnormal electrocardiograms, 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 online supplemental appendix C.

**Interventions**

A variety of interventions was included for analysis, which were divided into four categories based on Lambe et al: checklists, computerised decision support systems, instructions at test (ie, 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, computerised 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 hypothesised to reduce errors. Finally, reflective reasoning tools were based on the deliberate reflection procedure designed by Mamede et al. In some cases, similar procedures were named differently, for example, Ilgen et al 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 diagnoses 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).

**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, Impact of diagnostic checklists on the interpretation of normal and abnormal electrocardiograms, 2021). Only one study was preregistered. Three studies were assessed as high risk of bias. First, O’Sullivan and Schofield29 had a medium risk of bias due to a large drop-out rate during the study. Second, Shimizu et al43 was scored at high risk because of their quasi-random participant allocation. Third, Cairns et al48 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 six participants completed two or more ECGs. Inter-rater reliability for the total risk of bias score could not be calculated using Cohen’s kappa, but overall agreement was high (online supplemental appendix B). See online supplemental appendix D for the overall risk of bias assessment score.

**Main analysis**

Data on diagnostic accuracy were available for 29 studies. This resulted in analysable 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 to 0.43, p<0.001). There was evidence of considerable heterogeneity in this estimate (I²=70%, χ²(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 al49, Talebian et al50 and Thompson et al51 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 ($I^2=38\%$, $\chi^2(25)=40.22$, $p=0.028$) sufficiently to interpret the meta-analysis. The effect estimate was slightly reduced (0.20, 95% CI 0.10 to 0.29, $p<0.001$), although the effect magnitude and direction remained unchanged (figure 2). A more elaborate exploration of the heterogeneity is presented in online supplemental appendix E.

Publication bias
A funnel plot was drawn to check for small study effects due to publication bias and to further explore heterogeneity (online supplemental 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 online supplemental appendix G. Only the type of diagnostic task seemed to moderate the effect of clinical reasoning tools: studies using real or standardised 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 to diagnose real

<table>
<thead>
<tr>
<th>Category</th>
<th>Characteristics</th>
<th>Overall, n=29* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>Asia</td>
<td>4 (17)3</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
<td>11 (38)</td>
</tr>
<tr>
<td></td>
<td>North America</td>
<td>12 (41)</td>
</tr>
<tr>
<td></td>
<td>South America</td>
<td>2 (7) 7</td>
</tr>
<tr>
<td>Setting</td>
<td>Clinical practice</td>
<td>2 (7)4</td>
</tr>
<tr>
<td></td>
<td>Experiment</td>
<td>27 (93)</td>
</tr>
<tr>
<td>Specialism†</td>
<td>Emergency medicine</td>
<td>5 (24)</td>
</tr>
<tr>
<td></td>
<td>Family medicine, general practice</td>
<td>4 (19)</td>
</tr>
<tr>
<td></td>
<td>Internal medicine</td>
<td>9 (42)</td>
</tr>
<tr>
<td></td>
<td>Osteopathy</td>
<td>1 (5)</td>
</tr>
<tr>
<td></td>
<td>Paediatry</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Intervention</td>
<td>Checklist</td>
<td>13 (45)</td>
</tr>
<tr>
<td></td>
<td>Computerised decision support</td>
<td>2 (7)</td>
</tr>
<tr>
<td></td>
<td>Instructions at test (eg, red flag)</td>
<td>2 (7)</td>
</tr>
<tr>
<td></td>
<td>Guided reflection</td>
<td>12 (41)</td>
</tr>
<tr>
<td>Comparator</td>
<td>Non-analytical instructions</td>
<td>8 (28)</td>
</tr>
<tr>
<td></td>
<td>Diagnosis without the tool</td>
<td>20 (69)</td>
</tr>
<tr>
<td></td>
<td>Usual care</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Expertise</td>
<td>Novice</td>
<td>13 (35)</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
<td>17 (48)</td>
</tr>
<tr>
<td></td>
<td>Experts</td>
<td>7 (19)</td>
</tr>
<tr>
<td>Intervention type</td>
<td>Content focus</td>
<td>29 (85)</td>
</tr>
<tr>
<td></td>
<td>Process focus</td>
<td>5 (15)</td>
</tr>
<tr>
<td>Intervention moment</td>
<td>During initial diagnosis</td>
<td>12 (41)</td>
</tr>
<tr>
<td></td>
<td>Verification after initial diagnosis</td>
<td>17 (59)</td>
</tr>
<tr>
<td>Intervention items</td>
<td>Acknowledge</td>
<td>12 (41)</td>
</tr>
<tr>
<td></td>
<td>Report</td>
<td>17 (59)</td>
</tr>
<tr>
<td>Case difficulty</td>
<td>Simple</td>
<td>7 (21)</td>
</tr>
<tr>
<td></td>
<td>Complex</td>
<td>10 (59)</td>
</tr>
<tr>
<td>Diagnostic task</td>
<td>Patient diagnosis</td>
<td>2 (7)</td>
</tr>
<tr>
<td></td>
<td>Visual diagnosis</td>
<td>9 (31)</td>
</tr>
<tr>
<td></td>
<td>Written case diagnosis</td>
<td>18 (62)</td>
</tr>
<tr>
<td>Same cases used with and without intervention</td>
<td>Yes</td>
<td>13 (45)</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>16 (55)</td>
</tr>
<tr>
<td>Study intention</td>
<td>Evaluate</td>
<td>25 (86)</td>
</tr>
<tr>
<td></td>
<td>Fix</td>
<td>4 (14)</td>
</tr>
</tbody>
</table>

*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 specialised.
or virtual patients, 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.

Descriptively, participants of an intermediate level (ie, residents and fellows) seemed to benefit more from using cognitive reasoning tools than novices (ie, 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 four studies induced and then remedied errors.

### GRADE assessment

Finally, overall evidence was qualified for the meta-analysis excluding studies with extensive training \(^{49–51}\) \((table 4)\). 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 prespecified analysis plans were available for only one study (Staal et al, Impact of diagnostic checklists on

<table>
<thead>
<tr>
<th>Study</th>
<th>Category</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berbaum et al(^{66})</td>
<td>Checklists</td>
<td>Checklist listing separate anatomical regions to assist visual search of chest radiographs.</td>
</tr>
<tr>
<td>Caims et al(^{48})</td>
<td>Checklists</td>
<td>Interactive system (Interactive Progressive based Interpretation model, or IPI) to guide ECG interpretation via a series of systematic subtasks (eg, rhythm interpretation).</td>
</tr>
<tr>
<td>Chartan et al(^{47})</td>
<td>Instructions at test</td>
<td>Prompt to identify ‘red flags’ in clinical cases (I-RED strategy).</td>
</tr>
<tr>
<td>Costa Filho et al(^{47})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (based on Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Dinardo et al(^{48})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (adapted from Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Ely et al(^{46})</td>
<td>Checklist</td>
<td>Checklist providing differential diagnoses depending on the patient’s complaint.</td>
</tr>
<tr>
<td>Graber et al(^{49})</td>
<td>Computerised decision support system</td>
<td>ISABEL, a computerised system that produces a ranked list of diagnoses based on a set of clinical findings.</td>
</tr>
<tr>
<td>Ilgen et al(^{47})</td>
<td>Guided reflection</td>
<td>Directed search instructions (based on Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Ilgen et al(^{48})</td>
<td>Guided reflection</td>
<td>Directed search instructions (based on Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Kämmer et al(^{44})</td>
<td>Checklists</td>
<td>General debiasing checklist, which provided prompts to carefully consider the diagnosis, or a checklist which presented differential diagnoses based on the patient's complaint.</td>
</tr>
<tr>
<td>Kilian et al(^{70})</td>
<td>Checklists</td>
<td>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).</td>
</tr>
<tr>
<td>Kok et al(^{41})</td>
<td>Checklists</td>
<td>Checklist for diagnosing chest radiographs focusing on anatomy, potential pitfalls and frequently missed diagnoses.</td>
</tr>
<tr>
<td>Lambe et al(^{71})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tools (based on Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Li et al(^{49})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (based on Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Mamede et al(^{45})</td>
<td>Guided reflection</td>
<td>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.</td>
</tr>
<tr>
<td>Mamede et al(^{44})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (as used in Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Mamede et al(^{45})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (as used in Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Mamede et al(^{45})</td>
<td>Instructions to test</td>
<td>Instructions to write down findings that favoured the initial diagnosis (confirmatory), contradicted the initial diagnosis (contradictory) or both.</td>
</tr>
<tr>
<td>Martinez-Franco et al(^{49})</td>
<td>Computerised decision support system</td>
<td>DXPplain, a computerised system that produces a ranked list of diagnoses based on a set of clinical findings.</td>
</tr>
<tr>
<td>Myung et al(^{47})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (similar to Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Nickerson et al(^{72})</td>
<td>Checklists</td>
<td>Checklist of conditions that can cause syncope (for ECG diagnosis).</td>
</tr>
<tr>
<td>O'Sullivan and Schollfield(^{29})</td>
<td>Checklists</td>
<td>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?).</td>
</tr>
<tr>
<td>Schmidt et al(^{42})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (similar to Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Schmidt et al(^{43})</td>
<td>Guided reflection</td>
<td>Reflective reasoning tool (similar to Mamede et al(^{7})).</td>
</tr>
<tr>
<td>Shimizu et al(^{44})</td>
<td>Guided reflection</td>
<td>General debiasing checklist, which provided prompts to carefully consider the diagnosis, or a checklist which presented differential diagnoses based on the patient’s complaint.</td>
</tr>
<tr>
<td>Sibbald et al(^{45})</td>
<td>Checklists</td>
<td>Checklist that breaks up the process of performing a cardiac physical examination in systematic subtasks.</td>
</tr>
<tr>
<td>Staal et al (Impact of diagnostic checklists on the interpretation of normal and abnormal electrocardiograms, 2021)</td>
<td>Checklists</td>
<td>General debiasing checklist, which provided prompts to carefully consider the diagnosis, or a checklist which breaks ECG diagnosis into systematic subtasks.</td>
</tr>
<tr>
<td>Talebian et al(^{70})</td>
<td>Checklists</td>
<td>ECG checklist (DECKlist) for interpreting ECGs based on 12 items (such as rhythm and heart axis).</td>
</tr>
<tr>
<td>Thompson et al(^{51})</td>
<td>Checklists</td>
<td>Mnemonic checklist listing separate anatomical regions to assist visual search of lateral chest radiographs.</td>
</tr>
</tbody>
</table>

the interpretation of normal and abnormal electrocardiograms, 2021).

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 to 0.43, p<0.001). This estimate exhibited substantial heterogeneity (I²=70%), which was largely attributable to three studies that offered training with their tool before measuring performance.49–51 Removing these studies resulted in a lower, but more precise effect size (0.20, 95% CI 0.10 to 0.29, p<0.001) and reduced heterogeneity (I²=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 to 0.49) than for written (0.16, 95% CI 0.05 to 0.28) or visual cases (0.16, 95% CI 0.05 to 0.28). However, because only two 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 53 clinical decision support systems,12 19 20 54 cognitive forcing strategies7 12 and checklists.12 20 53 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 53 A more direct comparison can be made with Graber et al5 and Lambe et al,7 who specifically examined

---

**Table 4** GRADE certainty of evidence assessment

<table>
<thead>
<tr>
<th>Certainty assessment</th>
<th>No of studies (participants)</th>
<th>Risk of bias*</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Imprecision</th>
<th>Publication bias</th>
<th>Overall quality of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>261 (2539)</td>
<td>Serious</td>
<td>Not serious</td>
<td>Not serious</td>
<td>Not serious</td>
<td>Not detected</td>
<td>Moderate</td>
<td></td>
</tr>
</tbody>
</table>

*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 Martínez-Franco et al,49 Talebian et al,50 and Thompson et al.51

GRADE, Grading of Recommendations Assessment, Development, and Evaluation.
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., who reported a modest improvement of diagnostic decision-making when using reflection strategies (0.38, 95% CI 0.23 to 0.52, I²=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 (ie, studies that tested interventions with the aim to teach someone how to solve cases in the future) and workplace-oriented studies (ie, 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, 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. 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 (ie, aimed at preventing flaws in reasoning processes) were often less effective than content-focused interventions (ie, aimed at providing or triggering relevant knowledge). 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. 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 maximised 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 to determine the effects in different subgroups. Indications for potentially interesting factors may be taken from descriptive differences in our subgroup comparisons (online supplemental appendix G), which suggest diagnostic task and intervention type (content-focused or process-focused intervention) as factors of interest. Furthermore, the excluded subset of studies 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 or checklists) 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 three studies that allowed participants to train with the tool before using it. 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 takes extra variability 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 generalisable 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,\textsuperscript{37-39} 56 whether the correct diagnosis was included in the differential\textsuperscript{57,58} or whether a new diagnostic plan was made for a patient based on the leading diagnosis.\textsuperscript{59} 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 generalisable 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,\textsuperscript{3,7,12,54} for this review only two studies could be identified that were performed outside of an artificial setting.\textsuperscript{28,50} 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 optimised.

Author affiliations
1Institute of Medical Education Research Rotterdam, Erasmus Medical Center, Rotterdam, The Netherlands
2Public and Occupational Health, Amsterdam Public Health Research Institute, Amsterdam UMC, Locatie VUmc, Amsterdam, The Netherlands
3Medical Library, Erasmus Medical Center, Rotterdam, The Netherlands
4Department of Psychology, Erasmus School of Social and Behavioural Sciences, Erasmus University Rotterdam, Rotterdam, The Netherlands
5Department of Neuroscience, Erasmus Medical Center, Rotterdam, The Netherlands
6Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands

Twitter Laura Zwaan @laurazwaan81

Contributors All authors had full access to all the study data and take responsibility for the integrity of the data and the accuracy of the analysis. All authors read and approved the final manuscript. Guarantor: JS and LZ. Study conception and design: JS, JH, SM and LZ. Development of study materials: JS, JH, STGG and LZ. Acquisition of data: JS, JH, STGG and LZ. Analysis or interpretation of the data: JS, JH, SM and LZ. Drafting of the manuscript: JS and LZ. Critical revision of the manuscript for important intellectual content: JS, JH, STGG, SM, MAF, WWvdB, JA and LZ. Statistical analysis: JS and LZ. Administrative, technical or material support: JS, STGG and LZ. Supervision: JS and LZ.

Funding The authors are supported by a VENI grant from the Dutch National Scientific Organization (NOW; 45116032) and an Erasmus Medical Center Fellowship.

Disclaimer 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.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. The study protocol was preregistered and is available online in the PROSPERO database.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Justine Staal http://orcid.org/0000-0003-3790-0072
Silvia Mamede http://orcid.org/0000-0003-1187-2392
Laura Zwaan http://orcid.org/0000-0003-3940-1699
REFERENCES

43. Shimizu T, Matsumoto K, Tokuda Y. Effects of the use of differential diagnosis checklist and general de-biasing


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