Review

Unassisted Clinicians Versus Deep Learning–Assisted Clinicians in Image-Based Cancer Diagnostics: Systematic Review With Meta-analysis

Peng Xue[1*], MPH; Mingyu Si[1*], MPH; Dongxu Qin[1*], BM; Bingrui Wei[1*], BS; Samuel Seery[2*], PhD; Zichen Ye[1], BM; Mingyang Chen[1], BM; Sumeng Wang[3], BM; Cheng Song[1], MPhil; Bo Zhang[1], BM; Ming Ding[1], BM; Wenling Zhang[1], BM; Anying Bai[1], BM; Huijiao Yan[1], MPH; Le Dang[4], PhD; Yuqian Zhao[5], PhD; Remila Rezhake[6], PhD; Shaokai Zhang[7], PhD; Youlin Qiao[8], PhD; Yimin Qu[1], PhD; Yu Jiang[1], PhD

[1]Department of Epidemiology and Biostatistics, School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Faculty of Health and Medicine, Division of Health Research, Lancaster University, Lancaster, United Kingdom

[3]Department of Cancer Epidemiology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[4]Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[5]Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science & Technology of China, Sichuan, China

[6]Affiliated Cancer Hospital, The 3rd Affiliated Teaching Hospital of Xinjiang Medical University, Xinjiang, China

[7]Henan Cancer Hospital, Affiliated Cancer Hospital of Zhengzhou University, Henan, China

[8]Center for Global Health, School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yu Jiang, PhD
Department of Epidemiology and Biostatistics
School of Population Medicine and Public Health
Chinese Academy of Medical Sciences and Peking Union Medical College
No 9 Dongdan Santiao
Dongcheng
Beijing, 100730
China
Phone: 86 10 8778 8489
Email: jiangyu@pumc.edu.cn

Related Article:
This is a corrected version. See correction statement in: https://www.jmir.org/2023/1/e49146

Abstract

Background: A number of publications have demonstrated that deep learning (DL) algorithms matched or outperformed clinicians in image-based cancer diagnostics, but these algorithms are frequently considered as opponents rather than partners. Despite the clinicians-in-the-loop DL approach having great potential, no study has systematically quantified the diagnostic accuracy of clinicians with and without the assistance of DL in image-based cancer identification.

Objective: We systematically quantified the diagnostic accuracy of clinicians with and without the assistance of DL in image-based cancer identification.

Methods: PubMed, Embase, IEEEXplore, and the Cochrane Library were searched for studies published between January 1, 2012, and December 7, 2021. Any type of study design was permitted that focused on comparing unassisted clinicians and DL-assisted clinicians in cancer identification using medical imaging. Studies using medical waveform-data graphics material and those investigating image segmentation rather than classification were excluded. Studies providing binary diagnostic accuracy

data and contingency tables were included for further meta-analysis. Two subgroups were defined and analyzed, including cancer type and imaging modality.

Results: In total, 9796 studies were identified, of which 48 were deemed eligible for systematic review. Twenty-five of these studies made comparisons between unassisted clinicians and DL-assisted clinicians and provided sufficient data for statistical synthesis. We found a pooled sensitivity of 83% (95% CI 80%-86%) for unassisted clinicians and 88% (95% CI 86%-90%) for DL-assisted clinicians. Pooled specificity was 86% (95% CI 83%-88%) for unassisted clinicians and 88% (95% CI 85%-90%) for DL-assisted clinicians. The pooled sensitivity and specificity values for DL-assisted clinicians were higher than for unassisted clinicians, at ratios of 1.07 (95% CI 1.05-1.09) and 1.03 (95% CI 1.02-1.05), respectively. Similar diagnostic performance by DL-assisted clinicians was also observed across the predefined subgroups.

Conclusions: The diagnostic performance of DL-assisted clinicians appears better than unassisted clinicians in image-based cancer identification. However, caution should be exercised, because the evidence provided in the reviewed studies does not cover all the minutiae involved in real-world clinical practice. Combining qualitative insights from clinical practice with data-science approaches may improve DL-assisted practice, although further research is required.

Trial Registration: PROSPERO CRD42021281372; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=281372

(J Med Internet Res 2023;25:e43832) doi: 10.2196/43832

KEYWORDS

deep learning; cancer diagnosis; systematic review; meta-analysis

Introduction

Cancer is a leading cause of death, with an estimated 19.3 million new cases and 10 million deaths in 2020 worldwide [1]. One of the reasons for this high burden is delayed diagnosis due to inconspicuous symptoms at an early stage [2]. Cancer identification is often only feasible once serious symptoms manifest or a lesion (or tumor) is large enough to be identified using conventional diagnostic imaging techniques [3]. State-of-the-art medical imaging technologies make early diagnosis possible and instill optimism; however, subjectivity in cancer imaging diagnosis influences the application of these technologies for individual patients. Of course, specialists tend to be more accurate, but such expertise is not widely available [4]. The emergence of deep learning (DL) algorithms in medical artificial intelligence (AI) provides a way forward, despite potentially causing disruptions to standardized, established practice [5].

As a subfield of AI, DL is formally defined as "computational models, composed of multiple processing layers, to learn representations of data with multiple levels of abstraction" [6]. In medical imaging practice, DL algorithms extract representative imaging features for classification purposes, irrespective of personal experience and underlying assumptions [7]. Over the past decade, we have witnessed a growing interest in DL algorithms, specifically in cancer diagnostics. Numerous studies have reported the diagnostic performance of DL, and it is considered comparable to, or in some circumstances better than, that of clinicians [8-10]. However, medical DL is plagued with issues, including inherent biases due to limited training data, an absence of cross-population generalizability, and a lack of transparency and accountability for clinical practice [11]. Therefore, the evidence required to change policy and implement DL techniques is insufficient. In fact, many appear preoccupied with the debate around medical AI replacing human physicians. However, these technologies are designed to assist clinicians and improve diagnostics within existing clinical workflows. Human-computer collaboration can provide benefits above and beyond what either clinicians or DL algorithms can do in isolation [12-14]. This paradigm shift means that while the advantages of DL algorithms are necessary, so too are those of clinicians, who will need to fill gaps with personal knowledge of clinical histories. However, before implementing DL assistance into clinical practice, we must also assess current evidence in terms of methods and risk of bias. DL technologies must be subjected to the same rigorous assessment as any other technology in modern, evidence-based medicine. Doing anything else would impact patient acceptance and would impair the development of best practices for medical AI. There is also the wider issue of the impact on public trust in medicine, which cannot be overlooked. Therefore, it is imperative to systematically review the diagnostic performance of DL-assisted clinicians versus unassisted clinicians. Critically reviewing the evidence base is necessary to ensure these methods are both safe and effective. The synthesized evidence may also provide insights into the human factors involved in the use of DL assistance in cancer identification.

Methods

Search Strategy and Selection Criteria

The study protocol was registered with PROSPERO (CRD42021281372) and was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) 2020 guidelines [15].

Keywords, including "cancer," "AI/DL," "performance," and "image" were used to identify comparative studies that assessed the performance of unassisted clinicians and DL-assisted clinicians in image-based cancer diagnosis. Multimedia Appendix 1 provides an outline of the full search strategy. Records from PubMed, Embase, IEEEXplore, and the Cochrane Library from January 1, 2012, to December 7, 2021, were systematically searched with no language restrictions. The start date was chosen based on a recognized step change in the development of DL approaches [16].

The inclusion and exclusion criteria for related studies were jointly determined by 2 independent authors who screened titles and abstracts. All studies were then read in full and discrepancies were resolved by a third author. Studies were included if they focused on comparing the performance of unassisted clinicians and DL-assisted clinicians in image-based cancer diagnosis. Studies were excluded if they (1) examined medical waveform-data graphics material, (2) investigated image segmentation rather than cancer identification, (3) reported ternary diagnostic outcomes, or (4) did not study DL. Case reports, reviews, editorials, letters, comments, conference abstracts or proceedings, and duplicates were also excluded (Multimedia Appendix 1).

Data Extraction

Two authors independently extracted study characteristics, model-related details, and performance data. Uncertainties were resolved by another independent research associate. Binary diagnostic accuracy data, including true positives, false positives, true negatives, and false negatives, were extracted by 2 reviewers. Sensitivity and specificity data were then pooled for analysis. If a study provided a number of contingency tables for the same or different DL models, we assumed these were independent, unless otherwise stated.

Quality Assessment

The Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool was used to assess the risk of bias and applicability concerns of the included studies [17].

Statistical Analysis

Hierarchical summary receiver operating characteristic (HSROC) curves were used to estimate overall accuracy. In the HSROC curves, 95% CIs and 95% prediction regions of the summary operating points, including averaged sensitivity, specificity, and the area under the curve (AUC), were provided under a random effects model. Heterogeneity was assessed using the I^2 statistic.

Relative sensitivity and specificity were pooled for meta-analysis. Cancer type and imaging modality were established for subgroup analysis. Potential sources of heterogeneity were assessed across relative sensitivity and specificity for both DL-assisted clinicians and unassisted clinicians. Additional efforts were made to identify sources of heterogeneity in 2 separate subgroup meta-analyses: (1) according to cancer type, which included breast, lung, gastrointestinal, and endocrine cancers, and (2) according to imaging modality, which included ultrasound, X-ray, endoscopy, and magnetic resonance imaging (MRI).

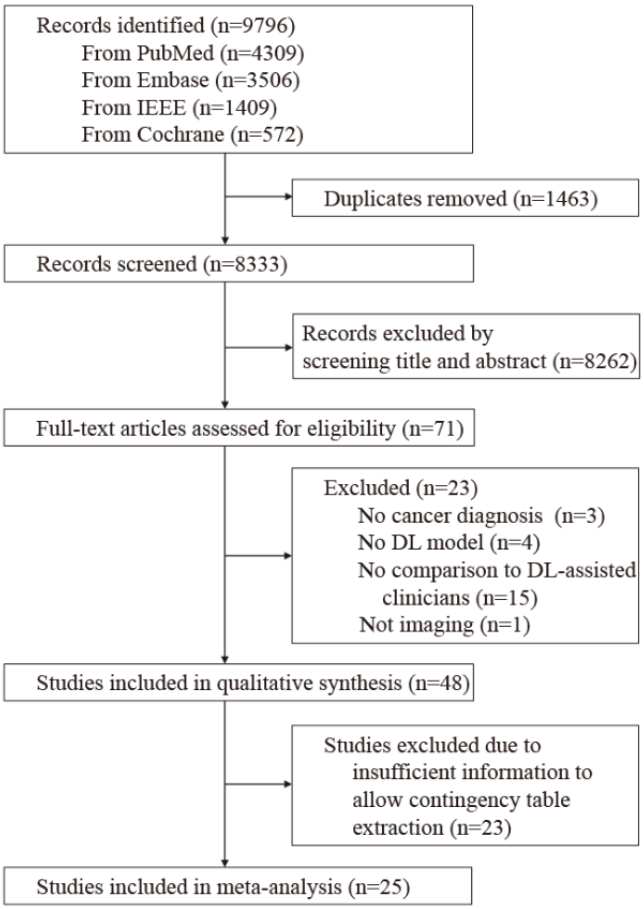
The random effect model was implemented because of inherent differences within the evidence base. Publication bias was visually assessed using funnel plots. Only studies with N≥4 were included for statistical pooling. All analyses were conducted with Stata (version 15.1; Stata Corp) and SAS (version 9.4; SAS Institute). Two-sided *P* values of less than .05 were considered statistically significant.

Results

Study Selection

Online searching was last updated on December 7, 2021, and 9796 studies were retrieved (Figure 1). After removing duplicates, 8333 publications were screened. After screening and selection, 71 full texts were considered eligible, although a further 23 were excluded due to insufficient information. This left 48 studies for systematic review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of study selection. DL: deep learning.



Characteristics of the Included Studies

Of the 48 enrolled studies [18-65], 52% (n=25) provided comprehensive inclusion and exclusion criteria while 12% (n=6) provided no information about participants (Multimedia Appendix 1). Breast and gastrointestinal cancer accounted for half the studies. The top 4 conditions were breast cancer (n=13, 27%), gastrointestinal cancer (n=11, 23%), lung cancer (n=8, 17%), and endocrine cancer (n=7, 15%). The top 4 imaging modalities were ultrasound (n=8, 17%), X-ray (n=8, 17%), and computerized tomography (n=7, 15%), with MRI and whole-slide imaging both used in 12% (n=6) of studies. Dermoscopy and endoscopy had 5 studies each, with each representing 10% of the sample. Multimedia Appendix 1 provides summaries of study characteristics. In total, 98% (n=47) of the studies were based on retrospectively collected data. Only 1 study could be considered prospective. Only 4 studies reported a prespecified sample size calculation. Meanwhile, 29% (n=14) of studies used data from open-access repositories. In addition, 54% (n=26) of studies performed external validation, whereas the remaining studies relied upon internal validation. Moreover, 33% (n=16) reported exclusion of low-quality images, and 35% (n=17) used heat maps. Transfer learning was applied in 21% of the studies (n=10) during the training phase.

Reference standards were wide ranging but in line with cancer types and imaging modalities, although some adopted multiple methods. Most of the studies (n=35, 73%) used histopathology with (or without) a follow-up period as the ground truth or gold standard control for image-based cancer diagnostics. The remainder implemented an expert consensus or medical record–based approach.

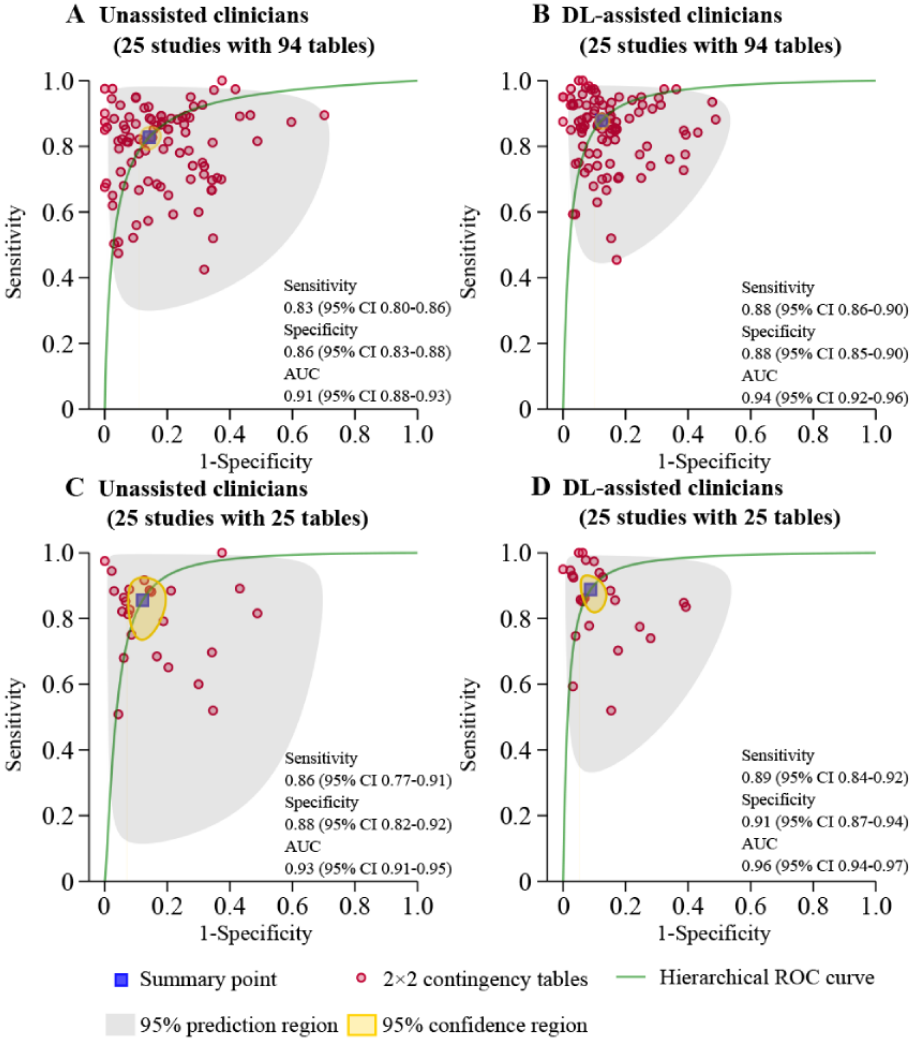
Diagnostic Performance of Unassisted Clinicians Versus DL-Assisted Clinicians

In total, 52% (n=25) of the included studies provided sufficient data to construct contingency tables, calculate diagnostic performance, and perform meta-analysis. HSROC curves generated using 25 studies (with a total of 94 contingency tables) are shown in Figure 2. When averaging across studies, the pooled sensitivity and specificity values for unassisted clinicians were 83% (95% CI 80%-86%) and 86% (95% CI 83%-88%), respectively, with an AUC of 0.91 (95% CI 0.88-0.93). By contrast, the pooled sensitivity and specificity values for DL-assisted clinicians were 88% (95% CI 86%-90%), and 88% (95% CI 85%-90%), respectively, with an AUC of 0.94 (95% CI 0.92-0.96). The pooled sensitivity and specificity values for DL-assisted clinicians were higher than those for unassisted clinicians at ratios of 1.07 (95% CI 1.05-1.09) and 1.03 (95% CI 1.02-1.05), respectively.

Most studies reported more than one DL algorithm for assessing diagnostic performance, and only the highest performance in each study was chosen for 25 contingency tables. The pooled sensitivity was 86% (95% CI 77%-91%) for unassisted clinicians and 89% (95% CI 84%-92%) for DL-assisted clinicians. The pooled specificity was 88% (95% CI 82%-92%) for unassisted

clinicians and 91% (95% CI 87%-94%) for DL-assisted clinicians. The clustered AUCs for unassisted and DL-assisted clinicians were 0.93 (95% CI 0.91-0.95) and 0.96 (95% CI 0.94-0.97), respectively (Figure 2). The pooled sensitivity and specificity values for DL-assisted clinicians were higher than those for unassisted clinicians at ratios of 1.07 (95% CI 1.03-1.10) and 1.03 (95% CI 1.01-1.06), respectively.

Figure 2. Hierarchical receiver operator characteristic curves of all studies included in the meta-analysis. A and B: ROC curves of all studies included in the meta-analysis (25 studies with 94 tables); C and D: ROC curves of studies reporting the highest accuracy (25 studies with 25 tables). AUC: area under the curve; DL: deep learning; ROC: receiver operator characteristic curve.



Subgroup Meta-analyses Comparing Diagnostic Performance

Cancer Type

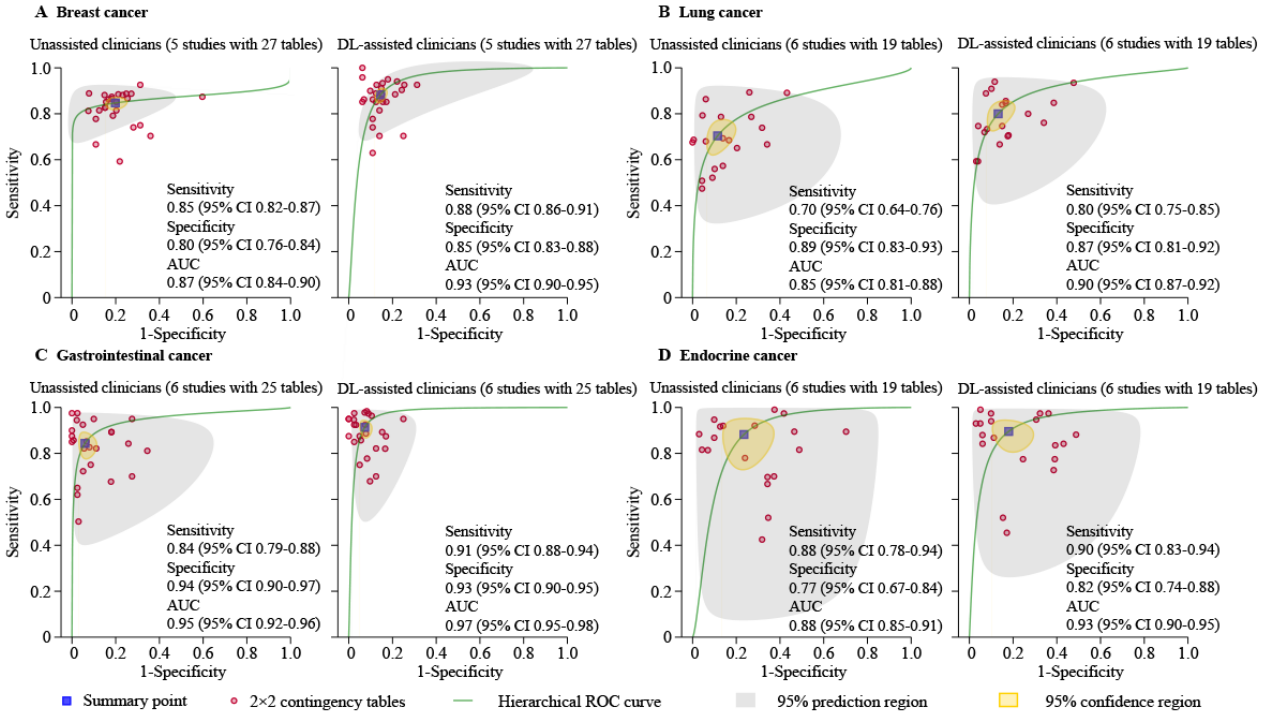
Five studies were used to create 27 contingency tables for breast cancer. The pooled sensitivity was 85% (95% CI 82%-87%) and specificity was 80% (95% CI 76%-84%), with an AUC of 0.87 (95% CI 0.84-0.90) for unassisted clinicians. For DL-assisted clinicians, we found a pooled sensitivity of 88% (95% CI 86%-91%) and specificity of 85% (95% CI 83%-88%), with an AUC of 0.93 (95% CI 0.90-0.95) (Figure 3A).

Six studies were used to develop 19 contingency tables for lung cancer. The pooled sensitivity was 70% (95% CI 64%-76%) for unassisted clinicians and 80% (95% CI 75%-85%) for DL-assisted clinicians. Pooled specificity was 89% (95% CI 83%-93%) for unassisted clinicians and 87% (95% CI 81%-92%) for DL-assisted clinicians, with an AUC of 0.85 (95% CI 0.81-0.88) for unassisted clinicians and 0.90 (95% CI 0.87-0.92) for DL-assisted clinicians (Figure 3B).

Six studies were used to generate 25 contingency tables for gastrointestinal cancer. The pooled sensitivity was 84% (95% CI 79%-88%), with a specificity of 94% (95% CI 90%-97%) and an AUC of 0.95 (95% CI 0.92-0.96) for unassisted clinicians. Pooled sensitivity was 91% (95% CI 88%-94%) and specificity was 93% (95% CI 90%-95%) with an AUC of 0.97 (95% CI 0.95-0.98) for DL-assisted clinicians (Figure 3C).

Another six studies were used to generate 19 tables for endocrine cancer. The pooled sensitivity was 88% (95% CI 78%-94%) for unassisted clinicians and 90% (95% CI 83%-94%) for DL-assisted clinicians. The pooled specificity was 77% (95% CI 67%-84%) for unassisted clinicians and 82% (95% CI 74%-88%) for DL-assisted clinicians, with an AUC of 0.88 (95% CI 0.85-0.91) for unassisted clinicians and 0.93 (95% CI 0.90-0.95) for DL-assisted clinicians (Figure 3D).

Figure 3. Hierarchical ROC curves of studies using different cancer types for comparing the performance of unassisted clinicians and DL-assisted clinicians. A: ROC curves of studies for detecting breast cancer (5 studies with 27 tables); B: ROC curves of studies for detecting lung cancer (6 studies with 19 tables); C: ROC curves of studies for detecting gastrointestinal cancer (6 studies with 25 tables); D: ROC curves of studies for detecting endocrine cancer (6 studies with 19 tables). AUC: area under the curve; DL: deep learning; ROC: receiver operator characteristic curve.



Imaging Modalities

Six studies were used to generate 27 tables for ultrasound, which displayed a pooled sensitivity of 83% (95% CI 79%-86%) and specificity of 79% (95% CI 74%-82%), with an AUC of 0.88 (95% CI 0.85-0.91), for unassisted clinicians; the pooled sensitivity was 87% (95% CI 83%-90%) and specificity was 86% (95% CI 83%-88%), with an AUC of 0.92 (95% CI 0.90-0.94), for DL-assisted clinicians (Figure 4A).

For 4 X-ray studies, 14 tables were generated, which revealed a pooled sensitivity of 70% (95% CI 63%-76%) and specificity of 90% (95% CI 84%-94%), with an AUC of 0.86 (95% CI 0.83-0.89), for unassisted clinicians. Pooled sensitivity was 78% (95% CI 71%-83%) and specificity was 90% (95% CI 86%-93%), with an AUC of 0.91 (95% CI 0.89-0.94), for DL-assisted clinicians (Figure 4B).

Four endoscopy studies were used to create 10 tables that highlighted a pooled sensitivity of 80% (95% CI 72%-86%) for unassisted clinicians and 94% (95% CI 89%-96%) for DL-assisted clinicians. Pooled specificity was 93% (95% CI 86%-97%) for unassisted clinicians and 91% (95% CI 88%-94%) for DL-assisted clinicians, with AUCs of 0.92 (95% CI 0.89-0.94) for unassisted clinicians and 0.97 (95% CI 0.95-0.98) for DL-assisted clinicians (Figure 4C).

Additionally, there were 4 studies with 14 tables using MRI. The pooled sensitivity was 81% (95% CI 74%-86%) for unassisted clinicians and 86% (95% CI 80%-90%) for DL-assisted clinicians. Pooled specificity was 77% (95% CI 66%-85%) for unassisted clinicians and 82% (95% CI 72%-88%) for DL-assisted clinicians, with an AUC of 0.86 (95% CI 0.83-0.89) for unassisted clinicians and 0.91 (95% CI 0.88-0.93) for DL-assisted clinicians (Figure 4D). Detailed comparisons of subgroup meta-analyses (for cancer type and image modality) of the relative sensitivity and specificity of DL-assisted clinicians versus unassisted clinicians are shown in Table 1.

Figure 4. Hierarchical ROC curves of studies using different imaging modalities for comparing performance between unassisted clinicians and DL-assisted clinicians. A: ROC curves of studies using ultrasound (6 studies with 27 tables); B: ROC curves of studies using X-rays (4 studies with 14 tables); C: ROC curves of studies using endoscopy (4 studies with 10 tables); D: ROC curves of studies using MRI (4 studies with 14 tables). AUC: area under the curve; DL: deep learning; MRI: magnetic resonance imaging; ROC: receiver operator characteristic curve.

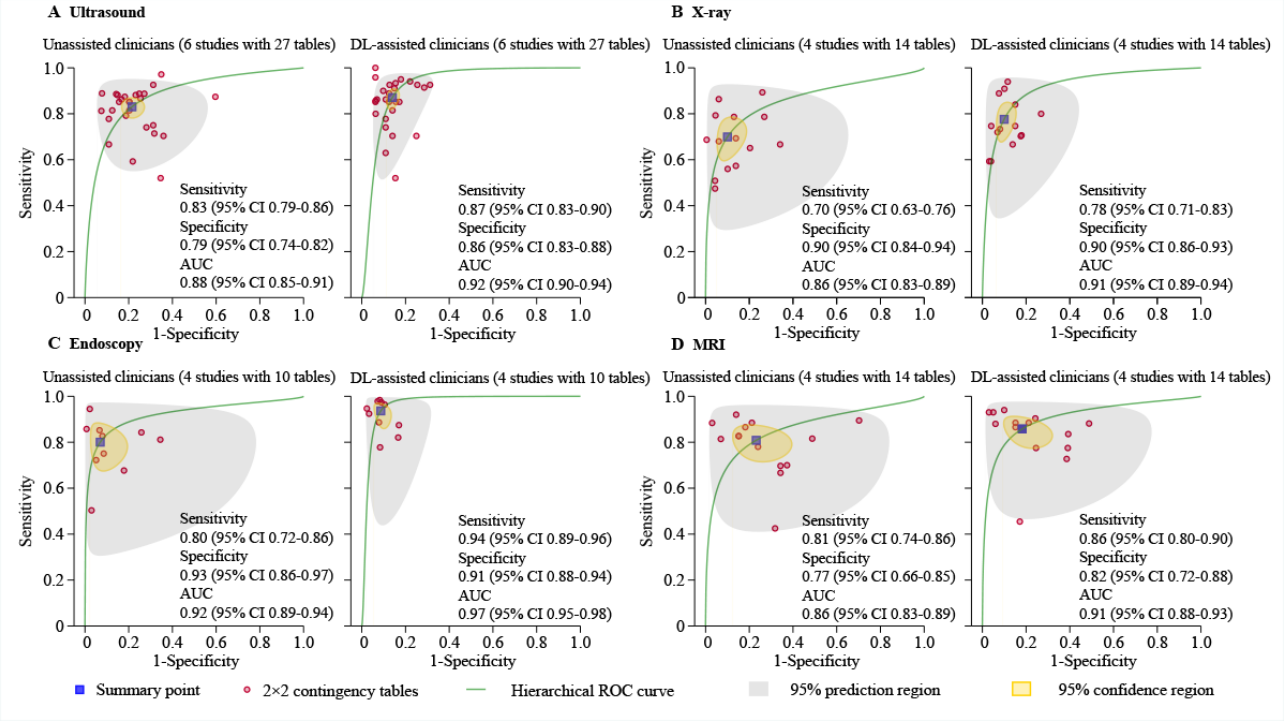


Table 1. Meta-analyses of the relative sensitivity and specificity of deep learning–assisted clinicians versus unassisted clinicians. *P* values represent a statistically significantly difference from unity (1 excluded from 95% CI, *P*<.05).

Deep learning–assisted clinicians versus unassisted clinicians	Studies, n	Tables, n	Relative sensitivity (95% CI)	*P* value	Relative specificity (95% CI)	*P* value
Overall	25	94	1.07 (1.05-1.09)	<.001	1.03 (1.02-1.05)	<.001
Studies reporting the highest performance[a]	25	25	1.07 (1.03-1.10)	<.001	1.03 (1.01-1.06)	.02
Cancer type						
Breast cancer	5	27	1.05 (1.02-1.08)	<.001	1.08 (1.03-1.13)	.003
Lung cancer	6	19	1.13 (1.08-1.18)	<.001	1.01 (0.99-1.03)	.36
Gastrointestinal cancer	6	25	1.09 (1.04-1.14)	<.001	1.02 (0.99-1.04)	.26
Endocrine cancer	6	19	1.01 (0.99-1.03)	.35	1.05 (1.01-1.09)	.02
Imaging modality						
Ultrasound	6	27	1.04 (1.02-1.07)	<.001	1.10 (1.05-1.16)	<.001
X-ray	4	14	1.11 (1.06-1.17)	<.001	1.01 (0.99-1.04)	.20
Endoscopy	4	10	1.17 (1.09-1.25)	<.001	1.02 (0.99-1.06)	.27
Magnetic resonance imaging	4	14	1.04 (1.00-1.09)	.08	1.06 (1.00-1.12)	.07

[a]Most studies reported more than one deep learning algorithm to assess diagnostic performance; only the highest-performing algorithm in each study was chosen for the 25 contingency tables.

Heterogeneity Analysis

All included studies found that DL-assisted clinicians' diagnostic performance appeared to be better than unassisted clinicians at image-based cancer identification; however, extreme heterogeneity was observed (I^2 for sensitivity was 90.3%, I^2 for specificity was 86.8%, *P*<.001) (Figures 5 and 6). This is discussed in more detail in the Limitations section.

A funnel plot was generated to assess publication bias. The studies appeared symmetrically distributed around the regression line, and a *P* value of .14 suggests no publication bias (Multimedia Appendix 1). To identify the source or sources of

heterogeneity, we conducted a subgroup analysis. Although the heterogeneity for both sensitivity and specificity within several subgroups decreased to an acceptable range after grouping, the I^2 values targeting overall diagnostic performance were still unsatisfactory. Therefore, cancer types and imaging modalities are likely to have confounded the image-based cancer diagnostic performance of unassisted clinicians versus DL-assisted clinicians.

Figure 5. Pooled relative sensitivity (A) and specificity (B) of DL-assisted clinicians versus unassisted clinicians for different imaging modalities in image-based cancer detection. The data are presented as forest plots for all studies with different imaging modalities included in the meta-analysis (n=25 studies). If a study provided multiple contingency tables for DL-assisted clinicians versus unassisted clinicians, it is listed more than once and labelled alphabetically. The data are presented as forest plots for all studies with imaging modalities included in the meta-analysis (n=25 studies). CT: computed tomography; DL: deep learning; MRI: magnetic resonance imaging; ROC: receiver operator characteristic curve; WSI: whole-slide imaging.

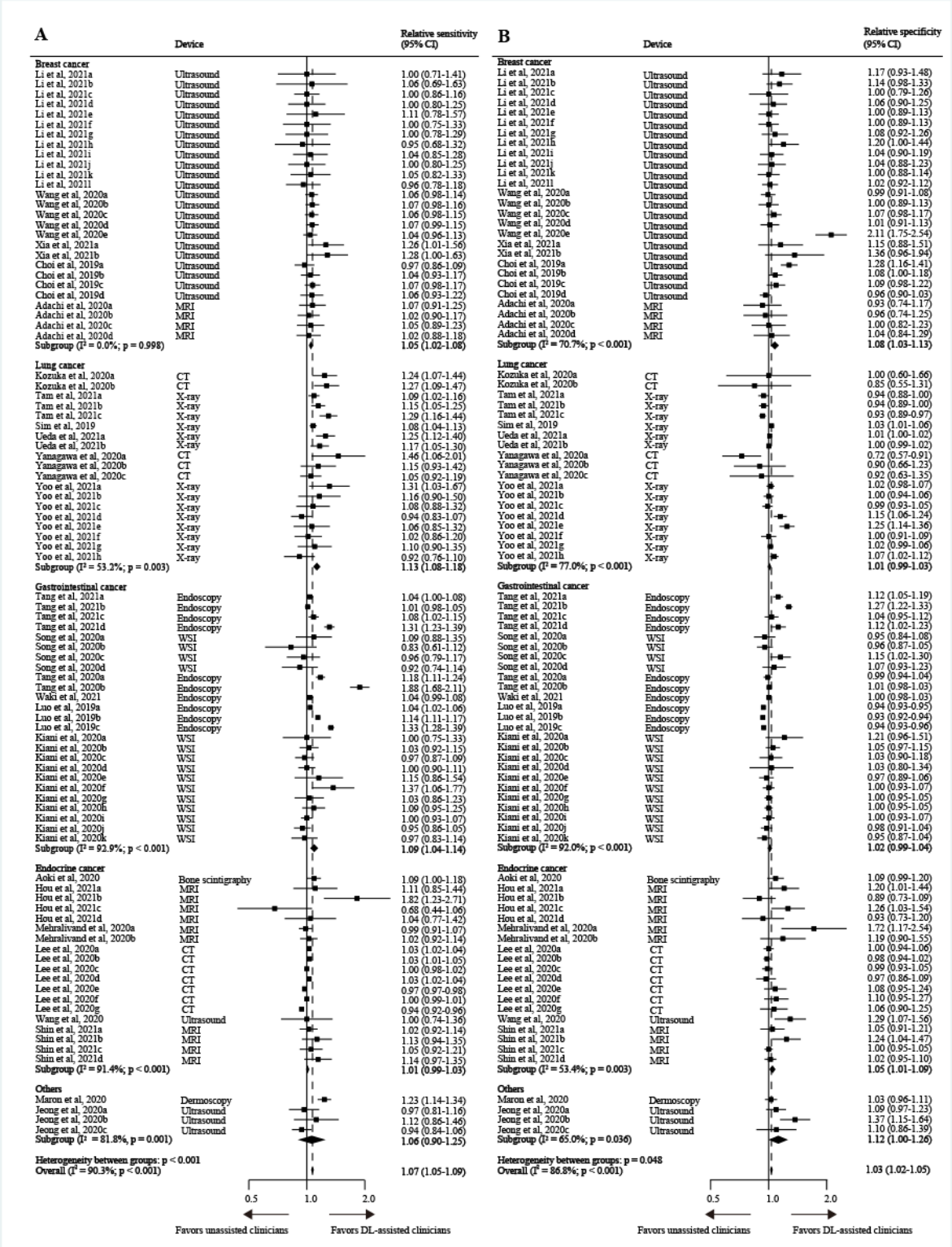
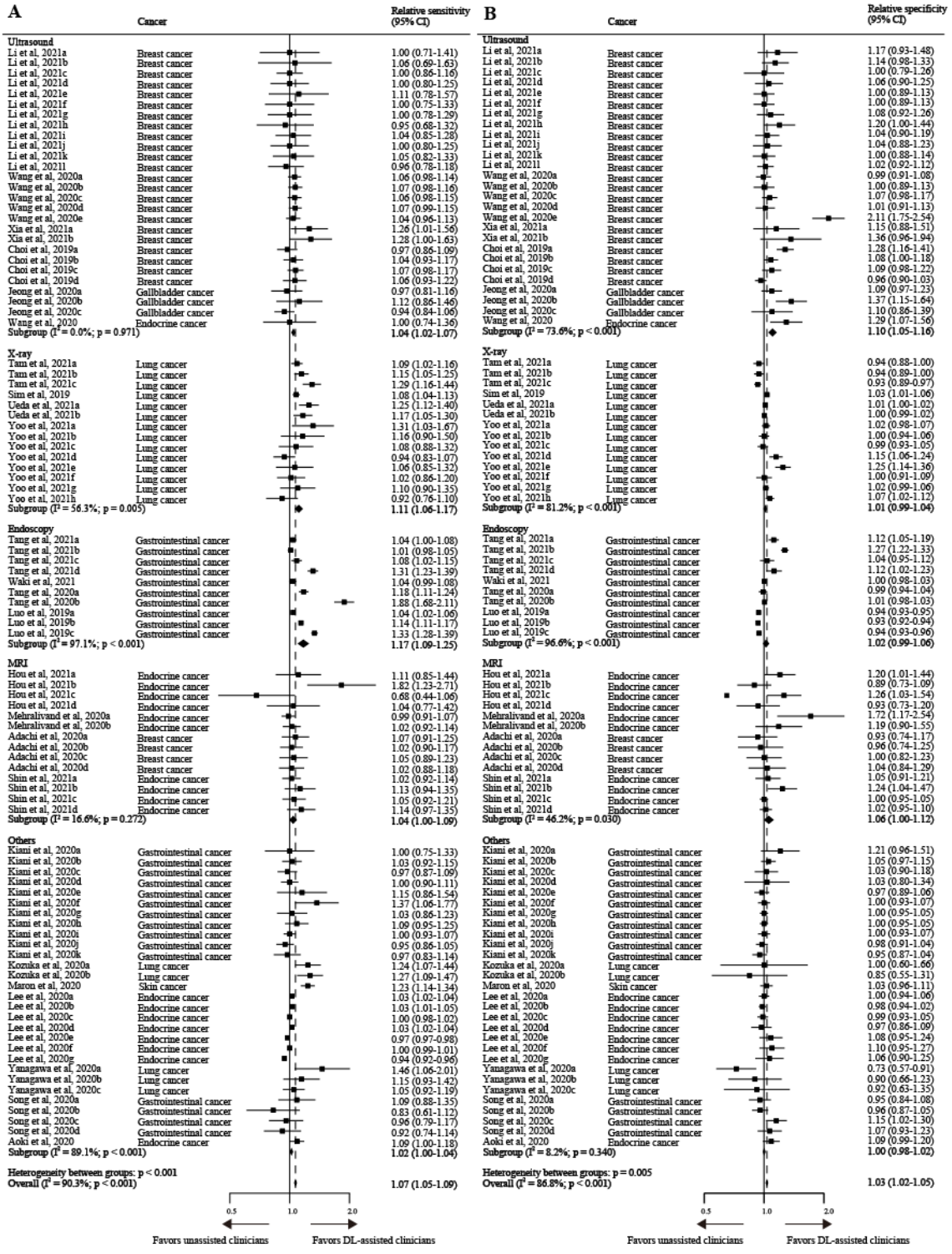


Figure 6. Pooled relative sensitivity (A) and specificity (B) of DL-assisted clinicians versus unassisted clinicians for different cancer types in image-based cancer detection. The data are presented as forest plots for all studies with different cancer types included in the meta-analysis (n=25 studies). If a study provided multiple contingency tables for DL-assisted clinicians versus unassisted clinicians, it is listed more than once and labelled alphabetically. DL: deep learning.



Quality Assessment

Quality was assessed using the QUADAS-2 tool; findings are summarized in Multimedia Appendix 1. The risk of bias and concerns of applicability for each study are also outlined in Multimedia Appendix 1. Nine and 25 studies were considered to have high and unclear risk in the patient-selection domain, respectively. Selection criteria in these studies were unreported, unclear, or considered inappropriate. The overall methodological quality was fair, and the applicability concerns were deemed acceptable.

Discussion

Principal Findings

We performed the first reported systematic review with a meta-analysis to assess the diagnostic accuracy of unassisted clinicians versus DL-assisted clinicians across distinct cancer types and imaging modalities. Evidence suggests that DL-assisted clinicians perform better at cancer identification than unassisted clinicians. DL-assisted clinicians also appeared to be superior across all cancer types and imaging modalities analyzed here. This suggests that DL assistance can be applied across different fields of image-based cancer identification.

Overall, the pooled sensitivity and pooled specificity values for DL-assisted clinicians were higher than for unassisted clinicians, at ratios of 1.07 (95% CI 1.05-1.09) and 1.03 (95% CI 1.02-1.05), respectively. Meta-analytical findings also support Budd et al [66] and Maadi et al [67], who acknowledged that a practical collaboration between humans and AI would improve clinical practice. Evidence is continually emerging that DL-assisted clinicians outperform unassisted clinicians in the diagnosis of breast and endocrine cancer, in terms of both sensitivity and specificity. Similar superiority occurs when using ultrasound. These results are consistent with previous research [68,69] that reported improvements in diagnostic performance. As for endocrine cancer, increments in specificity have been observed from 77% to 82%, despite there being no significant increases in sensitivity. Conversely, in lung and gastrointestinal cancer, increments in sensitivity have been observed from 70% to 80%, and from 84% to 91%, respectively, despite there being no significant increases in specificity. Similarly, when analyzing X-rays and endoscopic findings, increments in sensitivity have also been observed from 70% to 78% and from 80% to 94%, respectively, but there was no significant increase in specificity. The absence of a significant increase in specificity with DL assistance might be the result of a threshold effect. In other words, the lack of an improvement in specificity may have been a trade-off against an improvement in sensitivity. Among cancer types and imaging modalities, we observed that the sensitivity of DL-assisted clinicians ranged from 80% for lung cancer to 91% for gastrointestinal cancer, and from 78% with X-rays to 94% with endoscopy. The specificity of DL-assisted clinicians ranged from 82% for endocrine cancer to 93% for gastrointestinal cancer, and from 82% using MRI to 91% using endoscopy. Diagnostic performance disparities may be attributable to the participant composition of studies, designs, disease prevalence, clinical end points, cancer stage or histology type, and device type. This helps us to understand diagnostic gaps and to promote more accurate image-based cancer diagnoses.

Analysis of the Main Aspects

These observations suggest more balanced cutoff points may be necessary to train DL models to augment diagnostic sensitivity. This may also mean that DL assistance should be matched with an accepted level of diagnostic specificity that has yet to be determined. Future work may look to focus on reducing the number of false-positive results, but there does appear to be a reduction in the number of false negatives that can be attributed to DL assistance. Another possibility for improving specificity while retaining high sensitivity might be to combine DL assistance models with advanced screening or diagnostic technologies. However, this will require a more detailed health-economics analysis, and regulatory bodies will need to consider the affordability of these new workflows. From a system perspective, DL assistance will have to be adapted to new, more advanced technologies while being trained to adapt to changing workflows, similar to human physicians.

Our findings support the notion that human-computer collaboration represents an improvement over (or is at least equivalent to) clinicians working without assistance to identify cancer cases. However, the knowledge base suffers from broad methodological deficiencies and poor reporting, although this could be overcome during training. This review shows that research in this area is still in the early stages of development. Less than half of the included studies were eligible for meta-analysis. Many studies were excluded at the initial screening stage because they only assessed the diagnostic performance of human intelligence versus machine intelligence, rather than human intelligence with DL assistance, which does not reflect a logical progression. Nevertheless, we acknowledge that assessing the accuracy of DL algorithms in isolation compared to human clinicians is often the first step for new technologies, although this does not represent a real-world situation, and conducting such polarized research may be the cause of anxieties within the profession. We hold the opinion that technologies are created to assist medical professionals, rather than being developed as replacements. Misunderstandings should be avoided and research should aim for human-in-the-loop DL for optimal integration. Therefore, we should work to integrate data science training into clinical training (and vice versa) to ensure research, at a fundamental level, is truly interdisciplinary and practicable.

While it is encouraging to see that DL assistance improves cancer case identification, caution should be exercised when applying our findings to clinical practice, because the studies under review were generally based upon in silico research. Reporting standards, which are essential to assess study quality, may not be considered as important by data scientists. There are certainly divergences in how research is conducted and reported that have impacted this emerging evidence base. In clinical research, comparative studies should be considered for the primary technical assessment of DL-assisted clinicians. This is not just about the quantification of diagnostic effects, as if it were, we would overestimate the benefit by overfitting. Of course, in silico research will continue to play an important role in simulating DL-assisted clinical practice and is useful for ensuring safety, effectiveness, and patient acceptability. However, these studies must be used to pave the way for large-scale clinical trials and then on to real-world studies. There are many gaps in this knowledge base, and as we anticipate a great deal of future research, the architecture of the knowledge base should be considered in more detail. Implementing DL assistance in clinical settings will require a blend of research methods, overlapping disciplines, and more sophisticated collaboration. This has implications for project and program leadership, although these topics are beyond the scope of this

study. Suffice to say, if we are going to advance clinical practice, we ought to design DL assistance with reflective clinical advice and through mixed methods analysis, which demands improved reporting.

Recently, the DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence) reporting guidelines for the early-stage clinical evaluation of decision support systems were published [70]. The guidelines address issues in AI and DL development through exploratory research before large-scale efficacy testing is conducted. The majority of the studies included in this paper were probably conceived (and performed) before the DECIDE-AI guidelines were published. Therefore, it is reasonable to assume that the design features and reporting used to assess the diagnostic performance of DL-assisted clinicians will improve, as will transparency. However, with the future in mind, we should also assume that the DECIDE-AI guidelines are a prototype and will continue to be developed. Researchers will be at fault if they adopt these parameters without considering clinicians' perspectives. Some have already commented that the use of human-in-the-loop AI or DL to support image-based cancer diagnostics might represent the optimal strategy for real-world clinical practice [66,67,71]. Nevertheless, we need to acknowledge that this is a rapidly evolving research area that will require subsequent updates. We are seeing the emergence of decentralized and hybridized trials that will, like interdisciplinary medical-AI research, need to accept increasingly sophisticated reporting standards.

Another major problem we encountered when considering discussions within each study was that of "cooperation." Several researchers and theorists have already commented on the problems with possible cooperation modes between clinicians and DL algorithms [72]. For example, junior clinicians, who lack practical experience, may overly rely on AI while more experienced clinicians may find it difficult to make judgments and may find DL assistance forces them into a prolonged state of cognitive dissonance. This may also lead experienced experts and senior clinicians to distrust DL assistance, which may (in the long run) mean they are more likely to reject DL assistance. One possible approach to improving human-computer collaboration would be to leverage the advantages of DL algorithms (eg, rapid, automated detection) while having clinicians situated at various "checkpoints" to fill gaps where algorithms are not assured, or where they may fall short due to underlying biases. Cases are likely to vary, some with high confidence or, more concisely, with a high probability of the presence (or absence) of cancer relative to the probabilistic threshold for cancer detection, while others will be associated with less confidence. In cases with less confidence, outputs could be considered more closely by clinicians to generate combined decisions. This means we may need to weight DL-based decisions around algorithmic confidence, which may encourage symbiosis between human clinicians and DL algorithms.

We also note that approximately one-third of the included studies used data from open-access repositories, while the remainder made use of nonpublic data sets. These sources do not provide a valid within-source benchmark for comparison. Researchers have referred to the limited availability of open-access data and codes and the risk of bias and overfitting in existing DL research [9,73-75]. Our review supports these assertions and echoes these concerns. Lacking public data sets is a fundamental cause of the growing digital health divide [76]. Therefore, we encourage the DL and health care communities to collaborate and increase the number of studies that compare DL-assisted clinicians and unassisted clinicians. This will ensure clarity when designing interdisciplinary research and interpreting data, and it will also encourage acceptance by both clinicians and patients.

Limitations and Recommendations

Before providing recommendations, we should discuss the limitations of this study. Our search strategy might have unintentionally excluded some pertinent DL-assisted studies and potentially useful non-English references. There were also substantial differences in patient characteristics, cancer types, imaging modalities, diagnostic thresholds, DL algorithms, and clinician experiences. These likely impacted measures of heterogeneity, although the fundamental purpose of this study was to systematically review evidence. We focused specifically on studies of DL-assisted clinicians and image-based cancer diagnostics. Therefore, the goal was not really to obtain generalizable findings but to identify gaps in our current knowledge. While we provide a quality assessment for transparency, the QUADAS-2 tool is suboptimal for assessing AI diagnostic research. Given the importance of AI and DL technologies, we would encourage global stakeholders to develop a new QUADAS-AI/DL tool that assesses the risk of bias and applicability.

Finally, we must emphasize that reliable estimates for performance can only be achieved through well-designed and well-executed studies that minimize bias in conduct and reporting. There remains uncertainty around the estimates of diagnostic performance provided in this exploratory meta-analysis, which should be considered before implementation.

Conclusions

Human-machine cooperation in cancer diagnosis using medical imaging holds enormous potential. We found that the diagnostic accuracy of DL-assisted clinicians appeared better than unassisted clinicians. This area warrants further investigation, and we acknowledge that more rigorously designed, transparently reported, higher-quality studies are needed. This may help facilitate the transition of DL assistance into clinical practice, although further interdisciplinary mixed methods research is required.

Acknowledgments

This work was financially supported by the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (CAMS 2021-I2M-1-004).

Authors' Contributions

PX, Y Qiao, YJ, and Y Qu contributed to the conception and design of the study. MS, DQ, BW, ZY, MC, SW, BZ, CS, MD, WZ, and AB contributed to the literature search and data extraction. PX, MS, SS, HY, LD, YZ, RR, and SZ contributed to data analysis and interpretation. PX, SS, Y Qiao, and YJ contributed to critical revision of the manuscript. PX, MS, and SS contributed to writing the manuscript, and all authors approved the manuscript. PX, Y Qiao, and YJ guarantee the integrity of the work. PX, MS, DQ, BW, and SS contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplement document.
[DOCX File , 5012 KB-Multimedia Appendix 1]

References

1. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. Int J Cancer 2021 Apr 05;149(4):778-789 [FREE Full text] [doi: 10.1002/ijc.33588] [Medline: 33818764]
2. Crosby D, Bhatia S, Brindle KM, Coussens LM, Dive C, Emberton M, et al. Early detection of cancer. Science 2022 Mar 18;375(6586):eaay9040 [doi: 10.1126/science.aay9040] [Medline: 35298272]
3. Dillekås H, Rogers MS, Straume O. Are 90% of deaths from cancer caused by metastases? Cancer Med 2019 Sep;8(12):5574-5576 [FREE Full text] [doi: 10.1002/cam4.2474] [Medline: 31397113]
4. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015 Oct;35(6):1668-1676 [doi: 10.1148/rg.2015150023] [Medline: 26466178]
5. Coiera E. The fate of medicine in the time of AI. Lancet 2018 Dec 01;392(10162):2331-2332 [doi: 10.1016/S0140-6736(18)31925-1] [Medline: 30318263]
6. Kleppe A, Skrede O, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer 2021 Mar;21(3):199-211 [doi: 10.1038/s41568-020-00327-9] [Medline: 33514930]
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444 [doi: 10.1038/nature14539] [Medline: 26017442]
8. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021 Apr 07;4(1):65 [FREE Full text] [doi: 10.1038/s41746-021-00438-z] [Medline: 33828217]
9. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019 Oct;1(6):e271-e297 [FREE Full text] [doi: 10.1016/S2589-7500(19)30123-2] [Medline: 33323251]
10. Zheng Q, Yang L, Zeng B, Li J, Guo K, Liang Y, et al. Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: A systematic review and meta-analysis. EClinicalMedicine 2021 Jan;31:100669 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100669] [Medline: 33392486]
11. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: 10.1001/jamainternmed.2018.3763] [Medline: 30128552]
12. Knop M, Weber S, Mueller M, Niehaves B. Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence-enabled clinical decision support systems: literature review. JMIR Hum Factors 2022 Mar 24;9(1):e28639 [FREE Full text] [doi: 10.2196/28639] [Medline: 35323118]
13. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 09;293(10):1223-1238 [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
14. Grote T, Berens P. How competitors become collaborators-Bridging the gap(s) between machine learning algorithms and clinicians. Bioethics 2022 Feb;36(2):134-142 [doi: 10.1111/bioe.12957] [Medline: 34599834]
15. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]

16. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015 Apr 11;115(3):211-252 [doi: 10.1007/s11263-015-0816-y]
17. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
18. Aoki Y, Nakayama M, Nomura K, Tomita Y, Nakajima K, Yamashina M, et al. The utility of a deep learning-based algorithm for bone scintigraphy in patient with prostate cancer. Ann Nucl Med 2020 Dec;34(12):926-931 [doi: 10.1007/s12149-020-01524-0] [Medline: 32955663]
19. Aresta G, Ferreira C, Pedrosa J, Araujo T, Rebelo J, Negrao E, et al. Automatic lung nodule detection combined with gaze information improves radiologists' screening performance. IEEE J Biomed Health Inform 2020 Oct;24(10):2894-2901 [doi: 10.1109/JBHI.2020.2976150] [Medline: 32092022]
20. Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J Invest Dermatol 2020 Sep;140(9):1753-1761 [FREE Full text] [doi: 10.1016/j.jid.2020.01.019] [Medline: 32243882]
21. Harmon SA, Sanford TH, Brown GT, Yang C, Mehralivand S, Jacob JM, et al. Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. JCO Clin Cancer Inform 2020 Apr;4:367-382 [FREE Full text] [doi: 10.1200/CCI.19.00155] [Medline: 32330067]
22. Hou Y, Zhang Y, Bao J, Bao M, Yang G, Shi H, et al. Artificial intelligence is a promising prospect for the detection of prostate cancer extracapsular extension with mpMRI: a two-center comparative study. Eur J Nucl Med Mol Imaging 2021 Nov;48(12):3805-3816 [doi: 10.1007/s00259-021-05381-5] [Medline: 34018011]
23. Li H, Ye J, Liu H, Wang Y, Shi B, Chen J, et al. Application of deep learning in the detection of breast lesions with four different breast densities. Cancer Med 2021 Jul;10(14):4994-5000 [FREE Full text] [doi: 10.1002/cam4.4042] [Medline: 34132495]
24. Polónia A, Campelos S, Ribeiro A, Aymore I, Pinto D, Biskup-Fruzynska M, et al. Artificial intelligence improves the accuracy in histologic classification of breast lesions. Am J Clin Pathol 2021 Mar 15;155(4):527-536 [doi: 10.1093/ajcp/aqaa151] [Medline: 33118594]
25. Li C, Li J, Tan T, Chen K, Xu Y, Wu R. Application of ultrasonic dual-mode artificially intelligent architecture in assisting radiologists with different diagnostic levels on breast masses classification. Diagn Interv Radiol 2021 May;27(3):315-322 [FREE Full text] [doi: 10.5152/dir.2021.20018] [Medline: 34003119]
26. Neal Joshua ES, Bhattacharyya D, Chakkravarthy M, Byun Y. 3D CNN with visual insights for early detection of lung cancer using gradient-weighted class activation. J Healthc Eng 2021;2021:6695518 [FREE Full text] [doi: 10.1155/2021/6695518] [Medline: 33777347]
27. Liu S, Zhang C, Liu R, Li S, Xu F, Liu X, et al. CT texture analysis for preoperative identification of lymphoma from other types of primary small bowel malignancies. Biomed Res Int 2021;2021:5519144 [FREE Full text] [doi: 10.1155/2021/5519144] [Medline: 33884262]
28. Kwiatkowska D, Kluska P, Reich A. Convolutional neural networks for the detection of malignant melanoma in dermoscopy images. Postepy Dermatol Alergol 2021 Jun;38(3):412-420 [FREE Full text] [doi: 10.5114/ada.2021.107927] [Medline: 34377121]
29. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med 2020;3:23 [FREE Full text] [doi: 10.1038/s41746-020-0232-8] [Medline: 32140566]
30. Kozuka T, Matsukubo Y, Kadoba T, Oda T, Suzuki A, Hyodo T, et al. Efficiency of a computer-aided diagnosis (CAD) system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography. Jpn J Radiol 2020 Nov;38(11):1052-1061 [doi: 10.1007/s11604-020-01009-0] [Medline: 32592003]
31. Lucius M, De All J, De All JA, Belvisi M, Radizza L, Lanfranconi M, et al. Deep neural frameworks improve the accuracy of general practitioners in the classification of pigmented skin lesions. Diagnostics (Basel) 2020 Nov 18;10(11):969 [FREE Full text] [doi: 10.3390/diagnostics10110969] [Medline: 33218060]
32. Mehralivand S, Harmon SA, Shih JH, Smith CP, Lay N, Argun B, et al. Multicenter multireader evaluation of an artificial intelligence-based attention mapping system for the detection of prostate cancer with multiparametric MRI. AJR Am J Roentgenol 2020 Oct;215(4):903-912 [FREE Full text] [doi: 10.2214/AJR.19.22573] [Medline: 32755355]
33. Lee JH, Ha EJ, Kim D, Jung YJ, Heo S, Jang Y, et al. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. Eur Radiol 2020 Jun;30(6):3066-3072 [doi: 10.1007/s00330-019-06652-4] [Medline: 32065285]
34. Mango VL, Sun M, Wynn RT, Ha R. Should we ignore, follow, or biopsy? Impact of artificial intelligence decision support on breast ultrasound lesion assessment. AJR Am J Roentgenol 2020 Jun;214(6):1445-1452 [FREE Full text] [doi: 10.2214/AJR.19.21872] [Medline: 32319794]
35. Zhao Y, Xue D, Wang Y, Zhang R, Sun B, Cai Y, et al. Computer-assisted diagnosis of early esophageal squamous cell carcinoma using narrow-band imaging magnifying endoscopy. Endoscopy 2019 Apr;51(4):333-341 [doi: 10.1055/a-0756-8754] [Medline: 30469155]

36. Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng 2021 Jun;5(6):522-532 [doi: 10.1038/s41551-021-00711-2] [Medline: 33875840]
37. van Winkel SL, Rodríguez-Ruiz A, Appelman L, Gubern-Mérida A, Karssemeijer N, Teuwen J, et al. Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. Eur Radiol 2021 Nov;31(11):8682-8691 [FREE Full text] [doi: 10.1007/s00330-021-07992-w] [Medline: 33948701]
38. Yue M, Zhang J, Wang X, Yan K, Cai L, Tian K, et al. Can AI-assisted microscope facilitate breast HER2 interpretation? A multi-institutional ring study. Virchows Arch 2021 Sep;479(3):443-449 [doi: 10.1007/s00428-021-03154-x] [Medline: 34279719]
39. Tam MDBS, Dyer T, Dissez G, Morgan TN, Hughes M, Illes J, et al. Augmenting lung cancer diagnosis on chest radiographs: positioning artificial intelligence to improve radiologist performance. Clin Radiol 2021 Aug;76(8):607-614 [doi: 10.1016/j.crad.2021.03.021] [Medline: 33993997]
40. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: a retrospective evaluation. Radiology 2021 Jul;300(1):57-65 [doi: 10.1148/radiol.2021203555] [Medline: 33944627]
41. Wang Y, Choi EJ, Choi Y, Zhang H, Jin GY, Ko S. Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. Ultrasound Med Biol 2020 May;46(5):1119-1132 [doi: 10.1016/j.ultrasmedbio.2020.01.001] [Medline: 32059918]
42. Xia Q, Cheng Y, Hu J, Huang J, Yu Y, Xie H, et al. Differential diagnosis of breast cancer assisted by S-Detect artificial intelligence system. Math Biosci Eng 2021 Apr 27;18(4):3680-3689 [FREE Full text] [doi: 10.3934/mbe.2021184] [Medline: 34198406]
43. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020 Aug;159(2):512-520.e7 [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]
44. Maron RC, Utikal JS, Hekler A, Hauschild A, Sattler E, Sondermann W, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: web-based survey study. J Med Internet Res 2020 Sep 11;22(9):e18091 [FREE Full text] [doi: 10.2196/18091] [Medline: 32915161]
45. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, The DM DREAM Consortium, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020 Mar 02;3(3):e200265 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0265] [Medline: 32119094]
46. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology 2020 Jan;294(1):199-209 [doi: 10.1148/radiol.2019182465] [Medline: 31714194]
47. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. Korean J Radiol 2019 May;20(5):749-758 [FREE Full text] [doi: 10.3348/kjr.2018.0530] [Medline: 30993926]
48. Adachi M, Fujioka T, Mori M, Kubota K, Kikuchi Y, Xiaotong W, et al. Detection and diagnosis of breast cancer using artificial intelligence based assessment of maximum intensity projection dynamic contrast-enhanced magnetic resonance images. Diagnostics (Basel) 2020 May 20;10(5):330 [FREE Full text] [doi: 10.3390/diagnostics10050330] [Medline: 32443922]
49. Jeong Y, Kim JH, Chae H, Park S, Bae JS, Joo I, et al. Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: Preliminary results. Sci Rep 2020 May 07;10(1):7700 [FREE Full text] [doi: 10.1038/s41598-020-64205-y] [Medline: 32382062]
50. Fink C, Blum A, Buhl T, Mitteldorf C, Hofmann-Wellenhof R, Deinlein T, et al. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. J Eur Acad Dermatol Venereol 2020 Jun;34(6):1355-1361 [doi: 10.1111/jdv.16165] [Medline: 31856342]
51. Zhang Y, Wang Z, Zhang J, Wang C, Wang Y, Chen H, et al. Deep learning model for classifying endometrial lesions. J Transl Med 2021 Jan 06;19(1):10 [FREE Full text] [doi: 10.1186/s12967-020-02660-x] [Medline: 33407588]
52. Ueda D, Yamamoto A, Shimazaki A, Walston SL, Matsumoto T, Izumi N, et al. Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: a retrospective clinical validation study. BMC Cancer 2021 Oct 18;21(1):1120 [FREE Full text] [doi: 10.1186/s12885-021-08847-9] [Medline: 34663260]
53. Sui H, Ma R, Liu L, Gao Y, Zhang W, Mo Z. Detection of incidental esophageal cancers on chest CT by deep learning. Front Oncol 2021;11:700210 [FREE Full text] [doi: 10.3389/fonc.2021.700210] [Medline: 34604036]
54. Shin I, Kim H, Ahn SS, Sohn B, Bae S, Park JE, et al. Development and validation of a deep learning-based model to distinguish glioblastoma from solitary brain metastasis using conventional MR images. AJNR Am J Neuroradiol 2021 May;42(5):838-844 [FREE Full text] [doi: 10.3174/ajnr.A7003] [Medline: 33737268]

55. Tang D, Zhou J, Wang L, Ni M, Chen M, Hassan S, et al. A novel model based on deep convolutional neural network improves diagnostic accuracy of intramucosal gastric cancer (with video). Front Oncol 2021;11:622827 [FREE Full text] [doi: 10.3389/fonc.2021.622827] [Medline: 33959495]
56. Yanagawa M, Niioka H, Kusumoto M, Awai K, Tsubamoto M, Satoh Y, et al. Diagnostic performance for pulmonary adenocarcinoma on CT: comparison of radiologists with and without three-dimensional convolutional neural network. Eur Radiol 2021 Apr;31(4):1978-1986 [doi: 10.1007/s00330-020-07339-x] [Medline: 33011879]
57. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol 2021 Jan;22(1):132-141 [doi: 10.1016/S1470-2045(20)30535-0] [Medline: 33387492]
58. Waki K, Ishihara R, Kato Y, Shoji A, Inoue T, Matsueda K, et al. Usefulness of an artificial intelligence system for the detection of esophageal squamous cell carcinoma evaluated with videos simulating overlooking situation. Dig Endosc 2021 Nov;33(7):1101-1109 [doi: 10.1111/den.13934] [Medline: 33502046]
59. Yoo H, Lee SH, Arru CD, Doda Khera R, Singh R, Siebert S, et al. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. Eur Radiol 2021 Dec;31(12):9664-9674 [doi: 10.1007/s00330-021-08074-7] [Medline: 34089072]
60. Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020 Aug 27;11(1):4294 [FREE Full text] [doi: 10.1038/s41467-020-18147-8] [Medline: 32855423]
61. Wang F, Liu X, Yuan N, Qian B, Ruan L, Yin C, et al. Study on automatic detection and classification of breast nodule using deep convolutional neural network system. J Thorac Dis 2020 Sep;12(9):4690-4701 [FREE Full text] [doi: 10.21037/jtd-19-3013] [Medline: 33145042]
62. Tang D, Wang L, Ling T, Lv Y, Ni M, Zhan Q, et al. Development and validation of a real-time artificial intelligence-assisted system for detecting early gastric cancer: A multicentre retrospective diagnostic study. EBioMedicine 2020 Dec;62:103146 [FREE Full text] [doi: 10.1016/j.ebiom.2020.103146] [Medline: 33254026]
63. Zhao K, Wang C, Hu J, Yang X, Wang H, Li F, et al. Prostate cancer identification: quantitative analysis of T2-weighted MR images based on a back propagation artificial neural network model. Sci China Life Sci 2015 Jul;58(7):666-673 [FREE Full text] [doi: 10.1007/s11427-015-4876-6] [Medline: 26025283]
64. Zhao L, Jia K. Multiscale CNNs for brain tumor segmentation and diagnosis. Comput Math Methods Med 2016;2016:8356294 [FREE Full text] [doi: 10.1155/2016/8356294] [Medline: 27069501]
65. Luo H, Xu G, Li C, He L, Luo L, Wang Z, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol 2019 Dec;20(12):1645-1654 [doi: 10.1016/S1470-2045(19)30637-0] [Medline: 31591062]
66. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med Image Anal 2021 Jul;71:102062 [doi: 10.1016/j.media.2021.102062] [Medline: 33901992]
67. Maadi M, Akbarzadeh Khorshidi H, Aickelin U. A review on human-AI interaction in machine learning and insights for medical applications. Int J Environ Res Public Health 2021 Feb 22;18(4):2121 [FREE Full text] [doi: 10.3390/ijerph18042121] [Medline: 33671609]
68. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, et al. A survey of deep-learning applications in ultrasound: artificial intelligence-powered ultrasound for improving clinical workflow. J Am Coll Radiol 2019 Sep;16(9 Pt B):1318-1328 [doi: 10.1016/j.jacr.2019.06.004] [Medline: 31492410]
69. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 2019 May;29(2):102-127 [FREE Full text] [doi: 10.1016/j.zemedi.2018.11.002] [Medline: 30553609]
70. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ 2022 May 18;377:e070904 [FREE Full text] [doi: 10.1136/bmj-2022-070904] [Medline: 35584845]
71. Bruckert S, Finzel B, Schmid U. The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell 2020;3:507973 [FREE Full text] [doi: 10.3389/frai.2020.507973] [Medline: 33733193]
72. Hah H, Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: mixed methods approach. J Med Internet Res 2021 Dec 16;23(12):e33540 [FREE Full text] [doi: 10.2196/33540] [Medline: 34924356]
73. Cook TS. Human versus machine in medicine: can scientific literature answer the question? Lancet Digit Health 2019 Oct;1(6):e246-e247 [FREE Full text] [doi: 10.1016/S2589-7500(19)30124-4] [Medline: 33323246]
74. Ter Riet G, Bachmann LM, Kessels AGH, Khan KS. Individual patient data meta-analysis of diagnostic studies: opportunities and challenges. Evid Based Med 2013 Oct;18(5):165-169 [doi: 10.1136/eb-2012-101145] [Medline: 23704701]
75. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]

76. Wen D, Khan SM, Ji Xu A, Ibrahim H, Smith L, Caballero J, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit Health 2022 Jan;4(1):e64-e74 [FREE Full text] [doi: 10.1016/S2589-7500(21)00252-1] [Medline: 34772649]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence
DL: deep learning
HSROC: hierarchical summary receiver operating characteristic
MRI: magnetic resonance imaging
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2

Edited by A Mavragani; submitted 26.10.22; peer-reviewed by L Guo, K Gupta, N Jiwani; comments to author 11.01.23; revised version received 19.01.23; accepted 13.02.23; published 02.03.23

Please cite as:
Xue P, Si M, Qin D, Wei B, Seery S, Ye Z, Chen M, Wang S, Song C, Zhang B, Ding M, Zhang W, Bai A, Yan H, Dang L, Zhao Y, Rezhake R, Zhang S, Qiao Y, Qu Y, Jiang Y
Unassisted Clinicians Versus Deep Learning–Assisted Clinicians in Image-Based Cancer Diagnostics: Systematic Review With Meta-analysis
J Med Internet Res 2023;25:e43832
URL: https://www.jmir.org/2023/1/e43832
doi: 10.2196/43832
PMID: 36862499