

Artificial intelligence tools in dermatology education: A scoping review on their application, efficacy, and limitations

Chun Hian John Yap¹ MRCP, Ziying Vanessa Lim² MPH

ABSTRACT

Introduction: Artificial Intelligence (AI) is increasingly explored for medical education, and dermatology, with its visual diagnostic focus, holds promise for AI-enhanced learning. However, evidence on its educational effectiveness remains limited and fragmented. This scoping review aimed to assess the evidence base for applications and limitations of AI-based educational tools in dermatology.

Methods: A systematic search of PubMed, Embase, Web of Science, Scopus, and PsycINFO up to July 2025 was conducted. Data were synthesised narratively, considering types of AI interventions and their assessed outcomes. These were analysed with the Cost, Usability, Credibility, Fairness, Accountability, Transparency, Explainability (CUC-FATE) framework. Study quality was assessed with ROBINS-I and a COSMIN-informed checklist.

Results: A total of 827 records were screened, with 360 duplicates removed, yielding 467 studies. Six full-length studies and 1 conference abstract met inclusion criteria, mostly from 2023 to 2025. These explored AI-generated clinical images, Large Language Model-generated vignettes, intelligent tutoring systems, and clinical decision support tools. Content validation studies generally reported favourable ratings for accuracy, clarity, and educational utility, while intervention studies suggested possible benefits for learning performance and diagnostic accuracy. Usability and credibility were commonly assessed, whereas cost, accountability, fairness, transparency, and explainability were rarely examined.

Conclusions: Most of the studies reviewed had high risk of bias, small sample sizes, and limited methodological rigour, with significant heterogeneity in examined educational outcomes limiting synthesis. While current initial studies on AI hold promise, this scoping review underscores the need for more robust studies with standardised evaluation frameworks, prioritising ethical principles such as fairness, explainability, and accountability for safe integration in training.

Keywords: artificial intelligence, dermatology, large language models, medical education, scoping review

CLINICAL IMPACT

What is New

- This scoping review focuses on the emerging evidence of the direct educational impact of AI-based educational tools in dermatology and identifies major gaps in methodological quality, standardised evaluation, and ethical oversight.
- As the use of AI in healthcare and medical education expands, these findings provide a preview of opportunities and challenges in AI-based dermatology training.

Clinical Implications

- The findings support the need for rigorous and ethically grounded frameworks to guide safe integration of AI into dermatology education, with emphasis on educational effectiveness, fairness, explainability, and accountability.

INTRODUCTION

Artificial Intelligence (AI) technologies—particularly deep learning algorithms and Large Language Model

(LLM) computer systems—have progressed at an unprecedented pace in recent years, driving innovation across a wide range of healthcare applications.¹ These advancements have resulted in a corresponding surge in research on the application of AI in medical education,² including the development of virtual reality simulators³ and intelligent tutoring systems.⁴ Such tools enable the provision of individualised, innovative and adaptive learning experiences, allowing students and physicians to practise medical reasoning.^{5,6} Moreover, integrating AI into medical education allows for distance learning, making medical education more accessible in resource-constrained areas with limited resources⁷ or time due to clinical demands.⁸

In the field of dermatology, its inherent visual diagnostic process makes it an attractive possibility for AI to be applied as an educational tool,⁹ with similar promise being noted in other visually intensive specialties, such as radiology, for the enhancement of clinical competencies and optimising learning outcomes.¹⁰ AI applications—ranging from machine learning-based diagnostic simulators to image-based

The Annals is an open access journal, allowing non-commercial use under CC BY-NC-SA 4.0.

¹ Ministry of Health Holdings, Singapore

² National Skin Centre, Singapore

Correspondence: Dr Ziying Vanessa Lim, National Skin Centre, 1 Mandalay Road, Singapore 308205. Email: drvanessalim@gmail.com

Accepted: 30 June 2026

Published online first: 14 July 2026

learning platforms¹¹—have been proposed to improve diagnostic accuracy, engagement, and learning efficiency.¹² However, the evidence evaluating the effectiveness of these technologies in dermatology education—such as in assessing improvement in diagnostic accuracy, knowledge retention, clinical decision-making—remains fragmented and under-explored, with robust evidence for their use in dermatology education currently lacking.¹³

As such, this scoping review aims to address this gap, examining the current methods by which AI has been applied to dermatological education. By charting the existing research landscape and evaluating current evidence, this scoping review aims to highlight promising applications and identify limitations in existing studies. In doing so, this study aims to suggest future innovations for more efficacious AI usage, including areas of learner and faculty development necessary to integrate AI into dermatological education.

METHODS

The protocol detailing the search strategies of this review was formulated by the authors, who registered this on the International Prospective Register of Systematic Reviews (PROSPERO, registration number CRD420251058835) on 22 May 2025 before the study's initiation. While initially designed as a systematic review, in view of the small number of highly heterogeneous studies identified, the study was redesigned as a scoping review after article identification to provide a broader map of existing evidence. With a narrow evidence base, assessment on efficacy would be limited, and hence research questions were refined to define the applications and limitations of the selected studies.

Search strategy

The search included the databases of MEDLINE (through PubMed), Embase, Web of Science, SCOPUS, and PsycINFO, with the search strategies described in Supplementary Table S1. All studies published from the inception of the respective database till 1 July 2025 were included. Based on titles and abstracts, duplicate studies were first excluded. The remaining studies were then evaluated based on titles and abstracts for their relevance to the use of artificial intelligence in dermatological education.

Selection of articles

Studies published in any language, from all countries, were considered for review. Randomised controlled trials, non-randomised controlled studies, pre-post intervention studies, and observational studies were included if they referred to direct application of AI in dermatological education. Efforts were made to

obtain full-length studies for findings highlighted in conference abstracts. Studies were considered at various levels of healthcare education, including that of medical students, residents (especially dermatology and general medical trainees), and non-dermatology clinicians involved in dermatology learning. Tools used solely for clinical diagnosis without an educational aim, editorials, opinion pieces, protocols without results, and studies focused solely on AI tool development without educational evaluation were excluded. As title and abstract screening was done jointly, formal recording of reviewer disagreements and calculation of inter-rater reliability were not applicable.

Data collection

Data were extracted via a standardised collection form, with authors' names, publication date, country of publication (as defined by the setting of data collection), type of study, and level of healthcare education of the intended learner audience. Following this, categorisation was performed of the types of artificial intelligence educational tools assessed, any specific learning outcomes examined and overall conclusions.

Data analysis

In this review, the articles identified displayed significant heterogeneity, preventing meaningful pooling and synthesis of educational outcomes. As such, a scoping review was conducted according to the methodological framework proposed by Arksey and O'Malley¹⁴ and refined by Levac et al.,¹⁵ and was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist (Supplementary Table S2).¹⁶ With the extracted data, a narrative account of findings was produced, considering the types of AI interventions and the educational outcomes assessed. The limitations and gaps in these studies were also identified.

The studies were categorised based on various aspects of AI technologies specifically measured by each study. Utilising an approach developed by Quttainah et al.¹⁷ examining enablers for the use of LLMs in medical education, an assessment was made on whether the parameters of Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability (CUC-FATE) were considered by each included study.

Quality assessment

Only non-randomised studies were obtained, and their methodological quality was assessed. Study quality of quasi-experimental studies was assessed using the Cochrane Collaboration's Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I)

tool,¹⁸ modified for usage in before-after studies. As no established appraisal tool exists for validation of AI-based educational content, content validation studies were assessed based on a framework devised by the authors, informed by the CONsensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist for content validity.¹⁹ This framework examined core tenets of the COSMIN checklist, including relevance, comprehensiveness and clarity, as well as recommendations regarding expert involvement and qualitative evaluation during instrument development.

There are also currently no suitable frameworks for assessing the quality of AI-powered interventions in medical education. The commonly used Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) criteria, while modified recently to account for machine learning methods,²⁰ work to evaluate predictive models by their performance, and cannot be directly applied to effectiveness in medical education. As such, a framework previously utilised by Bates et al.,²¹ also referenced in other reviews on artificial intelligence in healthcare education,²² was utilised.

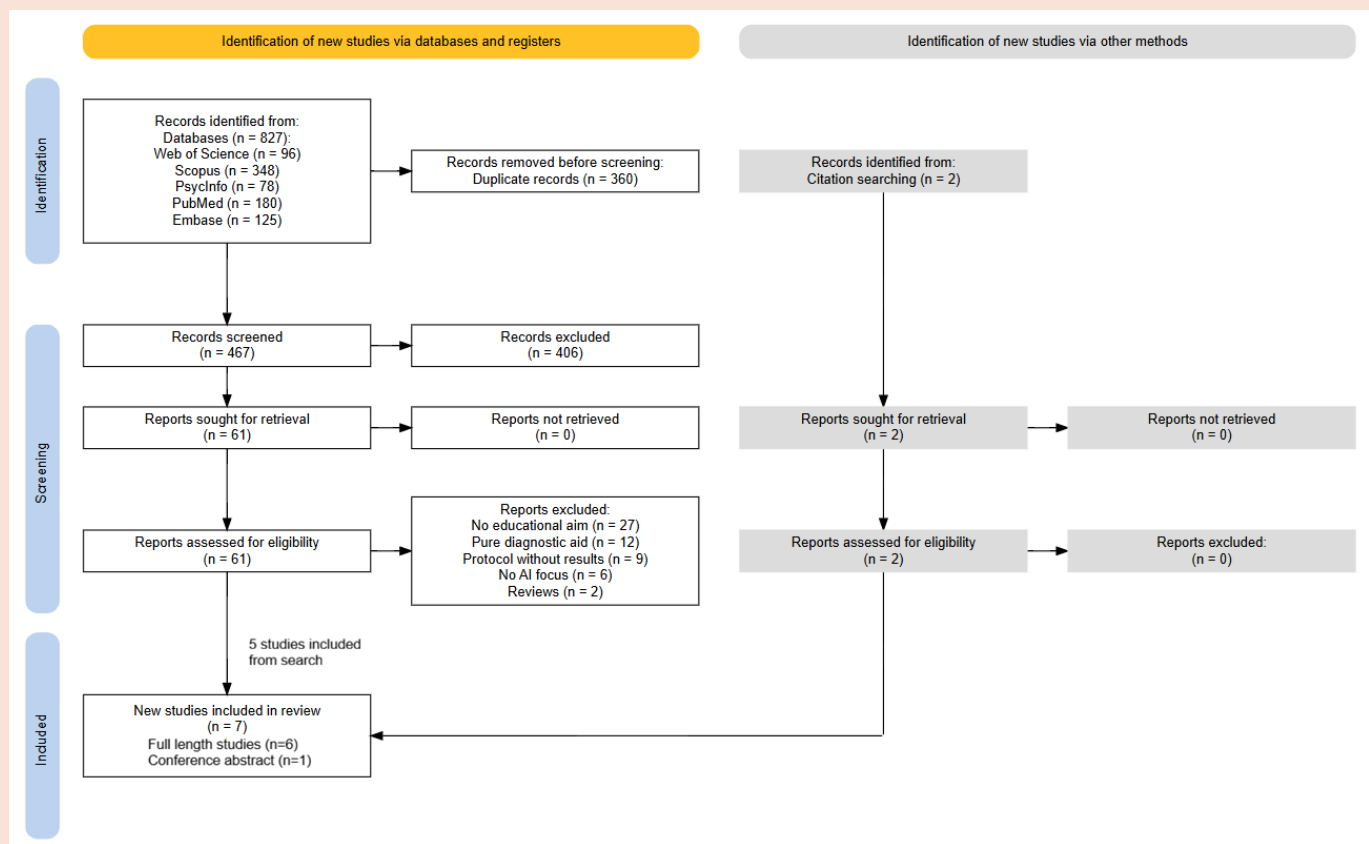
RESULTS

Search results

The comprehensive search of PubMed, Embase, Web of Science, Scopus, and PsycINFO yielded 827 studies, of which 360 duplicates were excluded (Fig. 1). From the remaining 467 studies, 406 were excluded based on titles and abstracts. In total, 61 studies were retrieved and assessed for eligibility, of which four full-length studies and one conference abstract were identified. Further search of citations yielded two other relevant articles,²³⁻²⁴ with a total of 6 full-length studies and 1 conference abstract²³⁻²⁹ selected for review.

A decision was made to include a conference abstract²⁷ to allow comprehensive mapping of the dermatological educational landscape, in accordance with a proposal by Scherer et al.³⁰ An attempt was made to contact the authors and examine study registers and published protocols to obtain further information on methods and results, but without success. However, due to its brevity, there was limited ability to appraise the abstract for its design, risk of bias, and assessed educational outcomes, which is noted in the subse-

Fig. 1. Database search.



quent discussion. The PRISMA flowchart for this study was created with the aid of an app designed by Haddaway et al.³¹

Description of included studies

The general characteristics of studies are summarised in Table 1, with specific outcomes examined summarised in Table 2 and further laid out in a frequency table for content validation studies (Supplementary Table S3). Out of the 6 identified studies and 1 conference abstract (henceforth considered with the rest of the studies), 6 of these were conducted over the past 3 years (2023–2025)^{23–28}—with 1 study focusing on an intelligent tutoring system in 2008.²⁹ Four of the studies were conducted in the US,^{24,27–29} and 1 each in Italy, Australia, and Germany.^{23,25,26}

Three of the studies (including the conference abstract) were assessed as quasi-experimental, defined here as studies analysing the effect of interventions without the assignment of experimental and control groups^{26–27,29}; while the remaining 4 studies^{23–25,28} were designed as cross-sectional studies for content validation, defined here as studies which focused on rating the quality and accuracy of AI output. Three studies had a stated targeted audience who were residents,^{25,28,29} three studies had a stated targeted audience who were medical students,^{24,26,27} and one study general practitioners.²³

AI technologies examined

Many studies focused on the utilisation of generative AI, including 2 for image generation^{25,26} and 3 for text generation.^{23,24,28} All studies examining text generation utilised ChatGPT in its various models, with 2 studies^{24,28} attempting to generate questions for use in clinical examinations. The last study²³ attempted to assess the LLM's ability to answer questions about acne vulgaris for education and was unique in assessing the LLM's ability to identify appropriate reference literature for learners.

Out of the 2 remaining studies, the first evaluated an intelligent tutoring system in dermatopathology serving as a precursor to modern AI today.²⁹ The second, a conference abstract, provided a different approach, utilising a clinical decision support system for medical students in an outpatient setting, creating a list of differential diagnoses during patient review.²⁷

Educational outcomes examined

Both full-length quasi-experimental studies^{26,29} attempted to examine the improvement of clinical competence, with 1 image-generation study²⁶ looking at the improvement in learning accuracy (albeit in a small sample of students, which cannot be interpreted as evidence of educational benefit), while another

study on intelligent tutoring²⁹ looked at the frequency of errors. The conference abstract²⁷ evaluating a clinical decision support system appeared to only look at patient perception and utilisation time of the AI tool, although the absence of a full-text limits assessment. The significant heterogeneity of these educational outcomes prevented further synthesis and comparison.

All studies designed for content validation attempted to break down the characteristics of AI-generated output important for educational use. The 3 studies focused on LLMs^{23,24,28} looked at accuracy and clarity of material generated, though they diverged significantly in aspects they found important for education. The content validation study for AI-generated images²⁵ looked similarly at accuracy and utility for clinical education.

Quality assessment

The 3 quasi-experimental trials (including the conference abstract) were analysed with the ROBINS-I tool with all studies assessed having a serious overall risk of bias (Table 3). None of the trials provided data on the participant recruitment or measures to control for confounders. In the conference abstract,²⁷ while there was limited information to assess for the risk of bias, the authors only recruited two medical students to test the intervention.

In assessing content validation studies on AI-generated materials,^{23–25,28} a content validity checklist informed by the COSMIN approach was adopted¹⁹ (Table 4). All studies had clear descriptions of measured constructs and assessed relevant (but different) characteristics of AI-generated materials. They also all used clear Likert scales for assessment, involving experts in its development and reporting on inter-rater agreement. Three out of 4 studies were significantly lacking in having a small number of individuals assessing the suitability of teaching material, and only 1 out of 4 studies reported qualitative insights.

When analysed using the approach by Bates et al.²¹ (Supplementary Table S4), the accuracy of the AI-based educational technologies was often unvalidated, and information with regard to uncertainty and data-handling was not provided in all studies.

Enablers of AI in medical education

Studies were assessed for examination of the key features of the CUC-FATE framework¹⁷ (Table 5) for effective and safe usage of AI in medical education, but many of these elements were not assessed in identified studies. While all studies examined the usability of AI technology, and 6 out of 7 studies examined their credibility, none of the studies assessed issues regarding cost.

Table 1. Description of included studies.

Reference/ country of study	Study design/ intended learner audience	AI technology assessed	Sample size/number of assessors	Study overview	Outcomes assessed
Cevik et al. (2025) ²⁵ Australia	Content validation study Residents	AI image generation	Images assessed by 3 doctors, namely, 2 plastic surgery residents and 1 specialist plastic surgeon.	Three prominent AI models were assessed on their ability to generate authentic looking images of skin cancers (squamous cell carcinoma, basal cell carcinoma and melanoma) as educational resources. Images underwent an accuracy check using AI Dermatologist’s machine learning algorithm.	Adequacy and accuracy in representing real-world pathology Utility in clinical education
Naldi et al. (2024) ²³ Italy	Content validation study General practitioners	LLM	The answers and references were independently evaluated by a panel of 5 doctors, including 4 residents and 1 dermatologist	Open AI GPT-4 was instructed to generate information for a CME on the treatment of acne vulgaris. A 23-item questionnaire was developed by the authors from National Institute of Clinical Excellence (NICE) guidelines and prompted into the LLM. These were evaluated via a Likert scale, with room for qualitative judgement. The LLM was also instructed to quote 3 references from biomedical literature to support the answers provided. After these were validated, the references were evaluated by the same 5 doctors via a Likert scale, with room for qualitative judgement.	Questions were assessed on 5 domains: quality, readability, accuracy, thoroughness, and consistency with guidelines References were assessed on 3 domains: pertinence, relevance, and whether they were updated
Pielage et al. (2025) ²⁶ Germany	Quasi-experimental study Medical students	AI image generation	93 students were involved in the 3D simulation. The small user study involved 7 students for comparative analysis	AI Diffusion models were used to generate images of malignant melanoma and benign nevi. These lesions were embedded into a 3D model of a virtual patient and integrated into a fully interactive virtual 3D simulation for a group of students. A small user study was also conducted with seven students, with a learning group of four students provided with AI-generated images of nevi and melanoma, and a control group of three students not provided with these images. They were then tasked to classify 20 selected real images of nevi and melanoma.	Perceived competence of students after training Realism of lesions Improvement in learning accuracy
Rao et al. (2025) ²⁴ US	Content validation study Medical students	LLM	Assessed by 3 practising attending physicians	The Open AI GPT 4.0 model was tasked to generate 20 clinical vignettes in dermatology for medical student education, and these were assessed on a Likert scale	Alignment to scientific consensus Comprehensiveness Risk of clinical harm Demographic bias Clinical quality

Table 1. Description of included studies. (Cont'd)

Reference/ country of study	Study design/ intended learner audience	AI technology assessed	Sample size/number of assessors	Study overview	Outcomes assessed
Lee et al. (2024) ²⁷ US	Quasi-experimental study (Conference Abstract) Medical students	Clinical decision support system	Two medical studies were asked to evaluate the decision support system	The medical students evaluated 80 patients using the machine learning clinical decision support application in a dermatology clinic without an attending, with the utilisation time of the application to generate a list of differential diagnoses being assessed. Patients were also surveyed regarding their perceptions of technology during this visit.	Time of utilisation of application Perception of utilisation of system in medical education
Ayub et al. (2023) ²⁸ US	Content validation study with qualitative elements Residents (preparation for board examinations)	LLM	The questions were independently assessed by two board-certified dermatologists	CME articles from the Journal of the American Academy of Dermatology (JAAD) were uploaded into Open AI GPT 3.5, and the LLM was instructed to generate multiple choice questions.	Accuracy Complexity of questions Clarity of wording and structure
Payne et al. (2009) ²⁹ US	Quasi-experimental study Residents	Intelligent tutoring	44 pathology residents were involved in the intelligent tutoring system	Pathology residents used Slide Tutor—a Medical Intelligent Tutoring System in dermatopathology using commonly used diagnostic algorithms. This prototype provided feedback including error analysis and confirmation of correct actions, providing students with context-specific hints based on the current problem state.	Frequency of errors Analysis of goals differences and solution path

CME: continuing medical education; LLM: large language model

Table 2. Results of included studies.

Content validation studies		
Reference	Descriptive statistics	Conclusions
Naldi et al. (2024) ²³	The 23 answers were evaluated on: - Quality (101/115 or 87.8%) - Readability (109/115 or 94.8%) - Accuracy (87/115 or 75.7%) - Thoroughness (98/115 or 85.2%) - Consistency (73/95^a or 76.8%) The 69 references provided were evaluated on: - Pertinence (246/260 or 94.6%) - Relevance (237/260 or 91.2%) - Updatedness (162/260 or 62.3%)	ChatGPT demonstrated potential as a resource for dermatology continued medical education, producing appropriate responses in terms of quality and thoroughness, and high readability. However, responses were sometimes inaccurate or inconsistent with guidelines, indicating areas for improvement. The references provided by AI were positively evaluated in pertinence and relevance, but issues in currency suggest gaps in the availability of papers ChatGPT had been trained.
Rao et al. (2025) ²⁴	The questions reflected high levels of: - Alignment with scientific consensus (4.45, 95% CI 4.28–4.62) - Comprehensiveness (4.3, 95% CI: 4.11–4.89^b) - Overall quality (4.28, 95% CI: 4.10–4.47) The questions reflected low levels of: - Clinical harm (1.6, 95% CI: 1.38–1.81) - Demographic bias (1.52, 95% CI: 1.31–1.72)	Mean ratings of assessors reflected high alignment with scientific consensus, comprehensiveness and overall quality, with low suspicion of clinical harm and demographic bias. Errors relating to factual accuracy, including potential AI hallucinations, were indirectly captured in ratings for scientific alignment and clinical harm.
Cevik et al. (2025) ²⁵	Initial ratings for DALL-E had the authors rate on the following: - Resemblance to real world pathology (3/5) - Adequate representation (4/5) - Easily visible and discernible details (4/5) - Quality (4/5) - Beneficial for educational purposes (4/5)	Only DALL-E had the ability to generate distinctive features associated with skin cancer, while images created by Midjourney and BlueWillow had exaggerated fictionalised features.
Ayub et al. (2023) ²⁸	Out of the questions: 10/40 (25%) were low complexity 9/40 (22.5%) were vague or unclear 5/40 (12.5%) were inaccurate 16/40 (40%) were accurate and at an appropriate level of complexity	The model was assessed to have a limited domain-specific knowledge, with a limited ability to understand context and generate high-quality distractor options
Quasi-experimental studies		
Reference	Descriptive statistics	Conclusions
Pielage et al. (2025) ²⁶	- Skin lesions were assessed as mostly realistic (average of 6.17/10) - Subjective competence of the students increased (from 2.28 to 3.12/5) - Out of the 20 images assessed, the 4 students in the learning group (exposed to AI generated images) achieved a mean accuracy of 0.64, as compared to 0.56 for the 3 students in the control group.	The skin lesions were assessed as mostly realistic and the subjective competence of the students increased after completing the dermatology scenario in VR in the study. In the small group, a small increase in learning accuracy was observed after learning from AI-generated images, suggesting potential for medical education
Lee et al. (2024) ²⁷	89% of patients supported the use of such technology in medical student education - Average utilisation time of the application was 67.0 seconds	Such machine learning clinical decision support technology has the potential to actively engage students during their clinical rotations without being disruptive to clinic flow. Overall, patients supported the use of the technology.
Payne et al. (2009) ²⁹	- A decrease in errors over time was noted from approximately 50% to 20% - An increase in Correct-BNS (optimal solution path) actions over time was noted from approximately 20% to 50%	During the tutoring, the frequency of errors gradually declined. While student performance frequently differed from the tutor-defined optimal path, students approached the optimal solution path as they continued to be tutored.

AI: artificial intelligence; BNS: best-next-step; VR: virtual reality

^a Only 95 out of the 115 items were deemed eligible for assessment of consistency by the authors.

^b This confidence is not symmetric about the point estimate with a much higher upper bound (4.89), which may be a possible error in the original manuscript.

Table 3. Cochrane risk-of-bias assessment tool for non-randomised studies.

Bias domains	Pielage et al. (2025) ²⁶	Lee et al. ^a (2024) ²⁷	Payne et al. (2009) ²⁹
Pre-intervention			
1.1 Bias due to confounding	PY	PY	PY
1.2 Bias in selection of participants into the study	PY	PY	PY
Risk-of-bias judgment	Serious	Critical	Serious
At intervention			
2.1 Bias in classification of interventions	PY	PY	PY
Risk-of-bias judgment	Serious	Serious	Serious
Post-intervention			
3.1 Bias due to deviations from intended interventions	PN	NI	PN
3.2 Bias due to missing data	NI	NI	NI
3.3 Bias in measurement of outcomes	PY	PY	PY
3.4 Bias in selection of the reported result	PY	PY	PN
Risk-of-bias judgment	Serious	Serious	Moderate
Overall bias	Serious	Serious	Serious

NI: no information; PN: probably no; PY: probably yes

^a Assessment for this study was limited as a conference abstract

Table 4. COSMIN-inspired assessment for content validation studies.

COSMIN-inspired concept	Criterion	Cevik et al. (2025) ²⁵	Naldi et al. (2024) ²³	Rao et al. (2025) ²⁴	Ayub et al. (2023) ²⁸
Relevance/ construct definition	Clear description of construct being measured	Met	Met	Met	Met
Item relevance	Relevance of generated items to construct	Met	Met	Met	Met
Content coverage	Comprehensiveness of construct	Unmet —only a few images of skin cancers	Unmet —only used for acne vulgaris	Partially met —randomly selected vignettes under USMLE	Partially met —utilised CME articles from JAAD
Comprehensibility	Comprehensibility of rating scale	Met	Met	Met	Met
Expert review	Involvement of experts in development	Met	Met	Met	Met
Reliability evidence	Inter-rater agreement reported	Met	Met	Met	Met
Methodological quality of expert evaluation	Number of experts appropriate	Not met —only 3 individuals, with poor rigor	Met —5 individuals with 4 residents, 1 specialist	Not met —only 3 individuals	Not met —only 2 individuals
Qualitative validation evidence	Qualitative insights reported	Not reported	Reported	Not reported	Reported

CME: continuing medical education; JAAD: Journal of the American Academy of Dermatology; USMLE: United States Medical Licensing Examination

Table 5. Assessment of studies using the CUC-FATE framework.

Study	Cost	Usability	Credibility	Fairness	Accountability	Transparency	Explainability
Cevik et al. (2025) ²⁵	NR	R	R	NR	NR	NR	NR
Naldi et al. (2024) ²³	NR	R	R	NR	R	NR	R
Pielage et al. (2025) ²⁶	NR	R	R	NR	NR	R	R
Rao et al. (2025) ²⁴	NR	R	R	R	NR	NR	NR
Lee et al. ^a (2024) ²⁷	NR	R	NR	NR	NR	NR	NR
Ayub et al. (2023) ²⁸	NR	R	R	NR	NR	NR	NR
Payne et al. (2009) ²⁹	NR	R	R	NR	NR	R	R
No. of studies examining	0	7	6	1	1	2	3

NR: not reported; R: reported

^a Assessment for this study was limited as a conference abstract.

Two of the studies^{26,29} reported AI interventions created by the authors, allowing assessment of their transparency, but studies that utilised proprietary LLMs, such as the content validation studies, lacked the ability to do so. Author-created AI interventions in these studies also allowed assessment of explainability. Including a study²³ which asked the LLM to generate reference literature for provided information, a total of 3 studies specifically assessed explainability.

This same study also highlighted the presence of inaccuracies and errors in the LLM's responses, urging the necessity of expert review to prevent misinformation, thus being the only study assessing accountability.²³ Only 1 study that looked at AI-generated questions considered fairness, specifically assessing the risk of demographic bias in AI-generated questions and concluding that the risk of this is low.²⁴

DISCUSSION

This scoping review has identified a small but emerging body of literature examining the application of AI in dermatological education. Though the included studies examined a range of AI-based interventions, such as AI-generated images of dermatological conditions and LLM-generated educational content, the overall evidence base remains limited. Most studies were preliminary, small-scale and heterogeneous in design, and current evidence is insufficient in supporting conclusions regarding its effect on improving learner performance and clinical competence.

The review also highlights several questions that are still unanswered. It remains unclear which AI-based interventions are the most effective for different groups, such as medical students, residents, and general practitioners, and whether these interventions necessarily provide meaningful added value over

existing traditional educational methods. Important concerns on implementation, such as transparency and explainability, are also inconsistently addressed in studies, and represent critical concerns for educators.

Since completion of the search by this review, a broader scoping review of AI and dermatological education has been published by Lau et al.,³² also highlighting educational applications, including the use of LLMs as well as generative and adaptive imaging tools. In contrast, this present review focused on studies examining direct educational applications, deliberately excluding the substantial number of studies solely assessing AI performance in answering dermatology examination questions. By adopting this narrower educational focus, this review provides a more targeted appraisal of the current evidence for the use of AI in dermatological education, concluding that despite growing interest in the field, current evidence is sparse and limits strong conclusions.

Nevertheless, this field continues to evolve rapidly. For example, a study published after the search period examining the usage of LLMs as a dermatology case narrator for medical students³³ would have been eligible for this review. As this study was published after the prespecified search period, it was not included in the formal synthesis, but its relevance suggests that future updated reviews may identify a larger body of eligible evidence.

The findings of this review are also consistent with emerging evidence from the broader medical education literature, with a rapidly increasing number of AI-related publications over the past few years.⁶ A scoping review of AI medical image generation across health professions education (including dermatology) by Gupta et al.³⁴ also concluded that generative AI should be treated as an experimental adjunct requiring

rigorous human review. Similarly, in radiology education (a similarly visually intensive specialty), a scoping review of AI-based educational approaches by Hui et al.³⁵ identified a range of possible applications, from personalised curriculum generation and diagnostic support tools to automated evaluation systems.

While similar themes were observed in the present review, the evidence base identified for dermatology was considerably smaller, with fewer studies evaluating educational outcomes and learner performance. Together, these reviews suggest that while dermatology shares characteristics with more visual specialties—making it potentially well-suited for AI-enhanced learning, robust evidence demonstrating educational effectiveness is lacking, making further research with rigorous validation necessary to identify the most effective AI tools in education.³⁶ Other potential interventions not included in this review, such as personalised learning platforms,³⁷ integrating AI with virtual-reality based training for skin cancer screening³⁸ as well as AI-assisted learning for knowledge and milestone assessment,³⁹ have been proposed by other authors, and would benefit from implementation studies to assess educational outcomes.

Key limitations

There was considerable heterogeneity noted in study design and evaluation metrics for educational efficacy in quasi-experimental studies, with a diversity of intervention types explored. These studies were also small-scale with low sample sizes (ranging from 2 to 44), with few participants and inadequate control for confounding variables causing serious risk of bias. None of the studies employed validated instruments to assess learning outcomes, often relying on self-developed Likert scales or informal expert ratings. Similarly, none of the content validation studies used standardised measures to assess content validity of AI-generated information. This lack of standardisation prevents meaningful synthesis.

None of the studies examined compared AI-based interventions with traditional teaching methods. This absence of a control group limited conclusions about added educational value and generalisability, and such comparative analysis is essential in evaluating AI-based interventions' educational effectiveness.⁷

All studies were conducted in pilot or isolated research settings, outside the realm of formal medical school or residency training curricula, suggesting institutional adoption of AI in dermatology education is currently limited. This is consistent with existing research, which suggests that adoption of AI into medical education is still in its early developmental stages.⁴⁰ A key challenge is the alignment of existing educational frameworks and curricula with rapidly evolving AI

capacities,⁴¹ with educational centres often struggling with infrastructure and policy support required.⁴²

A further limitation relates to the review process itself, where study screening was conducted through consensus between 2 reviewers rather than independent duplicate reviews. While this collaborative approach allowed for discussion and agreement at each stage, it may have increased susceptibility to shared reviewer bias.

CUC-FATE framework

This scoping review utilised the CUC-FATE framework to examine crucial enablers for AI in medical education, demonstrating that most dimensions were unassessed by included studies. Credibility and utility were frequently assessed, and while an earlier content validation study²⁸ using an older LLM noted high levels of inaccuracy, this was not noted in newer LLM based content validation studies^{23,24} utilising newer models of the same LLM. This underscores how LLMs are rapidly evolving and improving, and older studies with less positive results may be less relevant with newer generative AI models.

Most studies lacked reporting of assessments of transparency and explainability. Content validity is limited by the inability to trace presented evidence to its origins, which is especially pertinent when proprietary technologies, such as the LLMs ChatGPT or Gemini, are used.^{43,44} These deficiencies may undermine clinicians' trust in AI-assisted dermatological education, particularly when verification of accuracy in clinical contexts is essential.⁴⁵

Encouragingly, an included image-generation study²⁶ described the creation of a custom diffusion model, providing detailed methodological transparency. While other studies focusing on deep learning algorithms and clinical decision support systems for diagnosis of melanoma^{46,47} and other skin diseases⁴⁸ were excluded from review because of the absence of educational focus, their transparent designs would allow effective adaptation for educational purposes.

Ethical concepts of accountability and fairness were rarely assessed in included studies. With the use of AI in dermatological education rapidly expanding, these omissions are concerning, given that the identification of ethical challenges is paramount for responsible development and implementation.⁴⁹ The issue of accountability, establishing liability for harmful results sustained from incorrect educational content,⁵⁰ was assessed in only 1 study²³ which emphasised the need for human oversight to ensure clinical accuracy. For fairness, AI algorithmic bias can lead to error in diagnosis of under-represented populations such as individuals of colour,⁵¹ causing erroneous education of medical professionals. This was briefly assessed in 1

study on AI-generated dermatological examination questions,²⁴ but should be more widely evaluated.

Implications for practice and future research

Despite promising innovation in the application of AI in dermatological education, this review has demonstrated that significant gaps remain, highlighting key areas for future development. Current quasi-experimental studies had a high rate of bias, emphasising the need for larger, robust, multicentre studies involving diverse participant cohorts in institutional contexts for both medical student and healthcare professional training in dermatology.

Significant methodological heterogeneity across studies severely limited comparison, highlighting the need for standardised reporting frameworks and verified evaluation tools to ensure replicability. This review has demonstrated a framework for content validation studies devised by the authors and informed by the COSMIN approach, and the CUC-FATE framework, which incorporates not just practical but important ethical considerations. Adopting such standardised frameworks would improve the quality and transparency of AI-related educational research in dermatology, allowing appropriate and effective implementation. There is a need for studies to explicitly incorporate ethical principles in their design, ensuring AI-based educational tools are equitably and responsibly implemented.

Singapore has emerged as an early regional adopter of AI in healthcare training, resulting in an increased impetus to equip faculty and students with the necessary foundational knowledge in AI in a forward-looking medical education landscape.⁵² In this aspect, efforts have been made to design curricula locally to integrate AI into medical education,⁵³ with local medical educators highlighting the benefits in equipping medical learners with skills in digital technology.⁵⁴ By using dermatology as a targeted case study, this review provides timely evidence that can inform curriculum design for AI-enhanced medical education in Singapore and similar healthcare systems.

Given the diversity of AI interventions, we recommend that further research involve learners in their design, which can improve engagement and reduce barriers to adoption. Multiple studies have consistently demonstrated that dermatology specialists,⁵⁵ trainees,⁵⁶ and students,^{40,57} while agreeing that AI will play a key role in dermatological education, expressed that they lacked training to utilise it effectively. A mixed-methods approach, combining quantitative outcomes of efficacy with qualitative data of learner experience, would be useful in examining these perceptions to enhance usage of AI in dermatological education.

Beyond the design of learner-facing interventions, successful implementation of AI in dermatological education will also need investment in faculty development. Recent medical education literature has emphasised the importance of equipping educators with AI literacy through structured faculty development models, which can include competency-based certification and interdisciplinary training.^{58,59} As AI becomes increasingly integrated into educational practice, frameworks for faculty development have been devised in broader medical education contexts,⁶⁰ although these have not been explored specifically in dermatology. Future work should therefore examine not only the effectiveness of AI interventions for learners, but also the competencies and support required for educators to implement these tools safely and effectively.

CONCLUSIONS

This review suggests that AI is an evolving tool for dermatology education, with potential benefits for learners. Existing literature is characterised by methodological heterogeneity and limited validation, and there is insufficient evidence to determine whether AI-based educational tools improve learning outcomes compared with standard educational approaches. Establishing standardised frameworks for the development, validation, and evaluation of AI-based educational technologies is crucial to generate higher-quality evidence, allowing for effective and ethical implementation in dermatology training.

Supplementary materials

Supplementary Table S1. Search strategy.

Supplementary Table S2. Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist.

Supplementary Table S3. Frequency table of outcomes of included content validation studies.

Supplementary Table S4. Assessment of studies using the framework proposed by Bates et al.

Ethics statement

This research did not involve the collection of patient data, human participants, or biological samples. All analyses were conducted on publicly available, published literature.

Declaration

The authors declare they have no affiliations or financial involvement with any commercial organisation with a direct financial interest in the subject or materials discussed in the manuscript.

REFERENCES

- Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. *Informatics* 2024;11:57.
- Mir MM, Mir GM, Raina NT, et al. Application of artificial intelligence in medical education: current scenario and future perspectives. *J Adv Med Educ Prof* 2023;11:133-40.
- Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ* 2019;76:1681-90.
- Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw Open* 2022;5:e2149008.
- Wartman SA, Combs CD. Reimagining medical education in the age of AI. *AMA J Ethics* 2019;21:E146-52.
- Zarei M, Mamaghani HE, Abbasi A, et al. Application of artificial intelligence in medical education: a review of benefits, challenges, and solutions. *Med Clin Pract* 2024;7:100422.
- Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ* 2019;5:e13930.
- Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. *IEEE Access* 2020;8:75264-78.
- Hogarty DT, Su JC, Phan K, et al. Artificial intelligence in dermatology: where we are and the way to the future: a review. *Am J Clin Dermatol* 2020;21:41-7.
- Li R, Liu G, Zhang M, et al. Integration of artificial intelligence in radiology education: a requirements survey and recommendations from faculty radiologists, residents, and medical students. *BMC Med Educ* 2025;25:380.
- Liopyris K, Gregoriou S, Dias J, et al. Artificial intelligence in dermatology: challenges and perspectives. *Dermatol Ther* 2022;12:2637-51.
- Perrin J, Petronic-Rosic V. The potential role and restrictions of artificial intelligence in medical school dermatology education. *Clin Dermatol* 2024;42:477-9.
- Nahm WJ, Sohail N, Burshtein J, et al. Artificial intelligence in dermatology: a comprehensive review of approved applications, clinical implementation, and future directions. *Int J Dermatol* 2025;64:1568-83.
- Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8:19-32.
- Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci* 2010;5:69.
- Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews: checklist and explanation. *Ann Intern Med* 2018;169:467-73.
- Quttainah M, Mishra V, Madakam S, et al. Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education. *JMIR AI* 2024;3:e51834.
- Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.
- Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res* 2018;27:1159-70.
- Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.
- Bates DW, Auerbach A, Schulam P, et al. Reporting and implementing interventions involving machine learning and artificial intelligence. *Ann Intern Med* 2020;172:S137-44.
- Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ* 2025;25:129.
- Naldi L, Bettoli V, Santoro E, et al. Application of ChatGPT as a content generation tool in continuing medical education: acne as a test topic. *Dermatol Rep* 2024;17:10138.
- Rao AS, Kim J, Mu A, et al. Synthetic medical education in dermatology leveraging generative artificial intelligence. *NPJ Digit Med* 2025;8:247.
- Cevik J, Seth I, Rozen WM. Artificial intelligence-generated images in skin cancer education: an experimental study. *Australas J Plast Surg* 2025;8:118181.
- Pielage L, Schmidle P, Marschall B, et al. Interactive high-quality skin lesion generation using diffusion models for VR-based dermatological education. In: *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 2025:878-97.
- Lee G, Lopez D, Singh P, et al. Machine learning technology in medical student education. Conference abstract, *J Invest Dermatol* 2024;144:S59.
- Ayub I, Hamann D, Hamann CR, et al. Exploring the potential and limitations of ChatGPT in generating board-style dermatology questions: a qualitative analysis. *Cureus* 2023;15:e45934.
- Payne VL, Medvedeva O, Legowski E, et al. Effect of a limited-enforcement intelligent tutoring system in dermatopathology on student errors, goals, and solution paths. *Artif Intell Med* 2009;47:175-97.
- Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? *Syst Rev* 2019;8:264.
- Haddaway NR, Page MJ, Pritchard CC, et al. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams. *Campbell Syst Rev* 2022;18:e1230.
- Lau LD, Tran V, Chapman W, et al. Applications of machine learning, natural language processing, and generative artificial intelligence in dermatology education and research: a scoping review. *Int J Dermatol* 2026;65:1600-10.
- Karampinis E, Bozi Tzetzis DA, Pappa G, et al. Use of a large language model as a dermatology case narrator: exploring the dynamics of a chatbot as an educational tool in dermatology. *JMIR Dermatol* 2025;8:e72058.
- Gupta K, Latinovich MF, Latinovich MF, et al. Generative artificial intelligence for medical image creation in health professions education: a scoping review. *J Med Syst* 2026;50:19.
- Hui ML, Sacoransky E, Chung A, et al. Exploring the integration of artificial intelligence in radiology education: a scoping review. *Curr Probl Diagn Radiol* 2025;54:332-8.
- Tozsin A, Ucmak H, Soyuturk S, et al. The role of artificial intelligence in medical education: a systematic review. *Surg Innov* 2024; 31:415-23.
- Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach* 2024;46:446-70.
- Mergen M, Will L, Graf N, et al. Feasibility study on virtual reality-based training for skin cancer screening: bridging the gap in dermatological education. *Educ Inf Technol* 2025; 30:5251-82.
- Ramachandran V, Jairath N, Cheraghlou S, et al. Revolutionizing dermatology residency: artificial intelligence for knowledge and clinical milestones assessment. *Clin Exp Dermatol* 2024; 49:732-3.

40. Lee J, Wu AS, Li D, et al. Artificial intelligence in undergraduate medical education: a scoping review. *Acad Med* 2021;96:S62-70.
41. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. *Lancet Digit Health* 2020;2:e111-2.
42. Kolachalama VB, Garg PS. Machine learning and medical education. *NPJ Digit Med* 2018;1:54.
43. Boscardin CK, Gin B, Golde PB, et al. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. *Acad Med* 2024;99:22-7.
44. Goktas P, Grzybowski A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. *J Clin Med* 2024;13:5909.
45. Zbrzezny AM, Krzywicki T. Artificial intelligence in dermatology: methods, clinical applications, and perspectives. *Appl Sci* 2025;15:7856.
46. Maron RC, Utikal JS, Hekler A, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma classification. *J Med Internet Res* 2020;22:e18091.
47. Díaz-Ramón JL, Gardeazabal J, Izu RM, et al. Melanoma clinical decision support system: an artificial intelligence-based tool. *Cancers (Basel)* 2023;15:2174.
48. Han SS, Park I, Chang SE, et al. Augmented intelligence dermatology. *J Invest Dermatol* 2020;140:1753-61.
49. Gordon ER, Trager MH, Kontos D, et al. Ethical considerations for artificial intelligence in dermatology. *Br J Dermatol* 2024;190:789-97.
50. Ferreira AL, Lipoff JB. The complex ethics of applying ChatGPT in dermatology. *J Am Acad Dermatol* 2023;89:e157-8.
51. Rinderknecht FA, Nwandu L, Daneshjou R, et al. The impact of artificial intelligence on health equity in dermatology. *Curr Dermatol Rep* 2024;13:148-55.
52. Tan JA, Ng IKS, Hong A, et al. Empowering the next generation: integrating artificial intelligence education into medical training. *Postgrad Med J* 2026;102:185-8.
53. Cai C, Jong M, Ng YY, et al. Developing a postgraduate program for AI in medicine with Kern's six-step curriculum development approach in Singapore. *Proc AAAI Conf Artif Intell* 2025;39:28979-87.
54. Zainal H, Xin X, Thumboo J, et al. Medical school curriculum in the digital age: perspectives of clinical educators and teachers. *BMC Med Educ* 2022;22:428.
55. Patrzyk S, Bielecki W, Woźniacka A. Attitudes among Polish dermatologists regarding modern technologies. *Adv Dermatol Allergol* 2022;39:531-7.
56. Polesie S, Gillstedt M, Kittler H, et al. Attitudes towards artificial intelligence within dermatology. *Br J Dermatol* 2020;183:159-61.
57. Cho SI, Han B, Hur K, et al. Perceptions and attitudes of medical students regarding artificial intelligence in dermatology. *J Eur Acad Dermatol Venereol* 2021;35:e1-5.
58. Ahsan Z. Integrating artificial intelligence into medical education: a narrative systematic review of current applications, challenges, and future directions. *BMC Med Educ* 2025;25:1187.
59. Sánchez-Mendiola M, Anakin M, Findyartini A, et al. Applied insights for using generative artificial intelligence in faculty development in health professions education. *MedEdPublish* 2025;15:279.
60. Association of American Medical Colleges. CGEA Artificial Intelligence Framework for Educators. Washington (DC): Association of American Medical Colleges; 2024. <https://www.aamc.org/about-us/mission-areas/medical-education/advancing-ai-resource-collection/cgea-artificial-intelligence-framework-educators>. Accessed 7 June 2026.