Author/s:
Jones, OT;Matin, RN;van der Schaar, M;Prathivadi Bhayankaram, K;Ranmuthu, CKI;Islam, MS;Behiyat, D;Boscott, R;Calanzani, N;Emery, J;Williams, HC;Walter, FM

Title:
Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review

Date:
2022-06-01

Citation:
Jones, O. T., Matin, R. N., van der Schaar, M., Prathivadi Bhayankaram, K., Ranmuthu, C. K. I., Islam, M. S., Behiyat, D., Boscott, R., Calanzani, N., Emery, J., Williams, H. C. & Walter, F. M. (2022). Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. Lancet Digital Health, 4 (6), pp.e466-e476. https://doi.org/10.1016/S2589-7500(22)00023-1.

Persistent Link:
https://hdl.handle.net/11343/322506

# Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review

*O T Jones, R N Matin, M van der Schaar, K Prathivadi Bhayankaram, C K I Ranmuthu, M S Islam, D Behiyat, R Boscott, N Calanzani, J Emery, H C Williams, F M Walter*

**Skin cancers occur commonly worldwide. The prognosis and disease burden are highly dependent on the cancer type and disease stage at diagnosis. We systematically reviewed studies on artificial intelligence and machine learning (AI/ML) algorithms that aim to facilitate the early diagnosis of skin cancers, focusing on their application in primary and community care settings. We searched MEDLINE, Embase, Scopus, and Web of Science (from Jan 1, 2000, to Aug 9, 2021) for all studies providing evidence on applying AI/ML algorithms to the early diagnosis of skin cancer, including all study designs and languages. The primary outcome was diagnostic accuracy of the algorithms for skin cancers. The secondary outcomes included an overview of AI/ML methods, evaluation approaches, cost-effectiveness, and acceptability to patients and clinicians. We identified 14 224 studies. Only two studies used data from clinical settings with a low prevalence of skin cancers. We reported data from all 272 studies that could be relevant in primary care. The primary outcomes showed reasonable mean diagnostic accuracy for melanoma (89·5% [range 59·7–100%]), squamous cell carcinoma (85·3% [71·0–97·8%]), and basal cell carcinoma (87·6% [70·0–99·7%]). The secondary outcomes showed a heterogeneity of AI/ML methods and study designs, with high amounts of incomplete reporting (eg, patient demographics and methods of data collection). Few studies used data on populations with a low prevalence of skin cancers to train and test their algorithms; therefore, the widespread adoption into community and primary care practice cannot currently be recommended until efficacy in these populations is shown. We did not identify any health economic, patient, or clinician acceptability data for any of the included studies. We propose a methodological checklist for use in the development of new AI/ML algorithms to detect skin cancer, to facilitate their design, evaluation, and implementation.**

*Lancet Digit Health* 2022; 4: e466–76

**Department of Public Health & Primary Care** (O T Jones MPhil, N Calanzani PhD, Prof J Emery DPhil, Prof F M Walter MD) **and Department of Applied Mathematics and Theoretical Physics** (Prof M van der Schaar PhD), **University of Cambridge, Cambridge, UK; Department of Dermatology, Churchill Hospital, Oxford, UK** (R N Matin PhD); **School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK** (K Prathivadi Bhayankaram BA, C K I Ranmuthu MBBChir, M S Islam BA, D Behiyat BA, R Boscott BA); **Centre for Cancer Research and Department of General Practice, University of Melbourne, Melbourne, VIC, Australia** (Prof J Emery, Prof F M Walter); **Centre of Evidence Based Dermatology, School of Clinical Medicine, University of Nottingham, Nottingham, UK** (Prof H C Williams DSc); **Wolfson Institute of Population Health, Faculty of Medicine and Dentistry, Queen Mary University of London, London, UK** (Prof F M Walter)

Correspondence to:
Dr Owain T Jones, Department of Public Health & Primary Care, University of Cambridge, Cambridge, Cambridge CB1 8RN, UK
**otj24@medschl.cam.ac.uk**

## Introduction

Melanoma is a serious skin cancer and has a rapidly rising incidence in many populations.[1] The incidence in White populations has increased by 3–5% per annum since the mid-20th century, with rates currently at 20–60 cases per 100 000 people per annum.[1] More commonly occurring non-melanoma skin cancers include squamous cell carcinoma and basal cell carcinoma, and together are increasingly referred to as keratinocyte carcinomas.[2] Nearly 152 000 new cases of keratinocyte carcinoma were diagnosed in the UK in 2017,[3,4] and the age-standardised incidence rates in Germany in 2017 ranged from 147·8 to 391·4 per 100 000.[1] Similar to melanoma, the incidence of keratinocyte carcinomas is rising steeply.[1,3] Nonetheless, an earlier diagnosis of skin cancer leads to better outcomes. For example, the 1-year survival rate for melanoma when diagnosed at American Joint Cancer Committee stage 1 is 100%, compared with only 53% when diagnosed at American Joint Cancer Committee stage 4.[5]

In gatekeeper health-care systems such as that in the UK, most people first present with concerns about a skin lesion in primary care,[6] where general practitioners need to be able to distinguish all suspicious lesions requiring a biopsy (including melanoma) from common benign lesions (ie, a diagnostic triage). A more accurate assessment of suspicious skin lesions by general practitioners could lead to fewer onward referrals and unnecessary biopsies, improving the patient experience and reducing demand and costs to dermatology specialist services. Furthermore, increased accurate assessments could potentially lead to an earlier diagnosis of any skin cancer, thereby improving patient outcomes.

There is accumulating evidence that artificial intelligence and machine learning (AI/ML) can assist clinicians to make better clinical decisions, or even replace human judgement. Studies have shown that AI/ML algorithms can perform on par with or better than consultant dermatologists,[7–9] and that AI/ML algorithms can assist clinicians in the diagnosis of skin cancers.[10–12] If these findings could be replicated in primary care settings where there is a low prevalence of skin cancer, AI/ML algorithms could have a substantial effect on diagnostic services. There are a few existing market-approved technologies aimed at the diagnosis of skin cancer, but in the UK there are no AI/ML algorithms currently in routine clinical use for detecting or triaging suspicious skin lesions. There are several possible reasons for this absence of AI/ML use, but of these the most notable one is that there is a need for robust evidence on the diagnostic accuracy of AI/ML algorithms in relevant populations, to support the decision making of policy makers and commissioners on the appropriate implementation of AI/ML in clinical practice.[13–17]

The CanTest framework[17] (figure 1) was developed in 2019 for the evaluation of diagnostic tests and approaches, building on a systematic review of existing frameworks and the work of a consensus group of international

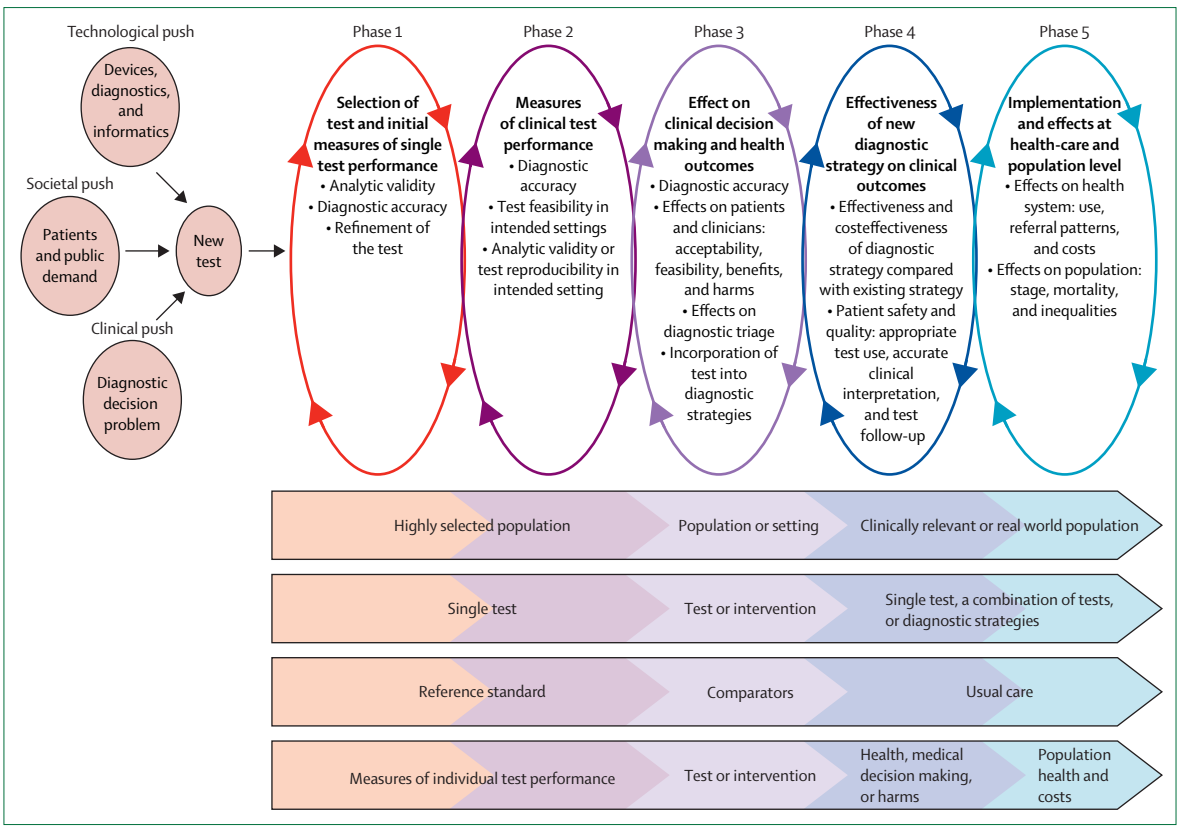


***Figure 1:*** **The CanTest Framework**
Factors driving the development of a new diagnostic test are shown on the far left-hand side. The phases of development of the diagnostic test, and the notable features of these phases are then shown from left to right. Shown at the bottom is a summary of the differences between early and late phase development studies in terms of the setting of the studies, the intervention design, the comparators used, and the outcomes assessed. Adapted from Walter et al.[17]

experts. This framework establishes the developmental phases required to ensure new diagnostic tests or technologies are fit for purpose when introduced into clinical practice, and provides a roadmap for developers and policy makers to bridge the gap from the development of a diagnostic test or technology to its successful implementation. We used this framework to guide the assessment of the studies identified in this Review, including assessing their eligibility and phase of development.

The aim of this Systematic Review was to evaluate the accuracy and safety of AI/ML technologies that could facilitate the early detection of skin cancer in primary and community care settings. We deliberately focused this Review on the applicability of diagnostic algorithms to primary and community care (hereafter referred to as primary care), where the prevalence of skin cancer is lower than in specialist clinics. This setting might be where AI/ML technologies can have the greatest benefit, because it is where the initial assessment of most suspicious skin lesions takes place. We analysed the quality of the evidence, the phase of development the AI/ML technologies had reached, the evidence gaps, and the potential for use in primary care. This Review complements our previous Review on the use of AI/ML techniques applied to electronic health record data in primary care to facilitate the earlier diagnosis of cancer.[18]

## Methods

### Search strategy and selection criteria

This Review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines,[19] and the protocol was registered with PROSPERO before conducting the Review (CRD42020176674).[20]

Preliminary searches identified few AI/ML algorithms that were developed or tested in primary care settings. Therefore, we broadened our search strategy to avoid missing relevant studies, including studies using data from secondary care, patients that have already been assessed and referred, or other populations with a high prevalence of skin cancer. All primary research articles published in peer-reviewed journals, without language restrictions, from Jan 1, 2000, to Aug 9, 2021, were considered for inclusion. Studies were included if they provided evidence around the accuracy, utility, acceptability, or cost-effectiveness of applying AI/ML algorithms to the triage or diagnosis of skin cancer. All types

| | Roffman and colleagues (2018)[28] | Udrea and colleagues (2020)[29] |
|---|---|---|
| Origin of the study authors | The USA | Romania, Netherlands, and the USA |
| Skin lesions included | NMSC | Melanomas, keratinocyte carcinomas, and benign lesions |
| Method of AI/ML | Artificial neural network | Support vector machine |
| Control test | Reported NMSC | Histopathology result, expert opinion, and previously developed AI/ML methods |
| Origin of data | Non-referred population | Mixed (referred and non-referred populations included) |
| Database(s) used | Data from the National Health Interview Survey | Datasets from previous studies at Munich University Hospital and Eindhoven hospital, and SkinVision user database of smartphone images |
| Type of skin lesion images | No images | All images taken with mobile phones; SkinVision user database images taken by patients |
| Disease-positive population | 2506 people with NMSC | Approximately 65 937 people with high-risk lesions from the SkinVision smartphone dataset, 178 with confirmed melanoma; 40 people with melanoma in the Munich dataset; and 155 people with skin cancers or pre-cancer in the Eindhoven database |
| Disease-negative population | 460 574 people who had never had cancer | 65 936 people with low-risk lesions from the smartphone dataset |
| Training set | 324 156 people | 131 873 people |
| Test set | 138 924 people | Sensitivity based on 285 people (with and without skin cancers), specificity based on 6000 cases (images from all datasets) |
| Validation set | 28 058 respondents (2016 survey data) | .. |
| Prospective or clinical evaluation | .. | .. |
| Outcome measures | | |
| Sensitivity (recall) | 0·862 | 0·95 overall for malignancy, 0·928 for melanoma, 0·973 for keratinocyte carcinoma |
| Specificity | 0·627 | 0·783 overall |
| Area under the receiver operating characteristic curve | 0·81 | .. |

We identified no data on positive predictive value (precision), negative predictive value, accuracy index, cost, acceptability, health economic analysis, or implementation barriers. We also identified no data from prospective or clinical evaluations of these algorithms. AI/ML=artificial intelligence and machine learning. NMSC=non-melanoma skin cancer. *Accuracy index=(true positives + true negatives)/(true positives + true negatives + false positives + false negatives).

***Table 1:*** **Outcomes measures for the two studies that included data from unreferred or low risk populations**

of study design were included because we anticipated that there would be a range of methods used depending on the phase of development of the diagnostic test. We evaluated evidence from any health-care system and assessed the applicability of the evidence in primary care populations. Unlike the Cochrane review of smartphone apps to assess the risk of skin cancer,[21] our Review focused on evaluating AI/ML-based algorithms that could specifically be used in primary care settings to facilitate the early diagnosis of skin cancers.

Studies that only described the development of an AI/ML algorithm without undertaking any testing or evaluation, or that did not incorporate an element of ML (ie, with training and testing or validation steps), that used AI/ML for biomarker discovery alone, and studies that had sample sizes of less than 50 people with cancer and 50 people in the control group were excluded. We chose to start our search from the year 2000, because this was when the earliest research describing the new extensions and developments of ML techniques emerged.[22] Although ML techniques and neural networks have been described since the 1960s,[23,24] they were initially limited by computing power and data availability, and their clinical applications did not begin to appear until the 21st century.

We searched MEDLINE, Embase, Scopus, and Web of Science bibliographic databases, using keywords relating to AI/ML, skin cancer, and early detection (appendix p 1). The study authors were contacted via email where required. Where studies were not published in English, we identified suitably qualified native speakers to help assess these studies. In addition to our published protocol, we extended these systematic searches through a search of Google Scholar using the same keywords for all databases, searching the first 100 studies ordered by relevance for studies that met our inclusion criteria. Many commercially developed AI/ML technologies do not have published data in academic journals; therefore, we used scoping review methods to identify currently available AI/ML technologies that might not have been identified through systematic searches. This method included a manual search of commercial research archives and networks (eg, Arxiv, Google, Microsoft, IBM, Apple, NHS digital, and International Skin Imaging Collaboration [ISIC], a structured Google search (using combinations of the keywords: "skin cancer", "melanoma", "artificial intelligence", "machine learning", and "diagnosis"), and review of websites including skin diagnostic tools.

After duplicate removal, one author (OTJ) screened titles and abstracts to identify studies that met the inclusion criteria. Of the titles and abstracts, 1838 (16%) of 11 296 were checked by two members of the research team (Smiji Saji and NC); inter-assessor reliability was good at 1769 (96%) of 1838. Any disagreements were discussed by the research team (OTJ, Smiji Saji, KPB, CKIR, NC, and FMW) and a consensus was reached. Three reviewers (OTJ, 355 [56%] of 638 articles; KPB, 337 [53%] of 638 articles; and

For **Arxiv** see https://arxiv.org/

For the **Google AI research** see https://ai.google/research/

For **Microsoft research** see https://www.microsoft.com/en-us/research/

For **IBM research** see https://www.research.ibm.com/artificial-intelligence/

For the **Apple machine learning network** see https://machinelearning.apple.com

For the **NHS digital website** see https://digital.nhs.uk

For the **ISIC website** see https://www.isic-archive.com/

See **Online** for appendix

| | Sensitivity | Specificity | Positive predictive value | Negative predictive value | Area under the receiver operating characteristic curve | Accuracy* | F1-score† |
|---|---|---|---|---|---|---|---|
| **Melanoma (197 studies provided outcome measures for melanoma alone, 2000–21)** | | | | | | | |
| Mean (95% CI) | 0·842 (0·816–0·868) | 0·891 (0·871–0·910) | 0·814 (0·769–0·859) | 0·929 (0·909–0·949) | 0·898 (0·882–0·915) | 89·5% (88·2–90·8%) | 0·807 (0·732–0·882) |
| Median (IQR) | 0·894 (0·792–0·950) | 0·920 (0·850–0·965) | 0·846 (0·720–0·955) | 0·930 (0·900–0·960) | 0·910 (0·849–0·950) | 91·3% (86·0–95·0%) | 0·850 (0·748–0·960) |
| Range | 0·13–1·00 | 0·36–1·00 | 0·280–1·000 | 0·86–1·00 | 0·71–1·00 | 59·7–100% | 0·280–0·975 |
| Number of studies | 146 | 127 | 49 | 17 | 64 | 141 | 24 |
| **Squamous cell carcinoma (ten studies provided outcome measures for squamous cell carcinoma alone, 2015–20)** | | | | | | | |
| Mean (95% CI) | 0·603 (0·396–0·810) | 0·933 (0·865–1·000) | 0·415 (0·247–0·582) | 0·951 (0·875–1·000) | 0·875 (0·777–0·973) | 85·3% (77·3–93·3%) | ... |
| Median (IQR) | 0·58 (0·394–0·799) | 0·965 (0·928–0·979) | 0·415 (0·372–0·457) | 0·951 (0·931–0·970) | 0·906 (0·859–0·922) | 86·0% (77·5–93·8%) | ... |
| Range | 0·256–1·000 | 0·800–0·995 | 0·329–0·500 | 0·912–0·989 | 0·730–0·958 | 71·0–97·8% | ... |
| Number of studies | 7 | 5 | 2 | 2 | 4 | 4 | 0 |
| **Basal cell carcinoma (29 studies provided outcome measures for basal cell carcinoma alone, 2012–20)** | | | | | | | |
| Mean (95% CI) | 0·837 (0·792–0·883) | 0·887 (0·783–0·990) | 0·834 (0·767–0·902) | 0·896 (0·743–1·000) | 0·923 (0·879–0·967) | 87·6% (80·7–94·6%) | 0·846 (0·783–0·909) |
| Median (IQR) | 0·880 (0·766–0·914) | 0·938 (0·893–0·988) | 0·877 (0·785–0·930) | 0·978 (0·939–0·988) | 0·946 (0·912–0·970) | 91·1% (77·5–97·5%) | 0·875 (0·845–0·913) |
| Range | 0·580–0·996 | 0·342–1·000 | 0·541–0·986 | 0·510–0·992 | 0·76–0·99 | 70·0–99·7% | 0·61–0·93 |
| Number of studies | 26 | 12 | 17 | 6 | 10 | 11 | 10 |
| **Benign versus malignant (33 studies involved more than two lesion types and provided outcome measures for benign vs malignant, 2018–20)** | | | | | | | |
| Mean (95% CI) | 0·870 (0·843–0·897) | 0·864 (0·820–0·908) | 0·859 (0·804–0·914) | 0·892 (0·832–0·951) | 0·883 (0·840–0·926) | 88·8% (86·3–91·3%) | 0·888 (0·817–0·959) |
| Median (IQR) | 0·851 (0·828–0·928) | 0·892 (0·842–0·923) | 0·871 (0·834–0·906) | 0·902 (0·874–0·939) | 0·895 (0·855–0·934) | 89·5% (83·8–93·1%) | 0·833 (0·830–0·957) |
| Range | 0·720–0·995 | 0·535–0·981 | 0·582–0·994 | 0·761–0·970 | 0·742–0·975 | 75·9–99·5% | 0·826–0·994 |
| Number of studies | 28 | 23 | 14 | 6 | 12 | 24 | 5 |

*Accuracy index=(true positives + true negatives)/(true positives + true negatives + false positives + false negatives). †F1 score=2 × (positive predictive value × sensitivity)/(positive predictive value + sensitivity).

*Table 2:* **Outcome measures reported in the included studies for melanoma, squamous cell carcinoma, basal cell carcinoma, and for studies that assessed the classification of benign versus malignant categories (in studies that included more than two lesion types; n=272)**

NC, 56 [9%] of 638 articles) independently assessed full-text articles for inclusion in the Review; inter-assessor reliability was good at 91 (83%) of 110 full text papers agreed on by more than one author. Any disagreements were resolved by consensus-based decisions.

### Data analysis

Data extraction was undertaken independently by at least two reviewers (any two of OTJ, KPB, CKIR, MSI, DB, RB, and NC) into a predesigned data extraction spreadsheet. Where studies stated that they had used a specific database to obtain the data, but not included the specific details of the included lesion types, we sought information from the database website wherever possible. The research team (OTJ, KPB, CKIR, MSI, DB, RB, NC, and FMW) resolved differences in data extraction through consensus agreement and a clinician group (FMW, RNM, and OTJ) reached consensus regarding clinical questions arising from studies. One author (OTJ) amalgamated the data extraction spreadsheets and summarised the data.

The primary outcome was diagnostic accuracy of the AI/ML algorithms for melanomas and keratinocyte carcinomas; the main summary measures collected included sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic (AUROC) curve. The secondary outcomes included the type of AI/ML used, the type and external applicability of the data used to develop the algorithm (including the clinical origin of the data and the prevalence of skin cancer in the data), and the methods of algorithm evaluation. We aimed to collect data, where available, on cost-effectiveness and patient or clinician acceptability of the algorithm.

A risk-of-bias assessment was undertaken for all included studies using the QUADAS-2 critical appraisal tool,[25] with 190 (70%) of 272 papers assessed by at least two independent researchers (any two of OTJ, KPB, CKIR, MSI, DB, and RB). This tool is more discriminative of the studies identified than the Joanna Briggs critical appraisal tools that we specified in our protocol.[26] In addition to the standard QUADAS-2 critical appraisal tool, we included an overall assessment of whether each paper was at a high, medium, or low risk of bias. Any disagreements in assessment were resolved by consensus among the research team.

Studies identified were heterogeneous, using different AI/ML techniques and evaluating the algorithms in various ways using different outcome measures. A meta-analysis was therefore not considered to be meaningful, and we instead used a narrative synthesis approach, following established guidance on the method of this approach.[27] We also conducted simple descriptive statistical analyses to summarise the diagnostic accuracy data where available, to provide an overview of the quantitative outcomes. Microsoft excel version 16.59 was used for analyses.

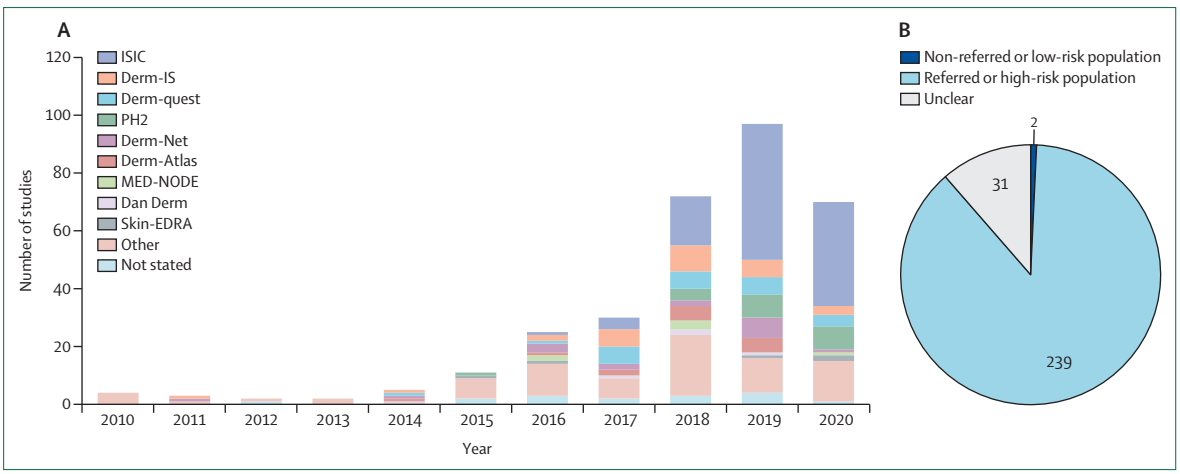


***Figure 2:*** **Number of studies from each of the image databases used and the type of population in the included studies for 2010–20 (N=272)**
(A) The number of included studies using each image database in each year from 2010 to 2020. (B) The proportion of included studies using data from different clinical settings. Because 2000–09 had less than five studies published per year, these years have been excluded from this figure. The cumulative totals for 2000–09 were as follows: two published studies from the Derm-IS database, ten from the Derm-Net database, and ten from other databases. The totals from 2021 were also excluded from this figure because data were not available for the full year so a visual comparison with other years would not be accurate. Totals for Jan 1, 2021, to Aug 9, 2021, were as follows: 53 published studies from the ISIC database, one from the Derm-Quest database, one from the Derm-IS database, one from the Derm-Net database, six from the PH2 database, eight from other databases, and one study did not state which database it was from. Note: many studies used multiple databases in sub-studies within the overall article. The database bar shown represents the proportion of studies that used a particular database within the total number of studies published that year. ISIC=International Skin Imaging Collaboration.

## Results

The searches identified 14 224 studies, with 14 additional studies identified from other sources (ie, identified from references in the included studies; appendix p 24). After removing duplicates, the titles and abstracts were screened for 11 296 studies, with subsequent full-text screening of 638 studies. Only two studies used data originating from unreferred populations with a low prevalence of skin cancer to develop and test their AI/ML algorithms (table 1). Roffman and colleagues[28] used data from the National Health Interview Survey in the USA for non-melanoma skin cancer risk prediction, and differed from most studies in this Review because they did not use image data. Udrea and colleagues[29] used data from previous specialist care studies in referred populations, and also from the SkinVision user database that contains images of skin lesions taken on smartphones by non-referred SkinVision users. We therefore chose to review the data for all 272 studies (appendix p 5; table 2) identified that applied AI/ML techniques to the evaluation of skin lesions, and although these studies have not been developed using data from low-prevalence populations, they still have relevance for the application of these technologies in primary care settings.

In the real world, the distinction between primary and secondary care clinical settings is often not completely clear. We addressed this by classifying patient populations into high-prevalence or referred populations, versus low-prevalence or non-referred populations. This approach is not perfect and does not produce two completely independent groups, but we felt it was the best approach to answer the question we were interested in—namely, whether any algorithms had been developed or tested in primary care or similar clinical settings. Images acquired from dermatology clinics were classified as being representative of high-prevalence populations, although we recognise that in some countries, patients in dermatology clinics might be a non-referred population.

### Primary outcomes

*Diagnostic accuracy of the AI/ML algorithms*

Although the reported measures of diagnostic accuracy were generally high, there was a wide range for most measures in the studies that reported relevant data (table 2). For example, the mean sensitivity for melanoma diagnosis was 0·842 (95% CI 0·816–0·868), but the range was 0·13–1·00. Most studies investigated melanoma, with fewer studies examining accuracy for keratinocyte carcinomas; only three studies reported the diagnostic accuracy results for keratinocyte carcinomas before 2018. Studies of squamous cell carcinomas in particular showed a lower sensitivity and diagnostic accuracy. Tschandl and colleagues[30] varied the proportion of malignant cases in test sets and showed that expert clinicians had a higher mean number of correct diagnoses than AI/ML algorithms on test sets with more malignant cases, whereas AI/ML algorithms had a higher mean number of correct diagnoses than expert and non-expert clinicians on random test sets or those containing more benign cases.

### Secondary outcomes

*Intended purpose of the AI/ML algorithm*

Differentiating melanoma from benign skin lesions (naevi) was the most common task given to the AI/ML

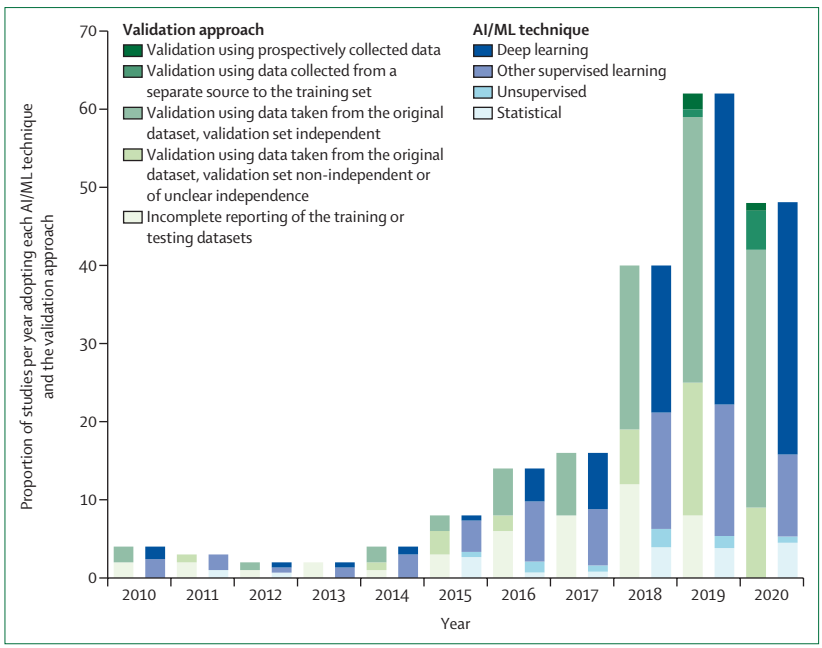


***Figure 3:*** **AI/ML techniques and validation approaches adopted for 2010–20 (N=272)**
The validation approaches used in the included studies are shown in the green bar for each year, with the AI/ML techniques adopted shown in the blue bar for each year. For each year validation approach is the left bar and AI/ML technique is the right bar. Because 2000–09 had less than five studies published per year, these years have been excluded from this figure. The cumulative totals for 2000–09 were as follows: for the AI/ML technique used, one study used a statistical technique, two used unsupervised techniques, six used other supervised learning techniques, and eight used deep learning techniques; and for the validation approach, six studies had incomplete reporting of training or testing datasets, two studies were validated using data taken from the original dataset and the validation set was non-independent or of unclear independence, three studies were validated using data taken from the original dataset and the validation set was independent, one study was validated using data collected from a separate source to the training set, and no studies were validated using prospectively collected data. The totals from 2021 were also excluded because data for the full year were not available, so a visual comparison with other years would not be accurate. The totals for Jan 1, 2021 to Aug 9, 2021, were as follows: for the AI/ML technique used, four studies used a statistical technique, no studies used an unsupervised technique, three studies used other supervised learning techniques, and 57 studies used deep learning techniques; and for the validation approach, 11 studies had incomplete reporting of the training or testing datasets, six studies were validated using data taken from the original dataset and the validation set was non-independent or of unclear independence, 38 studies were validated using data taken from the original dataset and the validation set was independent, two studies were validated using data collected from a separate source to the training set, and no studies were validated using prospectively collected data. Note: many studies used multiple AI/ML techniques in sub-studies within the overall article. The AI/ML technique bar here represents the proportion of studies that used a particular AI/ML technique from the total number of studies for that year. AI/ML=artificial intelligence and machine learning.

For the **ISIC archive** see https://www.isic-archive.com/

algorithms (n=100 [37%] of 272). 66 of the included studies (24%) aimed to differentiate malignant from benign lesions, and 31 studies (11%) aimed to differentiate lesions into one of three classes: melanoma, keratinocyte carcinoma, and benign lesions. Only six studies (2%) aimed to classify lesions into suspicious or non-suspicious categories (or into those that needed a biopsy). Such a diagnostic triage approach more closely mirrors the task that clinicians perform in primary care settings worldwide.

*Description of methods used to develop the AI/ML algorithm*

Descriptions of the methods used to develop the AI/ML algorithm were poor. A risk of bias assessment using QUADAS-2 was completed for all included studies (appendix p 25),[25] showing a wide range in the quality of reporting, with study methods frequently not prespecified nor described in sufficient detail. A wide range of study designs were used. The combination of these factors meant that the risk of bias assessment was often not discriminatory. However, despite a wide range of low, high, and unclear risk of bias in the included studies, no studies were excluded based on quality alone.

*Appropriateness of datasets used to develop the AI/ML algorithm*

Several datasets were used to develop the AI/ML algorithms. Figure 2 shows that datasets were often small and curated by the study authors before 2016. After 2016, there was an increase in the use of large independent datasets, the most frequently used being the International Skin Imaging Collaboration (ISIC) datasets. Melanoma images were included in 206 studies (76%), keratinocyte carcinomas in 77 studies (28%), in-situ carcinomas in 25 studies (9%), pre-malignant skin lesions (eg, actinic keratoses and dysplastic naevi) in 93 studies (34%), and benign lesions in 196 studies (72%). The included lesion types were unclear in 75 studies (28%). 44 studies (16%) used datasets that contained an equal number of or more malignant lesions than benign lesions.

Images of skin lesions that originated from referred populations and from individuals that were at a high risk of skin cancer were used in 239 studies (88%), from non-referred or low-risk populations in two studies (1%), and the population was unclear in 31 studies (11%). Comprehensive information regarding the clinical setting where the images and data originated from was often unclear, including for many larger independent datasets (appendix p 30), which made categorisation for this Review difficult.

Histopathology was used as the ground truth or gold standard control for images analysed in 16 studies (6%), an expert opinion from a clinical specialist (eg, a dermatologist) was used in 13 studies (5%), and a mixture of histopathology results and expert clinical opinion were used in 198 studies (73%). The expert clinical opinion category included long-term digital monitoring of skin lesions. Ground truth or gold standard reporting was unclear in 46 studies (17%).

Figure 3 shows that almost a quarter of studies (62 [23%]) did not fully report information on their training, testing, and validation datasets, and a further approximately fifth of studies (48 [18%]) used datasets that were not independent or of unclear independence. Figure 4 shows that many studies (186 [68%]) used databases containing dermoscopic images. However, in 210 studies (77%) it was unclear how the images were obtained or digitised, and in 105 studies (39%), details of the image resolution were not provided.

*AI/ML algorithm design*

Various AI/ML algorithm designs were used in the included studies (figure 3). Over time, there was an

increased use of supervised learning approaches, mostly involving neural-network-based deep learning techniques; this began around 2016 and increased each year.

Figure 4 highlights key issues regarding the interpretability of the AI/ML algorithms. Most studies (200 [74%] of 272 included studies) inadequately described the statistical methods used, for example they did not describe measures of spread or the statistical significance of their results. Most studies (178 [65%]) did not examine the interpretability of their results. 57 studies (21%) correlated the features identified as important by the AI/ML algorithm to existing clinical diagnostic features or diagnostic checklists (eg, ABCD score, modified Glasgow 7-point checklist, total dermoscopic score or telangiectasia in basal cell carcinoma). 26 studies (10%) had some form of feature selection to choose the most discriminative and clinically relevant features to include in the algorithm. 15 studies (6%) did interpretability analysis through techniques such as feature maps, heat maps, and saliency analysis, and 15 studies (6%) provided a measure of the algorithm's diagnostic confidence.

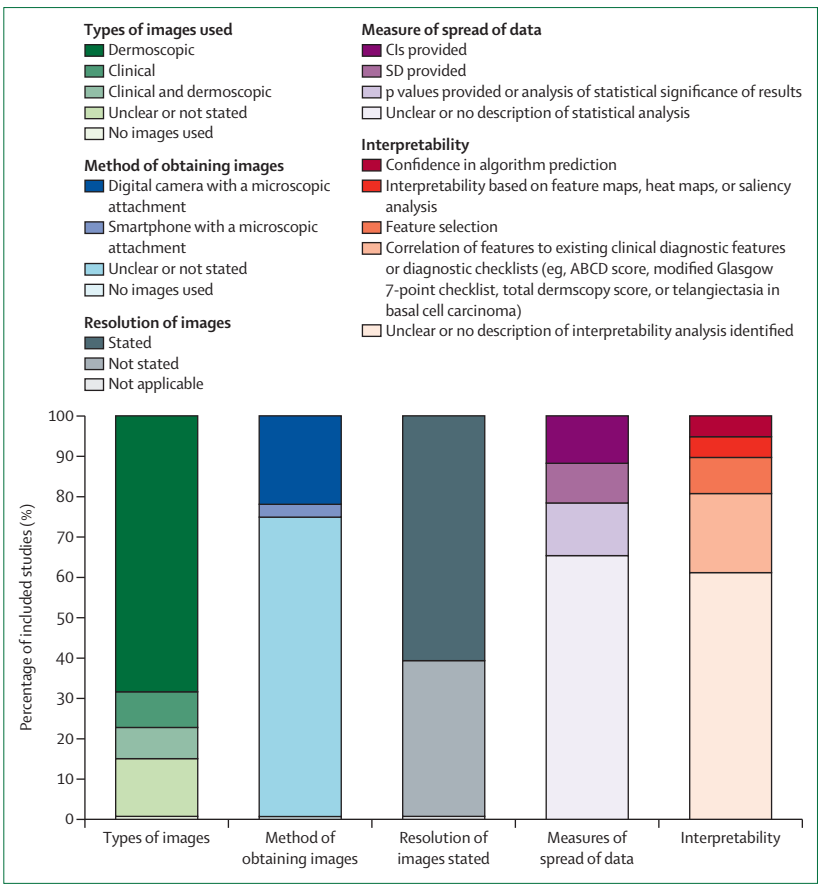


***Figure 4:*** **Characteristics of included study methods (N=272)**
The approaches used by the included studies in key methodological areas. Each bar is segmented to show the different approaches used for each area. In the bars referring to measures of spread of data and interpretability, studies could use multiple techniques so the numbers in the text might add up to more than the 272 included studies. No images used refers to studies that used electronic health record data to predict the risk of melanoma but did not include any images in their study.

### *AI/ML algorithm evaluation*

Tables 1 and 2 show heterogeneous approaches to evaluate the AI/ML algorithms, specifically in the given outcome measures. There was a focus on sensitivity, accuracy score, specificity, and AUROC, but few reports of negative predictive value measures. The results of the included studies are likely to have little external validation because most studies did not publish sufficient information about their algorithm design or dataset to allow replication of the results. Although 158 (58%) studies used the ISIC datasets, it was often unclear whether the algorithms had been entered into the ISIC annual skin lesion image classification challenge, and thus had undergone testing by an independent organisation, or had been developed and tested by the authors themselves using the ISIC datasets.

### *AI/ML algorithm implementation*

According to the CanTest Framework (figure 1),[17] most of the included studies would be classified as early-stage research, with little evidence of validation in prospective or real-world clinical settings (figure 3). Few studies considered implementation of the algorithms they had developed, including how their algorithms would fit into clinical practice and existing diagnostic pathways. We did identify studies that considered how an AI/ML algorithm could be used to support clinical decision making,[11,12,31–33] examined the accuracy of AI/ML algorithms in teledermatology settings,[34,35] used images collected by patients through a smartphone application,[34,36] and considered potential issues arising from using AI/ML algorithms on darker skin.[37] One study explored the effect of clinical decision support from an AI/ML algorithm on the diagnostic accuracy of Argentinian general practitioners using images from the ISIC archive.[32] We did not identify any health economic, patient, or clinician acceptability data for any of the included studies. We did not identify any prospective trials of algorithms in real-world clinical settings.

We are aware of several commercial technologies that might have unpublished efficacy data. We therefore also used scoping review methods to identify commercially developed AI/ML technologies aimed at the triage or diagnosis of skin lesions (appendix p 31). In summary, we identified 18 technologies, five of which had already been identified through the systematic bibliographic database searches; Fotofinder Moleanalyser Pro,[8,12,38–41] DERM,[34,42] SkinVision,[29] and model derm.[10,43] We additionally identified studies evaluating an algorithm developed by Google.[35,44] Some of these technologies had multiple studies evaluating various aspects of their diagnostic accuracy, validated in independent datasets. For nine of the AI/ML commercial technologies we were not able to identify peer-reviewed, published studies, or unpublished data evaluating their accuracy.

## Discussion

This is the first Review, to our knowledge, to evaluate AI/ML algorithms for skin cancer detection as applied to a primary care setting. We identified only two studies that used data from low prevalence populations. In the full review of studies with relevance to primary care settings, we identified 272 studies but found a wide variation in study designs, outcome measures, and quality of reporting. We identified a rapid increase in the number of studies published each year, with 57 studies published at the time of writing in 2021. Diagnostic accuracy results showed reasonable sensitivity for melanoma but were notably lower for keratinocyte carcinomas, and all had wide ranges.

However, we identified areas of concern when considering whether the AI/ML algorithms were sufficiently accurate and safe for implementation in primary care settings. First, the datasets largely included lesions from patients recruited in specialist clinical settings, where the types and prevalence of skin lesions are different to those seen in primary care settings. Any diagnostic approach needs to be evaluated using data reflecting the patient population and disease prevalence of the intended setting, otherwise the diagnostic performance will be prone to spectrum bias.[45] Second, the datasets were frequently small in size and did not clearly split the datasets into training, testing, and validation sets, which would lead to falsely high accuracy results because of overfitting.

Third, there is a maxim in ML communities: garbage in, garbage out, which highlights the importance of the quality of data used to train AI/ML algorithms. We found a wide variation in the reported data, particularly around the origin and type of data used to train and test the AI/ML algorithms. This variation was true even for the large, frequently used image datasets, where image origin, inclusion criteria, and patient demographics (including skin colour or Fitzpatrick skin type; appendix p 30) were not consistently reported, raising questions about the quality and generalisability of the underlying data. AI/ML algorithms developed and tested on data that are not representative of the entire population in which they are intended to be used in might have inherent biases.[46] For example, in this Review, few studies included skin lesions from darker skin colours (appendix p 30), so the future performance of the AI/ML algorithms on darker skin types might be impaired. Fourth, the technology design and the tasks given to the algorithm were often inappropriate for primary care implementation. Most AI/ML algorithms aimed to diagnose melanoma and, although this a dangerous skin cancer, in clinical practice these technologies are likely to be used more frequently on more common keratinocyte carcinomas; therefore their ability to accurately diagnose these lesions should be considered during algorithm development and evaluation. Only 2% of studies adopted a triage approach to establish which lesions needed either referral to specialist care or biopsy, the approach that most closely represents the task facing clinicians in primary care.

The strengths of this Systematic Review include: a broad and inclusive search strategy; the guidance of an international expert panel in protocol development and on the search strategy; independent screening, quality assessment, and data extraction processes; adherence to PRISMA guidance; and benefiting from the use of scoping review methods to identify commercially developed AI/ML technologies. The Review was limited by substantial variability in the quality of the reporting and study design, which precluded a full meta-analysis of the data. Many of the image datasets were used across several studies, possibly causing a distortion of the reported mean outcome measures. During the data extraction process we did not collect data separately on which studies used long-term sequential monitoring of skin lesions to obtain a control diagnosis, and which studies used an expert clinical diagnosis based on a single review. The accuracy of these diagnostic methods might differ slightly, but we were not able to take that into account in our data analysis. Furthermore, we were not able to record and analyse the software used to build the AI/ML algorithms, nor were we able to thoroughly search the references and citations, given the size of the search and number of studies screened. In summary, this is a rapidly evolving research area, which will require subsequent updates.

Previous reviews have highlighted the difficulty in comparing AI/ML algorithms because of the use of non-public datasets, not fully disclosing the methods used for training,[47] the absence of prospective studies, the risk of bias and overfitting in the existing research,[48] and issues with the performance of these algorithms on out-of-distribution images. Our Review supports and echoes these findings. Obermayer and Topol[49] have highlighted the importance of AI/ML algorithms learning from a diverse training dataset to ensure the algorithm advises fairly across gender, race, and sociodemographic status;[50] many studies evaluated in this Review used datasets predominantly from Europe and the USA containing mostly Fitzpatrick skin types 1–3.

AI/ML algorithms have great potential to support clinicians in the accurate detection of skin lesions in primary care settings. However, this Review showed that research in this area is at an early stage of development and raised concerns as to whether the diagnostic performance would be maintained among populations with lower skin cancer prevalence such as that seen in primary care populations, or in settings with non-dermatoscopic or lower quality images, which is the case for many primary care clinics and images taken by patients. It is encouraging to see progressively more studies addressing implementation considerations and issues, including the use of AI/ML to assist clinicians in accurately assessing skin lesions. The use of AI/ML algorithms to support primary care clinicians in the triage of suspicious skin lesions might represent the optimum

***Panel*: Proposed checklist for the design, development, and evaluation of artificial intelligence and machine learning (AI/ML) algorithms aiming to support the triage or detection of possible skin cancers**

**(1) Intended purpose of the AI/ML algorithm**

(a) Does it address a real unmet clinical need?

(b) Is the algorithm designed appropriately to address that clinical need?

- Appropriate and relevant data provided
- Complexity of the clinical task reflected in the algorithm task

(c) Is the task clearly specified?

- Is the intended use triage or diagnosis?

(d) Is it clear how the algorithm will fit in with existing clinical practice and diagnostic pathways?

**(2) Description of the methods used to develop the AI/ML algorithm**

(a) Were the study methods prespecified?

(b) Were all aspects of the study methods described in sufficient detail?

- Study design, recruitment method, source, and type of data
- Inclusion and exclusion criteria for the datasets used
- Approach to algorithm development and training
- Validation approach
- Statistical analysis plan

**(3) Robustness and appropriateness of the datasets used to develop the AI/ML algorithm**

(a) Are data used to develop the algorithm appropriate for the intended use and are they adequately described?

(i) Country of origin

(ii) Clinical setting

- Primary or specialist care, referred or non-referred populations, clinical trial, image bank, or other

(iii) Content and type of data

- Is the dataset of an appropriate size?
- Images: how were they captured and digitised? What is their resolution and pixel depth?
- Are other appropriate data types included: demographic data, coded data, associated metadata, free text data, and other data types?

(iv) Inclusion and exclusion criteria for patients

- Skin types, age, ethnicity, sex, and skin lesion types

(b) Are the data representative of those which the algorithm will encounter in the intended real world clinical settings?

(i) Balance and prevalence of skin lesion types

(ii) Size of the dataset

(iii) Types of data the algorithm evaluates (eg, are the metadata similar to those a clinician would have access to for a similar task?)

(c) Have data been labelled accurately with an appropriate diagnostic reference standard?

- Histopathology, expert opinion, mixed, or other
- How many labellers were involved? What is their level of expertise?

(d) Is there a clear description of the split of the datasets into training, validation, and testing sets?

- Are partitioned datasets independent?

**(4) Design of the AI/ML algorithm**

(a) Is the architecture clearly described?

(b) Are the data available to replicate the experimental results?

- Is the algorithm code publicly available?
- Is there a clear description of data and algorithm?

(c) Trustworthiness and reliability of the algorithm

(i) Is the output interpretable?

- Feature selection, clinical correlation of features, saliency analysis, heat maps, and other interpretability analyses

(ii) Are CIs provided for the strength of the algorithm's predictions?

**(5) Evaluation of the AI/ML algorithm**

(a) Has the diagnostic accuracy been fully reported?

- Minimum requirement: 2 × 2 table or true positives, true negatives, false positives, and false negatives

(b) Has the algorithm undergone training and validation steps, been tested in an independent test set, or been evaluated in a prospective clinical trial in the intended population or in a clinical setting?

(c) Has performance been assessed in the intended real world clinical setting?

- Have all relevant performance measures been evaluated (accuracy, safety, usability, and cost-effectiveness)?[57,58]

**(6) Implementation of the AI/ML algorithm**

(a) Are plans for ongoing development clearly articulated?

(b) Does the algorithm have the appropriate regulatory approvals needed for deployment? (eg, Conformité Européene or UK Conformity Assessed mark, US Food and Drug Administration approvals, and Digital Technology Assessment Criteria)

(c) Are there clear plans for post-market evaluation in real world clinical settings?

positioning of these technologies in primary care clinical settings. We also identified some good examples of AI/ML algorithms that have been developed commercially and subsequently evaluated in academic studies, but validation in real-world clinical settings is still required to prove their safety and effectiveness before they can be recommended for clinical use. Before implementation, consideration also needs to be given to: the use of interpretability analysis to help build the trust of clinicians and patients in these algorithms, understanding the acceptability of AI/ML algorithms to clinicians and patients, how AI/ML algorithms would best be incorporated into clinical workflows, and health-care system economic perspectives.

Although many dermatologists[51] and patients[52] hold optimistic attitudes towards AI/ML algorithms, these algorithms need to be evaluated carefully to ensure that they are accurate, effective, cost-effective, and safe enough for clinical use, and that increased access to skin lesion assessment will not add to the biopsy burden on specialist care providers or contribute to an overdiagnosis of melanoma.[53] There are currently guidelines in development for the reporting of studies assessing AI/ML interventions in diagnostic accuracy studies, prediction models, and clinical trials.[54–56] To address some of the issues highlighted in this Systematic Review we have developed a proposed checklist for the design, development, and evaluation of AI/ML algorithms aiming to support the triage or detection of possible skin cancers in primary care (panel). If widely adopted, we expect that this checklist will enable meaningful comparison between studies, increase the clinical relevance of AI/ML algorithms, and improve the likelihood of these promising technologies being successfully implemented. Although the checklist is aimed at the development of AI/ML algorithms for the early detection of skin cancers in primary care settings, many of the issues raised reflect more general issues in ML algorithm development, and could be applied in other disease areas and clinical settings as well.

#### Contributors

OTJ developed the protocol, completed the search, screened the articles for inclusion, extracted the data, synthesised the findings, interpreted the results, and drafted the manuscript. RNM and MvdS interpreted the results and critically revised the manuscript. KPB screened articles for inclusion, extracted the data, completed the risk of bias assessments, led the work using a scoping review method, and critically revised the manuscript. CKIR screened articles for inclusion, extracted the data, completed the risk of bias assessments, and critically revised the manuscript. MSI extracted the data, completed the risk of bias assessments, synthesised the findings, and interpreted the results. DB and RB extracted the data, completed the risk of bias assessments, and critically revised the manuscript. NC screened the articles for inclusion, extracted the data, and critically revised the manuscript. JE and HCW critically revised the manuscript. FMW developed the protocol, synthesised the findings, interpreted the results, and critically revised the manuscript. All authors had access to the data presented in the manuscript. The raw data were collected and verified by OTJ, KPB, CKIR, MSI, DB, RB, and NC. All authors approved the final version.

#### Declaration of interests

We declare no competing interests.

#### Acknowledgments

This systematic review was funded by the National Institute for Health Research Policy Research Programme, conducted through the Policy Research Unit in Cancer Awareness, Screening, and Early Diagnosis (PR-PRU-1217–21601). The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health and Social Care. The first author (OTJ) was also supported by the CanTest Collaborative funded by Cancer Research UK (C8640/A23385), of which FMW is Director, JE is an Associate Director, and NC is Research Fellow. During protocol development, this Review benefited from the advice of an international expert panel from the CanTest collaborative, including Willie Hamilton (University of Exeter, Exeter, UK), Greg Rubin (University of Newcastle, Newcastle, UK), Hardeep Singh (Baylor College of Medicine, Houston, TX, USA), and Niek de Wit (University Medical Center Utrecht, Utrecht, Netherlands). The research was also supported by a Cancer Research UK Cambridge Centre Clinical Research Fellowship for OTJ, and a National Health and Medical Research Council Investigator Fellowship (APP1195302) for JE. The funding sources had no role in the study design, data collection, data analysis, data interpretation, writing of the report, or in the decision to submit for publication. The authors would like to thank Isla Kuhn (Reader Services Librarian, University of Cambridge Medical Library, Cambridge, UK) for her help in developing the search strategy. We also thank Smiji Saji, who assisted with the early stages of the Review, Haruyuki Yanaoka, who assisted with the translation and assessment of papers that were written in Korean, and Steve Morris who assisted with the analysis of the data.

#### References

1 Garbe C, Keim U, Gandini S, et al. Epidemiology of cutaneous melanoma and keratinocyte cancer in white populations 1943–2036. *Eur J Cancer* 2021; **152:** 18–25.

2 Karimkhani C, Boyers LN, Dellavalle RP, Weinstock MA. It's time for "keratinocyte carcinoma" to replace the term "nonmelanoma skin cancer". *J Am Acad Dermatol* 2015; **72:** 186–87.

3 Cancer Research UK. Non-melanoma skin cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-melanoma-skin-cancer#heading-Zero (accessed July 27, 2021).

4 Karia PS. Epidemiology and outcomes of cutaneous squamous cell carcinoma. In: Schmults C (ed). High-risk cutaneous squamous cell carcinoma. Berlin, Heidelberg: Springer, 2016: 3–28.

5 Cancer Research UK. Melanoma skin cancer survival statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/survival (accessed July 27, 2021).

6 Hiom SC. Diagnosing cancer earlier: reviewing the evidence for improving cancer survival. *Br J Cancer* 2015; **112** (suppl 1)**:** S1–5.

7 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

8 Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018; **29:** 1836–42.

9 Brinker TJ, Hekler A, Enk AH, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer* 2019; **119:** 11–17.

10 Han SS, Park I, Chang SE, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol* 2020; **140:** 1753–61.

11 Maron RC, Utikal JS, Hekler A, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: web-based survey study. *J Med Internet Res* 2020; **22:** e18091.

12 Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med* 2020; **26:** 1229–34.

13 NHS. The Topol Review: preparing the healthcare workforce to deliver the digital future. February, 2019. https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf (accessed July 27, 2021).

14 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1:** e271–97.

15 Royal College of General Practitioners. Artificial intelligence and primary care. https://www.rcgp.org.uk/-/media/Files/CIRC/CIRC-AI-REPORT.ashx?la=en (accessed July 27, 2021).

16 Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019; **28:** 231–37.

17 Walter FM, Thompson MJ, Wellwood I, et al. Evaluating diagnostic strategies for early detection of cancer: the CanTest framework. *BMC Cancer* 2019; **19:** 586.

18 Jones OT, Calanzani N, Saji S, et al. Artificial intelligence techniques that may be applied to primary care data to facilitate earlier diagnosis of cancer: systematic review. *J Med Internet Res* 2021; **23:** e23483.

19 Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; **4:** 1.

20 Jones OT, Saji S, Prathivadi K, et al. Establishing which modalities of artificial intelligence (AI) for the early detection and diagnosis of cancer are ready for implementation in primary care: a systematic review. March 31, 2020. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020176674 (accessed March 30, 2022).

21 Freeman K, Dinnes J, Chuchu N, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ* 2020; **368:** m127.

22 House Of Lords: Select Committee on Artificial Intelligence. AI in the UK?: Ready, willing and able. April 16, 2018. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (accessed July 27, 2021).

23 McCarthy J, Minsky M, Rochester N, Shannon C. A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine* 2006; **27:** 12.

24 Muehlhauser L. What should we learn from past AI forecasts? Open Philanthropy Project. May, 2016. https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/what-should-we-learn-past-ai-forecasts (accessed July 27, 2021).

25 Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155:** 529–36.

26 Joanna Briggs Institute University of Adelaide. Joanna Briggs Institute critical appraisal tools, 2020. https://jbi.global/critical-appraisal-tools (accessed July 27, 2021).

27 Popay J, Roberts H, Sowden A, et al. Guidance on the conduct of narrative synthesis in systematic review: a product from the ESRC Methods Programme. January, 2006. https://www.researchgate.net/profile/Mark-Rodgers-3/publication/233866356_Guidance_on_the_conduct_of_narrative_synthesis_in_systematic_reviews_A_product_from_the_ESRC_Methods_Programme/links/02e7e5231e8f3a6183000000/Guidance-on-the-conduct-of-narrative-synthesis-in-systematic-reviews-A-product-from-the-ESRC-Methods-Programme.pdf (accessed July 27, 2021).

28 Roffman D, Hart G, Girardi M, Ko CJ, Deng J. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Sci Rep* 2018; **8:** 1701.

29 Udrea A, Mitra GD, Costea D, et al. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *J Eur Acad Dermatol Venereol* 2020; **34:** 648–55.

30 Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol* 2019; **20:** 938–47.

31 Lee S, Chu YS, Yoo SK, et al. Augmented decision-making for acral lentiginous melanoma detection using deep convolutional neural networks. *J Eur Acad Dermatol Venereol* 2020; **34:** 1842–50.

32 Lucius M, De All J, De All JA, et al. Deep neural frameworks improve the accuracy of general practitioners in the classification of pigmented skin lesions. *Diagnostics (Basel)* 2020; **10:** 969.

33 Sevli O. A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation. *Neural Comput Appl* 2021; **33:** 12039–50.

34 Phillips M, Marsden H, Jaffe W, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open* 2019; **2:** e1913436.

35 Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med* 2020; **26:** 900–08.

36 Veronese F, Branciforti F, Zavattaro E, et al. The role in teledermoscopy of an inexpensive and easy-to-use smartphone device for the classification of three types of skin lesions using convolutional neural networks. *Diagnostics (Basel)* 2021; **11:** 451.

37 Aggarwal P, Papay FA. Artificial intelligence image recognition of melanoma and basal cell carcinoma in racially diverse populations. *J Dermatolog Treat* 2021; published online June 30. https://doi.org/10.1080/09546634.2021.1944970.

38 Haenssle HA, Fink C, Toberer F, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol* 2020; **31:** 137–43.

39 MacLellan AN, Price EL, Publicover-Brouwer P, et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *J Am Acad Dermatol* 2021; **85:** 353–59.

40 Sies K, Winkler JK, Fink C, et al. Past and present of computer-assisted dermoscopic diagnosis: performance of a conventional image analyser versus a convolutional neural network in a prospective data set of 1,981 skin lesions. *Eur J Cancer* 2020; **135:** 39–46.

41 Winkler JK, Sies K, Fink C, et al. Melanoma recognition by a deep learning convolutional neural network–performance in different melanoma subtypes and localisations. *Eur J Cancer* 2020; **127:** 21–29.

42 Phillips M, Greenhalgh J, Marsden H, Palamaras I. Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatol Pract Concept* 2019; **10:** e2020011.

43 Muñoz-López C, Ramírez-Cornejo C, Marchetti MA, et al. Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study. *J Eur Acad Dermatol Venereol* 2021; **35:** 546–53.

44 Jain A, Way D, Gupta V, et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw Open* 2021; **4:** e217249.

45 Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* 2016; **353:** i3139.

46 Wen D, Khan SM, Xu AJ, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *Lancet Digit Health* 2022; **4:** e64–74.

47 Brinker TJ, Hekler A, Utikal JS, et al. Skin cancer classification using convolutional neural networks: systematic review. *J Med Internet Res* 2018; **20:** e11936.

48 Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. *JAMA Dermatol* 2019; **155:** 1291–99.

49 Obermeyer Z, Topol EJ. Artificial intelligence, bias, and patients' perspectives. *Lancet* 2021; **397:** 2038.

50 Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021; **3:** e260–65.

51 Polesie S, Gillstedt M, Kittler H, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol* 2020; **183:** 159–61.

52 Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front Med (Lausanne)* 2020; **7:** 233.

53 Welch HG, Mazer BL, Adamson AS. The rapid rise in cutaneous melanoma diagnoses. *N Engl J Med* 2021; **384:** 72–79.

54 Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat Med* 2020; **26:** 807–08.

55 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019; **393:** 1577–79.

56 Liu X, Rivera SC, Faes L, et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019; **25:** 1467–68.

57 Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020; **26:** 1364–74.

58 Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *BMJ* 2020; **370:** m3210.