Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma

Rajpara SM, Botello AP, Townend J, Ormerod AD

CRD summary
The review compared performance of different dermoscopic algorithms with each other and with digital dermoscopy/artificial intelligence for diagnosing melanoma and found that dermoscopy and artificial intelligence were useful tools, with no significant differences between them or between different dermoscopy algorithms. The authors’ conclusions reflected the data presented and are likely to be reliable.

Authors’ objectives
To assess accuracy of dermoscopy and digital dermoscopy/artificial intelligence in diagnosing melanoma and to compare diagnostic performance of different dermoscopic algorithms with each other and with digital dermoscopy/artificial intelligence.

Searching
MEDLINE and EMBASE were searched from inception to April 2007. Search terms were reported. ACP Journal Club, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews, DARE were searched (no dates reported). Ovid search engine and bibliographies of two systematic reviews were searched (no dates reported). No language restrictions were applied.

Study selection
Prospective and retrospective studies that used dermoscopy or digital dermoscopy/artificial intelligence to diagnose melanocytic skin lesions in any population in any health care setting were eligible for inclusion. Included studies were required to use histopathology of excised melanocytic skin lesions as the reference standard to confirm diagnosis. Studies that used any published algorithm for dermoscopy and artificial intelligence were considered. Methodological quality scores were used to exclude studies.

Most included studies were retrospective. The most common algorithm used in dermoscopy studies was pattern analysis. Most studies reported Breslow thickness (range 0.15mm to 4.5mm); other participant characteristics were poorly reported.

Two reviewers independently assessed studies for inclusion.

Assessment of study quality
Methodological quality of all articles retrieved for inclusion screening was assessed using a two-section 15-point quality assessment form adapted from the National Institute for Health and Clinical Excellence (NICE) guidelines on methodology checklist for diagnostic studies and the Standards for Reporting of Diagnosis Accuracy (STARD) initiative. The first part of the form assessed: description of the test; appropriateness of the reference standard; selection of participants; blinding; completeness of follow-up; and whether or not a pre-diagnosis was made and reported. The second section was considered to be specific to this study and assessed: clear statement of the study aim; reporting of inclusion criteria; definition of a diagnostic threshold; use of a trained expert; characteristics of the study population; reporting of 2x2 data; estimates of diagnostic accuracy; handling of indeterminate results; and clinical applicability of findings. Included studies had to fulfill at least four criteria from the first section and all criteria (except demographic data) from the second section.

Two reviewers independently assessed validity. Disagreements were resolved by a third reviewer.

Data extraction
Data were extracted on details of algorithms evaluated, numbers of patients and lesions, and estimates of sensitivity and
specificity, with 95% confidence intervals (CIs).

Data were extracted using a table adapted from NICE guidelines. The authors did not state how many reviewers performed the data extraction.

**Methods of synthesis**

Summary Receiver Operating Characteristic (sROC) curves and pooled estimates of sensitivities, specificities, positive and negative likelihood ratios (LRs) and diagnostic odds ratios, with 95% CIs, were calculated using Meta-DiSc software version 1.4.

sROC curves were used to check for threshold effect. Meta-DiSc used the Moses and Littenberg model to estimate sROC curves. $X^2$ and $I^2$ tests were used to assess between-study heterogeneity.

Sensitivities and specificities of different tests were compared using the binomial test and logs of likelihood ratios. Diagnostic odds ratios were compared using t-tests. Corrections for multiple comparisons were carried out using the Bonferroni method. For comparisons of dermoscopy algorithms (36 comparisons) $p<0.00139$ was taken as significant. For comparisons of individual algorithms with artificial intelligence (nine comparisons) $p<0.0056$ was taken as significant. For comparison of dermoscopy as a whole with artificial intelligence (one comparison) $p<0.05$ was considered significant.

Funnel plots were used to assess publication bias.

**Results of the review**

Thirty studies that assessed 9,784 lesions were included in the review. Numbers of participants were not reported for most studies. Eighteen studies assessed dermoscopy, seven studies assessed digital dermoscopy/artificial intelligence and five studies evaluated both. The most common algorithm used in dermoscopy studies was pattern analysis (10 studies).

Studies included in the review examined 8,045 lesions assessed using dermoscopy and 2,420 lesions assessed using artificial intelligence.

The pooled estimate of sensitivity for dermoscopy (any algorithm) was 0.88 (95% CI 0.87 to 0.89) and the pooled estimate of specificity was 0.86 (95% CI 0.85 to 0.86) from 23 studies (30 data sets).

The pooled estimate of sensitivity for artificial intelligence was 0.91 (95% CI 0.88 to 0.93) and the pooled estimate of specificity was 0.79 (95% CI 0.77 to 0.81) from 12 studies.

Sensitivity and specificity were calculated per lesion.

Pooled sensitivity for artificial intelligence was slightly higher than for dermoscopy ($p=0.076$) (not significant) and pooled specificity for dermoscopy was significantly higher than artificial intelligence ($p<0.001$). For individual dermoscopy algorithms, pattern analysis had significantly lower sensitivity than seven features for melanoma (7FFM), Menzies score and artificial intelligence. Pattern analysis showed significantly higher specificity than ABCD, ABCDE, seven-point checklist and artificial intelligence. ABCD had significantly lower specificity than 7FFM and three-point checklist. Artificial intelligence had significantly lower specificity than 7FFM.

The pooled estimates of diagnostic odds ratio were 51.52 (95% CI 38.02 to 69.82) for dermoscopy and 57.83 (95% CI 26.95 to 124.08) for artificial intelligence (no significant difference). There were no significance differences in diagnostic odds ratio between diagnostic algorithms.

Funnel plots showed slight evidence of publication bias, which was considered unlikely to have a major effect on results of the meta-analyses.
Authors' conclusions
The authors concluded that there was convincing evidence for the usefulness of dermoscopy and artificial intelligence as diagnostic tools for melanoma. There were no significant differences in overall diagnostic performance between dermoscopy and artificial intelligence, and between the various dermoscopy algorithms.

CRD commentary
The review addressed a clearly stated research question defined by appropriate inclusion criteria. A number of bibliographic databases and other sources were searched for relevant studies. No language restrictions were applied. Publication bias was assessed. Measures to minimise error and/or bias in the review process were reported for study selection and assessment of methodological quality, but it was unclear whether similar measures were applied to data extraction. Assessment of methodological quality formed part of the inclusion criteria of the review, but the full results of quality assessment for included studies were not reported. The methods used to generate pooled estimates of diagnostic performance were broadly appropriate and heterogeneity was considered, but bivariate/hierarchical models for sROC curves are now generally recommended over the model used. However, the dermoscopy data set contained multiple data points from the same studies and it appeared that this was not considered in the analysis. Similarly, in a per-lesion analysis, diagnostic status of lesions within individual patients was unlikely to be independent and this issue did not appear to have been considered in the analyses/interpretation. In general, the authors' conclusions were broadly reflective of data presented and are likely to be reliable.

Implications of the review for practice and research
Practice: The authors made no recommendations for practice.

Research: The authors stated that a study that compared dermoscopy and artificial intelligence for melanoma detection under real life clinical conditions would provide better evidence about the true usefulness of dermoscopy and artificial intelligence for melanoma diagnosis. Such a study should assess sensitivity and specificity of artificial intelligence in differentiating melanoma not only from preselected melanocytic lesions but also from other common benign lesions. Further studies were recommended on integration of artificial intelligence algorithms to reach higher sensitivity and combined performance of artificial intelligence and dermoscopy for diagnosing different lesions. An economic evaluation was recommended.

Funding
Not stated.

Bibliographic details

PubMedID
19302072

DOI
10.1111/j.1365-2133.2009.09093.x

Original Paper URL
http://onlinelibrary.wiley.com/journal/122267638/abstract

Indexing Status
Subject indexing assigned by NLM

MeSH
Algorithms; Dermoscopy /methods /standards; Humans; Image Processing, Computer-Assisted /standards; Melanoma /diagnosis /pathology; Sensitivity and Specificity; Skin Neoplasms /diagnosis /pathology

AccessionNumber
Record Status
This is a critical abstract of a systematic review that meets the criteria for inclusion on DARE. Each critical abstract contains a brief summary of the review methods, results and conclusions followed by a detailed critical assessment on the reliability of the review and the conclusions drawn.