JAMA Dermatology | Original Investigation | AI IN DERMATOLOGY

Prospective Evidence on Artificial Intelligence–Assisted Melanoma Diagnostics
A Systematic Review and Meta-Analysis

Sara Laiouar-Pedari, PhD; Arlene Kühn, MSc; Christoph Wies, MSc; Carina Nogueira Garcia, MD; Jana Therés Winterstein, MSc; Lukas Heinlein, MSc; Annemarie Hoffsommer, MSc; Tirtha Chanda, MSc; Sarah Haggenmüller, PhD; Titus J. Brinker, MD

IMPORTANCE Dermoscopy is a standard of care for melanoma diagnostics, and artificial intelligence (AI) systems are increasingly investigated as decision-support tools. Prospective evidence is essential to assess their performance compared to dermatologists.

OBJECTIVES To evaluate the diagnostic performance of dermatologists, AI systems, and dermatologists assisted by AI in prospective studies of melanoma detection, and to assess the readiness of AI for clinical use.

DATA SOURCES PubMed, Embase, Web of Science, and Google Scholar were searched from inception through July 9, 2025.

STUDY SELECTION Eligible studies were prospective, used dermoscopic images, and reported or allowed calculation of performance metrics for dermatologists, AI, or dermatologists assisted by AI against a histopathologic reference standard. Nondermoscopic comparators and retrospective designs were excluded. Studies with 20 or fewer histopathologically confirmed melanomas were excluded a priori from quantitative synthesis.

DATA EXTRACTION AND SYNTHESIS Two reviewers independently screened and extracted data and discrepancies or missing values were clarified among all authors. Risk of bias and applicability were assessed with QUADAS-2 and QUADAS-C. Study-level sensitivity and specificity were summarized and plotted; head-to-head comparisons were analyzed descriptively.

MAIN OUTCOMES AND MEASURES Diagnostic outcomes were sensitivity, specificity, accuracy, and balanced accuracy for melanoma detection.

RESULTS Eleven prospective studies with a total of more than 2500 patients and 50 participant-dermatologists were included in the analyses. Dermatologists achieved a pooled sensitivity of 78.6% (95% CI, 67.5%-88.1%) and specificity of 75.2% (95% CI, 63.3%-84.3%), whereas AI alone reached 80.9% (95% CI, 63.6%-94.5%) sensitivity and 75.6% (95% CI, 64.5%-85.6%) specificity. In the single study reporting AI-assisted dermatologists, sensitivity was 91.9% and specificity was 83.7%. In direct clinical comparisons, AI demonstrated higher specificity and similar sensitivity. Most studies were at high risk of bias in patient selection and index test domains, primarily due to the preselection of lesions suspected of melanoma and binary classifications.

CONCLUSIONS AND RELEVANCE In the systematic review and meta-analysis of prospective settings, AI systems perform at comparable levels to dermatologists for melanoma diagnostics and may enhance performance when used as a decision-support tool. However, the frequent risk of bias and limited generalizability of current studies highlight the need for broader validation in unselected patient populations in the clinical setting.

JAMA Dermatol. doi:10.1001/jamadermatol.2026.0217
Published online March 25, 2026.

Supplemental content

Author Affiliations: Division of Digital Prevention, Diagnostics and Therapy Guidance, German Cancer Research Center, Heidelberg, Germany (Laiouar-Pedari, Kühn, Wies, Garcia, Winterstein, Heinlein, Hoffsommer, Chanda, Haggenmüller, Brinker); Medical Faculty, University Heidelberg, Heidelberg, Germany (Wies, Winterstein, Heinlein, Hoffsommer, Chanda).

Corresponding Author: Titus J. Brinker, MD, Division of Digital Prevention, Diagnostics and Therapy Guidance, German Cancer Research Center, Im Neuenheimer Feld 223, 69120 Heidelberg, Germany (titus.brinker@dkfz.de).

Malignant melanoma is among the most aggressive forms of skin cancer. Early detection and timely intervention are crucial for improving patient outcomes.[1] Dermoscopy, a noninvasive diagnostic technique, has become an indispensable tool in dermatology by enabling improved visualization of subsurface skin structures. Compared to unaided visual inspection, it substantially increases diagnostic accuracy for melanoma[1] and helps to reduce unnecessary excisions of benign lesions.[1-5] Despite being the current standard of care, however, its diagnostic performance remains highly dependent on the clinician's level of experience.[6]

In recent years, artificial intelligence (AI) systems have shown promising results in the automated analysis and classification of dermoscopic images. Convolutional neural networks (CNNs), in particular, have become the most commonly used approach in this field and form the basis for most modern AI models for classifying skin lesions.[7,8] Numerous retrospective studies have reported diagnostic performances comparable to, or even exceeding, those of expert dermatologists.[9-11] These findings have fueled growing interest in the integration of AI as decision-support tools to enhance melanoma detection in clinical practice.

Nevertheless, most studies evaluating AI performance have been retrospective and rely on curated image datasets that may not reflect the complexity of everyday clinical practice. Retrospective analyses are limited in their ability to assess generalizability, risk of bias, and true diagnostic impact. To address this gap, we conducted a meta-analysis focusing exclusively on prospective studies that compared the diagnostic performance of dermoscopy alone with dermoscopy supported by AI, or AI alone, in melanoma detection. By synthesizing prospective evidence, this systematic review and meta-analysis aims to inform the current state of clinical readiness for AI in melanoma diagnostics and to identify key areas requiring further validation.

Key Points

Question How does the diagnostic performance of artificial intelligence (AI) for melanoma in prospective dermoscopy studies compare with that of dermatologists?

Findings Across 11 prospective studies including more than 2500 participants, AI and dermatologists showed comparable diagnostic performance. However, the evidence base remains small, and study designs are heterogeneous, with a high risk of bias in patient selection and index test domains.

Meaning Although current findings support the potential clinical application of AI, validation remains at an early stage because larger, multicenter, and methodologically rigorous prospective studies are required to confirm the safety and clinical utility of AI in routine practice.

Methods

This systematical review and meta-analysis used only previously published data and was therefore exempted from review by the Ethics Committee of the Faculty of Medicine, Heidelberg University. The study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) reporting guidelines[12] and the protocol was registered with PROSPERO on July 21, 2025 (CRD420251084932).[13]

Study Eligibility Criteria

The Population/Patient/Problem, Intervention, Comparison, and Outcome (PICO) framework[14] guided the definition of eligibility. Included studies assessed populations of adult patients (age ≥18 years) with skin lesions suspected of malignant melanoma; interventions involving the application of AI to dermoscopic images; compared diagnostic assessments by dermatologists using dermoscopy, with or without AI support; and their primary outcomes were diagnostic performance, expressed through sensitivity, specificity, accuracy, and/or balanced accuracy. Beyond the criteria defined by the PICO framework, studies were required to be peer-reviewed, be reported in English, and have a prospective design. The latter was defined as the collection and assessment of in vivo data for both dermatologist and AI evaluations. To ensure statistical robustness, included studies also had to enroll at least 20 histopathologically confirmed melanoma cases. Studies with only 1 relevant study group (eg, AI only or dermatologist only) were also eligible because the meta-analysis allowed for indirect comparisons across modalities. Interventional studies were only included if a diagnostic end point was reported.

A study was excluded if it reported combined malignant diseases without separate data for melanoma, described algorithm development without clinical validation (eg, on open databases or challenge datasets), or originated from computer science contexts without clinical application. Within the eligible studies, we excluded any study groups that did not meet our criteria, including retrospective groups and those using clinical images instead of dermoscopic images or that used nondermoscopic comparators.

Search Strategy

A systematic literature review was conducted of PubMed, Google Scholar, Embase, and Web of Science for studies published between January 1, 2000, and July 9, 2025. The search strategy combined terms related to melanoma, dermoscopy, AI, and prospective design. Full search strategies for each database are provided in eTables 1 to 3 in Supplement 1.

Reference Management and Study Selection

The reference management tool, Zotero,[15] was used to manage citations and remove duplicates. Title and abstract screening were performed independently by 2 reviewers (S.L.P. and A.K.). Potentially eligible full-text articles were retrieved and independently assessed for inclusion by the same reviewers.

Based on the predefined search criteria, a total of 308 publications were identified across the 4 databases. After removing duplicates and articles that were not peer-reviewed, 176

records remained and were independently screened. Disagreements were resolved through discussion. Reasons for excluding other prospective studies are provided in eTable 4 in Supplement 1.

Data Extraction and Synthesis

Two reviewers (S.L.P. and A.K.) independently extracted metric data and key features for each study. Eligible studies reported performance metrics for melanoma diagnostics validated against histopathologic test results. When metrics were not reported, they were calculated from available raw data. Using recorded true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the following measures were derived: sensitivity = TP/(TP + FN); specificity = TN/(FP + TN); accuracy = (TP + TN)/(TP + FN + FP + TN); and balanced accuracy = (sensitivity + specificity)/2.

In cases of inconsistencies or missing values, the study authors were contacted for clarification. If no response was received within 1 week, missing values were recalculated from the raw data (eMethods in Supplement 1).

Performance metrics were systematically extracted and categorized by diagnostic approach: dermatologists alone, AI alone, or dermatologists assisted by AI. A meta-analysis was subsequently performed to pool performance estimates for each diagnostic modality, based on data from all included studies.

Critical Appraisal

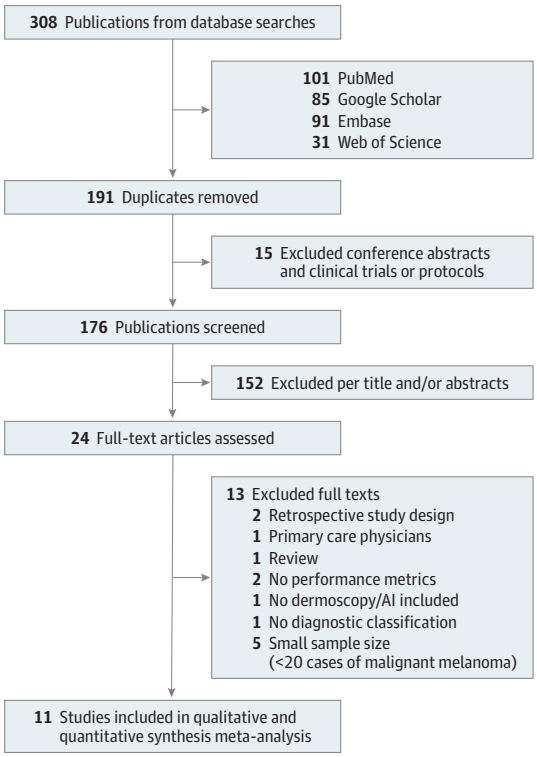
Risk of bias was evaluated using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool,[16] and, when appropriate, the QUADAS-Comparative (QUADAS-C) tool,[17] specifically designed for comparative diagnostic accuracy studies. Two reviewers (S.L.P. and A.K.) conducted the assessment independently, and any disagreements were resolved through consultation with a third reviewer (C.N.G.). The overall quality of studies was evaluated by considering the proportion of studies with high vs low risk of bias, and the consistency of reported performance metrics across studies.

Statistical Analysis

If not directly reported, sensitivity, specificity, accuracy, and balanced accuracy were calculated from 2 × 2 contingency tables provided in the studies. If 95% CIs for sensitivity or specificity were not reported, they were estimated using the bootstrap method; the same approach was applied to derive pooled-matrix CIs at study level.[18] For method-specific sensitivity and specificity, pooled metrics were calculated by averaging all metrics within a given study group. The summary receiver operating characteristic (SROC) curve was derived according to the approach proposed by Moses et al.[19]

Statistical analyses were performed using R, version 4.1.2 (R Foundation for Statistical Computing) using the boot package, version 1.3-28, and the base stats package. Forest plots were generated with the forest plot package, version 3.1.1, and SROC and the box plots were created using the ggplot2 package, version 2.4.4. Differences between groups were evaluated using the nonparametric Wilcoxon rank sum test.

Figure 1. PRISMA Flow Diagram of Included Studies



AI indicates artificial intelligence.

Results

Eleven studies met the predefined eligibility criteria (**Figure 1**), evaluating the diagnostic performance of dermatologists alone, AI alone, or dermatologists assisted by AI in prospective dermoscopic-image settings. Findings of histopathologic testing of melanoma-suspected lesions served as the reference standard in all studies. For benign lesions, the reference standard varied across studies and included histopathologic results, clinical follow-up, or expert consensus. Study populations differed considerably, with malignant melanoma cases ranging from 26 to 653 and nonmalignant cases, from 88 to 4495. Eight studies directly compared dermatologists' performance with that of AI.

Among the 11 eligible studies, 3 study groups were excluded due to retrospective design[20,21] or the use of clinical rather than dermoscopic images.[22] Additionally, some of the studies used comparators outside the scope of this meta-analysis, such as general physicians using AI,[23] confocal laser scanning microscopy,[24] naked-eye examination, and telespectrophotometry,[25] teledermatology, spectroscopy, or multispectral imaging[26]; therefore, these nondermoscopic study groups were excluded from the analysis. Furthermore, 1 study assessing dermatologist performance with AI support reported only area under the curve (AUC) without sensitivity or specificity metrics and thus, could not be included in the meta-analysis.[27]

Table. Study Characteristics of Included Studies[a]

Source	Cases/patients	Ground truth	Study group	Diagnostic categories or classification methods or reason for exclusion
Phillips et al,[20] 2019	MM (n = 125), others (n = 426)/(n = 514)	HPE for suspected lesions	Dermatologists group	MM, dysplastic nevi, or other lesions
			Not included: SkinAnalytics (AI algorithm)	Data for AI classification were prospectively collected but retrospectively analyzed
Heinlein et al,[21] 2024	MM (n = 653), others (n = 918)/(n = 1716)	HPE for all lesions	Dermatologists group	MM vs non-MM (dysplastic nevi or other lesions)
			Not included: "all data are external" (ADAE) algorithm (CNN models)	Data for AI classification were prospectively collected but retrospectively analyzed
Maier et al,[22] 2015	MM (n = 26), others (n = 119)/NA	HPE for all lesions	Dermatologists group (n = 2)	MM vs non-MM (dysplastic nevi + nevi)
			Not included: SkinVision app (CNN model)	Clinical images used for AI analysis
Dreiseitl et al,[23] 2009	MM (n = 27), others (n = 431)/(n = 511)	HPE for suspected lesions; follow-up for other lesions	Dermatologists group (n = 1)	MM vs non-MM
			Not included: nonexpert physicians + MoleMax II	Nonexpert physicians
Langley et al,[24] 2007	MM (n = 37) others (n = 88)/(n = 125)	HPE for all lesions	Dermatologists group (n = 1)	Melanocytic nevi vs MM
			Not included: Confocal scanning laser microscopy	No dermoscopic or AI system
Bono et al,[25] 2002	MM (n = 66), others (n = 247)/(n = 298)	HPE for all lesions	Dermatologists group (n = 1)	MM vs non-MM
			Not included: naked-eye telespectrophotometry	Naked eye and telespectrosphotometry are not dermoscopic techniques
MacLellan et al,[26] 2021	MM (n = 59), others (n = 150)/(n = 184)	HPE for all lesions	Dermatologists group (n = 2)	Management decision instead of diagnosis: excise, do not excise, observe
			Not included: teledermatologists, MelaFind, Verisante Aura	Teledermatology and for AI, no dermoscopic systems included
			AI group: FotoFinder Pro (CNN model)	Probability score between 0 and 1, with a threshold for melanoma of at least 0.5
			AI group: FotoFinder Tueb (CNN model)	Probability score between 0 and 1, with a threshold for melanoma of at least 0.5
Marchetti et al,[27] 2023	MM (n = 95), others (n = 508)/(n = 435)	HPE for suspected lesions	AI group: AI algorithm (ADAE-CNN models)	Probability malignancy score
			Not included: dermatologists (n = 11) + ADAE (CNN model)	Only AUC metrics
Thomas et al,[28] 2023	MM (n = 140), others (n = 4635)/NA	HPE for suspected lesions; CA or HPE for other lesions	AI group: DERM vA (CNN model) clinic 1	7-Class classification: MM, SCC, BCC, IEC, AK, AN, or benign
	MM (n = 58), others (n = 2527)/NA		AI group: DERM vB (CNN model) clinic 1	7-Class classification: MM, SCC, BCC, IEC, AK, AN, or benign
	MM (n = 33), others (n = 676)/NA		AI group: DERM vA (CNN model) clinic 2	7-Class classification: MM, SCC, BCC, IEC, AK, AN, or benign
	MM (n = 18), others (n = 624)/NA		AI group: DERM vB (CNN model) clinic 2	7-Class classification: MM, SCC, BCC, IEC, AK, AN, or benign
Menzies et al,[29] 2023	MM (n = 55), others (n = 117)/(n = 124)	HPE for suspected lesions	Dermatologists group (n = 5)	Single best diagnosis out of 7-class classification: MM, MN, BCC, pAK/IEC, BKL, BVL, and DF
			Dermatologists group: novice (n = 18)	
			AI group: MetaOptima; 7-class (CNN models)	The maximum probability class is returned as a prediction for the 7-class diagnostic algorithm
			AI group: MetaOptima (CNN models)	7-Class classification (results obtained within 1 h)
Winkler et al,[30] 2023	MM (n = 38), other (n = 190)/(n = 188)	HPE for suspected lesions; follow-up and/or EC for other lesions	Dermatologists group (n = 22)	Malignancy score between 0 and 1, with a threshold for malignancy of at least 0.5
			AI group: Fotofinder Pro (CNN model)	Malignancy score between 0 and 1, with a threshold for malignancy of at least 0.5
			Dermatologist + AI group: dermatologists + FotoFinder Pro (CNN model)	Malignancy score between 0 and 1, with a threshold for malignancy of at least 0.5

Abbreviations: ADAE, an open-source CNN-based melanoma detection algorithm; AI, artificial intelligence; AN, atypical nevus; AUC, area under the curve; BCC, basal cell carcinoma; BKL, benign keratotic lesion; BVL, benign vascular lesion; CA, clinical assessment; CNN, convolutional neural network; DF, dermatofibroma; EC, expert consensus; FotoFinder Pro, FotoFinder Moleanalyzer Pro; FotoFinder Tueb, FotoFinder Moleanalyzer Tuebinger; HPE, histopathologic examination; IEC, intraepithelial carcinoma; MM, malignant melanoma; MN, melanocytic nevus; NA, not applicable; pAK, pigmented actinic keratosis; vA, version A; vB, version B.

[a] Overview of key features of all studies meeting the inclusion criteria: 9 studies included dermatologist-only groups (dermatologist group), 5 AI-only groups (AI group), and 1 AI-assisted dermatologist group (dermatologist + AI group). Some study groups did not meet the eligibility criteria for specific setups and, therefore, were excluded from the quantitative synthesis (labeled as not included).

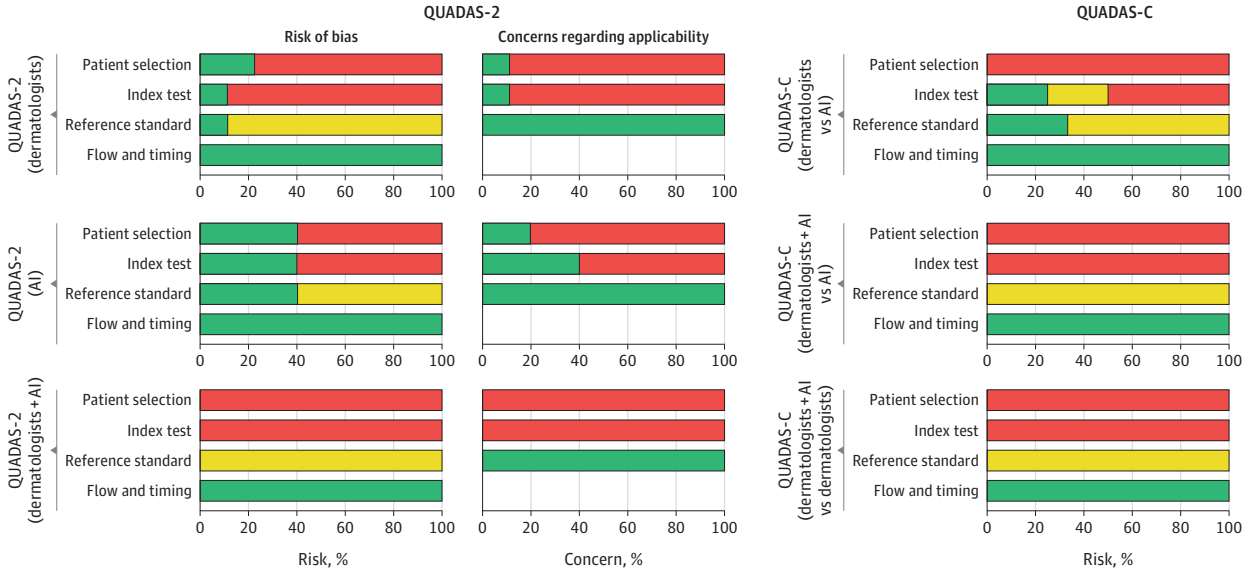
Figure 2. Heat Map and Bar Charts of QUADAS-2 and QUADAS-C Risk of Bias and Applicability Concerns Across All Eligible Studies (n = 11)

A Risk of bias and applicability concern plot

Test	Study	Risk of bias				Applicability concerns		
		Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
QUADAS-2 assessment								
Dermatologists	Phillips et al,[20] 2019	+	−	?	+	+	−	+
	Heinlein et al,[21] 2024	−	−	?	+	−	−	+
	Maier et al,[22] 2015	−	−	?	+	−	−	+
	Dreiseitl et al,[23] 2009	−	−	?	+	−	−	+
	Langley et al,[24] 2007	−	−	?	+	−	−	+
	Bono et al,[25] 2002	−	−	?	+	−	−	+
	MacLellan et al,[26] 2021	−	−	?	+	−	−	+
	Menzies et al,[29] 2023	−	+	+	+	−	+	+
	Winkler et al,[30] 2023	−	−	?	+	−	−	+
AI	MacLellan et al,[26] 2021	−	−	?	+	−	−	+
	Marchetti et al,[27] 2023	−	−	+	+	−	−	+
	Thomas et al,[28] 2023	+	+	?	+	+	+	+
	Menzies et al,[29] 2023	−	+	+	+	−	+	+
	Winkler et al,[30] 2023	−	−	?	+	−	−	+
Dermatologists plus AI	Winkler et al,[30] 2023	−	−	?	+	−	−	+
QUADAS-C assessment								
Dermatologists vs AI	MacLellan et al,[26] 2021	−	−	?	+			
	Menzies et al,[29] 2023	−	+	+	+			
	Winkler et al,[30] 2023	−	−	?	+			
Dermatologists plus AI vs AI	Winkler et al,[30] 2023	−	−	?	+			
Dermatologists plus AI vs dermatologists	Winkler et al,[30] 2023	−	−	?	+			

Low
Unclear
High

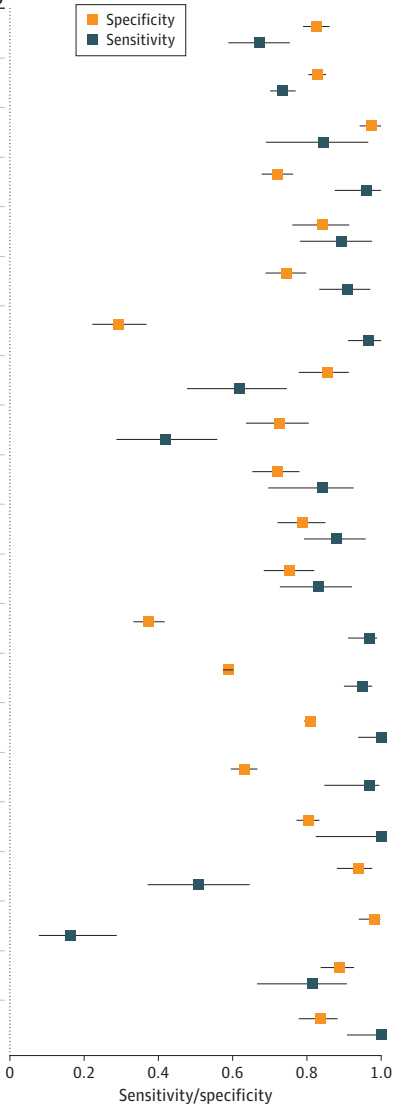
B Risk of bias and concern for QUADAS-2 and QUADAS-C



A, Risk of bias and applicability concern plot shows individual judgments across each QUADAS-2 and QUADAS-C domains. Boxes indicate risk of bias or concern. B, Bar charts show the proportion of studies rated at low, unclear, or high risk or concern for each QUADAS-2 and QUADAS-C domain. Study groups were assessed separately: dermatologists (n = 9), AI (n = 5), and dermatologists + AI (n = 1). For comparative designs (n = 3), QUADAS-C domains were applied: dermatologists vs AI (n = 3), dermatologists supported by vs AI (n = 1), and dermatologists supported by AI vs dermatologists (n = 1). AI indicates artificial intelligence; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies 2 tool; QUADAS-C, the Quality Assessment of Diagnostic Accuracy Studies Comparative tool.

Figure 3. Forest Plots of the Sensitivity and Specificity of the Eligible Studies

Study	Method	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Balanced accuracy
Phillips et al,[20] 2019	Derm[a]	84	74	352	41	0.672	0.826	0.791	0.749
Heinlein et al,[21] 2024	Derm	479	158	760	174	0.734	0.828	0.789	0.781
Maier et al,[22] 2015	Derm[a]	22	3	115	4	0.845	0.975	0.951	0.910
Dreiseitl et al,[23] 2009	Derm[a]	26	120	311	1	0.960	0.720	0.736	0.890
Langley et al,[24] 2007	Derm[a]	33	14	74	4	0.892	0.841	0.856	0.867
Bono et al,[25] 2002	Derm[a]	60	63	184	6	0.909	0.745	0.780	0.827
MacLellan et al,[26] 2021	Derm[a]	57	106	44	2	0.966	0.293	0.483	0.630
Menzies et al,[29] 2023	Derm-E	34	17	100	21	0.618	0.855	0.779	0.737
Menzies et al,[29] 2023	Derm-N	23	32	85	32	0.418	0.726	0.628	0.572
Winkler et al,[30] 2023	Derm	32	53	137	6	0.842	0.721	0.741	0.782
MacLellan et al,[26] 2021	AI-Pro[a]	52	32	118	7	0.881	0.788	0.813	0.835
MacLellan et al,[26] 2021	AI-Tueb[a]	49	37	113	10	0.831	0.753	0.775	0.792
Marchetti et al,[27] 2023	AI	92	318	190	3	0.968	0.374	0.468	0.671
Thomas et al,[28] 2023	AI-vB_c1	133	1852	2643	7	0.950	0.588	0.599	0.769
Thomas et al,[28] 2023	AI-vB_c1	58	482	2045	0	1.000	0.809	0.814	0.905
Thomas et al,[28] 2023	AI-vB_c2	32	249	427	1	0.970	0.632	0.647	0.801
Thomas et al,[28] 2023	AI-vB_c2	18	122	502	0	1.000	0.804	0.810	0.902
Menzies et al,[29] 2023	AI-7 class	28	7	110	27	0.509	0.940	0.802	0.725
Menzies et al,[29] 2023	AI-ISIC	9	2	115	46	0.164	0.983	0.721	0.574
Winkler et al,[30] 2023	AI	31	21	169	7	0.816	0.889	0.877	0.853
Winkler et al,[30] 2023	Derm + AI	38	31	159	0	1.000	0.837	0.864	0.919



Sensitivity, specificity, accuracy, and balanced accuracy were reported for each study. The results were stratified into dermatologists' performance, AI performance, and dermatologists' performance supported by AI. AI indicates artificial intelligence; AI-Pro, FotoFinder Pro; AI-Tueb, FotoFinder Tuebinger; AI-vA_c1, AI version A clinic 1; AI-vB_c1, AI version B clinic 1; AI-vA_c2, AI version A clinic 2; AI-vB_c2, AI version B clinic 2; Derm-E, expert dermatologists; Derm-N, novice dermatologists; FN, false negative; FP, false positive; ISIC, International Skin Imaging Collaboration; TN, true negative; TP, true positive.

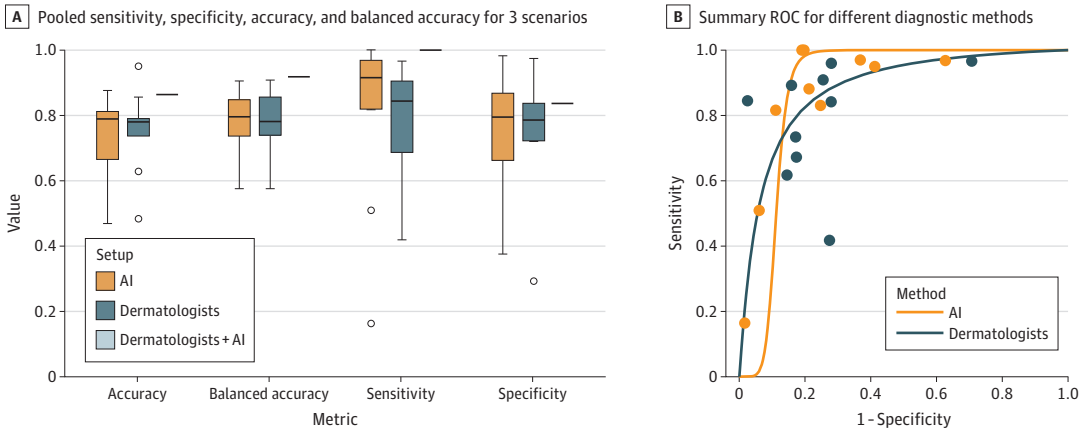
[a]Estimated 95% CI were used if they were not reported or were unreliable.

In total, 9 studies reported performance metrics for dermatologists, 5 for AI alone, and 1 for dermatologists supported by AI (the **Table**). All included AI studies used CNN-based methods. Thomas et al[28] reported 4 distinct AI performance outcomes—it was conducted across 2 clinical sites and evaluated 2 versions of the CNN algorithm (ie, DERM-vA and DERM-vB) following an update during the study period, and was designed as a reader study with a prospective AI group. Maclellan et al[26] and Menzies et al[29] each reported results for 2 different AI algorithms. In addition, Menzies et al[29] stratified dermatologist performance by level of experience (ie, expert vs novice; both trained in the evaluation of pigmented skin lesions). Given that the remaining studies did not stratify by experience, all dermatologist groups were pooled into 1 single group. Menzies et al[29] further noted that AI assessments were available within approximately 1 hour, which was considered sufficiently prospective for inclusion in our systematic review and meta-analysis.

Quality of Studies

We assessed the 11 eligible studies using the QUADAS-2 tool (**Figure 2**). Moreover, the QUADAS-C extension was applied in 3 studies, comparing diagnostic approaches (dermatologists vs AI).[26,29,30] In the study by Winkler et al,[30] a pairwise com-

Figure 4. Box Plots and Summary Receiver Operating Characteristic (ROC) Curves of the Eligible Studies



A, Box plots of all diagnostic metrics across all 3 setups, illustrating pooled sensitivity, specificity, accuracy, and balanced accuracy for dermatologists, AI, and dermatologists supported by AI. The horizontal line indicates the median; the upper and lower box limits denote the first and third quartiles. The ends extend to 1.5 times the IQR. The dots represent outliers that are beyond the third quartile. Dermatologists (n = 10), AI (n = 10), and dermatologists supported by AI (black line; n = 1). Outliers are represented as points. B, SROC for different diagnostic methods curves for different diagnostic methods. Study-level sensitivity and specificity estimates are shown, stratified by dermatologists and AI. AI indicates artificial intelligence.

parison was conducted among 3 groups (dermatologists, AI, and dermatologists plus AI), as the risk of bias and applicability domains differed between groups, and therefore, needed to be evaluated separately (Figure 2).

All 11 studies were conducted prospectively and enrolled consecutive patients in clinical practice. However, in 9 studies, patients were preselected, with inclusion restricted to melanoma-suspected, melanocytic, or pigmented lesions.[21-27,29,30] This design introduced patient selection bias given that many benign lesions typically encountered in routine practice were excluded. This limitation also contributed to applicability concerns in patient selection because in actual clinical settings, all patients would be considered.

In the index test domain, 1 study[23] adopted a patientwise rather than a lesionwise classification of melanoma. This approach inflates diagnostic accuracy because a patient is counted as correctly diagnosed if any melanoma lesion is identified, even if other lesions are misclassified.

Another study[26] defined the diagnostic outcome based on the dermatologists' management decision. Although all lesions were excised for verification, thus avoiding verification bias, this design introduced an artificial reduction in specificity given that many benign lesions not judged for excision were nonetheless removed and counted as false positives.

Nine studies applied a binary classification (malignant melanoma vs nonmalignant melanoma) rather than a multiclass approach (Table). This introduces index test bias because pooling all nonmalignant melanoma conditions can obscure differences in diagnostic performance across specific differential diagnoses, potentially over- or underestimating sensitivity and specificity. Moreover, the binary setup was deemed to have high applicability concerns because it does not reflect actual clinical practice, during which dermatologists must distinguish malignant melanoma from a variety of benign and malignant lesions. In contrast, studies using multiclass classifications[28,29] more closely reflected clinical decision-making and therefore were considered to have a lower risk of bias.

In 9 studies, it was not reported whether pathologists had prior knowledge of the dermatologists' diagnoses when conducting the histopathologic assessments.[20-26,28,30] This lack of information was classified as unclear risk of bias in the reference standard domain. However, this does not constitute an applicability concern because in routine clinical practice, pathologists typically have access to the dermatologists' clinical assessment (Figure 2).

Diagnostic Performance

The malignant melanoma diagnostic performance of the 11 eligible studies was assessed using sensitivity, specificity, accuracy, and balanced accuracy, either as reported or calculated from published data (eTable 2 in Supplement 1). Dermatologists' performance was reported in 9 studies, AI performance in 5 studies, and dermatologists' performance assisted by AI in 1 study.

In 1 study,[20] the sample size in each cohort appears to have been interchanged, leading us to recalculate the results with new values[22] (eTable 9 in Supplement 1 provides the details). Another study[26] provided specificity metrics that did not align with the reported diagnostic contingency table; therefore, we recalculated its specificity (eTable 12 in Supplement 1 provides the details). When further inconsistencies could not be resolved by contacting the original authors, we applied our predefined procedure of recalculating values based on available raw data.

Across the 9 studies reporting dermatologists' performance, sensitivity ranged from 41.8% to 96.6%, and specificity from 29.3% to 97.0% (**Figure 3**). The pooled estimates across all 10 investigators were 78.6% (95% CI, 67.5%-88.1%) sensitivity, 75.3% (95% CI, 63.3%-84.3%) specificity, 75.3% (95% CI, 67.6%-82.3%) accuracy, and 77.4% (95% CI, 70.8%-83.6%) balanced accuracy (**Figure 4**A; eTable 5 in Supplement 1).

For the 5 studies reporting AI performance, sensitivity ranged from 16.4% to 100.0%, and specificity, from 37.4% to 98.3% (Figure 3), with pooled estimates across all 10 investigators of 80.9% (95% CI, 63.6%-94.5%) sensitivity, 75.6% (95% CI, 64.5%-85.6%) specificity, 73.3% (95% CI, 65.4%-80.0%) accuracy, and 78.3% (95% CI, 72.0%-84.1%) balanced accuracy (Figure 4A; eTable 5 in Supplement 1). The single study reporting dermatologists' performance assisted by AI achieved a sensitivity of 91.9% and a specificity of 83.7%, corresponding to 86.4% accuracy and 87.8% balanced accuracy (eTable 5 in Supplement 1; Figure 3). Figure 4B shows the SROC curves with the trade-off between sensitivity and specificity across studies, providing a comprehensive summary of overall diagnostic performance. Although neither the statistical analysis (eTable 5 in Supplement 1) nor Figure 4 indicate significant differences between dermatologists and AI, the point estimates of the aggregated metrics uniformly favored AI.

Across the 3 prospective head-to-head studies,[26,29,30] AI systems consistently demonstrated higher specificity but similar or lower sensitivity compared with dermatologists (Figure 3; eTable 6 in Supplement 1 provides the head-to-head value comparisons).

Discussion

This systematic review and meta-analysis assessed prospective studies comparing dermatologists, AI alone, and AI-assisted dermatologists to evaluate diagnostic performance and clinical readiness. Prospective evidence indicates that AI performs at a level comparable to dermatologists in melanoma diagnosis, with similar pooled sensitivity (80.9% vs 78.6%), specificity (75.6% vs 75.3%), accuracy (73.3% vs 75.3%), and balanced accuracy (78.3% vs 77.4%) even under actual clinical conditions, suggesting its potential as a decision-support tool beyond retrospective benchmarks, which often overestimate performance.[10] Notably, models trained on larger and more diverse databases performed better in clinical practice,[29] highlighting the importance of dataset size and heterogeneity for real-world applicability.

Although pooled estimates suggest similar overall performance between AI and dermatologists, direct head-to-head comparisons within the same clinical setting[26,29,30] indicate higher specificity (94.0% and 98.3% vs 85.5%[29]; 78.8% and 75.3% vs 29.3%[26]; 88.9% vs 72.1%[30]) at comparable sensitivity for AI (50.9% vs 61.8%[29]; 88.1% and 83.1% vs 96.6%[26]; 81.6% vs 84.2%[30]). This observation may be explained by the fact that dermatologists tend to act cautiously and are more likely to recommend biopsy in cases of diagnostic uncertainty.[31,32] In contrast, AI-based assessments could help reduce unnecessary biopsies.

Compared with previous evidence syntheses, notable differences emerge. The pooled sensitivity (78.6%) and specificity (75.2%) for dermatologists in our meta-analysis (studies published between 2002 and 2024) were lower than the estimates reported by Vestergaard et al[3] in 2008 (sensitivity 0.87 and specificity 0.91, based on studies from 1993-2006). Notably, this meta-analysis compared dermoscopy findings with unaided clinical examination in studies conducted mostly in specialist referral clinics and often limited to lesions suspected of melanoma, with several using rule-based dermoscopic criteria, design features that may partly explain the higher reported accuracy. Earlier studies[33-40] often included more clinically evident melanomas and straightforward benign lesions, inflating accuracy via spectrum effects, a phenomenon where the performance of a diagnostic test varies across different patient populations. In contrast, more recent studies[21,29] tend to include diagnostically challenging lesion sets, use multicenter designs, and, in some cases, run head-to-head with AI systems—design choices that enrich equivocal cases. Taken together, our findings represent a more realistic reflection of contemporary dermatologists' performance across diverse clinical settings.

Limitations

While these findings highlight the potential clinical value of AI-assisted diagnosis, they must be interpreted with caution, considering several methodologic limitations. Notably, patient selection introduces a systematic bias: in 9 of the 11 included studies, participants were preselected to include only melanoma-suspected, melanocytic, or pigmented lesions. This does not reflect the broader spectrum of lesions encountered in daily clinical practice and likely results in overestimated sensitivity (fewer FNs) and underestimated specificity (given that benign routine lesions are underrepresented). Consequently, the reported diagnostic performance may not fully generalize to routine settings.

Other design-related biases also influence interpretation. Dreiseitl et al[23] calculated diagnostic accuracy on a patientwise rather than lesionwise basis, which inflates sensitivity because identifying a single melanoma lesion in a patient counts as a correct diagnosis, even if additional lesions are misclassified. In the study by MacLellan et al,[26] the dermatologist's management recommendation (excise, not excise, or watch) was treated as a melanoma classification. For verification, however, all lesions were excised, so many benign lesions initially considered for excision were counted as FPs. This design eliminated verification bias but led to an artificial reduction in specificity.

Similarly, several studies used binary classification (melanoma vs nonmelanoma), which oversimplifies the diagnostic task and may obscure differences in distinguishing between specific benign and malignant conditions. These design types reduce applicability given that clinical practice requires differentiation across multiple lesion types.

Another important limitation of the included studies is the imbalance in evidence: 9 studies reported dermatologist performance, 5 assessed AI alone, and only 1 investigated dermatologist with AI support. This makes it questionable to draw meaningful conclusions regarding the added value of AI support in routine care. In addition, most AI studies are still early in their prospective validation, and performance may be vulnerable to domain shifts between training and clinical application. Factors such as image quality, device heterogeneity, and patient selection can lead to deviations from retrospective performance benchmarks. Our inclusion of prospective designs

addresses some of these concerns, but the evidence base remains limited.

Conclusions

This systematic review and meta-analysis found prospective evidence indicating that AI achieves dermatologist-level performance for dermoscopic melanoma diagnosis, with no significant differences in pooled sensitivity or specificity. This finding is encouraging for clinical translation. Yet, the diversity of study designs, risks of bias, and the limited number of high-quality prospective datasets highlight that AI is still in the early phase of clinical validation. Larger, multicenter, and methodologically rigorous prospective studies with unselected, real-world patient populations will be essential to determine the reliability, safety, and added value of AI in routine clinical practice.

ARTICLE INFORMATION

Accepted for Publication: January 28, 2026.

Published Online: March 25, 2026. doi:10.1001/jamadermatol.2026.0217

Author Contributions: Dr Laiouar-Pedari and Ms Kühn had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis, and contributed equally.
Concept and design: Laiouar-Pedari, Kühn, Brinker.
Acquisition, analysis, or interpretation of data: Laiouar-Pedari, Kühn, Wies, Nogueira Garcia, Winterstein, Heinlein, Hoffsommer, Chanda, Haggenmüller.
Drafting of the manuscript: Laiouar-Pedari, Kühn.
Critical review of the manuscript for important intellectual content: All authors.
Statistical analysis: Laiouar-Pedari, Wies, Chanda.
Obtained funding: Laiouar-Pedari, Brinker.
Administrative, technical, or material support: Laiouar-Pedari, Haggenmüller, Brinker.
Supervision: Kühn, Brinker.

Conflict of Interest Disclosures: Dr Haggenmüller reported grants from reported holding a position in research and development at the HEINE Optotechnik GmbH & Co outside the submitted work. Dr Brinker reported ownership of a company that develops mobile apps (Smart Health Heidelberg GmbH, Heidelberg, Germany), receiving honoraria from Novartis, Roche, HEINE Optotechnik, and Merck outside the submitted work. No other disclosures were reported.

Funding/Support: This study was funded by the Ministry of Health, Social Affairs and Integration Baden-Württemberg, Stuttgart, Germany (No. 53-5400-77/1/2) awarded to Dr Titus J. Brinker (German Cancer Research Center, Heidelberg, Germany).

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Garbe C. Melanom: aktualisierte Therapieempfehlung der S3-Leitlinie zur Diagnostik, Therapie und Nachsorge des Melanoms, mit Fokus auf das fernmetastasierte stadium. *Hautarzt*; 2016;67(5):343-352. doi:10.1007/s00105-016-3925-8

2. Dinnes J, Deeks JJ, Chuchu N, et al; Cochrane Skin Cancer Diagnostic Test Accuracy Group. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database Syst Rev*. 2018;12(12):CD011902. doi:10.1002/14651858.CD011902.pub2

3. Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br J Dermatol*. 2008;159(3):669-676. doi:10.1111/j.1365-2133.2008.08713.x

4. Menzies SW. Cutaneous melanoma: making a clinical diagnosis, present and future. *Dermatol Ther*. 2006;19(1):32-39. doi:10.1111/j.1529-8019.2005.00054.x

5. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol*. 2002;3(3):159-165. doi:10.1016/S1470-2045(02)00679-4

6. Haggenmüller S, Wies C, Abels J, et al. Discordance, accuracy and reproducibility study of pathologists' diagnosis of melanoma and melanocytic tumors. *Nat Commun*. 2025;16(1):789. doi:10.1038/s41467-025-56160-x

7. Haggenmüller S, Maron RC, Hekler A, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *Eur J Cancer*. 2021;156:202-216. doi:10.1016/j.ejca.2021.06.049

8. Salinas MP, Sepúlveda J, Hidalgo L, et al. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *NPJ Digit Med*. 2024;7(1):125. doi:10.1038/s41746-024-01103-x

9. Brinker TJ, Hekler A, Hauschild A, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. *Eur J Cancer*. 2019;111:30-37. doi:10.1016/j.ejca.2018.12.016

10. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20(7):938-947. doi:10.1016/S1470-2045(19)30333-X

11. Brinker TJ, Hekler A, Enk AH, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur J Cancer*. 2019;113:47-54. doi:10.1016/j.ejca.2019.04.001

12. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Rev Esp Cardiol (Engl Ed)*. 2021;74(9):790-799. doi:10.1016/j.recesp.2021.06.016

13. UK National Institute for Health and Care Research. International prospective register of systematic reviews (PROSPERO). Accessed September 17, 2025. https://www.crd.york.ac.uk/PROSPERO/view/CRD420251084932

14. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 1995;123(3):A12-A13. doi:10.7326/ACPJC-1995-123-3-A12

15. Digital Scholar. Zotero. Accessed September 29, 2025. https://www.zotero.org

16. Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009

17. Yang B, Mallett S, Takwoingi Y, et al; QUADAS-C Group†. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Ann Intern Med*. 2021;174(11):1592-1599. doi:10.7326/M21-2234

18. Efron B. Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics. Springer. 1992:569-593. doi:10.1007/978-1-4612-4380-9_41

19. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12(14):1293-1316. doi:10.1002/sim.4780121403

20. Phillips M, Marsden H, Jaffe W, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open*. 2019;2(10):e1913436. doi:10.1001/jamanetworkopen.2019.13436

21. Heinlein L, Maron RC, Hekler A, et al. Prospective multicenter study using artificial intelligence to improve dermoscopic melanoma diagnosis in patient care. *Commun Med (Lond)*. 2024;4(1):177. doi:10.1038/s43856-024-00598-5

22. Maier T, Kulichova D, Schotten K, et al. Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. *J Eur Acad Dermatol Venereol*. 2015;29(4):663-667. doi:10.1111/jdv.12648

23. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. *Melanoma Res*. 2009;19(3):180-184. doi:10.1097/CMR.0b013e32832a1e41

24. Langley RGB, Walsh N, Sutherland AE, et al. The diagnostic accuracy of in vivo confocal scanning laser microscopy compared to dermoscopy of benign and malignant melanocytic lesions:

a prospective study. *Dermatology*. 2007;215(4):365-372. doi:10.1159/000109087

25. Bono A, Bartoli C, Cascinelli N, et al. Melanoma detection. A prospective study comparing diagnosis with the naked eye, dermatoscopy and telespectrophotometry. *Dermatology*. 2002;205(4):362-366. doi:10.1159/000066436

26. MacLellan AN, Price EL, Publicover-Brouwer P, et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *J Am Acad Dermatol*. 2021;85(2):353-359. doi:10.1016/j.jaad.2020.04.019

27. Marchetti MA, Cowen EA, Kurtansky NR, et al. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). *NPJ Digit Med*. 2023;6(1):127. doi:10.1038/s41746-023-00872-1

28. Thomas L, Hyde C, Mullarkey D, Greenhalgh J, Kalsi D, Ko J. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front Med (Lausanne)*. 2023;10:1264846. doi:10.3389/fmed.2023.1264846

29. Menzies SW, Sinz C, Menzies M, et al. Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. *Lancet Digit Health*. 2023;5(10):e679-e691. doi:10.1016/S2589-7500(23)00130-9

30. Winkler JK, Blum A, Kommoss K, et al. Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: human with machine. *JAMA Dermatol*. 2023;159(6):621-627. doi:10.1001/jamadermatol.2023.0905

31. Elder DE, Eguchi MM, Barnhill RL, et al. Diagnostic error, uncertainty, and overdiagnosis in melanoma. *Pathology*. 2023;55(2):206-213. doi:10.1016/j.pathol.2022.12.345

32. Kommoss KS, Winkler JK, Mueller-Christmann C, et al. Observational study investigating the level of support from a convolutional neural network in face and scalp lesions deemed diagnostically 'unclear' by dermatologists. *Eur J Cancer*. 2023;185:53-60. doi:10.1016/j.ejca.2023.02.025

33. Papachristou P, Söderholm M, Pallon J, et al. Evaluation of an artificial intelligence-based decision support for the detection of cutaneous melanoma in primary care: a prospective real-life clinical trial. *Br J Dermatol*. 2024;191(1):125-133. doi:10.1093/bjd/ljae021

34. Marsden H, Kemos P, Venzi M, et al. Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service. *Front Med (Lausanne)*. 2024;11:1302363. doi:10.3389/fmed.2024.1302363

35. Jahn AS, Navarini AA, Cerminara SE, et al. Over-detection of melanoma-suspect lesions by a ce-certified smartphone app: performance in comparison to dermatologists, 2D and 3D convolutional neural networks in a prospective data set of 1204 pigmented skin lesions involving patients' perception. *Cancers (Basel)*. 2022;14(15):3829. doi:10.3390/cancers14153829

36. Bono A, Bartoli C, Baldi M, Tomatis S, Bifulco C, Santinami M. Clinical and dermatoscopic diagnosis of small pigmented skin lesions. Euro J Dermatol. 2002;12(6). https://pubmed.ncbi.nlm.nih.gov/12459531/

37. Perrinaud A, Gaide O, French LE, Saurat JH, Marghoob AA, Braun RP. Can automated dermoscopy image analysis instruments provide added benefit for the dermatologist? A study comparing the results of three systems. *Br J Dermatol*. 2007;157(5):926-933. doi:10.1111/j.1365-2133.2007.08168.x

38. van der Rhee JI, Bergman W, Kukutsch NA. The impact of dermoscopy on the management of pigmented lesions in everyday clinical practice of general dermatologists: a prospective study. *Br J Dermatol*. 2010;162(3):563-567. doi:10.1111/j.1365-2133.2009.09551.x

39. van der Rhee JI, Bergman W, Kukutsch NA. Impact of dermoscopy on the management of high-risk patients from melanoma families: a prospective study. *Acta Derm Venereol*. 2011;91(4):428-431. doi:10.2340/00015555-1100

40. Durdu M, Baba M, Seçkin D. Dermatoscopy versus Tzanck smear test: a comparison of the value of two tests in the diagnosis of pigmented skin lesions. *J Am Acad Dermatol*. 2011;65(5):972-982. doi:10.1016/j.jaad.2010.08.019