



eISSN 2036-7406

Dermatology Reports

<https://www.pagepress.org/journals/index.php/dr/index>



SIDCO

Società Italiana di Dermatologia
Chirurgica, Oncologica, Correttiva ed Estetica

Publisher's Disclaimer. E-publishing ahead of print is increasingly important for the rapid dissemination of science. **Dermatology Reports** is, therefore, E-publishing PDF files of an early version of manuscripts that undergone a regular peer review and have been accepted for publication, but have not been through the copyediting, typesetting, pagination and proofreading processes, which may lead to differences between this version and the final one.

The final version of the manuscript will then appear on a regular issue of the journal.

E-publishing of this PDF file has been approved by the authors.

Please cite this article as:

Mwageni N, van Wijk R, Schoenmakers A, et al. The reliability of clinical judgments in dermatology: a descriptive review on their role as a reference standard for technological diagnostic research. *Dermatol Rep* 2026 [Epub Ahead of Print] doi: 10.4081/dr.2026.10064



© the Author(s), 2026
Licensee [PAGEPress](https://www.pagepress.org/), Italy

Submitted 11 June 2024 - Accepted 17 December 2025

Note: The publisher is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries should be directed to the corresponding author for the article.

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

The reliability of clinical judgments in dermatology: a descriptive review on their role as a reference standard for technological diagnostic research

Nelly Mwageni,¹ Robin van Wijk,^{2,3} Anne Schoenmakers,^{2,3} John Elisante Masenga,⁴ Christa Kasang,⁵ Liesbeth Mieras,² Stephen E. Mshana¹

¹Catholic University of Health and Allied Sciences, Mwanza, Tanzania; ²NLR | until No Leprosy Remains, Amsterdam, The Netherlands; ³Department of Public Health, Erasmus MC, University Medical Center Rotterdam, The Netherlands; ⁴Kilimanjaro Christian Medical Centre, Moshi, Tanzania; ⁵Deutsche Lepra- und Tuberkulosehilfe e.V., Würzburg, Germany

*Correspondence: Nelly Mwageni, Catholic University of Health and Allied Sciences, Mwanza, Tanzania.
E-mail: nellymwageni@yahoo.com*

Key words: dermatology; skin diseases; teledermatology; mobile applications; reproducibility of results; observer variation.

Contributions: all the authors made a substantive intellectual contribution. All of them read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors have no conflict of interest to declare.

Ethics approval and consent to participate: not required; no study participants were enrolled for this study specifically.

Availability of data and materials: all data generated or analyzed during this study are included in this published article.

Funding: this review is part of a PEP4LEP PhD trajectory. The PEP4LEP project was part of the second European and Developing Countries Clinical Trials Partnership Programme (EDCTP2) supported by the European Union (EU), awarded to NLR/LM (grant number RIA2017NIM-1839-PEP4LEP). PEP4LEP was also funded by the Leprosy Research Initiative (LRI; www.leprosyresearch.org) awarded to

NLR/LM (grant number 707.19.58.). The funders had no role in study design, data collection, and analysis, the decision to publish, or preparation of the manuscript. Because of COVID-19 delays, research work continued under PEP4LEP 2.0 (grant number 101145677-PEP4LEP 2.0). The PEP4LEP 2.0 project is part of the Global Health EDCTP3 Joint Undertaking, co-funded by the EU. Views and opinions expressed are those of the authors only and do not necessarily reflect those of EDCTP2, LRI, the EU, or the Global Health EDCTP3 Joint Undertaking. Neither of the funders can be held responsible for them.

Abstract

The expansion of dermatological mobile health (mHealth) and teledermatology has generated numerous diagnostic support tools, typically validated against traditional face-to-face assessments. Histopathological confirmation through skin biopsy is invasive, resource-intensive and, in many African settings, not readily available as a dermatological research reference standard. Yet, the reliability of clinical diagnoses used as such standards remains uncertain. The study objective was to assess the consistency (inter-rater reliability) of face-to-face clinical judgments among multiple dermatologists evaluating the same patient group. For this descriptive literature review, a comprehensive search of seven databases and information platforms was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) reporting guideline. In total, 1,700 studies were found after deduplication and screened using pre-set criteria. Eighteen studies were included: three compared clinical diagnoses directly and 15 assessed the reliability of dermatological scoring systems for disease severity and activity. All studies involved two or more dermatologists assessing the same patients. Reported agreement among raters ranged from 0.15 to 0.98 for studies evaluating clinical diagnosis and from 0.07 to 0.95 for studies assessing disease severity or activity scores, demonstrating wide variability in inter-rater reliability. In conclusion, clinical diagnosis alone, especially if made by one or a limited number of dermatologists, may not provide a sufficiently reliable reference standard for validating teledermatology, dermatological mHealth applications, or artificial intelligence (AI) tools. Strengthening evidence on the reliability of diagnostic reference standards is essential for evaluating emerging digital health innovations, particularly in regions with limited access to dermatologists and histopathology services. Future studies, with similarly experienced dermatological assessors, should have larger sample sizes, assess diagnostic agreement rather than disease severity outcomes, combine inter- and intra-rater assessments, and include diverse geographic contexts such as Africa, Asia and Latin America. The potential influence of assessors' levels of experience on agreement should also be explored. A final recommendation is to come to a consensus about skin color classification.

Introduction

Skin conditions contribute to 1.8% of the global disease burden and rank fourth among causes of disability.¹ They affect around 10% of patients seeking outpatient care in African countries.² Approximately 30-70% of people worldwide have skin-related health problems that need effective and efficient diagnosis and treatment.² Skin diseases impact quality of life, mental health and productivity,

and can indicate systemic diseases.^{3,4} Various factors influence the types of skin diseases, including geography, environmental factors, hygienic standards, cultural and social customs and genetics.⁵

Diagnoses of skin conditions are typically made clinically through visual observation,⁶ though laboratory investigations such as biopsies may be necessary for confirmation or specific conditions.⁷ The World Health Organization (WHO) emphasizes that skin diseases should be diagnosed by properly trained individuals.⁵

Mobile tele dermatology applications and other mobile health (mHealth) tools, such as dermatological apps including artificial intelligence (AI), are increasingly used for clinical decision-making, particularly where face-to-face assessment is challenging.⁸ Validation of these tools often involves comparison with clinical judgment by dermatologists.^{8,9} Assessing the reliability of face-to-face clinical judgment is crucial before adopting it as a standard for evaluating new digital diagnostic tools, especially in areas lacking access to dermatologists, trained laboratory technicians, pathologists and associated diagnostic resources.

This review builds on previous validation studies in which clinical (face-to-face) judgment was assessed through comparisons among multiple dermatologists. From the literature comparing face-to-face diagnosis with tele dermatology outcomes, we included only those studies in which in-person evaluations were used to assess disease activity and severity scores and where at least two dermatologists were involved in the assessment process.

Materials and Methods

Search strategy

The search was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) reporting guideline, which provides guidance for reporting systematic reviews.¹⁰ The research question was modelled according to the PICO model, whereby “P” (population) represents two or more dermatologists diagnosing or rating the same cohort of patients; “I” (intervention) represents a diagnostic judgment offered by raters; “C” (comparison) stands for the comparison between the raters’ judgment; and “O” (outcome), which, in this setting, is the agreement on the diagnostic judgment among the raters.¹¹

A comprehensive search on PubMed, Google Scholar, Embase, Medline, Cochrane Library, WHO.int, and ClinicalTrials.gov was conducted. The search syntax used was: “(dermatologist OR dermatology OR skin) AND (consultation OR diagnosis OR face-to-face OR clinical) AND (reliability OR inter-rater OR tele dermatology)”, using the search fields: “Title”, “Abstract”, “Key words”. Articles and documents

published from 1990 onwards were included. The search was performed between September 30, 2022, and December 10, 2022. We searched for reports regardless of their publication type.

Inclusion and exclusion criteria

Initial selection took place through title and abstract screening, and all studies discussing clinical accuracy regarding skin conditions in the dermatological field were included. All full-text original articles published in English and in scholarly and peer-reviewed academic journals from 1990 to 2019 that assessed the accuracy of dermatologists' judgment were included in the second stage of the review. Two independent researchers screened the relevant articles, and full agreement was reached on all studies that met the selection criteria.

The selected studies assessed the accuracy of judgment regarding skin conditions with a face-to-face approach. The criteria for inclusion were: i) the presence of two or more dermatologist raters providing their clinical judgment independently; ii) the presence of the same patient cohorts; and iii) at least two compared face-to-face consultations. All studies that did not meet these criteria were excluded. Assessment of the quality of studies was conducted according to the Newcastle-Ottawa Scale, and factors like population, intervention, and outcome were considered.¹²

Data extraction and synthesis

For this descriptive review, all data relevant to the research question were extracted using the Joanna Briggs Institute Model, which is a developmental framework for evidence-based practice situated in healthcare evidence.¹³ The data extraction form was used with specific consideration of the study details, objectives, participants' characteristics, study settings/context, methods of analysis, outcome assessed, and results. The narrative summary is described with emphasis on the study characteristics, and the findings are presented in Tables 1 and 2.

Results

Study selection

A total of 1,712 articles were identified from the literature search, with 12 being duplicates. In total, 1,610 were excluded because they did not include any face-to-face dermatological consultations. Thirteen articles were excluded because no full texts were available and two were not written in the English language. A total of 75 articles were reviewed in full text; 18 papers met the pre-set criteria and

were included for further analysis. A flowchart describing the study selection procedure is shown in Figure 1.

Study characteristics

Of the 18 articles included, two papers included a purely pediatric population,^{14,15} three papers had a mixed population regarding age,¹⁶⁻¹⁸ and the other 10 papers only included an adult population.¹⁹⁻²⁸ The majority of the included studies were performed in North America (n=12), of which 10 were from the United States,^{14,15,17,20-23,25,27,29} three were from Europe,^{21,28,30} three from Oceania, particularly Australia,^{16,18,31} and none from Africa, Asia, or Latin America. Almost all studies had an all-gender representation, except for one study by Pandya *et al.*, which comprised female participants only.¹⁹

Among the included studies, 16 involved more than two dermatologists who provided their observations of the same cohort of patients.^{14-28,31} Of those 16 studies, three involved a senior dermatology resident among the dermatologists who participated.^{16,19,20} Pandya *et al.* included five dermatologists and one senior dermatology resident; Rosenbach *et al.* included seven dermatologists and one senior dermatology resident; and Zhao *et al.* from 2017 included four dermatologists and one senior dermatology resident.^{16,20} Fifteen studies evaluated specific diseases by scoring their progression and the extent of damage. In these studies, dermatologists conducted face-to-face assessments of patients and rated the diseases according to predefined criteria.¹⁴⁻²⁸ Of these studies, three were performed in patients with atopic dermatitis,^{16,17,21} two in patients with cutaneous sarcoidosis,^{20,22} and two in patients with cutaneous lupus erythematosus.^{14,23} In all studies using scales, dermatologists assessed the same cohort of patients and a matched score or scale outcome from at least two dermatologists was considered as the “correct assessment”.

Of the included studies, only two had more than 100 participants; Whited *et al.* and Tan *et al.* from 2010 included patient samples of 126 and 200, respectively.^{29,31} In some studies, patient characteristic information was missing; for example, Hanifin *et al.*, Pandya *et al.*, Chandran *et al.*, and Lozzi *et al.* did not state the ethnicity and/or skin color characteristics of their participants, which is relevant in dermatology.^{17,19,24,30} The characteristics of the participants of the reviewed studies are summarized in Tables 1 and 2.

Inter-rater reliability from selected studies

The inter-rater agreement among the dermatologist raters for the selected studies showed some variation. Three of the selected studies used clinical diagnosis.²⁹⁻³¹ Among these, a high agreement was observed.

In the 2010 study by Tan *et al.*, in which three dermatologists in three pairs rated 200 patients with the aid of a handheld dermatoscope, an agreement between 0.95 and 0.98 per pair was found.³¹ Of the three studies, one had a moderate agreement of 0.54 for their single most likely diagnosis, and the agreement improved to 0.92 after including the differential diagnosis.²⁹ The lowest agreement was observed in the study by Lozzi *et al.*, in which two dermatologists rated 33 patients, and the same two dermatologists together diagnosed 5 out of 33 skin conditions the same, resulting in an agreement of 0.15.³⁰ The authors considered the low level of agreement to be due to the unusually difficult dermatoses among this group of 33 patients.

The remaining 15 studies used either severity or disease activity scale scoring for a particular disease. The diseases scored were cutaneous lupus,^{14,23} cutaneous sarcoidosis,^{20,22} atopic dermatitis,¹⁶⁻¹⁸ dermatomyositis,^{15,25} acne,²⁶ hand eczema,²¹ facial erythema (rosacea),²⁷ psoriasis,²⁴ melasma,¹⁹ and hidradenitis suppurativa.²⁸ Both studies that evaluated patients with cutaneous lupus erythematosus demonstrated good inter-rater agreement when dermatologists assessed disease severity using the Cutaneous Lupus Erythematosus Disease Area and Severity Index,^{14,23} but a moderate agreement was observed by Kushner *et al.* when using the Physician Global Assessment of the disease.¹⁴ Low agreement scores were reported in the studies by Rosenbach *et al.* and Yeung *et al.* when assessing patients with cutaneous sarcoidosis using the Cutaneous Sarcoidosis Activity and Morphology Instrument for damage caused by the disease, with agreement coefficients of 0.42 and 0.37, respectively.^{20,22} In contrast, substantial agreement was observed when the same instrument was used to score disease activity, with coefficients of 0.82 and 0.74, respectively. Twelve dermatologists in the 2014 study by Tan *et al.* scored facial erythema in patients with rosacea in two separate sessions on the same day and found a slightly lower agreement in session 2 compared to session 1 (0.58 and 0.60, respectively).²⁷ In the 2006 study by Tan *et al.*, six dermatologists assessed acne severity in patients. Agreement was lowest for non-inflammatory acne lesion counts (0.17), compared with substantially higher agreement for inflammatory lesion counts (0.84) and for the global severity assessment (0.71).²⁶ Two studies assessed patients with dermatomyositis; one involved the adult population,²⁶ and the other assessed the pediatric population.¹⁵ Lower agreement scores were observed in the study involving pediatric participants (ranging from 0.49 to 0.71) compared with the study involving adult participants (ranging from 0.58 to 0.86).^{15,25} In another study, six experienced dermatologists assessed 18 patients with chronic hand eczema using the Hand Eczema Extent Score and achieved substantial agreement (0.82).²¹ In the study by Hanifin *et al.*, 15 dermatologists evaluated patients with atopic dermatitis over 2 separate days. Patients were divided into two age-based cohorts: Cohort 1 (adults) and Cohort 2 (children).¹⁷ Agreement was higher in Cohort 1

on both day 1 and day 2 (0.72 and 0.82, respectively) compared with Cohort 2 (0.66 and 0.59, respectively). Thorlacius *et al.* reported a wide range of agreement scores for lesion counts among the expert dermatologists in hidradenitis suppurativa, with the lowest score of 0.07 when scoring the abscess and the highest score of 0.78 when scoring the Hidradenitis Suppurativa Severity Index during the assessment of the outcome measurement instruments and staging systems used in hidradenitis suppurativa.²⁸ Tables 1 and 2 summarize the findings from the reviewed studies.

Discussion

This study reviewed validation studies comparing face-to-face dermatological judgment between two or more dermatologists or doctors with/receiving dermatological training, finding a relatively wide range of inter-rater agreement scores. Overall, among the studies that involved disease scoring, the scores obtained when the raters assessed the disease activity were higher compared to when the dermatologists assessed the damage caused by the disease.^{17,20,22,27} This might be because disease activity could be easier to assess compared to damage caused by the disease, and/or because disease activity scales are easier to use. The study by Tan *et al.*, published in 2010, observed the highest agreement (0.97) of all studies when validating the face-to-face diagnosis of two out of three included dermatologists.³¹ A reason for this high score could be that the study only focused on selected patients with lesions suspicious of malignancy, using only three diagnostic disease categories: “melanoma”, “non-melanoma” and “others”. Another reason could be that the dermatologists diagnosed patients by examining them clinically with the aid of a hand-held dermatoscope. The presence and the use of the dermatoscope may have contributed to the high number of corresponding diagnoses. Not all studies described dermatoscope usage.^{29,30} Carlsson *et al.* found an agreement of 0.84 when six expert dermatologists evaluated the reliability of hand eczema extent score among patients with hand eczema.²¹ Moderate inter-rater agreement scores were reported by Pandya *et al.*, with values of 0.68 on the first day and 0.70 on the second day. Their study involved medical doctors with varying levels of dermatology experience: one dermatology resident, two highly experienced dermatologists, two with minimal experience, and two with no experience in dermatology.¹⁹ High agreement levels are generally expected when raters are especially experienced in a field, although the literature in other medical specialties shows mixed results.³²⁻³⁵ In contrast, the study by Lozzi *et al.* found significantly lower inter-rater reliability in studies relying on clinical diagnosis alone, reporting only 15% agreement.³⁰ The authors attributed this to the unusual difficulty of the dermatoses assessed, which may explain their low scores. It may also be because some diseases require laboratory support, such as histopathological examination, to reach a reliable diagnosis.³⁶

Among studies evaluating disease-specific scoring scales, Thorlacius *et al.* reported the lowest agreement score of 0.07 and a wide range of agreement values. This was despite the fact that the raters were experienced dermatologists specializing in hidradenitis suppurativa and had received introductory sessions on the scoring system.²⁸ The authors aimed to evaluate the “real-world” inter-rater reliability and agreement among experts in a scenario; however, bedside training sessions were not included in the training package. Incorporating bedside training sessions into the implementation of the scoring system could potentially enhance agreement when assessing the disease-specific scoring tool. Another explanation could be that the outcomes were related to lesion count, which can easily differ if many lesions are present.

Hanifin *et al.* demonstrated that agreement varied when dermatologists assessed two cohorts of the same individuals on two consecutive days.¹⁸ Re-assessment differences, although smaller, were also observed in the study by Tan *et al.* from 2014, in which 28 patients with rosacea were assessed in two different sessions on the same day.²⁷ The results show that re-assessment, when physicians assess the same cohort of patients on different days or in different sessions, is likely to introduce some discrepancy as a result of a second thought that may affect the judgment. It is therefore important to also take intra-rater reliability into account.

No clear difference in inter-rater agreement was observed between studies with two raters and those with multiple raters.^{24,26} Agreement from studies with a large patient sample size showed no difference from those with a small patient sample size.^{29,31}

There was little variation in inter-rater reliability when dermatology clinicians rated adult vs. pediatric dermatological conditions. For most dermatological manifestations, the diagnostic approach is likely similar between adults and children among experienced dermatology experts.^{14,15,18,23,25}

Unfortunately, no relevant studies from Africa, Asia, or Latin America were found to be available for this review. Geographical context is crucial in dermatology due to variations in skin color, socio-cultural factors, health systems, and disease prevalence.³⁷ Some studies included patients of diverse ethnicity or ancestry, but the majority described participants as “Caucasian” or “European”, which is vague given Europe’s diversity.^{14,15,18,23,25,27,29,31} The same accounts for other terminology used. Terms as “African American” and “Black” are not always interchangeable, and a wide variation, including in skin color, can similarly be seen within the other “ethnicity groups”, which were described as “Asian” and “Hispanic/Latino”.³⁸ We would also advise against using the grouping term “mixed”.^{14,15} Besides “ethnicity groups”, Zhao *et al.* from 2017 also used terminology such as “skin of color” and “white”, as well as the melanin index.^{16,39} The study by Tan *et al.*, published in 2014, reported that all their study

participants were Caucasians of Fitzpatrick skin type I-III.^{27,40} The concepts of ethnicity, ancestry, race and skin color are being hotly debated in epidemiology, public health, global health, dermatology, and also outside the medical-scientific field.^{36,38} It is important for scientists and clinicians to use appropriate consistent terminology, and continuously educate themselves about this topic, as descriptions of background or skin color are clinically relevant to be included in their work.⁴¹⁻⁴³ Zhao *et al.* (2017) also reported differences in inter-rater agreement across described skin color categories.¹⁶ Future studies should use a clear and appropriate skin color and/or type classification system when reporting dermatological findings; however, as no consensus currently exists on the most suitable scale,⁴³ developments in validated tools should be closely monitored and adopted as they emerge.^{42,43}

Only three included studies assessed inter-rater reliability for clinical diagnosis;²⁹⁻³¹ the other studies used agreement of judgment based on the use of a disease scoring tool, including pre-selected patients with a specific disease.^{14,16,17,19-24,27} This does not reflect the scenario of clinic-based consultation, where an undiagnosed (not pre-selected) patient comes for a clinical examination of a skin condition. The diversity in the methodology of the studies included makes meta-analysis challenging; therefore, a descriptive method was used to display results. In addition, most studies had relatively small sample sizes, impacting their findings (Tables 1 and 2). The factors named above may have also contributed to the wide variation in inter-rater reliability among dermatologists that was observed in our review.

While many skin conditions can be clinically diagnosed, laboratory investigations are necessary for some.³⁶ These procedures vary from routine and non- or minimal invasive (*e.g.*, swabs for bacterial cultures), to more invasive diagnostic procedures (*e.g.*, skin biopsies for histopathology). Histopathology is usually considered the gold standard for the diagnosis of most skin lesions,⁴⁴ but not always desired or practical because of the invasiveness, time consumption (follow-up contact is needed), and resource availability (trained staff, materials, funding, *etc.*). Consequently, this is not always feasible in dermatology within low-resource settings, even in more specialized or secondary healthcare services.^{45,46}

As the use of AI, teledermatology, and mHealth tools in dermatology expands, establishing reliable reference standards for clinical diagnoses is crucial to ensure patient safety and diagnostic accuracy.⁴⁷⁻⁴⁹

Without robust reference standards in research, there is a risk of misdiagnosis or suboptimal care, particularly in underserved or remote regions, highlighting the need for further studies to establish universally accepted benchmarks for diagnosis in dermatology. Additionally, conducting dermatological research across diverse populations, including patients with varying skin tones, is essential, as these differences can significantly influence study outcomes.^{38,49} The findings of this review are particularly timely following the adoption of the 2025 World Health Assembly (WHA) resolution recognizing skin

diseases as a global public health priority (WHA78.15), which highlights the need for strengthened dermatology services, improved diagnostic capacity, enhanced disease surveillance, and research and innovation to improve skin health.⁵⁰

Conclusions

Based on the findings of this study, the inter-rater reliability scores of face-to-face consultations by multiple dermatologists independently have a relatively wide range. Clinical judgment by multiple dermatologists with similar levels of experience may be used in settings where potentially more reliable comparison methods, such as histopathology, are not available or desirable, but it is important to take shortcomings into account. Nevertheless, it remains uncertain whether a clinical diagnosis by a dermatologist (especially if made by one or a limited number of dermatologists), without additional testing, serves as a reliable comparative standard for AI applications, mobile tele dermatology interventions, or (other) diagnostic mHealth tools in research settings. As research from the Asian, African, or Latin American context is lacking on this topic, we would advise the setup of new studies with larger sample sizes in these geographical contexts. In addition, more inter-rater reliability research, preferably combined with intra-rater assessments, is needed on the primary diagnosis of dermatological patients, instead of providing disease severity scores. A final recommendation is to come to a consensus about skin color and/or type classification.

References

1. Karimkhani C, Dellavalle RP, Coffeng LE, et al. Global Skin Disease Morbidity and Mortality: An Update From the Global Burden of Disease Study 2013. *JAMA Dermatol* 2017;153:406-12.
2. Hay RJ, Johns NE, Williams HC, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J Invest Dermatol* 2014;134:1527-34.
3. Hong J, Koo B, Koo J. The psychosocial and occupational impact of chronic skin disease. *Dermatol Ther* 2008;21:54-9.
4. Holmes CE, Massa MC. Skin signs of systemic disease: When the problem is more than skin-deep. *Postgrad Med* 1994;96:93-102.
5. World Health Organization. Epidemiology and Management of Common Skin Diseases in Children in Developing Countries. 2005. Accessed December 23, 2025. Available from: <https://www.who.int/publications/i/item/WHO-FCH-CAH-05.12>.
6. Whited JD, Hall RP. Diagnostic accuracy and precision in assessing dermatologic disease: problem or promise? *Arch Dermatol* 1997;133:1409-15.
7. Yang J, Sun X, Liang J, Rosin PL. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2018:1258-66.
8. Eedy DJ, Wootton R. Teledermatology: a review. *Br J Dermatol* 2001;144:696-707.
9. Warshaw EM, Lederle FA, Grill JP, et al. Accuracy of teledermatology for pigmented neoplasms. *J Am Acad Dermatol* 2009;61:753-65.
10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372.
11. Schardt C, Adams MB, Owens T, et al. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak* 2007;7:16.
12. Santos CM da C, Pimenta CA de M, Nobre MRC. The PICO strategy for the research question construction and evidence search. *Rev Lat Am Enfermagem* 2007;15:508-11.
13. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-analyses. Ottawa Hospital Research Institute. 2000. Accessed December 23, 2025. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp
14. Kushner CJ, Tarazi M, Gaffney RG, et al. Evaluation of the reliability and validity of the Cutaneous Lupus Erythematosus Disease Area and Severity Index (CLASI) in paediatric cutaneous lupus among paediatric dermatologists and rheumatologists. *Br J Dermatol* 2019;180:165-71.
15. Tiao J, Feng R, Berger EM, et al. Evaluation of the reliability of the Cutaneous Dermatomyositis Disease Area and Severity Index and the Cutaneous Assessment Tool–Binary Method in juvenile dermatomyositis among paediatric dermatologists, rheumatologists and neurologists. *Br J Dermatol* 2017;177:1086-92.
16. Zhao CY, Hao EY, Oh DD, et al. A comparison study of clinician-rated atopic dermatitis outcome measures for intermediate-to dark-skinned patients. *Br J Dermatol* 2017;176:985-92.
17. Hanifin JM, Thurston M, Omoto M, et al. The eczema area and severity index (EASI): assessment of reliability in atopic dermatitis. *Exp Dermatol* 2001;10:11-8.
18. Zhao CY, Tran AQT, Lazo-Dizon JP, et al. A pilot comparison study of four clinician-rated atopic dermatitis severity scales. *Br J Dermatol* 2015;173:488-97.
19. Pandya AG, Hynan LS, Bhore R, et al. Reliability assessment and validation of the Melasma Area and Severity Index (MASI) and a new modified MASI scoring method. *J Am Acad Dermatol* 2011;64:78-83.
20. Rosenbach M, Yeung H, Chu EY, et al. Reliability and convergent validity of the cutaneous sarcoidosis activity and morphology instrument for assessing cutaneous sarcoidosis. *JAMA Dermatol* 2013;149:550-6.
21. Carlsson A, Svensson Å, Anderson CD, et al. Scoring of hand eczema: good reliability of Hand Eczema Extent Score (HEES). *Acta Derm Venereol* 2017;97:193-7.
22. Yeung H, Farber S, Birnbaum BK, et al. Reliability and validity of cutaneous sarcoidosis outcome instruments among dermatologists, pulmonologists, and rheumatologists. *JAMA Dermatol* 2015;151.

23. Krathen MS, Dunham J, Gaines E, et al. The cutaneous lupus erythematosus disease activity and severity index: expansion for rheumatology and dermatology. *Arthritis Care Res (Hoboken)* 2008;59:338-44.
24. Chandran V, Gottlieb A, Cook RJ, et al. International multicenter psoriasis and psoriatic arthritis reliability trial for the assessment of skin, joints, nails, and dactylitis. *Arthritis Care Res (Hoboken)* 2009;61:1235-42.
25. Tiao J, Feng R, Bird S, et al. The reliability of the Cutaneous Dermatomyositis Disease Area and Severity Index (CDASI) among dermatologists, rheumatologists and neurologists. *Br J Dermatol* 2017;176:423-30.
26. Tan JKL, Fung K, Bulger L. Reliability of dermatologists in acne lesion counts and global assessments. *J Cutan Med Surg* 2006;10:160-5.
27. Tan J, Liu H, Leyden JJ, Leoni MJ. Reliability of clinician erythema assessment grading scale. *J Am Acad Dermatol* 2014;71:760-3.
28. Thorlacius L, Garg A, Riis PT, et al. Inter-rater agreement and reliability of outcome measurement instruments and staging systems used in hidradenitis suppurativa. *Br J Dermatol* 2019;181:483-91.
29. Whited JD, Hall RP, Simel DL, et al. Reliability and accuracy of dermatologists' clinic-based and digital image consultations. *J Am Acad Dermatol* 1999;41:693-702.
30. Lozzi GP, Soyer HP, Massone C, et al. The additive value of second opinion teleconsulting in the management of patients with challenging inflammatory, neoplastic skin diseases: a best practice model in dermatology? *J Eur Acad Dermatol Venereol* 2007;21:30-4.
31. Tan E, Yung A, Jameson M, et al. Successful triage of patients referred to a skin lesion clinic using teledermoscopy (IMAGE IT trial). *Br J Dermatol* 2010;162:803-11.
32. Fantoni M, Izzi I, Del Borgo C, et al. Inter-rater reliability of a modified Karnofsky Scale of Performance Status for HIV-infected individuals. *AIDS Patient Care STDS* 1999;13:23-8.
33. Fraga M, Taffé P, Méan M, et al. The inter-rater reliability of the Pulmonary Embolism Severity Index. *Thromb Haemost* 2010;104:1258-62.
34. Suriadi H, Imran AD. Reliability study of a new wound classification system for patients with diabetes. *Clin Pract* 2021;18:1672-7.
35. McKeith IG, Fairbairn AF, Bothwell RA, et al. An evaluation of the predictive validity and inter-rater reliability of clinical diagnostic criteria for senile dementia of Lewy body type. *Neurology* 1994;44:872.
36. Ruocco E, Baroni A, Donnarumma G, Ruocco V. Diagnostic procedures in dermatology. *Clin Dermatol* 2011;29:548-56.
37. Mosam A, Todd G. Dermatology training in Africa: successes and challenges. *Dermatol Clin* 2021;39:57-71.
38. Bhopal R. Glossary of terms relating to ethnicity and race: for reflection and debate. *J Epidemiol Community Health* (1978) 2004;58:441-5.
39. Takiwaki H, Overgaard L, Serup J. Comparison of narrow-band reflectance spectrophotometric and tristimulus colorimetric measurements of skin color: Twenty-three anatomical sites evaluated by the dermaspectrometer® and the chroma meter CR-200®. *Skin Pharmacol Physiol* 1994;7:217-25.
40. Fitzpatrick TB. Soleil et peau. *J Med Esthet* 1975;2:33-4.
41. Harvey VM, Brown SG, Burgess CM, et al. A call to action: The need to update population descriptors in dermatology research studies with the use of an inclusive skin classification system. *J Am Acad Dermatol* 2025;93:553-6.
42. Khoza N, Callender VD, Xiang LF, Alexis AF. Universal dermatology: bridging gaps in skin health. *JEADV Clinical Practice* 2025;4:S26-35.
43. Harvey VM, Alexis A, Okeke CA V, et al. Integrating skin color assessments into clinical practice and research: a review of current approaches. *J Am Acad Dermatol* 2024;91:1189-98.
44. Werner B. Biópsia de pele e seu estudo histológico: Por quê? Para quê? Como? Parte I. *An Bras Dermatol* 2009;84:391-5.
45. Fleming KA, Naidoo M, Wilson M, et al. An essential pathology package for low-and middle-income countries. *Am J Clin Pathol* 2017;147:15-32.

46. World Health Organization. Skin-related neglected tropical diseases: WHO launches new version of the WHO Skin NTDs mobile application. 2023. Accessed December 23, 2025. Available from: <https://www.who.int/news/item/09-10-2023-skin-related-neglected-tropical-diseases--who-launches-new-version-of-the-who-skin-ntds-mobile-application>
47. Giansanti D. Advancing Dermatological Care: A Comprehensive Narrative Review of Tele-Dermatology and mHealth for Bridging Gaps and Expanding Opportunities beyond the COVID-19 Pandemic. *Healthcare (Basel)* 2023;11:1911.
48. Liopyris K, Gregoriou S, Dias J, Stratigos AJ. Artificial Intelligence in Dermatology: Challenges and Perspectives. *Dermatol Ther (Heidelb)* 2022;12:2637-51.
49. Groh M, Badri O, Daneshjou R, et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nat Med* 2024;30:573-83.
50. World Health Organization. Skin Diseases as a Global Public Health Priority. 2025. Accessed December 23, 2025. Available from: https://apps.who.int/gb/ebwha/pdf_files/WHA78/A78_R15-en.pdf

Figure 1. Flow chart for study selection (1990-2022).

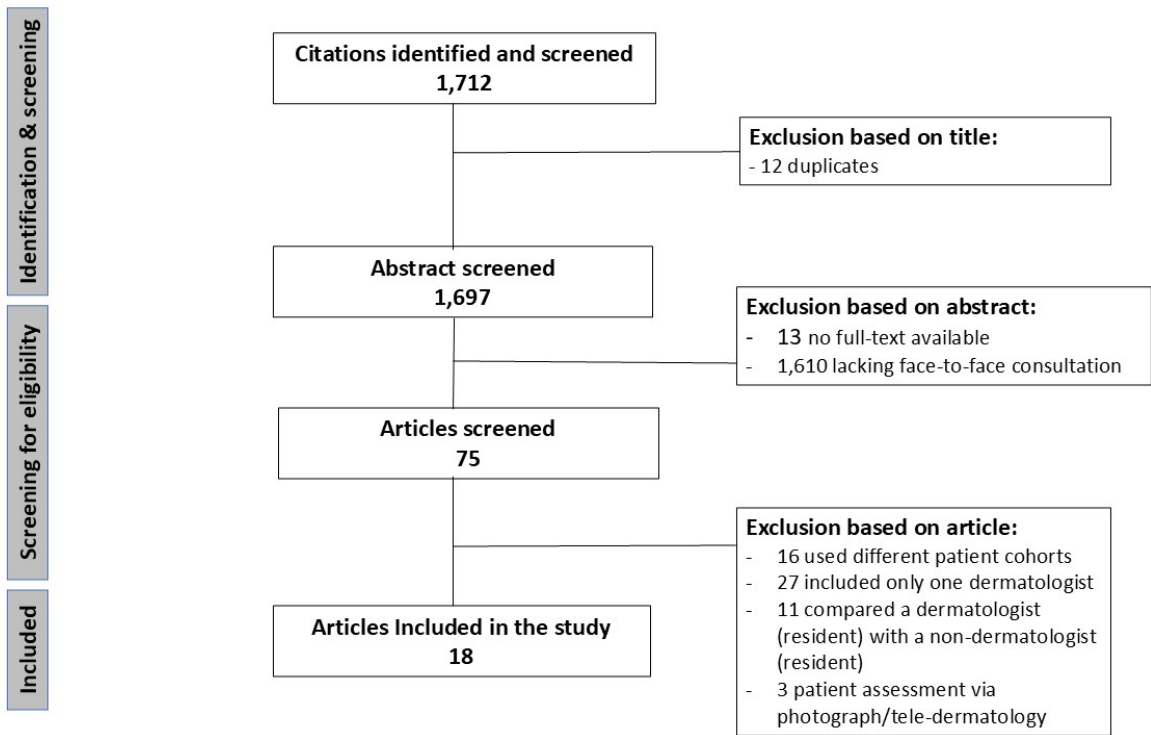


Table 1. Summary of the reviewed studies using clinical diagnosis.

Study (author, year)	Country	Patient sample size	Sex		Age (years)	Ethnicity/skin color/phototype as described in the article	Disease (category)	No. of different diseases	No. of MDs	Inter-rater agreement calculation	Inter-rater agreement (95% CI)
			F	M							
Lozzi <i>et al.</i> , 2007 ³⁰	Italy, Austria	33	15	18	Mean 51.1	-	Mixed	10 neoplastic diseases, 23 inflammatory diseases	2	Percentage agreement calculated	0.15
Tan <i>et al.</i> , 2010 ³¹	Australia, New Zealand	200	126	74	11-94	European ancestry 94%	Mixed	3 disease categories: melanoma, non-melanoma, and others	3	Weighted kappa	Dermatologists Pair 1: 0.97 (0.92-1.00); Dermatologists Pair 2: 0.98 (0.96-1.00); Dermatologists Pair 3: 0.95 (0.88-1.00)
Whited <i>et al.</i> , 1999 ²⁹	USA	126	3	123	Mean 61	Caucasian 100 (described as 79.8% in study)	Mixed	All referred skin lesions, including skin malignancy, basal cell carcinoma, squamous cell carcinoma, and melanoma	2	Weighted kappa	0.54 (0.45-0.61) for single most likely diagnosis; 0.92 (0.88-0.96) if including differential diagnosis

CI, confidence interval; F, female; M, male; MDs, medical doctors.

Table 2. Summary of the reviewed studies using scale scoring.

Study (author, year)	Country	Patient sample size	Sex		Age (years)	Ethnicity/skin color/phototype as described in the article	Disease (category)	Scale	No. of MDs	Inter-rater agreement calculation	Inter-rater agreement (95% CI)
			F	M							
Carlsson <i>et al.</i> , 2017 ²¹	Sweden	18	14	4	Mean 49	-	Hand eczema	HEES	6	ICC	0.82 (0.54-0.92)
Chandran <i>et al.</i> , 2009 ²⁴	Canada	20	9	11	Mean 51	-	Psoriasis and psoriasis arthritis	a) BSA b) PASI score c) LS-PGA d) NPF-PS e) mNAPSI f) NN g) PGA-PS	9	ICC	a) 0.65 (0.47-0.82) b) 0.74 (0.58-0.87) c) 0.73 (0.56-0.86) d) 0.66 (0.47-0.82) e) 0.78 (0.63-0.89) f) 0.77 (0.61-0.89) g) 0.66 (0.48-0.82)
Hanifin <i>et al.</i> , 2001 ¹⁷	USA	Cohort 1: 10 Cohort 2: 20			a) Cohort 1: adults, mean 43; b) Cohort 2: children, mean 4.3	-	Atopic dermatitis	EASI	15	Weighted kappa	a) Cohort 1: Day 1: 0.76; Day 2: 0.82 b) Cohort 2: Day 1: 0.66; Day 2: 0.59

Krathen <i>et al.</i> , 2008 ²³	USA	13	12	1	Mean 41.4	African American 10, White 3	Cutaneous lupus Erythematosus	a) CLASI activity b) CLASI damage	5	ICC	a) 0.92 (0.86-0.99) b) 0.82 (0.68-0.96)
Kushner <i>et al.</i> , 2019 ¹⁴	USA	11	10	1	Mean 16.5 (SD±4.2)	Hispanic/Latino 4, African American 3, White 2, Mixed 2	Cutaneous lupus Erythematosus	a) CLASI activity b) CLASI damage c) PGA activity d) PGA damage	6	ICC	a) 0.95 (0.90-0.99) b) 0.76 (0.56-0.92) c) 0.77 (0.57-0.92) d) 0.50 (0.25-0.78)
Pandya <i>et al.</i> , 2011 ¹⁹	USA	21	21			Hispanic 18, Asian 2, African 1	Melasma	MASI	6	Weighted kappa	Day 1: 0.68; Day 2: 0.70
Rosenbach <i>et al.</i> , 2013 ²⁰	USA	11	9	2	Mean 52.5 (SD±7.6)	African American 9, White 2	Cutaneous sarcoidosis	a) CSAMI activity b) Inflammation c) Induration/depression d) Surface changes e) Area f) CSAMI damage g) Modified facial SASI h) Erythema i) Induration j) Desquamation k) Area l) PGA	8	ICC	a) 0.82 (0.66-0.94) b) 0.85 (0.70-0.95) c) 0.59 (0.36-0.83) d) 0.74 (0.53-0.90) e) 0.76 (0.57-0.91) f) 0.42 (0.21-0.72) g) 0.40 (0.17-0.72) h) 0.55 (0.43-0.68) i) 0.57 (0.45-0.69) j) 0.54 (0.42-0.66) k) 0.81 (0.74-0.88) l) 0.40 (0.18-0.70)
Tan <i>et al.</i> , 2006 ²⁶	Canada	6	4	2	18-24	-	Acne	a) Inflammatory acne lesion count b) Non-inflammatory acne lesion count c) Global severity assessment	6	ICC	a) 0.84 b) 0.17 c) 0.71
Tan <i>et al.</i> , 2014 ²⁷	USA	28	25	3	45.0 (SD±10.1) range: 18-61	All Caucasians (Fitzpatrick skin type I-III)	Facial erythema (rosacea)	CEA	12	ICC	Overall session 1: 0.60; Overall session 2: 0.58
Tiao <i>et al.</i> , 2017 ²⁵	USA	15	10	5	Mean 56.4 (SD±16.0)	White 12, African American 2, Asian 1, Hispanic/Latino 1	Dermatomy ositis	a) CDASI activity b) CDASI damage c) PGA activity d) PGA damage	5	ICC	a) 0.73 (0.55-0.9) b) 0.86 (0.76-0.96) c) 0.61 (0.40-0.83) d) 0.58 (0.46-0.81)
Tiao <i>et al.</i> , 2017 ¹⁵	USA	14	8	6	Mean 12.6 (SD±5.6)	White 12, African American 1, Hispanic/Latino 1, Mixed 1	Dermatomy ositis	a) CDASI activity b) CDASI damage c) CAT-BM activity d) CAT-BM damage e) Skin PGA activity f) Skin PGA damage	5	ICC	a) 0.52 (0.44-0.61) b) 0.59 (0.51-0.67) c) 0.58 (0.50-0.66) d) 0.49 (0.41-0.58) e) 0.64 (0.41-0.58) f) 0.71 (0.65-0.83)
Thorlacius <i>et al.</i> , 2019 ²⁸	Denmark	24	15	9	42.7 (SD±11.9)	Most patients were Caucasians	Hidradenitis suppurativa	<u>Classification instruments:</u> a) Hurley axillae b) Hurley gluteal c) Hurley groin d) IHS4 classification	12	ICC	<u>Classification instruments:</u> a) 0.72 (0.63-0.81) b) 0.72 (0.62-0.80) c) 0.55 (0.44-0.67) d) 0.54 (0.39-0.71)

							e) Hurley staging refined <u>Classification/OMIs:</u> f) Physician global VAS g) HS-PGA <u>Physical signs OMIs:</u> h) MSS i) IHS4 total j) HiSCR baseline <u>Composite OMIs:</u> k) HSSI l) AISI <u>Lesion counts:</u> m) Total tunnel n) Draining tunnel o) Total nodules p) Inflammatory nodules q) Non-draining tunnel r) Non-inflammatory nodules s) Abscesses			e) 0.51 (0.35-0.68) <u>Classification/OMIs:</u> f) 0.72 (0.59-0.84) g) 0.64 (0.50-0.79) <u>Physical signs OMIs:</u> h) 0.66 (0.52-0.80) i) 0.47 (0.32-0.65) j) 0.44 (0.29-0.63) <u>Composite OMIs:</u> k) 0.78 (0.66-0.88) l) 0.40 (0.25-0.59) <u>Lesion counts:</u> m) 0.55 (0.41-0.72) n) 0.46 (0.31-0.64) o) 0.44 (0.29-0.59) p) 0.40 (0.26-0.59) q) 0.28 (0.17-0.47) r) 0.20 (0.10-0.36) s) 0.07 (0.02-0.17)	
Yeung <i>et al.</i> , 2015 ²²	USA	13	10	3	Mean 51.2 (SD±10.6)	African American 12, White 1	Cutaneous sarcoidosis	a) CSAMI Activity scale b) Inflammation c) Induration/depression d) Surface changes e) Area f) CSAMI Damage g) Modified facial SASI h) PGA	4	ICC	a) 0.74 (0.49-0.90) b) 0.76 (0.55-0.90) c) 0.64 (0.31-0.86) d) 0.77 (0.57-0.91) e) 0.58 (0.31-0.82) f) 0.37 (0.12-0.68) g) 0.81 (0.63-0.93) h) 0.65 (0.41-0.86)
Zhao <i>et al.</i> , 2015 ¹⁸	Australia	12	9	3	3 adults and 9 children; Mean 13 (SD±6.4)	Caucasian 8, Asian 3, African 1	Atopic dermatitis	a) EASI b) SASSAD c) TIS d) oSCORAD	5	ICC	a) 0.73 (0.50-0.90) b) 0.68 (0.44-0.88) c) 0.50 (0.23-0.79) d) 0.50 (0.23-0.79)
Zhao <i>et al.</i> , 2017 ¹⁶	Australia	25	7	18	Paediatric: 13; Adults: 12	11 lighter skin and 14 SOC; Caucasian 9, Asian 12, Indian 3, African 1; melanin index <150 4, 150-199 7, 200-400 11, >400 3	Atopic dermatitis	a) EASI lighter skin b) EASI SOC c) IGA lighter skin d) IGA SOC e) oSCORAD lighter skin f) oSCORAD SOC g) Greyscale SOC	5	ICC	a) 0.83 (0.66-0.94) b) 0.77 (0.60-0.91) c) 0.80 (0.62-0.93) d) 0.70 (0.49-0.87) e) 0.68 (0.44-0.88) f) 0.74 (0.54-0.89) g) 0.78 (0.60-0.91)

AISI, Acne Inversa Severity Index; BSA, body surface area involved by psoriasis; CAT-BM, cutaneous assessment tool-binary method; CDASI, Cutaneous Disease Area and Severity Index; CEA, clinician erythema assessment; CI, confidence interval; CLASI, Cutaneous Lupus Erythematosus Disease Area and Severity Index; CSAMI, Cutaneous Sarcoidosis Activity and Morphology Instrument; EASI, Eczema Area and Severity Index; F, female; HEES, Hand Eczema Extent Score; HiSCR, hidradenitis suppurativa clinical response; HSSI, Hidradenitis

Suppurative Severity Index; HS-PGA, hidradenitis suppurativa physician's global assessment; ICC, intraclass correlation coefficient; IGA, investigator's global assessment; IHS4, international hidradenitis suppurativa severity scoring system; LS-PGA, lattice system physician's global assessment; M, male; MASI, Melasma Area and Severity Index; MDs, Medical Doctors; mNAPSI, modified Nail Psoriasis Severity Index; MSS, modified sartorius score; NN, number of nails with nail changes; No., number; NPF-PS, National Psoriasis Foundation Psoriasis Score; OMI, outcome measurement instrument; oSCORAD, objective scoring atopic dermatitis; PASI, Psoriasis Area and Severity Index; PGA, physician's global assessment; PGA-PS, physician's global assessment of psoriasis activity; SASI, Sarcoidosis Activity and Severity Index; SASSAD, six area, six sign atopic dermatitis; SOC, skin of color; SD, standard deviation; TIS, three items severity score; VAS, Visual Analogue Scale.