Justine Staal*, Robert Zegers, Jeanette Caljouw-Vos, Sîlvia Mamede and Laura Zwaan

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

https://doi.org/10.1515/dx-2022-0092
Received August 19, 2022; accepted November 27, 2022;
published online December 9, 2022

Abstract

Objectives: 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.

Keywords: checklist; clinical reasoning; diagnostic error; ECG diagnosis; general practice.

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

***Corresponding author: Justine Staal**, Erasmus Medical Center Rotterdam, Institute of Medical Education Research Rotterdam, Dr. Molewaterplein 40, 3015GD Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands, E-mail: j.staal@erasmusmc.nl
Robert Zegers, Department of General Practice, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands
Jeanette Caljouw-Vos, General Practice Caljouw, Ridderkerk, The Netherlands
Sîlvia Mamede, Erasmus Medical Center Rotterdam, Institute of Medical Education Research Rotterdam, Rotterdam, The Netherlands; and Department of Psychology, Education and Child Studies, Erasmus School of Social and Behavioral Sciences, Rotterdam, The Netherlands. https://orcid.org/0000-0003-1187-2392
Laura Zwaan, Erasmus Medical Center Rotterdam, Institute of Medical Education Research Rotterdam, Rotterdam, The Netherlands. https://orcid.org/0000-0003-3940-1699

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.

Materials and 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].

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 Supplementary Material 1.

Table 1: Overview of checklist materials.

Debiasing checklist [38]
Please check your ECG diagnosis carefully considering each of the following: (1) Was I comprehensive? (2) Did I consider the inherent flaws of heuristic thinking? (3) Was my judgment affected by any cognitive bias? (4) Were any of the following biases present (anchoring, availability, confirmation, search satisficing, framing)? (5) What is the worst-case scenario?
Content-specific checklist [34]
Please check your ECG diagnosis carefully considering each of the following: (1) Frequency and rate (2) Axis (3) P-wave (4) PQ-interval (5) Q-wave (6) QRS-complex (7) ST-interval (8) T-wave (9) QT-interval (10) 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.

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 min 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.

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

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 min 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-min 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=0.05 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 Supplementary 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.

Diagnostic performance

Phase 1 vs. 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:

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: χ^2 (3)=6.93, p=0.074 Phase 2: χ^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: χ^2 (3)=8.18, p=0.042 Phase 2: χ^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: χ^2 (3)=2.36, p=0.501 Phase 2: χ^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], seconds				Phase 1: χ^2 (3)=2.08, p=0.556 Phase 2: χ^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).

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 (χ^2(2)=2.54, p=0.281), confidence (χ^2(2)=2.74, p=0.254), patient management (χ^2(2)=2.10, p=0.350) and time to diagnose (χ^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

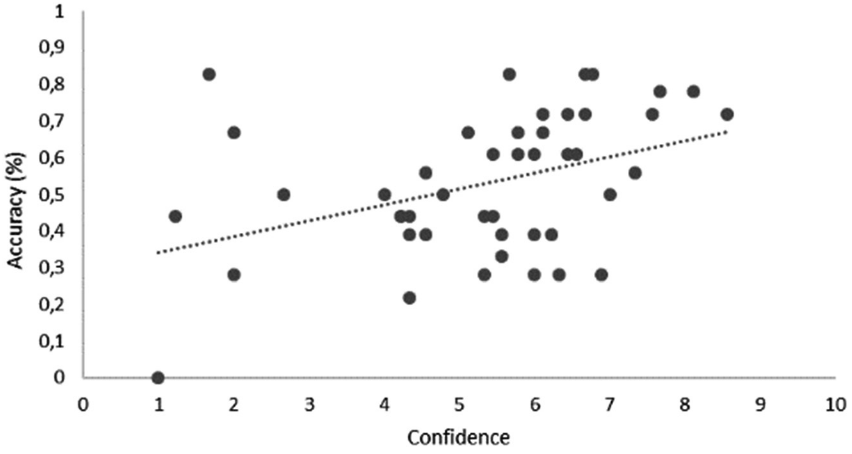

Figure 1: Scatterplot of residents' confidence-accuracy calibration in phase 1, R^2=0.14.

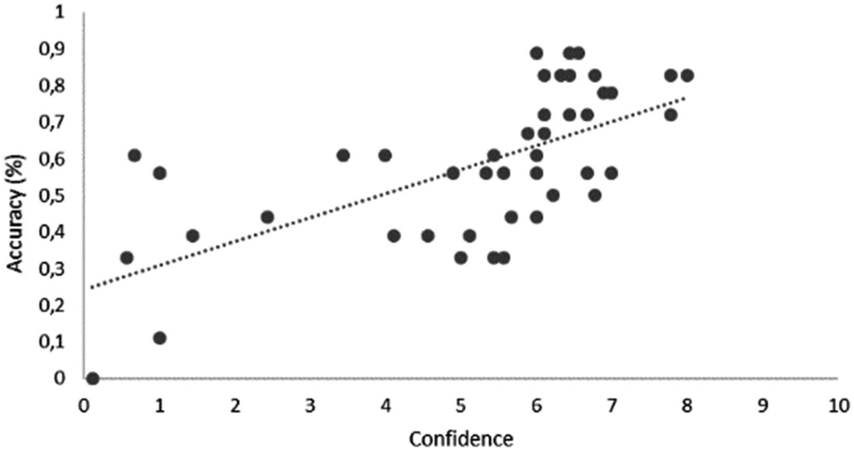

Figure 2: Scatterplot of residents' confidence-accuracy calibration in phase 2, R^2=0.40.

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).

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.

Acknowledgments: 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.

Research funding: Dr. Laura Zwaan is supported by a VENI grant from the Dutch National Scientific Organization (NWO; 45116032) and an Erasmus MC Fellowship. 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.

Author contributions: 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. Study conception and design: all authors. Development of study materials: JS, RZ, JCV. Acquisition of data: JS, RZ, JCV. Analysis or interpretation of the data: all authors. Drafting of the manuscript: JS, LZ. Critical revision of the manuscript for important intellectual content: all authors. Statistical analysis: JS, LZ. Administrative, technical or material support: JS, JCV, LZ. Supervision: JS, LZ.

Competing interests: Authors state no conflict of interest.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Ethical approval: Research involving human subjects complied with all relevant national regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration (as revised in 2013) and has been approved by the medical ethical committee of the Erasmus University Medical Center (MEC-2021-0284).

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request. The study protocol was preregistered and is available online at Open Science Framework (DOI: 10.17605/OSF.IO/W28QA).

References

1. Zia SMR, Zahid R, Ashraf H. The WHO surgical safety checklist: a systematic literature review. Arch Surg Res 2021;2:27–30.
2. Thomassen Ø, Storesund A, Søfteland E, Brattebø G. The effects of safety checklists in medicine: a systematic review. Acta Anaesthesiol Scand 2014;58:5–18.
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? Annu Rev Publ Health 2010;31:479–97.
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. N Engl J Med 2006;355: 2725–32.
5. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AHS, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med 2009;360: 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. West J Emerg Med 2020;21:125.
7. Gawande A. The checklist manifesto: how to get things right. J Nurs Regul 2011;1:64.
8. Gupta A, Graber ML. Annals for hospitalists inpatient notes-just what the doctor ordered—checklists to improve diagnosis. Ann Intern Med 2019;170:HO2–3.
9. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ Qual Saf 2016;25:808–20.
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 Qual Saf 2012;21:535–57.
11. Clinician Checklists [Internet]: Society to Improve Diagnosis in Medicine; 2020. Available from: https://www.improvediagnosis.org/clinician-checklists/ [Accessed 1 Jul 2021].
12. Wachter RM. Why diagnostic errors don't get any respect—and what can be done about them. Health Aff 2010;29:1605–10.
13. Balogh EP, Miller BT, Ball JR. Improving diagnosis in health care. Washington: National Academies Press; 2015.
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. Arch Intern Med 2010;170: 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. J Gen Intern Med 2021;36:1–9.
16. Phua DH, Tan NC. Cognitive aspect of diagnostic errors. Ann Acad Med Singapore 2013;42:33–41.
17. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003;78:775–80.
18. Croskerry P. From mindless to mindful practice—cognitive bias and clinical decision making. N Engl J Med 2013;368: 2445–8.
19. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Adv Health Sci Educ 2009; 14:27–35.
20. Elia F, Apra F, Verhovez A, Crupi V. "First, know thyself": cognition and error in medicine. Acta Diabetol 2016;53:169–75.
21. Monteiro S, Norman G, Sherbino J. The 3 faces of clinical reasoning: epistemological explorations of disparate error reduction strategies. J Eval Clin Pract 2018;24: 666–73.
22. Zwaan L, Staal J. Evidence on use of clinical reasoning checklists for diagnostic error reduction. In: AHRQ papers on diagnostic safety topics [Internet]; 2020.
23. Kämmer JE, Schauber SK, Hautz SC, Stroben F, Hautz WE. Differential diagnosis checklists reduce diagnostic error differentially: a randomized experiment. Med Educ 2021;55: 1172–82.

24. Kahneman D, Egan P. Thinking, fast and slow. New York: Farrar, Straus and Giroux; 2011.
25. Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Ann Emerg Med 2003;41:110–20.
26. Ely JW, Graber MA. Checklists to prevent diagnostic errors: a pilot randomized controlled trial. Diagnosis 2015;2:163–9.
27. Norman MSD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med 2017; 92:23–30.
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. Med Teach 2013;35:e1218–29.
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 Qual Saf 2013;22:333–8.
30. Sibbald M, Bruin ABHD, van Merrienboer JJG. Finding and fixing mistakes: do checklists work for clinicians with different levels of experience? Adv Health Sci Educ 2014;19:43–51.
31. Sibbald M, de Bruin ABH, van Merrienboer JJG. Checklists improve experts' diagnostic decisions. Med Educ 2013;47: 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. Adv Health Sci Educ 2015;20:1053–60.
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. Adv Health Sci Educ 2019; 24:427–40.
34. Konings K, Willemsen R. ECG 10+: systematisch ECG's beoordelen. Huisarts Wet 2016;59:166–70.
35. Berbaum K, Franken EA Jr., Caldwell RT, Schartz KM. Can a checklist reduce SOS errors in chest radiography? Acad Radiol 2006;13:296–304.
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? J Digit Imag 2017;30:726–31.
37. Krage R, Len LTS, 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;69:983–9.
38. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. Acad Med 2011;86:307–13.
39. Abimanyi-Ochom J, Mudiyanselage SB, Catchpool M, Firipis M, Dona SWA, Watts JJ. Strategies to reduce diagnostic errors: a systematic review. BMC Med Inf Decis Making 2019;19:1–14.
40. Nedorost S. A diagnostic checklist for generalized dermatitis. Clin Cosmet Invest Dermatol 2018;11:545.
41. Rutten FH, Kessels AGH, Willems FF, Hoes AW. Is elektrocardiografie in de huisartspraktijk nuttig? Huisarts Wet 2001;44:179–83.
42. Faul F, Erdfelder E, Lang AG, Buchner A. G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175–91.
43. Boode BSP, Frijling BD, Heeringa J, Rutten FH, Van den Berg PJ, Zwietering PJ, et al. NHG-standaard atriumfibrilleren. NHG-standaarden 2009. Houten: Bohn Stafleu van Loghum; 2009: 67–86 pp.
44. Rutten FH, Grundmeijer H, Grijseels EWM, Van Bentum STB, Hendrick JMA, Bouma M, et al. NHG-standaard acuut coronair syndroom. NHG-standaarden 2009. Houten: Bohn Stafleu van Loghum; 2009:3–24 pp.
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? Adv Health Sci Educ 2021;27:1–12.
46. Eva KW, Hatala RM, Blanc VRL, Brooks LR. Teaching from the clinical reasoning literature: combined reasoning strategies help novice diagnosticians overcome misleading information. Med Educ 2007;41:1152–8.
47. Ark TK, Brooks LR, Eva KW. The benefits of flexibility: the pedagogical value of instructions to adopt multifaceted diagnostic reasoning strategies. Med Educ 2007;41:281–7.
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? Acad Med 2006;81:405–9.
49. Coderre S, Mandin H, Harasym PH, Fick GH. Diagnostic reasoning strategies and diagnostic success. Med Educ 2003;37:695–703.
50. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med 2008;121:S2–3.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/dx-2022-0092).