HIGHLY CONFIDENTIAL

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

Diagnostics Advisory Committee – Thursday 18 January 2024
Early Value Assessment

Artificial Intelligence technologies for assessing skin lesions selected for referral on the urgent suspected cancer pathway to detect benign lesions and reduce secondary care specialist appointments

The following documents are made available to the Committee:

1. **Overview**

2. **Updated Early Value Assessment Report (16/01/2024)** produced by CRD and CHE Technology Assessment Group, University of York.
   Note, this report is an updated version to the one issued to stakeholders on 5 January 2024. The updates are listed on page 3 of the report.

3. **Patient organisation submission:**
   The British Association of Dermatologists

4. **External Assessment Report (EAR) consultation comments and responses**

Artificial Intelligence technologies for assessing skin lesions selected for referral on the urgent suspected cancer pathway to detect benign lesions and reduce secondary care specialist appointments: early value assessment

Diagnostics advisory committee

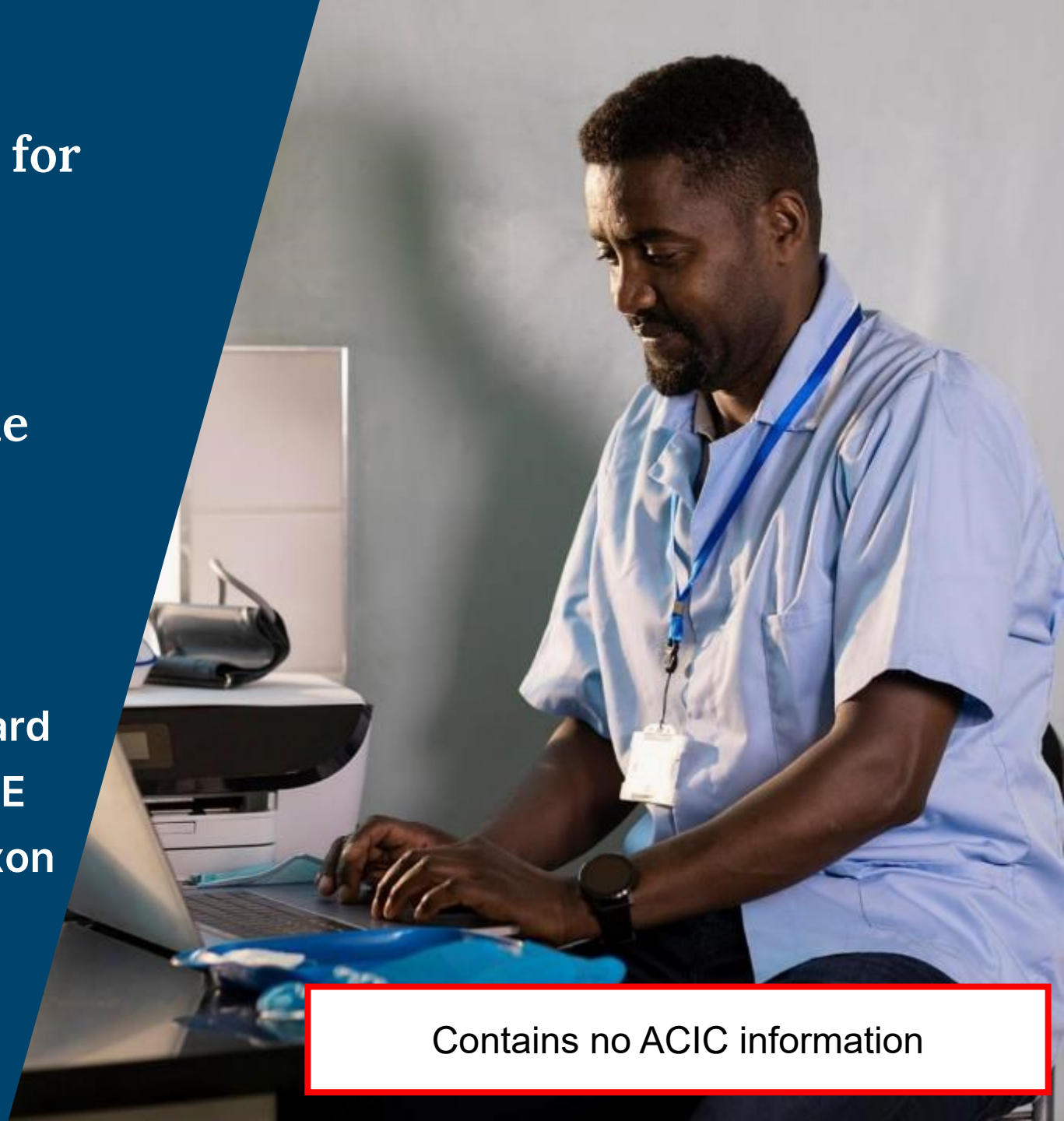
18 January 2024

Lead team: Sam Creavin, Stephanie Gallard

External assessment group: CRD and CHE

Technical team: Anam Khan, Frances Nixon

NICE National Institute for Health and Care Excellence



Contains no ACIC information

EVA decision making at committee

Based on the available evidence and expert opinion what is the likely impact of using the technology in the NHS on:

- Patients/carers
- System

If multiple technologies are being considered, each should be assessed independently unless specified differently in the scope.

Risks | Benefits

Risk of harm

Some benefits and risks

Does the assessment show sufficient promise to conditionally recommend for use with data collection?

Yes

Conditionally recommend for use while further evidence is generated

This outcome is likely for a technology which has

- some acceptance by healthcare professionals & patients
- acceptable risks for use in a data collection context
- acceptable evidence of potential benefit

Highlight any risks which could be mitigated in evidence generation

Identify key evidence gaps to be addressed by evidence generation

No

Is more research needed to better understand the potential risks and benefits of this technology?

Yes

Recommend for use only in a research context

This outcome is likely for a technology which has

- significant uncertainty about how the technology fits into NHS pathways
- significant uncertainty about the risks or benefits of the technology that needs exploration in a research context.

Describe key uncertainties for future research

Recommend not used

NICE

Early value assessment considerations

Unmet need and risk of harm?

- Do all the technologies have the potential to address the unmet need?
- Is there any potential risk of harm to patients with any of the technologies?

Suitable for use with data collection?

- Is the evidence of potential benefit sufficient to support use with data collection?
- Are the risks acceptable? Can any of the risks be mitigated?
- Can the technology be integrated into the NHS and is it likely to be acceptable to healthcare professionals and patients?

Suitable for research only?

- Is there significant uncertainty about the potential risks and/or benefits of using the technology?
- Are there concerns about integration of the technology or its acceptance in the NHS?

Possible recommendations

Conditionally recommended for use while further evidence is generated

- Likely that the technology will solve the unmet need and it is acceptable for the technology to be used in practice while further evidence is generated

Recommended only in a research context

- Uncertain if the technology has the potential to solve the unmet need, or it is not acceptable to be widely used in practice while further evidence is generated

Not recommended for use

- Unlikely that a technology has the potential to meet the unmet need, or where there are concerns about the potential harms associated with using the technology even in a research context

Background

NICE National Institute for Health and Care Excellence

Background on skin lesions suspicious of cancer

Need for appropriate referrals of skin lesions

- In the UK, dermatology services receive 1.2 million referrals a year. About 60% of these are suspected skin cancer pathway referrals, but only 6% are converted to a confirmed case of skin cancer.
- A significant proportion of primary care referrals to the urgent cancer pathway may not require face to face (F2F) appointments in secondary care.
- Shortages in the workforce and high numbers of unnecessary referrals are negatively impacting other dermatology services and delaying diagnosis and treatment of skin cancer patients. There is also a low threshold for referrals by GPs because they don’t receive in-depth dermatology training and many do not have access to dermatoscopes, which are essential for confidently identifying both benign skin lesions and skin cancer.
- This assessment covers all types of skin cancer. This includes three main types of skin cancer: melanoma, squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), as well as other, rarer, forms of skin cancer.

Referral criteria for GPs

		A urgent referral for suspected skin cancer is needed for:
Suspected melanoma	NICE – NG12	• a suspicious pigmented lesion that has a weighted 7-point checklist score of 3 or more, • dermoscopy suggests melanoma, • a pigmented or non-pigmented skin lesion that suggests nodular melanoma.
	BAD, BMJ, NICE – CSG8	• any new persistent skin lesion, especially if growing, pigmented, or vascular in appearance and the diagnosis is unclear, • a new pigmented line in the nail (especially if there is associated damage to the nail), or a lesion growing under the nail, • any doubt about the lesion, or there is a history of recent change, • a biopsy has confirmed the diagnosis of malignant melanoma. Note: if a lesion is suspected to be melanoma, an urgent referral to a dermatologist or other suitable specialist with experience of melanoma diagnosis should be made, and excision in primary care should be avoided. • a pigmented or non-pigmented skin lesion suggests nodular melanoma, • any major features in the 7-point checklist, or any features of the ABCDE system.
	BAD & RCP	• a new mole which is growing quickly over the age of puberty • any mole which has 3 or more colours or has lost its symmetry • a mole which has changed in appearance and is also itching or bleeding.
Suspected SCC	NICE – NG12	• a skin lesion that raises the suspicion of SCC.
		A routine referral for suspected skin cancer is needed for:
Suspected BCC	NICE – NG12	• a skin lesion that raises the suspicion of a BCC. • an urgent suspected skin cancer pathway referral should only be considered for a lesion that raises suspicion of BCC if there is a particular concern that a delay may have a significant impact, because of factors such as lesion site or size.

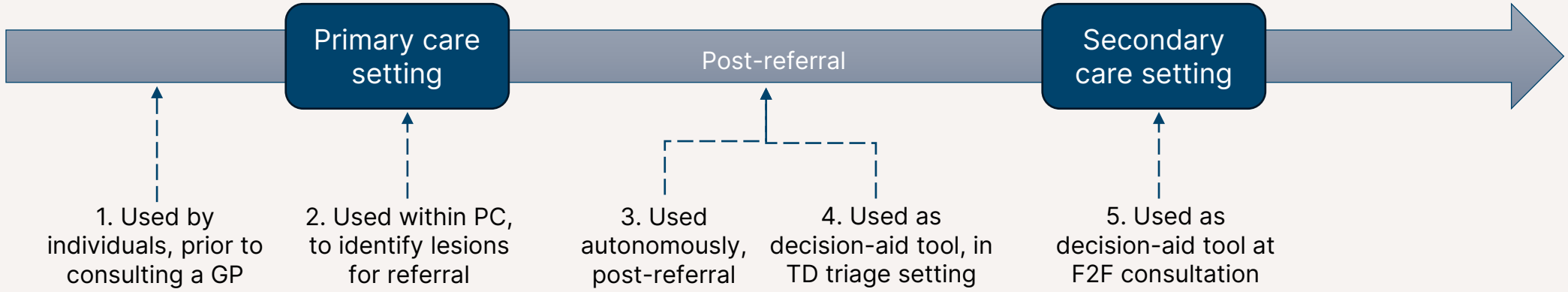
Technology purpose

Triage of cancer from benign skin lesions

Purpose of this assessment is to investigate the use of AI technologies for the analysis of skin lesions suspicious of cancer following a referral on the urgent suspected skin cancer pathway.

The assessment evaluates whether the technologies represent an effective and reliable means of triaging cancer from benign skin lesion, alongside current clinical practice.

Technologies may be used at various points in the diagnostic pathway:



This assessment focuses on the settings 3 and 4, but considered evidence from other settings where it informed understanding.

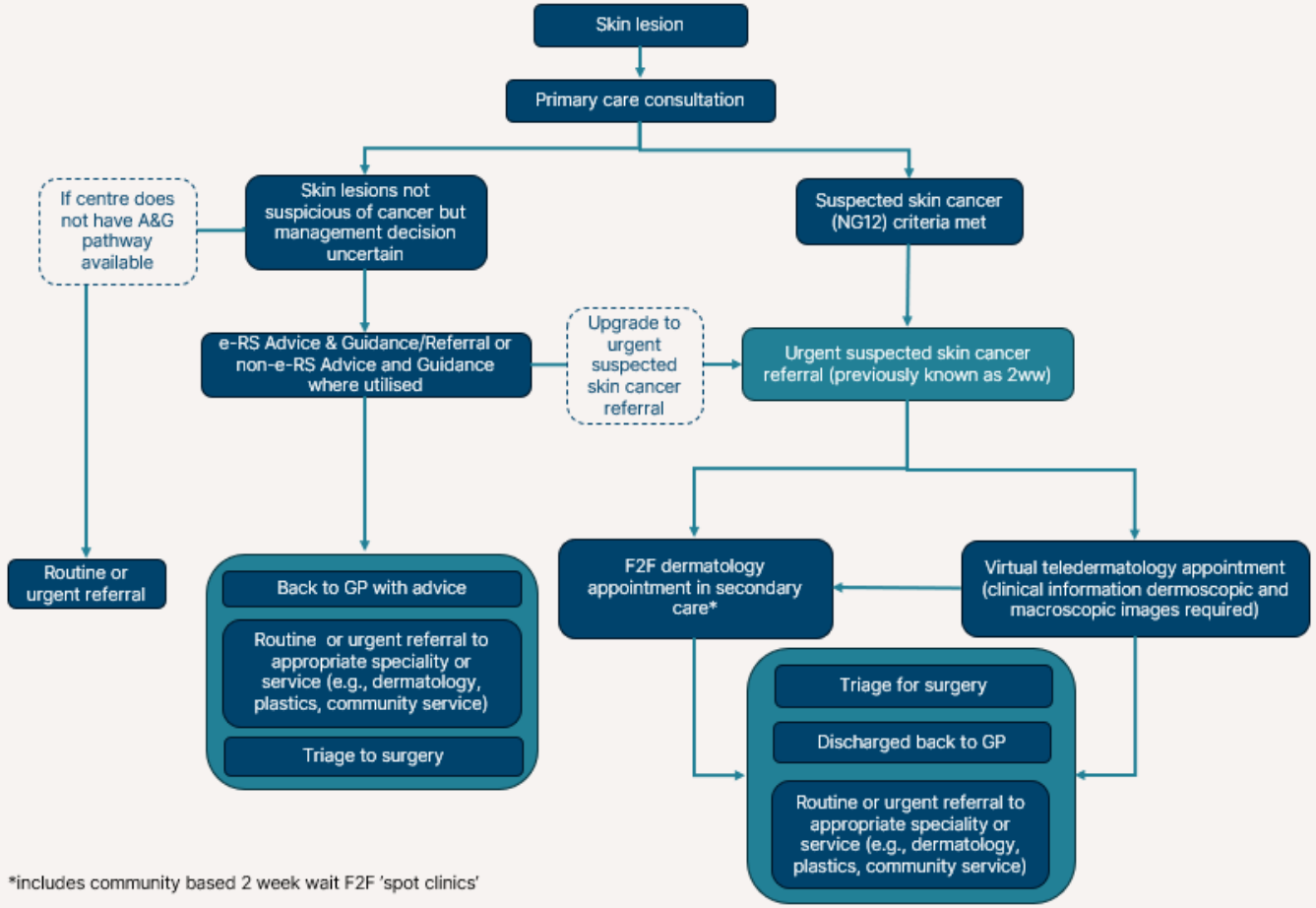
Current practice

Diagnostic pathway for suspected skin lesions

Options for GPs include:

- eRS Advice and Guidance (A&G)
- Routine referral
- Urgent suspected skin cancer referral pathway
    - F2F assessment
    - Teledermatology
        - Excludes: difficult sites, multiple lesions, children
        - Requires: clinical information, macroscopic and dermoscopic image taken ideally by a medical photographer
        - Image capture settings: GP surgery, community hub, teledermatology clinic in secondary care

There is variation within these pathways in availability and accessibility across the country.

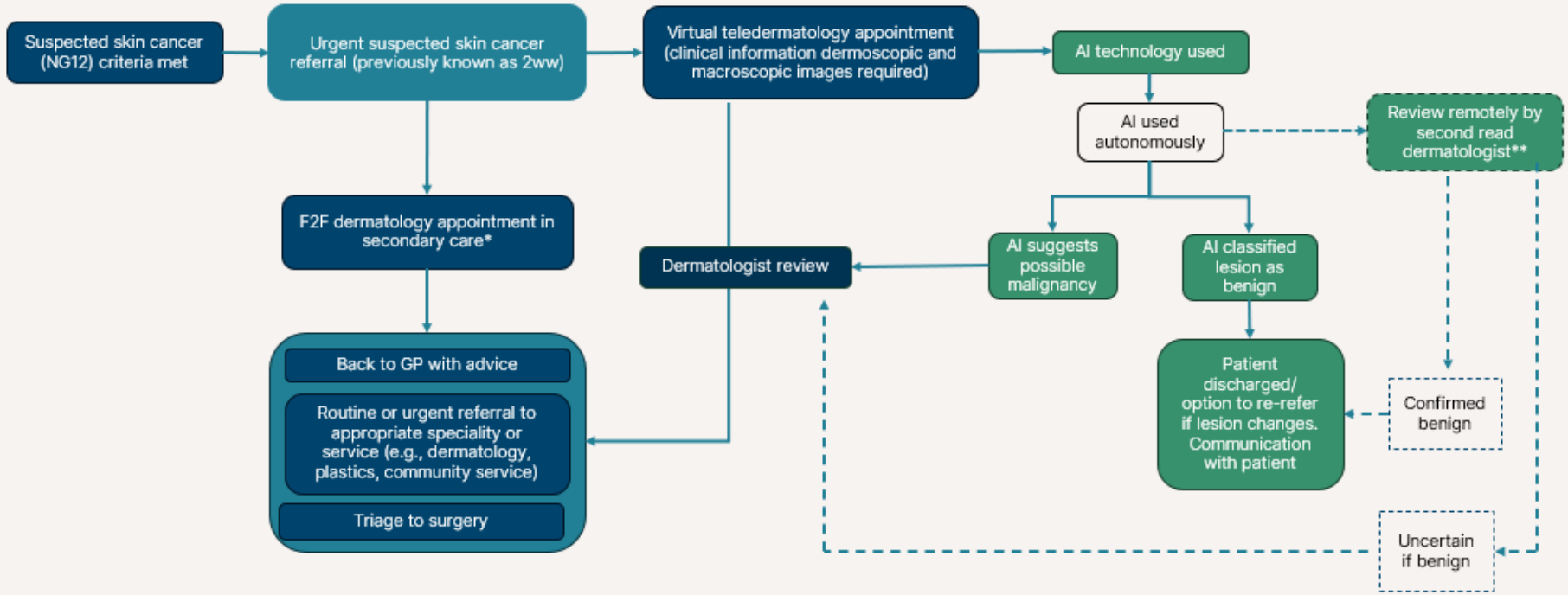


Unmet need

Pre-referral and post-referral settings

Current unmet need is in primary care within the pre-referral setting, however there is limited evidence in this setting.

Focus for this assessment is on the post-referral setting to identify those with benign lesions who can be discharged from the urgent suspected skin cancer pathway.



* includes community based 2 week wait F2F 'spot clinics'

** second read dermatology review currently in place for evaluation and safety; to be removed in due course

Decision problem (1)

1. **What are the cost and resource use implications of the use of AI technologies following an urgent suspected skin cancer referral to identify benign skin lesions?**
2. **What would a health economic model to estimate the cost-effectiveness of AI technologies to identify benign skin lesions in this setting look like, and what are the key evidence requirements necessary to populate such a model?**

Population	People with skin lesions suspicious of cancer, who have been referred from primary care for further evaluation
Interventions	DERM (Skin Analytics) MoleAnalyzer pro (FotoFinder systems)
Comparators	Clinical assessment and triage of suspicious lesions through the existing diagnostic pathway without use of AI. This can include assessment by specialist dermatologists either remotely or in person
Healthcare setting	The particular setting of interest was patients undergoing teledermatology assessments, but all settings after primary care referral were considered
Outcomes	Diagnostic accuracy, implementation, resource use and practicality, Clinical impact and patient benefit, and costs

Decision problem (2)

Key outcomes

Diagnostic accuracy	• Diagnostic test accuracy (sensitivity and specificity, area under ROC curve) ○ Where available, separately for each type of skin cancer (melanoma, BCC, SCC, rare skin cancers) • proportion of cancers missed and detected • proportion of benign lesions missed and detected • proportion of referrals confirmed to be skin cancer (positive predictive value)

Implementation, resource use, and practicality	• proportion of urgent cancer referrals:
	  ○ needing a face-to-face hospital appointment with a specialist for review of lesion
	  ○ converted to routine referral pathway
	  ○ resulting in a diagnostic biopsy
	  ○ booked for surgical procedure
	  ○ discharged back to GP
	• Time to:
	  ○ diagnosis
	  ○ discharge
	  ○ face-to-face consultant appointment
	  ○ treatment (surgery)
	• Cancer stage at detection
	• Ease of use/acceptability of AI software by healthcare professionals
	• Number of people consenting to use the technology
	• Test failure rates (with reasons, e.g. image capture issues)
	• proportion of suspicious skin lesions/patients excluded (with reasons, e.g. due to lesion location or scarring)

Decision problem (3)

Key outcomes

Clinical impact and patient benefit	• Clinical morbidities
	  ○ Including distant metastases and adverse outcomes of treatment
	• Mortality
	• Health-related quality of life
	• Non-clinical benefits to patients
	  ○ Reassurance that lesion is not cancerous.
	  ○ Anxiety associated with waiting for a diagnosis.
	  ○ Acceptability of AI technologies or processes

Costs	Costs were considered from an NHS and Personal Social Services perspective. Costs for consideration include: • Cost of annual subscription for AI software • Cost of training healthcare professionals to take images and to interpret AI software results • Cost of consultant dermatologist face-to-face appointments • Cost of staff time to upload images to AI software platforms and to interpret the results • Costs related to missed cancers • Costs of consultant dermatology triage team • Costs of teledermatology • Costs of new services required to support AI technologies (such as establishing new teledermatology services and setting up image capture)

Technologies under assessment (1)

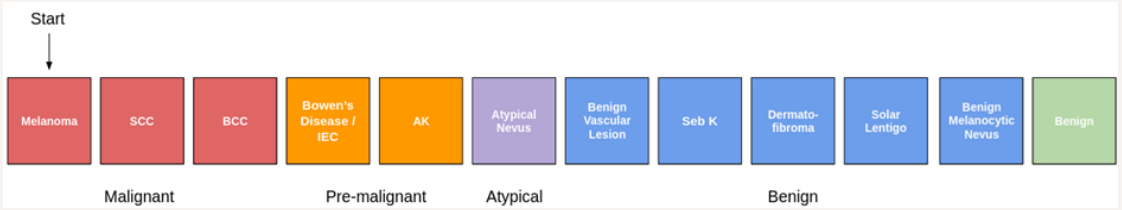
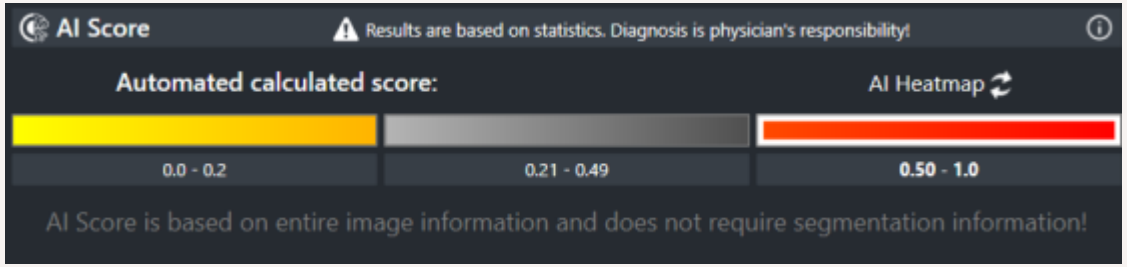
	DERM (Skin Analytics)	MoleAnalyzer pro (FotoFinder)

Key features	• Requires dermoscopic lens to take images
	• Granted phase 4 AAC AI in health and care award by NHS AI Lab
	• AI algorithm provides a suspected diagnosis of lesion and if applicable, a referral recommendation.
	• Uses fixed algorithm – does not update itself automatically
	• Algorithm trained on historical and prospectively collected images from populations in UK, USA and Italy
		• Requires FotoFinder dermatoscopes to take image
		• Uses CNN deep-learning algorithm to generate a risk score (AI score)
		  • AI score is based on comparison with images of skin cancers. It is a statistical estimate of the similarity to the malignant lesion images.
		  • AI score is displayed on a colour scale
		• 2 options: Online AI (updates continuously) and offline AI (updated annually)
		• Optional heatmap view available (shows which parts of the image of the lesion were used for the calculation of the AI score)
		• Optional second opinion service available upon purchase

Regulatory approval	UKCA Class IIa	CE mark Class IIa
Population	Adults (18 and above) and eligible	All ages, except where exclusions apply.
Intended use	Designed for screening, triage and assessment of suspicious skin lesions Skin Analytics states DERM can be used autonomously within clinical pathways under UKCA class IIa clearance. Where a lesion is marked as benign, there is an option for the case to be assessed either by the Trust Dermatologist or Skin Analytics Dermatologist (second read).	Intended for use by healthcare professionals for the assessment of single skin lesions. It should not be used to confirm a clinical diagnosis of melanoma, but rather support in decision making. The company states the technology can be used in primary care to reduce unnecessary referrals to secondary care and can also be used in the secondary care setting to obtain additional information to support decisions on whether to do a biopsy.

Technologies under assessment (2)

Benign pathway

DERM	MoleAnalyzer pro
Uses risk-based approach, where if the possibility of malignant is ruled out, then the algorithm returns a benign label, which follows the benign path. These include: • Benign vascular lesion • Seborrheic keratosis • Dermatofibroma • Solar lentigo • Melanocytic benign nevus • Generic benign label	Uses a risk score-based approach, which is a confidence score of the algorithm assessing the similarity of the lesion in question to malignant lesions. The score does not make any statement regarding the medical risk or malignancy of a lesion. Where a risk score between 0-0.2 indicates the 'lesion is inconspicuous' and would follow the benign path.
	

Artificial Intelligence technologies for assessing skin lesions

Kate Hawley and Susan Mountain
Specialist lay committee members
18 January 2023

NICE National Institute for Health and Care Excellence

Patient and carer considerations (1)

Impact of current assessment process

- Concern about skin lesions may adversely affect patients' mental health, including anxiety or depression.
- If skin cancer has been previously identified, any change in skin can cause panic and fear of spread to other organs.
- This may be exacerbated by:
  - Long waits for biopsy results
  - Constant GP or hospital appointments.

'I feel as though my life is a countdown for the next one I may miss.'

Patient and carer considerations (2)

Advantages and disadvantages of the software proposed

Potential benefits	• More people will use the technology as it will be more accessible than always seeing a physician in person • This will save medical professional's time as it could be done through an app/by the service user, and will also lessen the number of skin cancer referrals from GPs • Having no person present can be less embarrassing and avoids cultural sensitivities when removing clothing to display lesions

Potential concerns	• Will the decision to escalate or not escalate be communicated to the patient in a emotionally sensitive way by AI software? How will this be safeguarded to avoid excess anxiety?
	• Recognising that the software is not infallible, what safeguards will be put in place to ensure:
	  ○ an experienced dermatologist is always involved in reviewing images, and
	  ○ treatment decisions are always taken by a senior dermatologist
	• Software updates, delays and crashes could cause delays in access to treatment (if the clinician was used to getting reassurance from the software)
	• Possible delayed uptake of new AI-related technology if it is not trusted or understood

Patient and carer considerations (3)

Equality issues

AI technologies may lead to indirect discriminations/inequalities, for example:

- The technology may find it difficult to differentiate lesions from normal skin in certain skin tones, making it more or less effective for different ethnic groups;
- AI technology availability is likely to be clustered in urban areas, meaning people living in rural areas may not be able to be checked so regularly, or choose not to travel for imaging;
- The technology is ineligible for children under 18 years old;
- AI technologies are aimed at people with <3 lesions, but older individuals will likely have more than this, making them ineligible;
- Some people may prefer touch as opposed to technology - and may have an inherent distrust of the outcome.

Clinical perspectives

Submissions received from British Association of Dermatologists:

The pathway of care not clearly defined, and a clear use case definition is crucial for applications used in clinical pathways.

When second read is removed the key question will be how the settings are dialed… a high sensitivity for melanoma needs to match with a high specificity or there is a risk of increased referrals rerouted to secondary care, thus increasing workload

There is a need for rigorous clinical validation through studies involving diverse populations to confirm its effectiveness in various skin types, ethnicities, and lesions times

It is time-consuming, involves additional steps in the pathway, and unless deployed as a direct diagnostic tool, it doesn't contribute value to the pathways

AI will not give an alternative diagnosis, but a GP appointment may be able to 'see it, sort it' in the same appointment for an alternative diagnosis.

Challenging for patients who cannot leave their home or are immobile (nursing home patients) – how will they have their photo taken in the TD hub. This population may be more likely to have skin lesions of concern

EAG's analysis shows a 50% reduction in referral rates with DERM but human behaviour of GPs over-relying on AI may increase referral rates to secondary care. Consultants may end up reviewing more cases in the end.

NIC

Equality considerations

The EAG notes several equality concerns arising from their review

NICE is committed to promoting equality of opportunity, eliminating unlawful discrimination and fostering good relations between people with particular protected characteristics and others.

- The evidence base for both technologies included few patients with non-white ethnicity or darker skin tones. Since skin cancer may be harder to detect in these people this is of concern. It is unclear whether the AI tools have been properly validated in people with darker skin tone, and the resulting diagnostic accuracy. Differences in diagnostic accuracy could lead to inequalities due to different diagnostic pathways, such as if some people have to wait for a face-to-face appointment because an AI assessment was inconclusive.
- DERM could not be used for a substantial number of patients, due to lesions being too large to assess; lesions being in areas with tattoos, scarring or hair covering; or lesions being on parts of the body unsuited to assessment with a dermatoscope. This could potentially cause inequalities due to resulting differences in diagnostic pathways and access to diagnostic services.
- Use of AI could improve access to skin cancer diagnosis as it may reduce the need for face-to-face appointments, so reducing patient time commitment and need to travel to appointments.

Clinical effectiveness

NICE National Institute for Health and Care Excellence

Objectives

To investigate the clinical effectiveness of AI technologies as decision aids to triage and diagnose suspicious skin lesions, specifically for DERM and Moleanalyzer pro, the following objectives were proposed by the EAG:

Clinical effectiveness
• To perform a rapid systematic review, and if feasible a meta-analysis, of the diagnostic accuracy of the included AI technologies • To perform a rapid systematic review with a narrative synthesis of the clinical impact and practical implementation of the AI technologies • Based on the results of the rapid review, to identify evidence gaps and formulate recommendations for future research.

In addition to looking at diagnostic accuracy of benign lesions it is important to assess diagnostic accuracy to detect malignant lesions to ensure the technologies are safe to use and false negative results are minimised. There is a trade-off between high sensitivity to detect malignant lesions and high specificity to reduce unnecessary referrals of benign lesions.

Overview of included studies

Six studies were included in the review plus 13 in the evidence map

DERM

Included in the review			
Study	**Total cases (N)**	**Setting**	**Rationale**
DERM-003	572	Hospital (UK)	prospective study design
DERM-005 (unpublished)	[redacted]	[redacted]	
Thomas 2023 (UHBFT & WSFT)	14,500	Conducted in TD hub for triage within 2WW pathway (UK)	
Leicestershire (unpublished)	[redacted]		Before-and-after pilot

Included in evidence map only (excluded from main review)		
Phillips 2019	Hospital (UK)	Older version of DERM for detection of melanoma only
Phillips 2020	NA	Retrospective design and evaluated older version of DERM for evaluation of melanoma only

Other DERM evaluations ongoing across UK in post- and pre-referral settings. Outcome data expected in fourth quarter of 2024.

Characteristics of participants:

- Majority of participants were white and very few had darker skin (Fitzpatrick types IV-VI)
- Lesions mostly on face and scalp, followed by chest/back
- None reported for Leicestershire study

MoleAnalyzer pro

Included in the review			
Study	**Total cases (N)**	**Setting**	**Rationale**
MacLellan 2021	184	Secondary care (Canada)	prospective study design
Winkler 2023	188	Secondary care (Germany)	

Included in evidence map only (excluded from main review)	
Fink 2019	Retrospective study design
Haenssle 2018, 2020	
Kommoss 2023	
Sies 2020, 2021, 2022	
Winkler 2020, 2021a, 2021b, 2022	

Characteristics of participants:

- Majority had lighter skin colours (Fitzpatrick types II-III)
- Lesions were most often on the trunk, followed by extremities

Study characteristics:

- All studies assessed diagnostic accuracy for melanoma only (excluded all non-melanocytic lesions)

Study quality of DERM

Risk of bias and applicability concerns for DERM

Study	Test	Risk of bias				Applicability concerns		
		Patient selection	Index test	Reference standard	Flow and Timing	Patient selection	Index test	Reference standard
DERM 003	DERM	✗	✓	? Insufficient details	✓	✗	? Older version of DERM and outdated lens used	✓
DERM 005	DERM	✗	✓/✗ Pre-specified thresholds	✓	✓	✗	? Older version of DERM and outdated lens used	✓
Thomas 2023	DERM	✗	✓/✗ Post-specified thresholds	✓	✓	✗	? Older version of DERM used	✓
DERM 003	Dermato.	✓	✓	✓	✓	✗	✗	✓
DERM 005	Dermato.	✓	✓	✓	✓	✗	✗	✓
DERM 003	DERM vs. dermato.	✓	✓	?	✓	n/a	n/a	n/a
DERM 005	DERM vs. dermato.	✓	✓/✗	✓	✓	n/a	n/a	n/a

✓ indicates low risk; ✗ indicates high risk; ? indicates unclear risk; Dermato: dermatologist assessment

EAG: All studies have high risk of selection bias. This raises concern regarding applicability of their populations due to high exclusion rates.

NICE

EAG: Older versions of DERM used in all 3 studies, which among other elements, includes a different set of thresholds for sensitivity and specificity. Two studies also used outdated teledermoscopy lenses. Hence, applicability of diagnostic accuracy results for DERM to current practice is uncertain.

Study quality of MoleAnalyzer pro

Risk of bias and applicability concerns for MoleAnalyzer pro

Study	Test	Risk of bias				Applicability concerns		
		Patient selection	Index test	Reference standard	Flow and Timing	Patient selection	Index test	Reference standard
MacLellan 2021	Moleanalyzer	✗	? Threshold for positive diagnosis NR	✓	? Insufficient information	✗ Exclusion of non-melanocytic lesions	✓	✓
Winkler 2023	Moleanalyzer	✗	✓	✓	? Insufficient information	✗ Exclusion of non-melanocytic lesions	✓	✓
MacLellan 2021	Dermato.	✗	✓	✓	?	✗	✗ Model of dermatoscope out of date	✓
Winkler 2023	Dermato.	✗	✓	✓	?	✗	? Model of dermatoscope NR	✓
Maclellan	Teledermato.	✗	✓	✓	?	✗	✗	✓
Maclellan	Moleanalyzer vs. dermato & teledermato.	✗	?	✓	?	n/a	n/a	n/a
Winkler	Moleanalyzer vs. dermato.	✗	✓	✓	?	n/a	n/a	n/a

✓ indicates low risk; ✗ indicates high risk; ? indicates unclear risk; Dermato: dermatologist assessment

NICE

EAG: All studies have high risk of selection bias due to exclusion of participants with non-melanocytic lesions and Fitzpatrick skin types higher than III.

EAG: Model of dermatoscope NR in Winkler 2023 and was out of date in MacLellan 2023. This raises concerns on applicability of dermatologist assessments.

Diagnostic accuracy of autonomous DERM

Published and unpublished studies included in EAG's assessment report

DERM-003 (Marsden et al (2023))

Evaluated diagnostic accuracy of autonomous DERM, compared to standard of care (F2F dermatologist assessment without AI use).
Dermoscopic images taking with 3 different smartphone cameras with DL1 lens.

Thomas et al. (2023) (Birmingham and West Suffolk)

Evaluated diagnostic accuracy of DERM used as a triage tool within the teledermatology pathway at UHB and WSFT centres. Two versions of DERM (vA and vB) assessed.

DERM 005 (Chelsea & Westminster)

[redacted]

Leicestershire (UHL)

[redacted]

[redacted]

Diagnostic accuracy of autonomous DERM

DERM has higher sensitivity but lower specificity than dermatologists

Meta-analysis of sensitivity and specificity for malignant lesions performed for where 2 or 3 studies reported data.

- Sensitivity for autonomous use of DERM is higher than ‘standard diagnostic pathway without DERM*’ but some malignant lesions will still be missed
- Specificity for autonomous use of DERM is lower than dermatologists (standard practice)
    - Specificity particularly lower for detecting SCC and BCC, suggesting DERM has difficulty in distinguishing these from benign lesions
    - Lower specificity may be a result of using an older version of DERM but this is not certain

Study	Index test	Cancer	Sensitivity	Specificity
All	Autonomous DERM	Any malignancy	[redacted]	[redacted]
All	Autonomous DERM	Melanoma	[redacted]	[redacted]
DERM 003 & 005	Standard of care	Any malignancy	[redacted]	[redacted]

*‘Standard diagnostic pathway’ may refer to a mix of F2F dermatologist and teledermatologist review

Diagnostic accuracy of autonomous DERM

Diagnostic accuracy for benign lesions

In DERM-003, dermatologists achieved a much higher specificity for malignant lesions and were more accurate at identifying benign lesions than DERM.

Study	Index test	Lesions (N)	Outcome	Sensitivity	Benign misclassified as non-benign	Specificity	Non-benign misclassified as benign *(these were mostly pre-malignant lesions)*
DERM-003	DERM [redacted] (iPhone 11)	571	Benign	43.9% (95% CI 37.4–50.6)	56%	93.3% (95% CI 90.0–95.6)	7%
	DERM [redacted] (iPhone 6s)	578		39.3% (95% CI 33.0–46.0)			
	DERM [redacted] (Samsung)	578		48.3% (95% CI 41.7–54.9)			
	Dermatologist alone	581		73.9% (95% CI 67.6–79.4)	26%	93.7% (95% CI 90.5-95.9)	6%

In DERM-005, the DERM sensitivity threshold was set at [redacted]

Study	Index test	Lesions (N)	Outcome	Sensitivity
DERM-005 (Chelsea & Westminster) Under review for publication	[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]		[redacted]

NICE

Sensitivity for benign lesions: Benign lesions correctly diagnosed

Specificity for benign lesions: malignant lesions correctly diagnoses

Diagnostic accuracy of autonomous DERM

Diagnostic accuracy for benign lesions

In the Thomas et al (2023) published study, DERM vB had specificity greater than the previous version (DERM-vA).

The second-read reviewer overturned 40–50% of cases that DERM had marked as eligible for discharge; however, for DERM-vA, only 1.2% of these cases resulted in skin cancer diagnosis and with DERM-vB, none resulted in a skin cancer diagnosis.

Study		Index test	Lesions (N)	Outcome	Sensitivity	% of overturned cases by second reviewer out of all cases eligible for discharge by DERM	% of skin cancers found by hospital dermatologist among overturned cases
Thomas et al. (2023)	Birmingham (UHB)	DERM-vA - Teledermatologist	2851	Benign	49.4% [47.6–51.2%]	36%	2%
		DERM-vB - Teledermatologist	1815		73.4% [71.4–75.4%]	39.6%	0%
	West Suffolk (WSFT)	DERM-vA - Teledermatologist	376		40.7% [35.8–45.7%]	48.3%	1.2%
		DERM-vB - Teledermatologist	394		70.1% [65.4–74.4%]	49.2%	0%

Diagnostic accuracy of autonomous DERM (unpublished data)

DERM may misclassify BCC as more serious malignancies

Data overlaps with data from publications, but appear more up-to-date. In comparison to the published data, the EAG assumed this data is more detailed – enabling a more thorough analysis of autonomous use of DERM.

For detecting 'all malignant' lesion, DERM has a high sensitivity but low specificity.

For detecting 'exact' classification of lesions, the sensitivity for detecting melanoma remains near 95% but decreases substantially for SCC and BCC. This suggests both SCC and BCC lesions may be misclassified as more serious malignancies.

DERM classification	Sensitivity	Specificity
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	

[redacted]

Exact analysis: DERM result exactly matches ground truth diagnosis;
All malignant: A malignant diagnosis by DERM matches ground truth diagnosis of any malignant lesion, even if not in the correct category

Diagnostic accuracy of autonomous DERM (unpublished data)

Data on benign lesions from Birmingham & London centres

Study	Index test	Lesions (N)	Outcome	Sensitivity	Benign misclassified as non-benign	Specificity	Non-benign misclassified as benign *(these were mostly non-malignant lesions)*
[redacted]	[redacted]		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Note:

The reference standard in this analysis was usually a “ground truth” diagnosis made by dermatologists where the lesion was judged to be non-malignant. Therefore, the diagnostic accuracy of DERM may be slightly incorrect as some genuinely malignant lesions may have been incorrectly classified as benign by dermatologists. This also means that estimates of the diagnostic accuracy of dermatologists without DERM may not be reliable.

Diagnostic accuracy of DERM

Subgroup data by skin type

Two studies reported separate diagnostic data for Fitzpatrick skin types V and VI.

Thomas 2023:

- of 159 lesions assessed, 94 lesions had a final diagnosis, including BCC (n = 1) and IEC (n = 1), and actinic keratosis (n = 1), all correctly referred by DERM in both versions of DERM
- Three atypical nevus were pending face-to face assessment, and the remainder were benign with a benign specificity of 44.3% (39/88).

DERM 003 found no Fitzpatrick skin types V and VI.

Intraepidermal Carcinoma, IEC

Diagnostic accuracy of DERM

Diagnostic accuracy of DERM used in a full TD pathway is unknown

The unpublished Edge Health report on the Leicestershire study included some data DERM assessment, followed by dermatologist assessment (second read).

Data suggests:

- A second read classed as benign by DERM [redacted] when compared to using autonomous DERM
    - This suggests that using a second read for lesions classed as benign by DERM could [redacted] After a final teledermatology assessment this would [redacted]
- The sensitivity is uncertain due to lack of a perfect reference standard. However if the sensitivity of autonomous DERM is 95% then use of a second read [redacted] based on the Leicestershire data.

Diagnostic accuracy of MoleAnalyzer pro

Poorer sensitivity, but higher specificity for detecting melanoma than F2F

No data reported for other skin cancer types, or for premalignant and benign lesions.

Meta-analysis (2 studies)	Sensitivity	Specificity
Moleanalyzer pro alone	84.4% (95% CI 73.9 - 91.0)	84.5% (95% CI 72.0 - 92.1)

Winkler (2023)	Sensitivity	Specificity
Moleanalyzer pro alone	81.6 (66.6-90.8)	88.9 (83.7-92.7)
Dermatologist alone (F2F)	84.2 (69.9-92.6)	72.1 (65.3-78.0)
Moleanalyzer pro w/ dermatologist input	100.0 (90.8-100.0)	83.7 (77.8-88.3)

MacLellan (2021)	Sensitivity	Specificity
Moleanalyzer pro alone	88.1 (79.4-96.9)	78.8 (71.5-86.2)
Dermatologist alone (F2F)	96.6 (91.9-100)	32.2 (18.4-46.0)
Teledermatologist alone (remote)	89.8 (79.6-96.2)	66.0 (57.8-73.5)

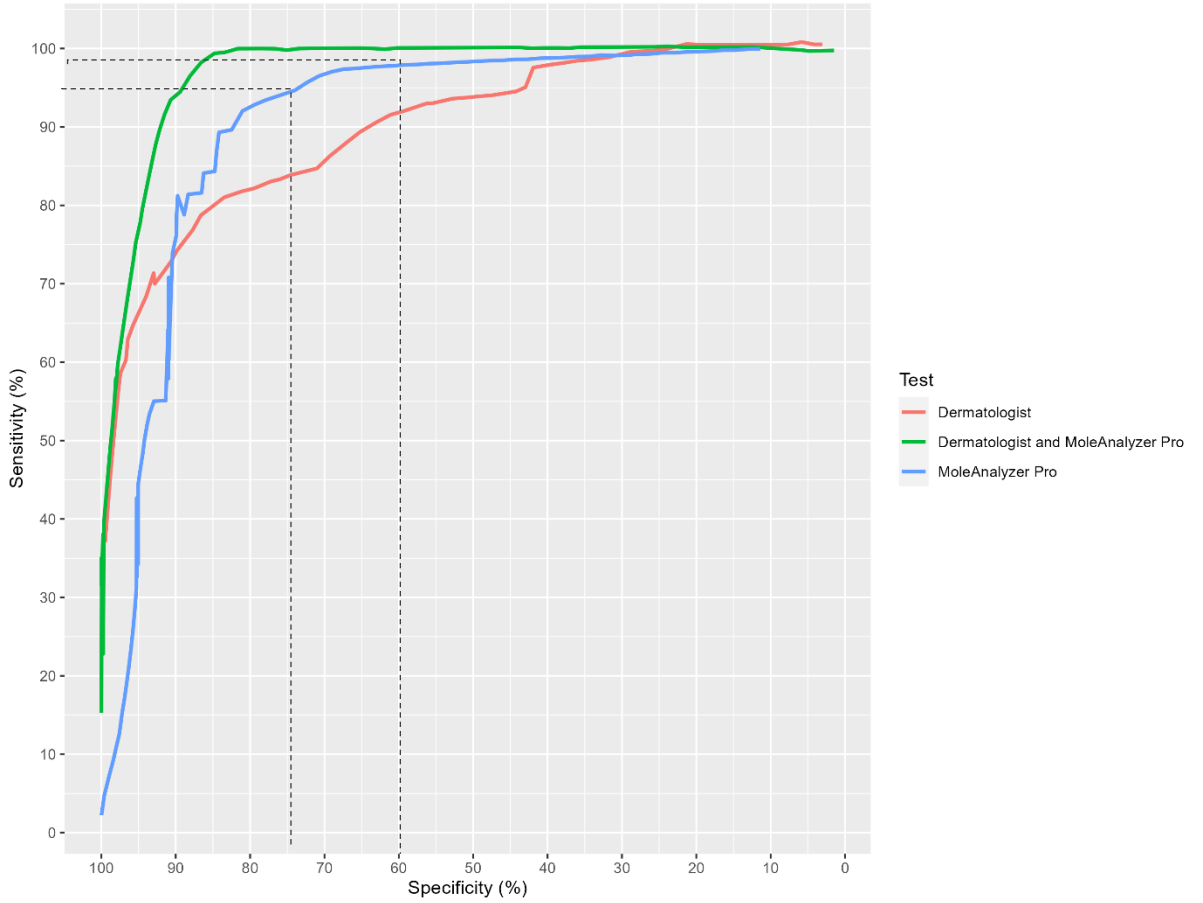
Diagnostic accuracy of MoleAnalyzer pro

MoleAnalyzer pro could achieve similar diagnostic accuracy as DERM

The estimated sensitivity is lower than observed for DERM, but the ROC suggests than Moleanalyzer pro could achieve a specificity of around 60 to 75% at a sensitivity of over 95%, which is similar to that observed for DERM.

Test	Cancer	Sensitivity	Specificity
DERM	Any malignant	96.1%	65.4%
DERM	Melanoma	~95%	~62.5%

ROC curves for Moleanalyzer pro for melanoma diagnosis (adapted from Winkler 2023)



NICE

Clinical outcomes of DERM

Eligibility rates

Thomas 2023 reported DERM-vB data on the diagnostic pathway at UHB and WSFT centres. Use of DERM varied by location (varied rate of use, and referral rates).

A substantial number of lesions could not be assessed by DERM (25% and 17%). These patients had to be referred directly to F2F assessment.

[redacted]

Clinical outcomes of DERM

Reduced referral rates with autonomous DERM

EAG's analysis estimated that autonomous use of DERM could halve (reduced by 47.7%) all referrals (urgent and routine combined) to a dermatologist (among lesions that can be assessed by DERM).

- Of all referrals made (urgent and routine), most would be false positive (64% of all referrals being benign lesions).
- Among urgent referrals, majority (~85%) would be false positive.
- Routine referrals would be uncommon (~9% of total DERM population), because BCC cases are overdiagnosed.

Referral group			% of total DERM population
All referrals (urgent and routine)	Total		47.7%
	Urgent	Total	39%
		False positives	33.1%
		False negatives	0.3%
		Underdiagnoses	0.1%
	Routine	Total	8.7%
		False positives	4.9
		False negatives	0.6%
		Overdiagnoses	7.4%

Thomas (2023) reported rates of referrals between two sites:

Of all patients/cases (including those not assessed by DERM), 44% and 62% were referred to be seen as a F2F dermatologist assessment in Birmingham and West Suffolk, respectively.

		Birmingham	West Suffolk
Referred to dermatologist by DERM	Total	44%	62%
	Malignant lesions	7.5%	9.7%

Impact on resource use and timings for DERM

DERM potentially increases waiting times for urgent skin cancer referrals

In the Leicestershire (UHL) study:

[redacted]

Findings reported from Leicestershire (UHL) study on AI-Teledermatology pilot:

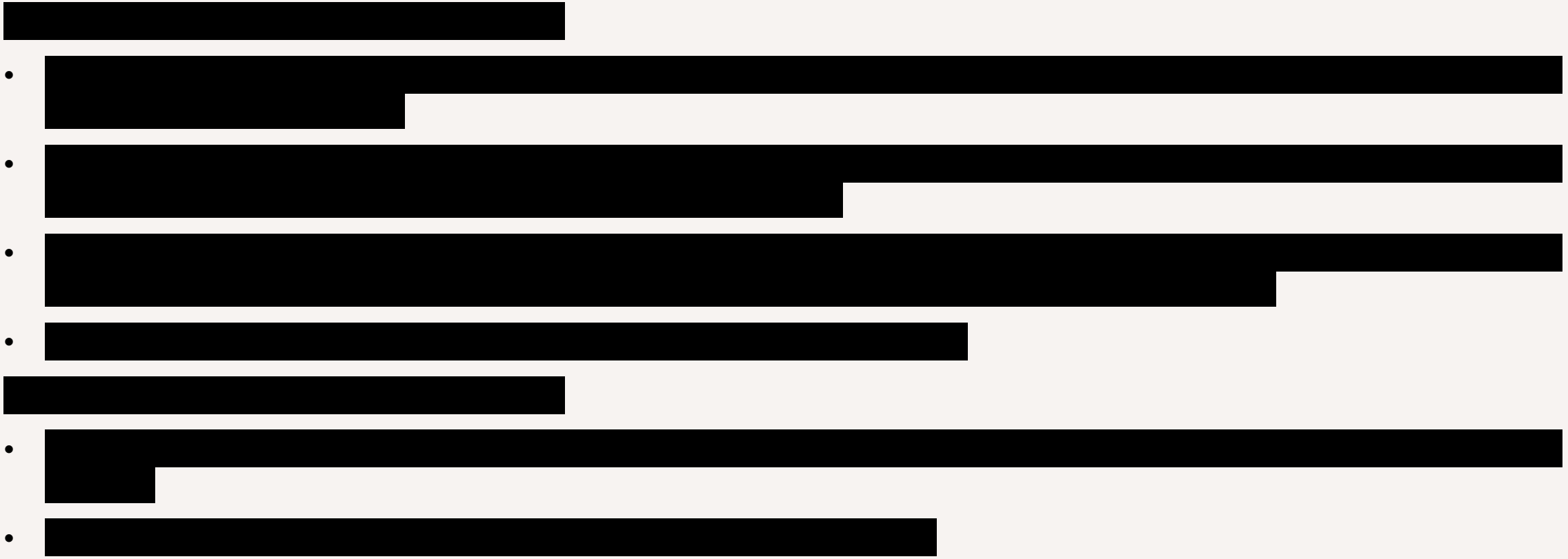
- [redacted]
- [redacted]
    - [redacted]
- [redacted]

Impact on resource use and timings for MoleAnalyzer pro

Impact of Moleanalyzer pro on referral rates

- In Winkler (2023), the integration of Moleanalyzer into decision making reduced the rate of unnecessary excisions of benign nevi by 19.2% ($p<0.001$).
- The rate of excision of malignant lesions was not significantly changed ($p>0.99$).
- The percentage of nevi managed by follow up examinations was increased with the integration of Moleanalyzer pro results into decision making (from 37.9% to 44.7%, $p=0.053$).
- The EAG did not identify any other evidence from Moleanalyzer pro studies included in the synthesis on implementation, resource use, and related outcomes.

Acceptability to healthcare professionals and patients



The use of Moleanalyzer pro was generally supported by both clinicians and patients, and its results were trusted, however, most patients indicated that they would like the opinion of an expert physician besides an AI-assisted diagnosis.

Acceptability to patients

DERM-005 patient survey responses:

- Participants generally responded positively when considering AI as a tool to help doctors, but more cautiously when considering the use of AI to replace a dermatologist
- most would rather have their lesion assessed by a computer than waiting weeks to see an in-person dermatologist
- patients generally indicated they felt comfortable with the use of AI and the dermoscopic images required, but there was a mixed response to a statement on preference for a face-to-face dermatologist appointment.

Key issues

Key issue 1: Limited evidence base on diagnostic accuracy

Background

Only 3 DERM and 2 MoleAnalyzer studies that prospectively evaluated diagnostic accuracy, and there are limitations in the evidence base to inform the relative value of implementing AI in the current pathway for identification of benign lesions.

- Lack of evidence on teledermatology +/- AI
- Unclear how autonomous use of DERM compares to DERM with 'second-read' dermatologist assessment
- Lack of evidence on latest version of DERM used in the UK
- Teledermoscopy lenses used in DERM studies are out of date
- MoleAnalyzer pro studies did not evaluate the accuracy of AI for detecting non-melanoma skin cancers
- MoleAnalyzer pro studies were also not explicitly in teledermatology settings nor based in the UK

Potential risks

The applicability of diagnostic accuracy results to current practice cannot be determined
Unclear whether the findings for melanoma could be adapted to detect all forms of skin cancer, or if not, how a melanoma-only tool would be used in practice.
Unclear what clinical benefits it could have within NHS practice or as part of a teledermatology programme.

Key issue 2: Lack of evidence on teledermatology

Background

- Comparable diagnostic accuracy data describing current service provision is lacking, particularly for the teledermatology pathway.
- Much of the evidence compared DERM as part of the teledermatology pathway to face-to-face dermatology.

Potential risks

Without comparative evidence on the diagnostic accuracy of AI technologies and teledermatology, their relative value for safe and cost-effective identification of benign lesions will remain unclear.

Key issue 3: DERM has lower specificity than dermatologists

Background

Specificity for autonomous use of DERM for detection of malignant lesions is lower than for dermatologists.

	Specificity from meta-analysis	Specificity from unpublished data
DERM alone	63.8% (95% CI 48.0 to 77.1)	65.4% (95% CI 64.7 to 66.1).
Dermatologists alone	85.7% (95% CI 66.7 to 94.7)	NR

Specificity was much lower for detecting SCC and BCC with autonomous use of DERM.

Potential risks

Many people with benign lesions will still be referred for a F2F appointment

Potential benefits

Fewer missed cancers with autonomous use of DERM than current practice.

Key issue 4: Fewer missed malignant cases with DERM use

Background

Autonomous use of DERM has a similar or higher sensitivity compared with dermatologists for detecting any malignant lesions.

	Sensitivity from meta-analysis	Sensitivity from unpublished data
DERM alone	97.8% (95% CI 93.1 to 99.3)	96.1%(95% CI 95.4 to 96.8)
Dermatologists alone	90.6% (95% CI 78.7 to 96.1)	NR

Potential risks

There is still a risk of missed malignant lesions with autonomous use of DERM.
Overdiagnoses of SCC and BCC.

Potential benefits

Fewer missed cancers with autonomous use of DERM than current practice.

Key issue 5: Identification of benign lesions for discharge

Background

Sensitivity and specificity of benign lesions in DERM 003 (published study)

	Sensitivity	Specificity
DERM alone	43.9% (95% CI 37.4 to 50.6)	93.3% (95% CI 90.0–95.6)
Dermatologists alone	73.9% (95% CI 67.6–79.4)	93.7% (95% CI 90.5-95.9

Sensitivity and specificity of benign lesions from unpublished data:

	Sensitivity	Specificity
DERM alone	71.5% (95% CI 70.7 to 72.3)	86.2% (95% CI 85.4 to 87.0)

Potential risks

Poor sensitivity for benign indicate large number of benign cases will be misclassified as non-benign cases.

Key issue 6: Proportion of patients unsuitable for DERM

Background

Considerable proportion of patients were unsuitable for DERM assessment, hence would need F2F referrals:

- Birmingham: 25%
- West Suffolk: 17%
- [redacted]

Note: virtual teledermatology cannot be used for lesions on difficult sites (palms, soles, scalp, intimate areas), for people with multiple lesions, and for children. This does have some overlap with the exclusions for DERM.

Potential risks

Different proportions of patients are suitable for each technology, with current eligibility criteria more restrictive for the use of AI than for teledermatology. Patients unsuitable for AI would need a F2F referral which has higher associated costs.

There are also additional costs associated with unsuccessful photography/ an indeterminate AI result.

Key issue 7: DERM use could half number of all 2WW referrals

Background

Unpublished data suggests that among eligible lesions, the autonomous use of DERM could approximately half the number of all referrals (routine and urgent) that would normally be sent for a F2F or teledermatology assessment. (see table 7 in assessment report)

Potential benefit

As a result of reducing referrals by 50%, a small number of malignant lesions (0.8%) would be incorrectly discharged, but these are mostly BCCs.

This may be beneficial as false negatives also occur in current practice, particularly if these false negatives are mostly BCC cases.

Key issue 8: Limited data on darker skin types & ethnicities

Background

It is unclear whether the AI tools have been properly validated in people with darker skin tones and non-white ethnicities, and the resulting diagnostic accuracy in these groups.

- DERM: Most patients included in the diagnostic accuracy data had light coloured skin (Fitzpatrick skin types II-III) and were from white ethnic backgrounds. There was limited accuracy data on darker skin types (Fitzpatrick IV-VI).

- MoleAnalyzer pro: One study (McLellan) excluded patients with Fitzpatrick skin types higher than III.

Potential risks

This raises concern about the applicability of AI technologies to practice and safety/effectiveness of AI diagnoses within these populations.
Differences in diagnostic accuracy could lead to inequalities due to different diagnostic pathways, such as if some people have to wait for a face-to-face appointment because an AI assessment was inconclusive.

Key issue 9: Limited data on DERM used with second read

Background

Data suggests that a second read of lesions classed as benign by DERM identifies an additional [redacted] (that would have been missed by autonomous DERM use [redacted]

This suggests [redacted]

After a final teledermatology assessment, [redacted]

Potential risks

Higher rate of false positives will occur with lowered specificity second read DERM, resulting in more costly F2F appointments.

Potential benefits

Second read could [redacted] meaning more cancers detected earlier

Key issue 10: [redacted]

Background

Median time from referral to face-to-face outpatient appointments was [redacted] for patients who were on the traditional pathway.

Potential risks

[redacted]

Key issue 11: Patients/clinicians resistant to autonomous use

Background

Patient opinion was broadly supportive of using DERM in some form as part of diagnosis. Clinicians were generally very resistant to using DERM in isolation without human assessment of lesions.

Potential risks

- Low acceptance rates by healthcare community may see hesitancy to adoption.
- Patients may feel more anxious about being discharged by autonomous AI.

Cost considerations

Objectives

To investigate the cost effectiveness of AI technologies as decision aids to triage and diagnose suspicious skin lesions, specifically for DERM and Moleanalyzer pro. The following objectives were proposed:

Cost effectiveness
• To perform a rapid systematic review of published cost-effectiveness studies of alternative diagnostic strategies used to aid the diagnosis of skin cancer. This will focus on the included AI technologies but will also include alternative strategies if no evidence is identified for the included technologies. • To develop a conceptual model that will identify likely drivers of health benefit, harms and cost associated with implementing the included AI technologies in the NHS. • If evidence and time allows: to develop a budget impact model capturing the direct resource implications of implementing the included AI technologies in the NHS. This may additionally include threshold analysis to explore how health effects or indirect costs may impact cost-effectiveness.

Cost effectiveness studies

Cost-effectiveness studies on skin cancer diagnosis in the NHS, relevant to the development and parameterisation of the conceptual model.

1. Wilson et al. (2013)
2. Edwards et al. (2016)
3. Wilson et al. (2018)

Unpublished reports for DERM

1. Leicestershire (UHL) pilot: evaluation of DERM implemented in 2WW pathway
2. U of Exeter: preliminary report describing *de novo* cost-utility model
3. NHSE AI Award group: draft report on economic analyses

Cost-utility model by Exeter Test Group for DERM

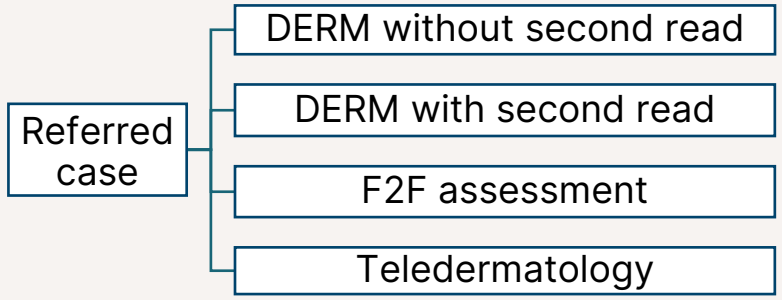
The company analysis aligned with the scope of the EVA

Decision problem: Triage of patients referred from primary care via the urgent skin cancer referral pathway.

Model assesses 2 ways of implementing of DERM in this setting:

1. DERM with a second read
2. DERM without a second read

Comparators used in model: F2F assessment and teledermatology.



There are 4 possible diagnostic pathways represented by decision trees.

5 Markov models represent the differing prognoses from each pathway of the decision tree.

Modelled population and structure used: 3 diagnostic categories used. Each has a distinct diagnostic accuracy and associated treatment costs.

- 5.9% of patients assumed to be high-risk cancers (melanoma, SCC and other high-risk cancers);
- 6.9% assumed to have BCC;
- 87.2% pf patients assumed to have low-risk lesions (precancerous or benign)

Cost-utility model results

Company model finds that autonomous DERM generates cost savings

	DERM without second read	DERM with second read	Teledermatology	Comments
Eligibility for assessment	81%	81%	90%	• DERM has higher cost than teledermatology. • Fewer lesions are eligible for DERM than for teledermatology. These lesions would need a more expensive F2F assessment.
First line assessment cost per patient (avg.)	£72	£72	£57	
Specificity	42%	Lower than DERM without second read	35% *(Cochrane review reported 84.3%)*	• Assumed specificity of teledermatology appears low compared with published sources
Effective discharge rate	36.9%	15.7%	30.9%	• Higher discharge rate lead to fewer F2F appointments and biopsies. • Second read dermatologists appear very cautious and overturn a large number of lesions marked as benign by DERM and re-route to see a F2F dermatologist. These would have otherwise been discharged by DERM without second read.
Total avg. cost of peri-referral pathway	£118	£172	£146	• The higher discharge rate for DERM without a second read offsets first line assessment costs, which generates cost savings compared with teledermatology. • DERM with a second read is the costliest approach but may be associated with non-cash releasing benefits related to outsourcing of teledermatology review to Skin Analytics consultants.

Cost-utility model by Exeter Test Group for DERM

Company model finds that DERM dominates teledermatology and usual care

			Incremental (vs usual care)		Incremental (vs teledermatology)		
Strategy	Cost (£)	QALY	Cost (£)	QALY	Cost (£)	QALY	ICER
DERM + second read	£465.84	11.1925	-£31.14	+0.0077	-£6.27	+0.0039	£24,655.23
DERM	£445.09	11.1917	-£51.89	+0.0069	-£27.02	+0.0031	-
Teledermatology	£472.11	11.1886	-£24.87	+0.0038	-	-	Strictly dominated
Usual care	£496.98	11.1848				N/A	Strictly dominated

EAG: It remains highly uncertain whether currently available diagnostic accuracy evidence is sufficient to reliably populate a cost-utility model, particularly with regards to the specificity of AI technologies compared with the specificity of an effectively implemented teledermatology service.

EAG: Whilst this analysis predicted that DERM with or without a second read would dominate all other options, this was highly dependent on the relative specificity of teledermatology. If the Cochrane diagnostic accuracy values are applied for teledermatology, DERM strategies become more costly than teledermatology. Teledermatology also becomes cost saving versus the traditional pathway.

Cost-utlility model by Exeter Test Group for DERM

Company model finds that DERM dominates teledermatology and usual care

DERM (with or without second read) vs teledermatology
Both DERM strategies are cost-saving relative to teledermatology. Fewer patients are eligible for assessment by DERM and require more expensive F2F assessment. However, this is offset by a higher effective discharge rate associated with: • Autonomous DERM (generates cost savings as fewer F2F appointments and fewer biopsies conducted) • DERM with second read (generates cost savings through avoidance of missed diagnoses)   • may be the most costly approach, but may be associated with non-cash releasing benefits related to outsourcing of teledermatology review to Skin Analytics consultants

DERM (with or without second read) vs F2F assessment
Both DERM strategies are cost-saving relative to F2F assessment. Cost savings driven by avoiding unnecessary referrals to F2F appointments and inappropriate biopsies.

DERM with second read vs autonomous DERM
Autonomous DERM is cost-saving relative to DERM with second read. Driven by avoiding costs associated with second read, however partially offset by the lower rate of missed diagnoses.

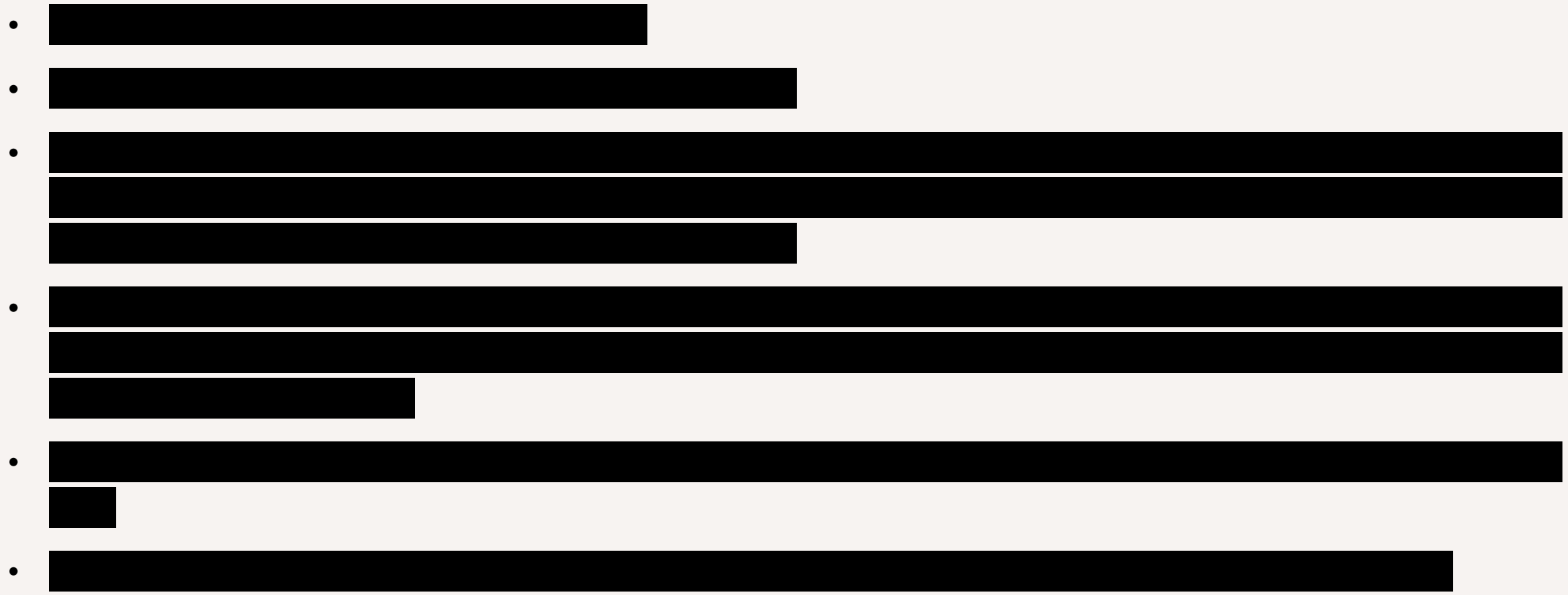
Cost-utility model by Exeter Test Group for DERM

EAG critique

Assumption	EAG critique
BCC does not progress if not diagnosed, i.e. no health consequences of a missed BCC	DERM is less sensitive than teledermatology and F2F assessment for BCC diagnosis. The model punishes correct BCC diagnoses, as excision is associated with accrual of costs and a QALY decrement. Counterintuitively means there is greater value generated by having lower sensitivity.
Specificity of teledermatology is 35%	Higher specificity means fewer unnecessary and expensive biopsies and F2F consultations. Specificity of teledermatology reported in published sources (Cochrane review) is substantially higher that that observed in pilot sites (which were largely not set up for teledermatology).
Biopsy and treatment costs	These appear high relative to other studies and were not consistently based on NHS Reference costs/ PSSRU costs. Punishes diagnostic strategies with lower specificity and may inflate the potential cost savings associated with a higher specificity.
The sensitivity of F2F assessment increases to 99% following triage with DERM or teledermatology.	The introduction of a triage step reduces missed diagnoses and avoids associated cost and health implications. This results in better cost-effective estimates for DERM and teledermatology. The plausibility of this assumption is not clear.

East Midlands Academic Health Science Network

Evaluation of a pilot of DERM with second read at Leicestershire sites



NHSE AI in Health and Care Economic Evaluations

[redacted]

[redacted]

- [redacted]
- [redacted]
- [redacted]

[redacted]

[redacted]

[redacted]

EAG's conceptual model

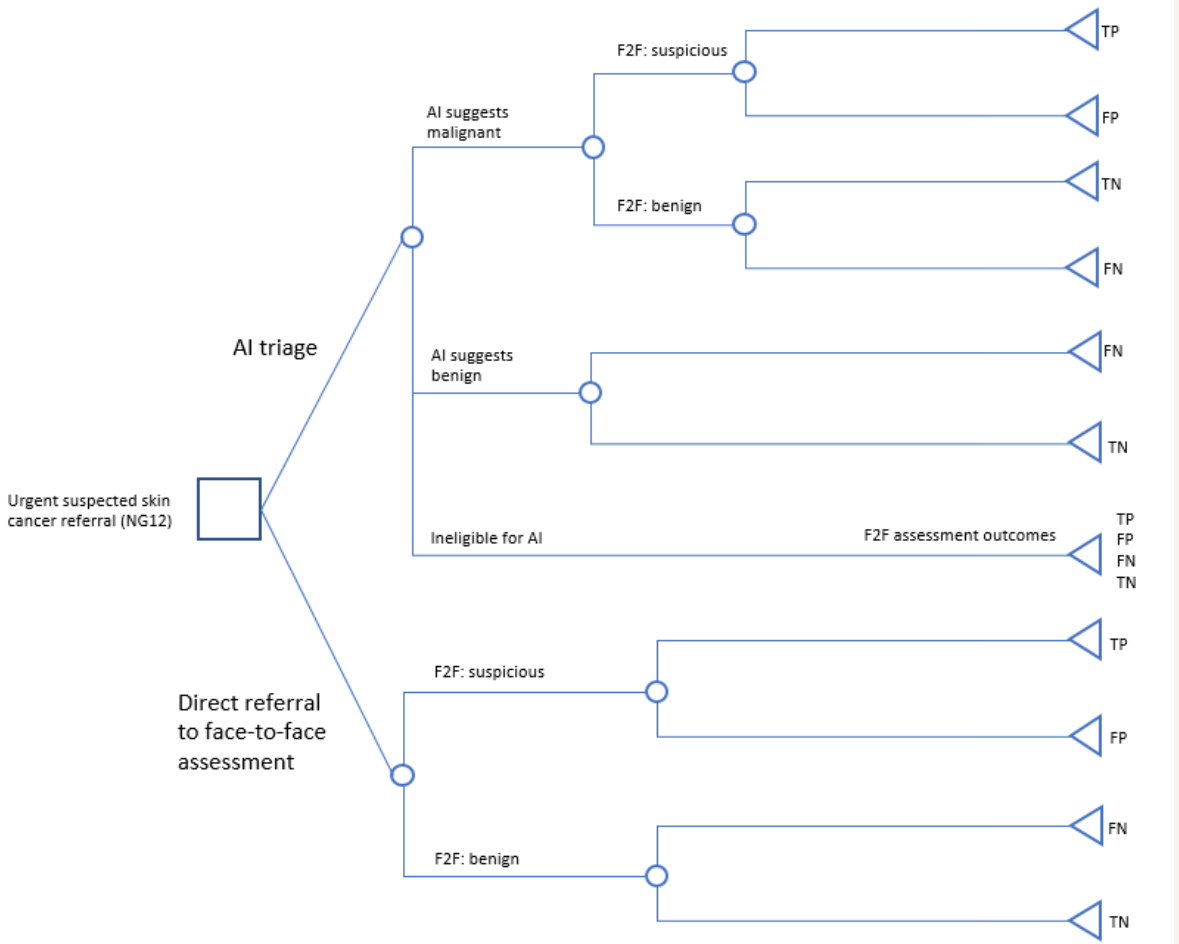
Model framework

Decision problem	The conceptual model considers the following use cases for the identification of benign lesions in all patients referred on the urgent skin cancer pathway from primary care: 1. Autonomous AI triage 2. AI triage with second read prior to discharge
Comparators	• Teledermatology model • Conventional referral to F2F assessment model
Markov model categories	Diagnostic categories have distinct long-term consequences (represented by Markov models): • 'High-risk cancers', including melanoma, squamous cell carcinoma, and other rare high-risk cancers; • basal cell carcinoma; • low-risk lesions (benign and precancerous). Concern regarding technologies having limited experience of rare cancers – uncertain if the high sensitivity for melanoma and SCC is maintained across the rarer indications. Treating them as a single diagnostic category is subject to uncertainty. Where possible, sensitivity analysis should be undertaken in which rare cancers are categorised separately.

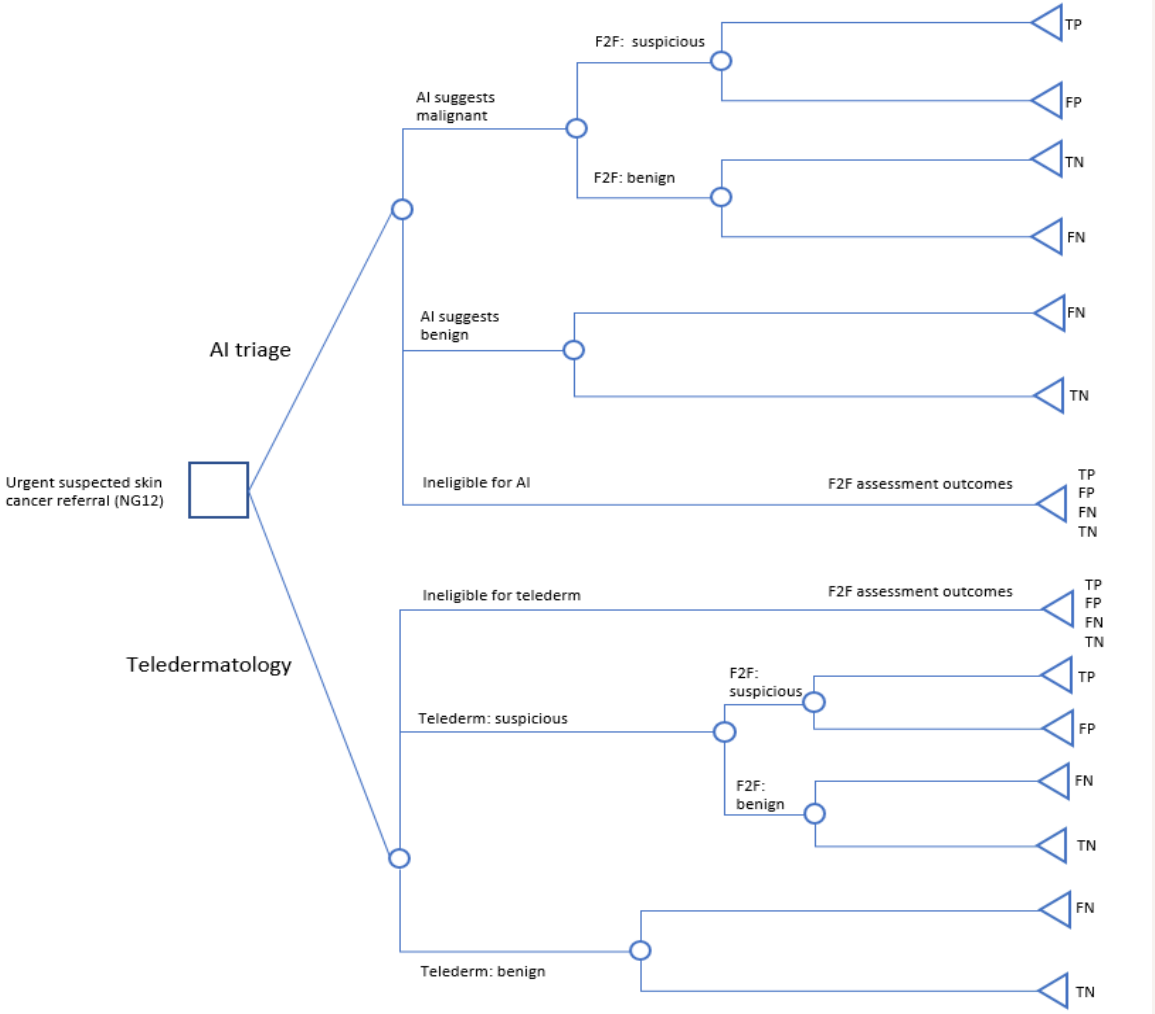
EAG's conceptual model (1)

proposed model decision tree directs to diagnostic classification

A) AI without second read vs F2F assessment



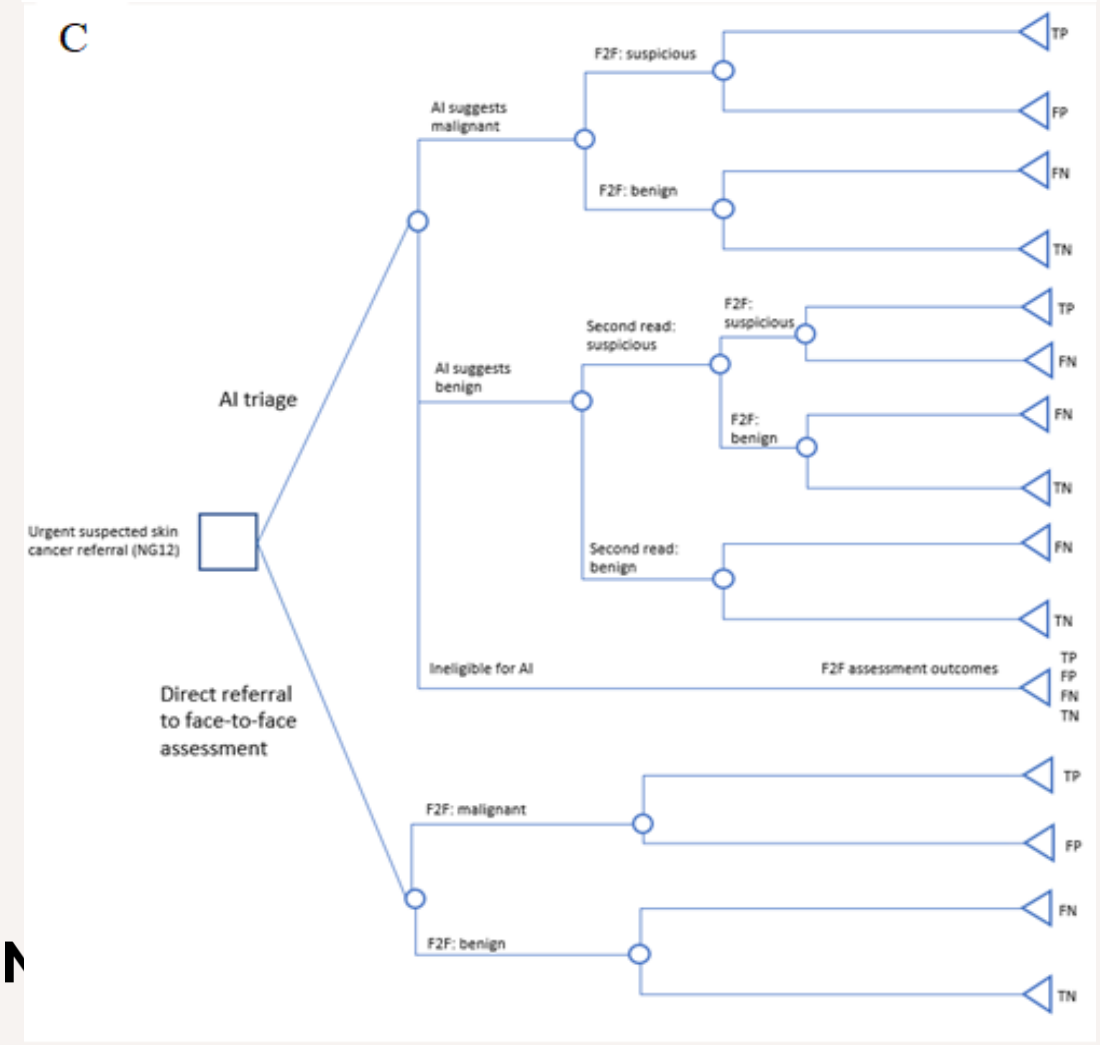
B) AI without second read vs teledermatology



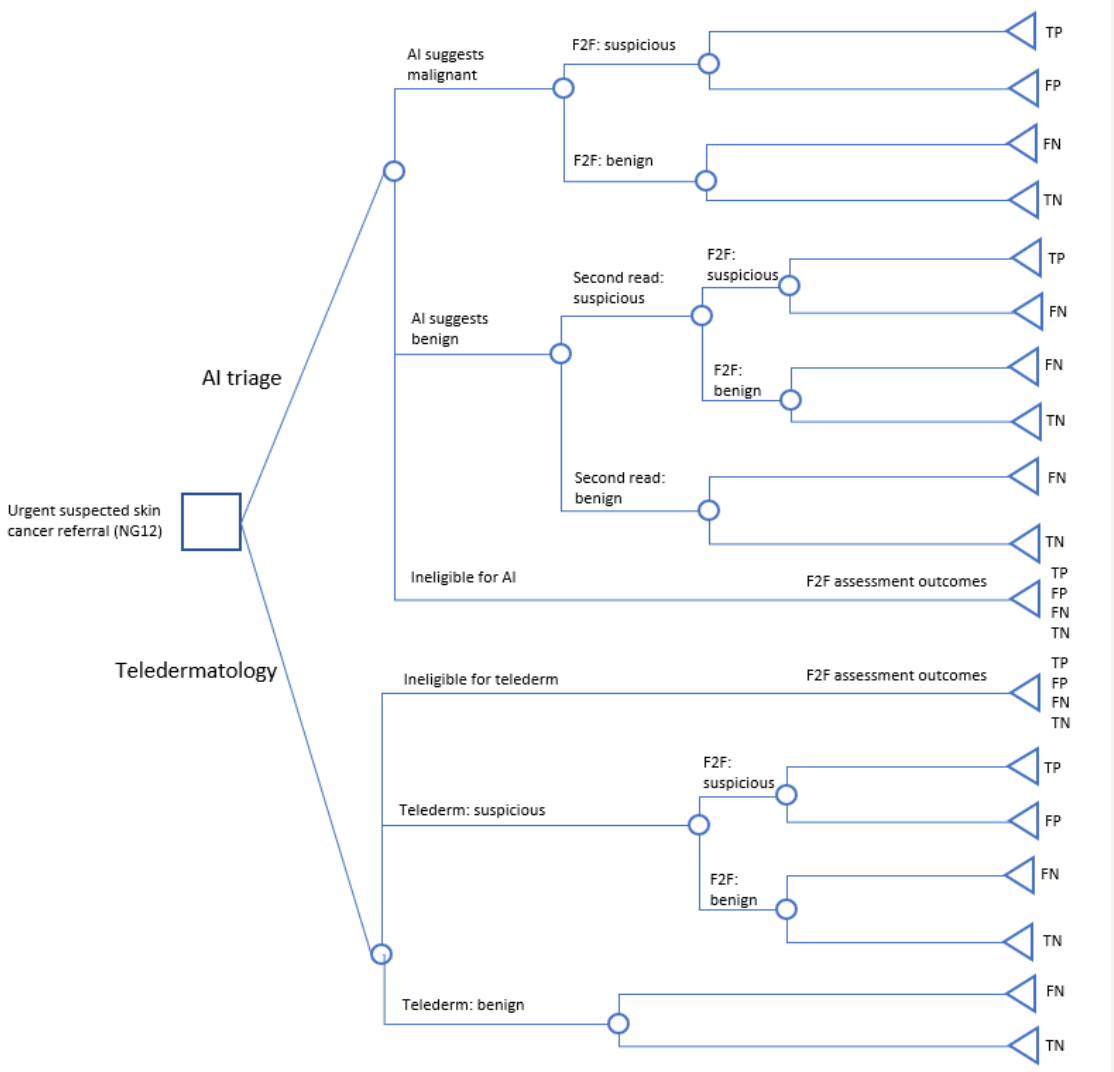
EAG's conceptual model (2)

proposed model decision tree directs to diagnostic classification

C) AI with second read vs F2F assessment



D) AI with second read vs teledermatology



EAG's conceptual model

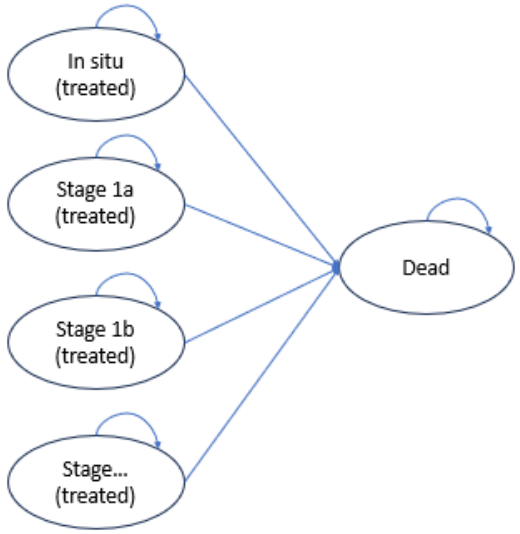
Markov model terminal nodes

True Positives: Patients correctly identified have ongoing mortality and HRQoL implications following treatment. Tunnel states represent post-treatment mortality risks.

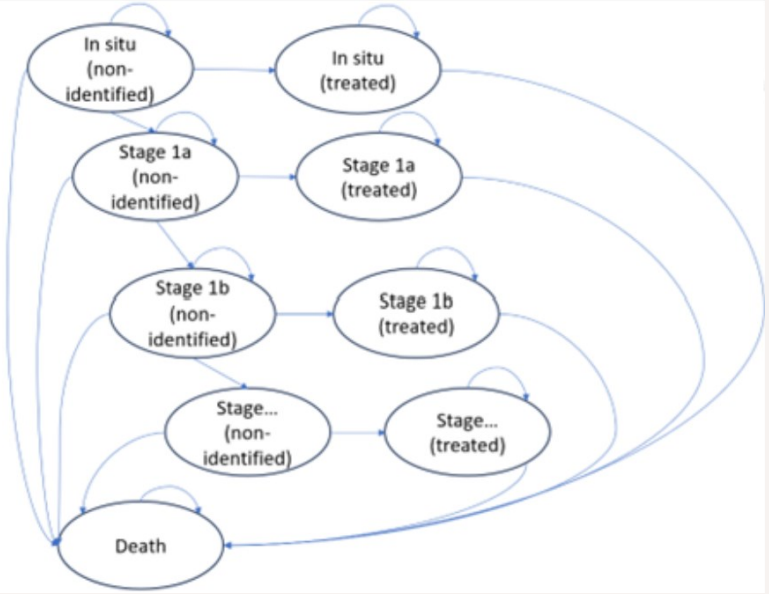
False Negatives: Risk of progression, mortality, and opportunisitic detection.

True Negative or False Positive: General population mortality and HRQoL outcomes. Utility decrements may be applied for inappropriate biopsy in FP patients.

True positives



False negatives



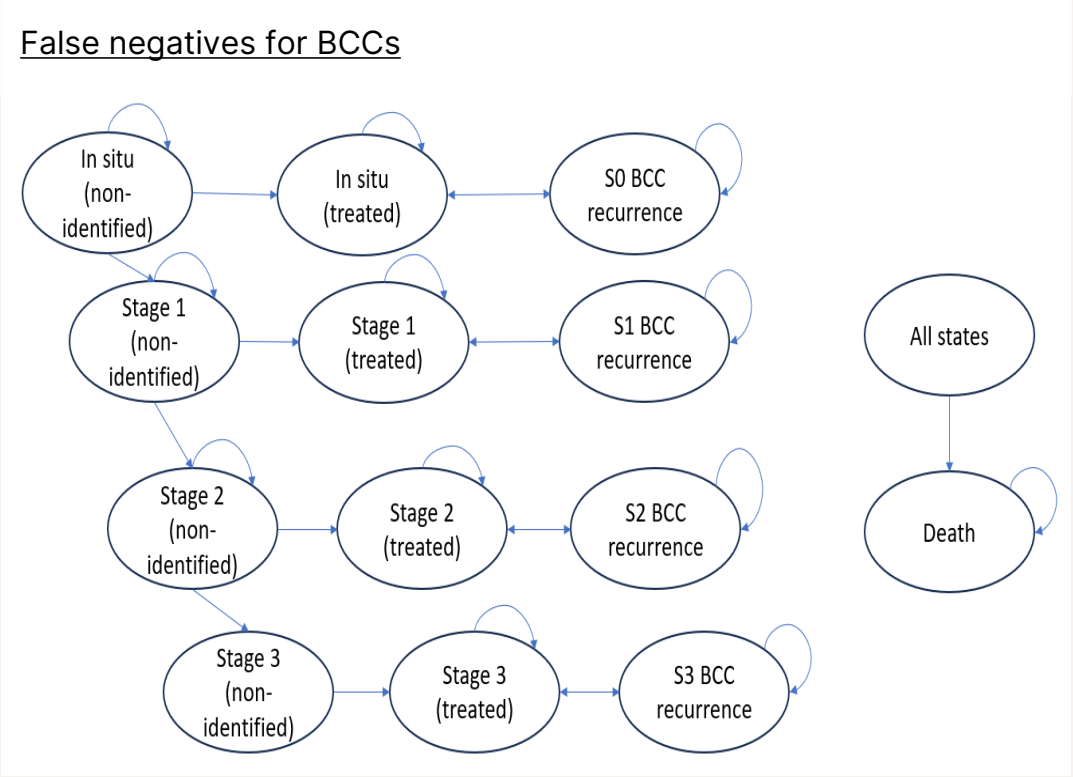
True negatives and false positives

EAG's conceptual model

Markov model terminal node for false negatives for BCCs

EAG proposed a separate Markov model for false negatives for BCCs to capture long-term impact of missed diagnoses of BCC:

- Untreated BCC may become more advanced over time and be prone to higher rates of recurrence.
- Whilst recurrence of BCC remains manageable, it is associated with additional treatment costs.
- This model intends to capture the slow development of non-identified BCC and its opportunistic detection.
- Following treatment, patients are then subject to a stage-specific recurrence rate.



EAG's conceptual model

Key points on clinical input parameters (1)

- Use case in the proposed model is identification of benign lesions to allow patients to be discharged following referral, but prior to F2F assessment.
- Estimates of prevalence and distribution of each disease by stage are required for the model.
- The key statistic is specificity – to estimate the capacity of a test to correctly identify true negative cases. Discharge of benign cases reduces the cost and health implications associated with unnecessary investigations.
- Sensitivity/specificity of each test cannot be pieced together from different sources to estimate whole-pathway sensitivity and specificity
- Data on the probability of FN outcomes (undiagnosed progression of disease or being opportunistically detected) is limited, hence the proposed model may need to adopt transition probabilities from other studies.

EAG's conceptual model

Key points on clinical input parameters (2)

- Parameters on probability of FN outcomes (undiagnosed progression of disease or being opportunistically detected) inform the transition between states in the long-term Markov components of the model.
- This data is limited, hence the proposed model may need to adopt transition probabilities from other studies.
- Mortality risks increase with disease stage, with mortality risk converging with that of the general population following successful treatment. Markov state-specific mortality rates are likely appropriate for the conceptual model.

EAG's conceptual model

Health-related quality of life

HRQoL is represented in the model through health utilities:

- Utility decrements representing the disutility associated with diagnostic and treatment procedures (e.g., anxiety associated with the wait for biopsy results, scarring as a result of biopsy/excision);
- Health-state utilities representing diagnostic status and disease stage (or presence of disease).

A systematic review of HRQoL values should seek to identify disutilities associated with BCC treatment. In the absence of disease-specific HRQoL data, it may also be appropriate to apply a one-off disutility equivalent to that applied for the treatment of melanoma in situ, which is typically managed using excision in a similar manner to BCC.

EAG's conceptual model

Cost and resource use parameters

Relevant costs in the proposed cost-effectiveness model include:

- Diagnosis related (cost of technologies, comparators, and clinical appointments)
- Treatment and investigation-related costs (e.g. biopsy, excision, imaging)
- Long-term state-dependent management costs based on treatment and disease-stage
- Implementation costs (e.g., staff training, establishing new medical photography infrastructure).

A technology which has a high sensitivity may generate value though avoiding the costs associated with missed diagnosis and delayed treatment.

A technology which has high specificity may generate value through avoid unnecessary referrals and further investigations.

Cost data should be based on the latest national sources in alignment with NICE methods guide for consistency with previous and future NICE decisions.

EAG's conceptual model

Cost and resource use parameters

A limitation of the proposed model is that it cannot capture one of the primary benefits of the system, namely non-cash-releasing benefits. The hybrid structure proposed (a decision tree and Markov extension) cannot meaningfully quantify the impact of reducing demand on services in terms of reducing waiting times (and potential improvements in quality of care) for a specialist consultation across all dermatological indications. A more complex modelling approach would be required to capture demand, capacity, and temporal dynamics.

The EAG have outlined a non-exhaustive list of unit costs in section 6.4.2 of the assessment report that could be adapted to implement into a future cost-effectiveness model, alongside a comparison with the values used by the Exeter/Skin Analytics model.

For implementation into a future cost-effectiveness model, unit costs should be updated based on the most recent published reference costs.

Key issue 1: Non-cash-releasing benefits not captured

Background

EAG's proposed conceptual model would not capture non-cash-releasing benefits associated with reducing demand on dermatologists' time. The hybrid structure proposed (a decision tree and Markov extension) cannot meaningfully quantify the impact of reducing demand on services for a specialist consultation across all dermatological indications. A more complex modelling approach would be required to capture demand, capacity, and temporal dynamics.

Potential risks

Unable to meaningfully quantify and determine benefits if they exist. Such as:

- impact of reducing demand on health services (i.e. reducing wait times)
- potentially improving quality of care

Question: Is the proposed structure of the EAG's conceptual model appropriate?

Key issue 2: Limited diagnostic accuracy data on rare cancers

Background

The company's model has placed rare cancers within the same diagnostic category as other high risk cancers (melanoma and SCC).
AI technologies may have limited experience of rare cancers, hence there is uncertainty as to whether their high sensitivity to melanoma and SCC is maintained across these rare conditions.

Potential risks

Placing rare cancers with other more common high-risk cancers in a single diagnostic category in terms of diagnostic accuracy, stage at diagnosis, rate of progression, and impact upon mortality may therefore be subject to uncertainty.

EAG: Where possible, sensitivity analysis should be undertaken in which rare cancers are categorised separately, and alternative sources of diagnostic and prognostic data are used to parameterise this sub-population in the model.

Key issue 3: EAG's critique of Exeter model

Background

1. Specificity of teledermatology reported in Exeter model differs with the specificity reported in Cochrane review.
2. The company's model for BCC punishes correct diagnoses, as excision is associated with accrual of costs and QALY decrement. Model assumes that undiagnosed BCC does not progress if not diagnosed, and undiagnosed BCC has no further health consequences.
3. Company's model implies the sensitivity of the pathway increases when a triage step is added (i.e. DERM assessment followed with teledermatology assessment) without sufficient evidence to support this.

Potential risks

1. Lower specificity of teledermatology makes DERM appear more cost saving.
2. The company model for BCC disincentivizes higher sensitivity, which makes DERM appear more cost saving for BCC diagnosis as its less sensitive than its comparators (teledermatology and F2F assessment).
3. Increased sensitivity with a triage step gives the impression of better cost-effective estimates when DERM and teledermatology are used in a pathway.

Accurate teledermatology specificity figures are vital for EAG to build their model in the future.

DERM costs

Component	£ per 10k catchment population	£ per urgent referral	Description
Base platform with DERM review	3,300	30.00	Image and medical history capture platform, DERM assessment, PDF report with suspected diagnosis and recommended next steps. Inclusive of training and data storage costs.
Teledermatology functionality add on (optional)	900	8.20	Allows clinical staff to virtually review patient's cases.
Discount if contributing outcome data (optional)	(250)	(2.30)	Discount provided if > 50% of biopsy results are shared with Skin Analytics.
Total cost per year (ex VAT) – with outcomes discount	3,950	35.90	
Total cost per year (ex VAT) – without outcomes discount	4,200	38.20	
Second read (Skin Analytics dermatologist)		£17 per case	

MoleAnalyzer pro costs

Pricing option	Cost
FotoFinder Moleanalyzer AIMEE scoring (flat per year)	£1,210
FotoFinder Moleanalyzer pro includes AIMME offline package (per year)	£1,750

AIMME: artificial intelligence mole examination and evaluation

No cost for training and the company offers a discount for multi-user access.

NICE resource impact assessment

Edgar Masanga

Business Analyst (Accountant) - Resource Impact Assessment

Finance, Strategy and Transformation directorate

NICE National Institute for Health and Care Excellence

NICE's resource impact assessment (RIA) overview

NICE performed a resource impact assessment based on the EAG's assessment report, national cancer registry data, and technology costs provided by companies.

Details	%	England	ICB
Total population		56,550,138	1,346,432
Total dermatology first appointments[1]	2.12	1,200,000	28,571
Proportion estimated to be on urgent suspected skin cancer pathway referrals[1]	60	720,000	17,143
Proportion with confirmed malignant melanoma of skin[2]	2.20	15,861	378
Proportion with confirmed other malignant neoplasms of skin[2]	26.59	191,459	4,559
Total confirmed cases		**207,320**	**4,936**
Total non-cancer cases		**512,680**	**12,207**

NICE RIA - Capacity savings

Technology may reduce secondary care appointments

Details	England			ICB		
% of referral appointments avoided	19%	40%	50%	19%	40%	50%
Cost of technology based on population catchment[1] (excludes training costs)	£20.6m	£20.6m	£20.6m	£0.5m	£0.5m	£0.5m
Average weighted cost of dermatology first appointment (fully absorbed cost[2])	£160	£160	£160	£160	£160	£160
Equivalent number of referrals	129,256	288,000	360,000	3,078	6,857	8,571
Dermatology appointment cost avoided (at cost collection unit rate)	£20.6m	£46.0m	£57.5m	£0.5m	£1.1m	£1.4m
Saving	**£0**	**£25.4m**	**£36.9m**	**£0**	**£0.6m**	**£0.9m**

[1] This is lower than if based on cancer referral numbers.

[2] Fully absorbed cost of a first appointment includes all hospital overheads/infrastructure costs.

NICE RIA - Capacity savings

Use of AI may reduce dermatology clinic sessions/ free up consultant hours

Details	England			ICB		
Duration of a dermatology first appointment *(assuming a 20-minute appointment)*	20	20	20	20	20	20
Total time for avoided referrals - hours	43,085	96,000	120,000	1,026	2,286	2,857
Duration per clinic session (hours)	4	4	4	4	4	4
Total potential clinic sessions avoided	**10,771**	**24,000**	**30,000**	**256**	**571**	**714**
Average clinical hours per consultant	1,376	1,376	1,376	1,376	1,376	1,376
Estimated equivalent wte consultant saved	31.31	69.77	87.21	0.75	1.66	2.08

Summary

NICE National Institute for Health and Care Excellence

Summary of clinical effectiveness

DERM

- High diagnostic accuracy for detection of malignant lesions [Sensitivity: 96.1% (95% CI 95.4 to 96.8) , Specificity: 65.4%(95% CI 64.7 to 66.1)]. Similar diagnostic accuracy found for melanoma or SCC
- Some evidence DERM overdiagnoses BCCs as SCC or melanoma.
- Sensitivity for detecting benign lesions was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0).
- Dermatologists had slightly lower sensitivity to detect any malignancy by higher specificity (Sensitivity: 90.6%, Specificity: 85.7%)
- The diagnostic accuracy of the whole teledermatology pathway including DERM could not be reliably assessed because of a lack of any independent reference standard of diagnosis.
- Autonomous use of DERM would result in half of all patients being discharged, and half referred for further assessment (either in person or through teledermatology). About 0.8% discharged with malignant lesion, mostly discharged with BCC.
- [redacted]
- [redacted]
- Patient opinion was broadly supportive of using DERM in some form as part of their diagnosis, but patients were divided on whether they preferred teledermatology to face-to-face appointments. Clinicians were generally very resistant to using DERM in isolation without human assessment of lesions.

Summary of clinical effectiveness

Moleanalyzer pro

- Two prospective studies of Moleanalyzer pro were identified; neither were performed in the UK.
- Moleanalyzer pro alone had a sensitivity of 84.4% (95% CI 73.9 to 91.0) and a specificity of 84.5% (95% CI 72.0 to 92.1) to detect melanoma from a meta-analysis of the studies. This appeared similar to the accuracy of dermatologists alone.
- No eligible evidence was found for the diagnosis of SCC, BCC or other cancers.
- The EAG did not identify any relevant evidence on the clinical impact of using Moleanalyzer pro.
- Patient and clinician opinion was generally supportive of using Moleanalyzer pro in some way to aid diagnosis. However, the overwhelming majority of patients indicated that they would like the opinion of an expert physician besides an AI-assisted diagnosis.

Summary of cost considerations

- No cost-effectiveness evidence available for Moleanalyzer pro.
- 3 published cost-effectiveness studies were identified evaluating any diagnostic technology for skin cancer in an NHS setting.
- The EAG considered the Exeter model structure to be largely appropriate to assess potential value of AI technologies for identifying benign lesions in a post-referral setting, but noted several issues where the main value drivers may not be appropriately characterised.
- It remains highly uncertain whether current evidence is sufficient to reliably use in a cost-utility model, particularly with regards to the comparative specificity of AI technologies to an effectively implemented teledermatology service. Therefore, whilst this analysis predicted that DERM with or without a second read would dominate all other options, this was highly dependent on the relative specificity of teledermatology.
- The EAG outlines a conceptual model to provide an alternative to the Exeter model, with replacing the Markov model for capturing the natural history of BCC in false negatives

Data gaps

Current evidence for both DERM and Moleanalyzer is lacking with regard to the diagnostic accuracy of the whole diagnostic pathway (i.e. inclusive of subsequent steps). Availability of this data is essential to understanding the likelihood of missed cases which cannot be inferred from the partial data currently available.

Comparable diagnostic accuracy data describing current service provision is lacking, particularly for the teledermatology pathway. Without comparative evidence on the diagnostic accuracy of AI technologies and teledermatology, their relative value for safe and cost-effective identification of benign lesions will remain unclear.

EAG also note a lack of robust data available to inform progression probabilities in undiagnosed disease, and a focus on expert-elicited parameters in previous cost-utility models

The prospective studies of Moleanalyzer pro were conducted outside of the UK, were not explicitly in a teledermatology setting, and did not evaluate the accuracy of AI for detecting non-melanoma cancer.

The DERM versions (in particular, the set sensitivity/specificity thresholds) and the dermatoscopes used for clinical assessments were out of date; there was restricted eligibility for both technologies, therefore, the applicability of the diagnostic accuracy results to current practice is uncertain.

Data gaps

There is a need for further research on the diagnostic accuracy of AI compared to standard teledermatology in specific patient subgroups:

- In individuals with darker skin types (Fitzpatrick IV-VI) and a broad range of ethnicity groups;
- For lesion types and lesions located in body sites and not currently covered by DERM and Moleanalyzer pro evidence;
- To identify rare skin cancers.

Future diagnostic studies should, where possible, examine and compare the diagnostic accuracy of:

- AI as a standalone device;
- AI in combination with human teledermatology (e.g. with a “second read” for all AI assessed lesions;
- Teledermatology without AI;
- Face-to-face assessment without teledermatology.

Future studies should evaluate the diagnostic accuracy of ruling out benign lesions, rather than triaging/prioritising.

NICE National Institute for Health and Care Excellence

Thank you

CONFIDENTIAL UNTIL PUBLISHED
Assessment Group's Report

Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments: Early Value Assessment

Produced by	CRD and CHE Technology Assessment Group, University of York, Heslington, York, YO10 5DD
Authors	Matthew Walton, Research Fellow, CRD Alexis Llewellyn, Research Fellow, CRD Eleonora Uphoff, Research Fellow, CRD Joseph Lord, Research Fellow, CRD Melissa Harden, Information Specialist, CRD Robert Hodgson, Senior Research Fellow, CRD Mark Simmonds, Senior Research Fellow, CRD
Correspondence to	Mark Simmonds Centre for Reviews and Dissemination University of York York, YO10 5DD
Date completed	22/12/2023

Source of funding

This report was commissioned by the NIHR Evidence Synthesis Programme as project number NIHR136014 [DG76].

Declared competing interests of the authors

None

Acknowledgements

The EAG would like to acknowledge the help and advice from all members of the scoping committee for this assessment. In particular, we acknowledge the advice provided by Dr Rubeta Matin (Consultant Dermatologist, Oxford University Hospitals).

Rider on responsibility for report

The views expressed in this report are those of the authors and not necessarily those of the NIHR Evidence Synthesis Programme. Any errors are the responsibility of the authors.

This report should be referenced as follows:

Walton M, Llewellyn A, Uphoff N, Lord J, Harden M, Hodgson R, Simmonds M. Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments: Early Value Assessment. York EAG, 2024.

Contributions of authors

M Walton conducted the cost-effectiveness review, developed the conceptual model and led the writing of the economic sections of the report (Sections 4-6).

A Llewellyn performed the diagnostic review and led the writing of the clinical sections of the report (Sections 1-3).

N Uphoff performed the diagnostic review and wrote parts of the clinical sections of the report.

J Lord contributed to the cost-effectiveness review and the development of the conceptual economic model.

M Harden acted as information specialist and performed database searches.

R Hodgson had overall responsibility for the economic sections of the report.

M Simmonds had overall responsibility for the clinical sections of the report.

Note on the text

All academic in confidence (AIC) data have been [redacted] All commercial-in-confidence (CIC) data have been [redacted]

Copyright statement

Changes to the original report

The original EAG report was submitted to NICE on 22 December 2023. Since that submission changes have been made to respond to some stakeholder comments, and to correct errors.

The changes are summarised in the table below.

Location in report	Edit made
Throughout	References to 2-week wait (2WW) replaced with "urgent referral pathway", except where 2WW is specifically intended.
	"West Sussex" corrected to "West Suffolk"
Scientific summary - Background	Reference to melanoma as deadliest form of cancer removed
	Clarification made on impact of high burden of referrals
Scientific summary - Conclusions	Clarification that the impact of DERM on clinical pathway and appointment burden is unclear added.
	Additional text added on lack of evidence in people with darker skin tones.
	Additional text added on need for evidence in people with darker skin tones or hard to diagnose lesions.
Section 1.4.2	Discussion of NICE guidance (NG12) removed.
	Incorrectly located text on DERM assessment at end of this section removed.
Section 2.3.1.1	Clarification on nature of diagnostic clinics added.
Section 3.2 generally	Evidence drawn from unpublished material from the DERM-005 study now marked as academic-in-confidence.
Section 3.2.6.1	Discussion of discharged lesions when using DERM vA removed.
Section 3.2.7.3	Cross-reference to figure 9 corrected.
Section 5.1.3	Point added regarding inappropriate inclusion of phototherapy in BCC treatment costs in Exeter model.
Section 5.1.3.1	Clarification added on cost savings of DERM with a second read.
Section 5.1.3.3	Text on DERM with a second read and missed diagnoses clarified.
Section 8.1	Clarification added on non-cash releasing benefits.

ABSTRACT

Background

Skin cancers are some of the most common types of cancer. Dermatology services receive about 1.2 million referrals a year, but only a small minority are confirmed cases of skin cancer. Artificial Intelligence (AI) may be helpful in the diagnosis of skin cancer by identifying lesions that are or are not cancerous using a high-quality photograph of the lesion.

Objectives

To investigate the clinical and cost-effectiveness of two AI technologies: DERM (Skin Analytics) and Moleanalyzer Pro (Fotofinder), as decision aids to triage and diagnose suspicious skin lesions following a primary care referral.

Methods

A rapid systematic review of published and unpublished evidence on the two technologies was conducted. A narrative synthesis was performed, with a meta-analysis of diagnostic accuracy data where feasible.

Published and unpublished cost-effectiveness evidence on the named technologies, as well as other diagnostic technologies considering a UK setting, were reviewed. Based on the evidence identified, a conceptual model was developed that could form the basis of a full economic evaluation.

Results

Four studies of DERM and two of Moleanalyzer Pro were subject to full synthesis. All recruited highly selected populations and raised concerns about bias and applicability. DERM had a sensitivity of 96.1% to detect any malignant lesion (95% confidence interval (CI) 95.4 to 96.8); at a specificity of 65.4% (95% CI 64.7 to 66.1). Diagnostic accuracy was similar for melanoma, squamous cell carcinoma (SCC) and basal cell carcinoma (BCC). For detecting benign lesions the sensitivity was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0). Moleanalyzer Pro had lower sensitivity, but higher specificity for detecting melanoma than face-to-face dermatologists.

Evidence on the clinical impact of the technologies was limited. Unpublished data suggested that DERM might lead to around half of all patients being discharged without assessment by a dermatologist, but a small number of malignant lesions would be missed. Patient and clinical opinions showed substantial resistance to using AI without any assessment of lesions by a dermatologist.

No published assessments of the cost-effectiveness of the technologies were identified; three assessments related to skin cancer more broadly in an NHS setting were identified. These studies employed similar model structures, but the mechanism by which diagnostic accuracy influenced costs and health outcomes differed. An unpublished cost-utility model was provided by Skin Analytics. Several issues with the modelling approach were identified, particularly the mechanisms by which value is driven and how diagnostic accuracy evidence was used.

The conceptual model presents an alternative approach, which aligns more closely with the NICE reference case and which more appropriately characterises the long-term consequences of BCC.

Conclusions

DERM shows promising diagnostic accuracy for triage and diagnosis of suspicious cancer lesions in selected patients referred from primary care. Its impact on the diagnostic pathway and patient care is, however, uncertain. Moleanalyzer Pro shows promising accuracy for diagnosing melanoma, but its evidence base is limited.

While AI has the potential to be cost-effective for the identification of benign lesions, further research addressing the limitations in the diagnostic accuracy evidence is necessary. Without comparative evidence on the diagnostic accuracy of AI technologies, their value will remain uncertain.

Word count: 521

SCIENTIFIC SUMMARY

Background

Skin cancers are some of the most common types of cancer. Over 16,000 cases of melanoma, and more than 210,000 cases of non-melanoma skin cancer are diagnosed every year in the UK. In current practice, patients with suspicious skin lesions are referred to secondary care through the urgent suspected skin cancer referral pathway, where people attend a secondary care dermatology department for a face-to-face appointment with a consultant dermatologist. As benign skin lesions and skin cancer are so common, this places a very high burden on dermatology clinics, which may lead to a reduction in capacity to handle other skin conditions.

Artificial Intelligence (AI) may be helpful in the diagnosis of skin cancer. An AI system could potentially identify which referred lesions are not cancerous using a high-quality photograph. An AI system could be used alone, or in combination with a dermatologist looking at the photograph. People judged not to have cancer could then be quickly discharged prior to secondary care consultation, while people whose lesion may be cancerous may be seen by a specialist in person. AI systems could therefore potentially speed up the diagnostic process and reduce the burden on the health service. AI systems are already used in the NHS in a research context, but there is a need to evaluate their clinical impact and value.

This project investigated whether two such AI technologies: DERM (Skin Analytics) and Moleanalyzer Pro (FotoFinder) can produce clinically meaningful benefits for skin cancer diagnosis, and whether they have the potential to be cost-effective.

Objectives

The aim of the project was to investigate the clinical and cost-effectiveness of the two AI technologies, DERM and Moleanalyzer Pro, as decision aids to triage and diagnosis of suspicious skin lesions following a referral on the urgent suspected skin cancer pathway. To achieve this, the following objectives were proposed:

- To perform a rapid systematic review, narrative synthesis, and where feasible a meta-analysis, of the diagnostic accuracy, clinical impact and practical implementation of the included AI technologies
- To perform a rapid systematic review of published cost-effectiveness studies of diagnostic strategies used to aid the diagnosis of skin cancer.

- To develop a conceptual model that will identify likely drivers of health benefit, harms and costs associated with implementing the included AI technologies in the NHS and identify areas for further research.

Methods

Data sources

MEDLINE, Embase, CENTRAL and the Association for Computing Machinery (ACM) Digital Library were searched in November 2023. Clinical trial registries were searched. Unpublished material supplied by the included companies was also assessed.

Inclusion criteria

Any clinical study evaluating DERM or Moleanalyzer Pro in people with skin lesions suspicious of cancer, presenting in primary care, rapid diagnostic clinic, teledermatology or secondary care settings were eligible for inclusion. Included studies must report either diagnostic accuracy, clinical outcomes, or evidence on implementation. The comparator was clinical judgement by dermatologists, but this did not need to be reported for a study to be eligible. The preferred reference standard for diagnosis was histology, but for unbiopsied lesions, clinical confirmation of non-malignancy was accepted.

The cost-effectiveness review included any economic evaluation including budget impact models, return on investment analysis, and other cost-only analyses of either DERM or Moleanalyzer Pro in the above population and setting. It was anticipated that no relevant studies would be identified for the name technologies therefore additional searches were also conducted to identify cost-effectiveness studies looking at any technology used to aid diagnosis of skin cancer in an NHS setting.

Data extraction

An initial scoping of studies was performed by extracting data on intervention, study location, size, setting, type of outcomes reported, and design and key quality indicators. Only studies with prospective recruitment of patients were taken forward for full data extraction and synthesis. For those studies, full data on the intervention, patient characteristics and all reported outcomes were extracted. Risk of bias was assessed using QUADAS-2 and QUADAS-C.

Identified economic evaluations were reviewed and discussed in detail, with the aim of informing the design and parameterisation of conceptual model. Material provided by submitting stakeholders pertaining to the value case for their product were also reviewed.

Synthesis

A scoping process was used to classify identified studies for relevance to the decision problem, based on study quality, setting, outcomes reported and relevance to the NHS. For studies taken forward from the scoping phase for full synthesis, a narrative synthesis was performed. Results are presented in structured tables and figures as appropriate, with a text summary. Random effects meta-analyses of sensitivity and specificity were performed to pool diagnostic accuracy estimates across studies.

Evidence related to cost-effectiveness studies was reviewed and synthesised narratively.

Modelling

The conceptual model described sought to provide an overview of the structure of a cost-utility model and key evidence required for the assessment of AI technologies for the identification of benign lesions amongst suspected cancer cases referred on the urgent referral pathway. The structure of the conceptual model was designed considering the strengths and limitations of previously published diagnostic models for skin cancer in an NHS setting, and evidence submitted by stakeholders.

Results

Diagnostic accuracy and clinical impact of DERM

Six studies of DERM were identified, of which four were considered for full synthesis. Those four studies were all conducted in the UK. All studies excluded a substantial proportion of participants from assessment, which may produce biased results.

Meta-analysis of diagnostic accuracy data supplied by the company suggested that DERM has a high sensitivity of 96.1% to detect any malignant lesion (95% confidence interval (CI) 95.4 to 96.8); at a specificity of 65.4% (95% CI 64.7 to 66.1). The diagnostic accuracy for detecting melanoma or squamous cell carcinoma specifically was similar. For the detection of benign lesions the sensitivity was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0). This appears to be comparable in diagnostic accuracy to that achieved by dermatologists without the use of DERM. The diagnostic accuracy of combining DERM with assessment by a dermatologist could not be assessed.

Data on the clinical impact of using DERM was limited, and mostly unpublished. Some trial data suggested that autonomous use of DERM would lead to approximately half of patients being referred to a dermatologist for further assessment, and half being discharged. However around 1% of people would be discharged with malignant lesions (mostly basal cell carcinomas). DERM use as part of a teledermatology service could potentially [redacted] compared to face-to-face assessment. However, use of DERM may slow progress to diagnosis.

Patient and clinical opinions of DERM were generally favourable towards accepting its use as part of the diagnostic pathway. However, there was very substantial resistance, particularly among clinicians, to using DERM without any assessment of lesions by a dermatologist.

Diagnostic accuracy and clinical impact of Moleanalyzer Pro

Seventeen publications of Moleanalyzer Pro were identified, but these were mostly retrospective reviews, and two prospective studies were eligible for full data extraction. The applicability of the evidence for Moleanalyzer Pro to practice is limited, notably due to the lack of studies from the UK and the lack of data for non-melanocytic lesions.

When pooled these studies found that Moleanalyzer Pro alone had a sensitivity of 84.4% (95% CI 73.9 to 91.0) and a specificity of 84.5% (95% CI 72.0 to 92.1) to detect melanoma. Moleanalyzer Pro had a lower sensitivity and higher specificity to detect melanoma when compared with face-to-face dermatologist and remote teledermatology. There was no evidence on the diagnostic accuracy of Moleanalyzer Pro to detect other skin cancers, and no evidence was found on its clinical impact.

Economic evidence

No economic studies relating to the named technologies were identified from searches of the literature. Broader searches for any technology used to aid diagnosis of skin cancer in an NHS setting identified three studies. Although relevant to this review, none related to the use of AI for the detection of skin cancer and considered populations which were not relevant to the decision problem. While all identified studies adopted similar model structures, the mechanisms by which diagnostic accuracy generated value (in terms of either cost savings or QALY gain) differed across these models. For instance, diagnostic sensitivity had less value in some models with value instead generated by the avoidance of unnecessary referral and diagnostic procedures.

Economic evidence on the cost-effectiveness of DERM was submitted by Skin Analytics and NHSE. This evidence was preliminary and did not include an executable model. The most relevant analysis was a cost-utility model developed by the Exeter Test Group and Skin Analytics. The EAG considered the model structure largely appropriate to capture important the direct cost and health implications of AI technologies for directing discharge in a post-referral setting. However, a lack of key comparative data meant the relative clinical and cost-effectiveness of alternative pathways was necessarily based on often optimistic assumptions. The model suggested DERM could be highly cost-effective in the NHS, but the EAG noted that results may be very sensitive to the use of alternative sources of diagnostic accuracy data. The EAG also noted several issues which may mean that the main value drivers were not appropriately characterised. Namely, the model imposed disincentives for the correct diagnosis and treatment of BCC; structurally imposed assumed sensitivity benefits for any

strategy incorporating a triage step; used costs associated with biopsy and treatment which were inconsistent with sources generally used in NICE appraisals, and may overvalue specificity in terms of generating cost-savings.

No economic evidence related to Moleanalyzer Pro was identified.

Conceptual model

The EAG developed a conceptual model aimed at providing an alternative to that presented in the Skin Analytics submission. While the proposed model retained the structure reported by Skin Analytics, the EAG propose an alternative structure for patients with BCC, aimed at better capturing the cost- and health-consequences of BCC, particularly with reference to disease recurrence.

The EAG consider the current evidence inadequate to fully address the decision problem. Current evidence for both DERM and Moleanalyzer Pro is lacking with regards to the diagnostic accuracy of the whole diagnostic pathway (i.e. inclusive of subsequent steps). Availability of this data is essential to understanding the likelihood of missed cases which cannot be inferred from the partial data currently available. Similarly, comparable diagnostic accuracy data describing current service provision is lacking, particularly for the teledermatology pathway.

Conclusions

Impact on practice

The diagnostic accuracy of DERM suggests that it has potential for use within a post-primary care referral setting. This could be either alongside assessment by dermatologists, or as an autonomous tool within the post-referral pathway within a subset of patients. However, the practical impact and clinical benefit of using DERM in a post-referral setting is currently unclear. In particular, the impact on referrals and secondary care appointments, the burden on clinicians and the subsequent clinical impact on patients is largely unclear. Although Moleanalyzer Pro shows promising accuracy for diagnosing melanoma, its evidence base is currently too limited to fully assess its clinical value.

Evidence on the diagnostic accuracy and clinical value of AI in people with darker skin tones or with lesions that are more difficult to assess (such as when versions are large, or obscured by scarring, tattooing or hair) was largely absent. Only a small number of people with darker skin tones were recruited to the included studies, and people with hard to assess lesions were often excluded. This raises concerns as to whether AI could be used in these people.

Current economic evidence supporting the cost-effectiveness of DERM is limited, and it is unclear whether the plausible advantages of DERM represent value for money relative to other strategies. Company-sponsored analyses suggested that DERM used autonomously and with a second read could be highly cost-effective compared to current 2WW diagnostic models. However, much of this value is generated through potentially optimistic assumptions around the diagnostic accuracy of comparators, and of the surrounding pathway.

[redacted] Notably, the magnitude of uncaptured non-cash releasing benefits remains unquantified.

There is currently no economic evidence supporting the use of Moleanalyzer Pro, but assuming a similar use case to DERM and appropriate data collection, the value of Moleanalyzer Pro could be assessed using the conceptual framework presented by the EAG.

Future research needs

The diagnostic accuracy of AI in a post-primary care referral pathway is uncertain and requires further evaluation. A lack of key comparative data on diagnostic accuracy means the relative clinical and cost-effectiveness of pathways incorporating AI technologies and teledermatology remains highly uncertain. Assessments of diagnostic accuracy of AI in people with darker skin tones, or with hard to assess lesions are urgently needed.

Directly comparable evidence on the diagnostic accuracy of AI technologies and teledermatology in a post-referral setting compared with unassisted teledermatology is required to assess the potential value of AI technologies. This would require studies comparing AI with dermatologists assessments, recruiting a representative population and case-mix, use up-to-date versions of AI and dermoscopy, and with a robust independent reference standard for all patients.

A better understanding of the clinical benefits and resource implications associated with the implementation of AI technologies will also require further research to set-up AI and teledermatology services in the NHS. Further research must also be undertaken to quantify the benefits to population health within skin cancer and other dermatological indications associated with any release of NHS consultant dermatologist resource, and understand the effects of these technologies on waiting times for final diagnosis.

This could potentially be achieved through continuations and extensions of existing ongoing pilot studies of DERM, but truly comparative evidence may also be required. Moleanalyzer Pro requires evaluation within a UK teledermatology setting.

The substantial resistance from both patients and clinicians to using AI without any human dermatological assessment means that if AI is to be used to direct discharge autonomously, more evidence is needed to demonstrate that it has clear benefits to patients, without sacrificing accuracy.

Word count: 2316

PLAIN ENGLISH SUMMARY

Skin cancers are and suspicious skin lesions are very common. People with moles or lesions that might be cancerous are referred to a skin cancer specialist (a dermatologist) to make a diagnosis. This places a very high burden on dermatology clinics and as a result, there can be delays in seeing a dermatologist and getting a diagnosis. Artificial Intelligence (AI) systems could potentially use a high-quality photograph to identify which lesions do not need to be seen by a specialist. This could be done by the AI system alone, or in combination with remote review by a dermatologist.

This project investigated whether two AI technologies: DERM (Skin Analytics) and Moleanalyzer Pro (FotoFinder) could be useful in reducing the burden on dermatology services whilst helping to identify skin cancer. The evidence was reviewed to investigate whether the technologies can accurately identify skin cancer cases, and whether their use might improve the diagnosis process for patients. We also designed a theoretical model in which the economic value of AI technologies for the diagnosis of skin cancer could be assessed. As part of this process, we sought to outline what further evidence would be needed to implement a full assessment.

The evidence we reviewed suggests DERM could potentially identify 95% of all skin cancers but would require about half of all patients to be referred to dermatologists. Moleanalyzer Pro could identify about 85% of malignant melanomas. This appears to be a similar accuracy to that achieved by dermatologists alone. How their use would impact diagnosis and treatment for patients in practice, and the burden on clinicians, is currently unclear.

Because of limitations in the evidence on the diagnostic accuracy of AI technologies, a full assessment of their economic value is not possible at this time. Further research should focus on better establishing the diagnostic accuracy of both AI technologies and current service provision.

Table of Contents

ABSTRACT 4
SCIENTIFIC SUMMARY 6
PLAIN ENGLISH SUMMARY 13
Table of Contents 14
List of abbreviations 21
Assessment Group Report 22
1 BACKGROUND AND DEFINITION OF THE DECISION PROBLEM 22
1.1 Purpose of the decision to be made 22
1.2 Interventions 22
1.2.1 Deep Ensemble for Recognition of Malignancy (DERM) (Skin Analytics) 22
1.2.2 Moleanalyzer Pro (FotoFinder systems) 23
1.3 Populations and relevant subgroups 23
1.4 Place of the intervention in the treatment pathway 23
1.4.1 Types of skin cancer 24
1.4.1.1 Melanoma 24
1.4.1.2 Squamous cell carcinoma (SCC) 25
1.4.1.3 Basal cell carcinoma (BCC) 25
1.4.1.4 Other rare skin cancers 25
1.4.2 Current diagnostic pathway 25
1.4.2.1 Urgent suspected skin cancer referral pathway 26
1.4.2.2 Teledermatology service 27
1.4.3 Potential positioning of AI technologies in the pathway 28
1.4.3.1 Adjunctive use of AI with dermatologist assessment 29
1.4.3.2 Autonomous use of AI 29
1.4.4 Treatment of confirmed skin cancer 29
1.5 Relevant comparators 30
1.6 Key outcomes addressed as part of the assessment 30
1.6.1 Diagnostic accuracy 30
1.6.2 Implementation, resource use, and practicality 30
1.6.3 Clinical impact and patient benefit. 31
1.6.4 Costs 31
1.7 Objectives 32
2 METHODS 33
2.1 Systematic review methods 33
2.1.1 Search strategy 33
2.1.2 Study selection 33

2.1.3 Inclusion criteria 34
2.1.3.1 Population 34
2.1.3.2 Interventions 34
2.1.3.3 Comparators 34
2.1.3.4 Reference standard 34
2.1.3.5 Outcomes 34
2.1.3.6 Study designs 35
2.1.4 Data extraction 35
2.2 Quality assessment 35
2.2.1 Methods of analysis and synthesis 35
2.2.1.1 Scoping review 35
2.2.1.2 Narrative synthesis 36
2.2.1.3 Meta-analysis 36
2.3 Methods for synthesising evidence of cost effectiveness 37
2.3.1 Identifying and reviewing published cost-effectiveness studies 37
2.3.2 Development of a conceptual cost-effectiveness model 37
2.4 Handling information from the companies 38
3 RESULTS: Diagnostic accuracy and clinical impact 39
3.1 Search results 39
3.2 DERM 40
3.2.1 Summary of DERM studies 40
3.2.2 Characteristics of studies 43
3.2.3 Risk of bias 44
3.2.4 Diagnostic accuracy data from publications 45
3.2.4.1 Subgroup data by skin type 47
3.2.5 Diagnostic accuracy from unpublished data 47
3.2.5.1 Diagnostic accuracy of full teledermatology pathway 50
3.2.5.2 Referral status 51
3.2.6 Implementation, resource use and related outcomes 52
3.2.6.1 Referral and exclusion rates 52
3.2.6.2 Impact on resource use 54
3.2.6.3 Timings 54
3.2.6.4 Cancer stage 54
3.2.6.5 Acceptability to healthcare professionals 54
3.2.7 Clinical impact and patient benefit 55
3.2.7.1 Clinical morbidities and mortality 55
3.2.7.2 Health-related quality of life 55

3.2.7.3 Non-clinical benefits to patients 55
3.3 Moleanalyzer Pro 57
3.3.1 Summary of Moleanalyzer Pro studies 57
3.3.2 Characteristics of Moleanalyzer Pro studies 59
3.3.3 Risk of bias 59
3.3.4 Diagnostic accuracy 60
3.3.5 Implementation, resource use and related outcomes 62
3.3.5.1 Referral rates 62
3.3.5.2 Acceptability to healthcare professionals 63
3.3.6 Clinical impact and patient benefit 63
3.3.6.1 Non-clinical benefits to patients 63
3.4 Conclusions 65
3.4.1 DERM 65
3.4.1.1 Diagnostic accuracy 65
3.4.1.2 Clinical outcomes 66
3.4.1.3 Patient and clinician perspectives 66
3.4.2 Moleanalyzer Pro 66
3.4.2.1 Diagnostic accuracy 67
3.4.2.2 Clinical outcomes 67
3.4.2.3 Patient and clinician perspectives 67
4 RESULTS: Cost-effectiveness Review 68
4.1 Results of literature searches 68
4.1.1.1 Economic studies relating to the named technologies 68
4.1.1.2 Economic studies related to diagnostics in skin cancer 68
4.1.1.3 Other identified studies 68
4.2 Summary of identified evidence 69
4.2.1 Wilson et al. (2013) 72
4.2.1.1 Model structure 72
4.2.1.2 Mechanism of cost consequences 72
4.2.1.3 Mechanism of health consequences 73
4.2.2 Edwards et al. (2016) 73
4.2.2.1 Model structure – diagnosis of melanoma 74
4.2.2.2 Mechanism of cost consequences – diagnosis of melanoma 75
4.2.2.3 Mechanism of health consequences – diagnosis of melanoma 76
4.2.2.4 Model structure – diagnosis of BCC 77
4.2.2.5 Mechanism of cost consequences – diagnosis of BCC 77
4.2.2.6 Mechanism of health consequences – diagnosis of BCC 78

4.2.3 Wilson et al. (2018) 78
4.2.3.1 Model structure 78
4.2.3.2 Mechanism of cost consequences 79
4.2.3.3 Mechanism of health consequences 79
4.3 Discussion 79
5 Economic Models Submitted by Skin Analytics 82
5.1 Cost-utility model (Exeter Test Group/Skin Analytics) 82
5.1.1 Modelled population 82
5.1.2 Model structure 82
5.1.3 Mechanism of cost consequences 84
5.1.3.1 DERM (with or without second read) versus teledermatology 86
5.1.3.2 DERM without second read versus face-to-face assessment 87
5.1.3.3 DERM with a second read versus DERM without a second read 87
5.1.4 Mechanism of health consequences 88
5.1.5 Summary of critique 89
5.2 East Midlands Academic Health Science Network (2023) 90
5.3 NHSE AI in Health and Care Economic Evaluations 91
6 Model Conceptualisation and Identification of Evidence Gaps 93
6.1 Model conceptualisation 93
6.1.1 Decision problem 93
6.1.2 Proposed model structure 94
6.2 Clinical input parameters 100
6.2.1 Prevalence of disease and distribution by disease stage 100
6.2.2 Diagnostic accuracy data 101
6.2.3 Progression and opportunistic detection parameters 102
6.3 Health-related quality of life 104
6.4 Cost and resource use parameters 105
6.4.1 Technology costs 106
6.4.1.1 Skin Analytics DERM 106
6.4.1.2 FotoFinder Moleanalyzer 107
6.4.2 Diagnosis, treatment and follow-up costs 107
6.5 Strengths and limitations of the proposed modelling approach 108
6.6 Summary of evidence requirements 109
7 DISCUSSION 110
7.1 Statement of principal findings 110
7.1.1 Diagnostic accuracy and clinical impact 110
7.1.1.1 DERM 110

7.1.1.2 Moleanalyzer Pro 110
7.1.2 Cost-effectiveness review and stakeholder submissions 111
7.1.3 Conceptual model 112
7.2 Strengths and limitations of the assessment 113
7.2.1 Strengths 113
7.2.2 Limitations 113
7.3 Key limitations of the evidence base 114
7.3.1 Diagnostic accuracy 114
7.3.2 Clinical impact and benefit 115
7.3.3 Resource use 115
7.3.4 Cost-effectiveness 115
7.4 Patient and public inclusion 115
7.5 Equality, diversity and inclusion 116
8 CONCLUSIONS 117
8.1 Implications for service provision 117
8.2 Suggested research priorities 118
8.2.1 Diagnostic accuracy 118
8.2.2 Clinical impact 119
8.2.3 Cost-effectiveness and resource use 119
9 References 121
Appendices 127
Appendix 1: Literature search strategies 127
Appendix 2: excluded studies at full text screening stage 137
Appendix 3: Ongoing studies 140
Appendix 4: Data extraction tables 142

Table of Tables

Table 1 Study priority hierarchy for scoping review 36
Table 2 Summary of DERM studies identified 42
Table 3 Participant characteristics of DERM studies included in the review 43
Table 4 Quality assessment of DERM diagnostic accuracy studies 44
Table 5 Meta-analysis of diagnostic accuracy from DERM publications 47
Table 6 Results of "second read" assessment in the Leicestershire study 50
Table 7 Percentages of patients by referral status with autonomous DERM use 51
Table 8 Diagnostic pathway for patients in Thomas 2023 when using DERM-vB 53
Table 9 Summary of Moleanalyzer (Fotofinder) studies included in the evidence map 58
Table 10 Participant characteristics of the Moleanalyzer Pro studies included in the review 59
Table 11 Quality assessment of Moleanalyzer Pro diagnostic accuracy studies 60
Table 12 Diagnostic accuracy in Winkler 2023 61
Table 13 Diagnostic accuracy in MacLellan 2021* 61
Table 14 Characteristics of identified studies 70
Table 15 Incremental costs associated with VivaScope pathway 76
Table 16 Incremental QALYs associated with VivaScope pathway 77
Table 17 Exeter model diagnostic costs 85
Table 18 Results of cost-effectiveness of DERM 87
Table 19 Skin Analytics DERM pricing 106
Table 20 Moleanalyzer pricing 107
Table 21 Cost items required for the proposed model 107
Table 22 DERM registered, ongoing studies 140
Table 23 Ongoing evaluations of DERM in the UK – post-referral 140
Table 24 Ongoing evaluations of DERM in the UK – pre-referral 140
Table 25 Selection criteria of studies included in the synthesis 142
Table 26 Full diagnostic accuracy results (DERM studies) 142
Table 27 Included studies reporting subtype, Breslow thickness, and stage of melanoma 144
Table 28 Included studies reporting subtype and stage of SCC, BCC, and other malignancies 144
Table 29 Diagnostic pathway outcomes for patients in Thomas 2023 145

Table of Figures

Figure 1 Current diagnostic pathway for suspect skin lesions (from NICE scope)..... 26
Figure 2 Proposed positioning of AI technologies in post-referral setting (from NICE scope)..... 29
Figure 3 Study selection process (PRISMA diagram)..... 40
Figure 4 Diagnostic accuracy results from DERM publications..... 46
Figure 5 Diagnostic accuracy of DERM from pooled Birmingham and London data..... 48
Figure 6 Diagnostic accuracy of DERM from separated Birmingham and London data..... 49
Figure 7 Summary of responses DERM-005 relating to confidence in AI diagnosis..... 56
Figure 8 Summary of responses DERM-005 relating to waiting for a diagnosis..... 56
Figure 9 Summary of responses DERM-005 relating to acceptability..... 57
Figure 10 ROC curves for Moleanalyzer Pro for melanoma diagnosis (adapted from Winkler 2023) 62
Figure 11 Reassurance offered by Moleanalyzer results, percentage agreement..... 64
Figure 12 Acceptability of Moleanalyzer in diagnostic process, percentage agreement..... 65
Figure 13 Proposed model structure: Decision tree schematic (A) AI without second read vs referral to face-to-face assessment; (B) AI without second read vs teledermatology; (C) AI with second read vs referral to face-to-face assessment; (D) AI with second read..... 96
Figure 14 Markov model components: (A) True positives; (B) False negatives; (C) True negatives and false positives..... 98
Figure 15 Markov model component capturing outcomes of basal cell carcinoma false negatives..... 99

List of abbreviations

2WW	two-week wait
A&G	advice and guidance
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
AK	actinitic keratosis
BCC	basal cell carcinoma
CDC	Community Diagnostic Centre
CI	confidence interval
DA	diagnostic accuracy
DERM	Deep Ensemble for Recognition of Malignancy
dermato.	dermatologist assessment
EAG	Evidence Assessment Group
F2F	face-to-face
FN	false negative
FP	false positive
FT	flow and timing
HRQoL	health-related quality of life
I	index test
ICER	incremental cost-effectiveness ratio
IEC	intraepidermal squamous cell carcinoma
IQR	interquartile range
MA	meta-analysis
MDT	multidisciplinary team meetings
NHS	National Health Service
NPV	negative predictive value
NR	not reported
P	patient selection
PPV	positive predictive value
prosp.	prospective
PSS	Personal Social Services
QALY	quality-adjusted life year
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies-2
QUADAS-C	Quality Assessment of Diagnostic Accuracy Studies-Comparative
R	reference standard
ROC	receiver operating characteristic
SCC	squamous cell carcinoma
SLR	systematic literature review
TD	teledermatology
TN	true negative
TP	true positive
UHBFT	University Hospital Birmingham Foundation Trust
UHL	University Hospital Leicestershire
VAS	visual analogue scale
WSFT	West Suffolk Foundation Trust

ASSESSMENT GROUP REPORT

1 BACKGROUND AND DEFINITION OF THE DECISION PROBLEM

1.1 Purpose of the decision to be made

The purpose of this assessment was to investigate the use of Artificial Intelligence (AI) technologies for the analysis of skin lesions suspicious of cancer following a referral on the urgent suspected skin cancer pathway. The assessment considered the use of two technologies: Deep Ensemble for Recognition of Malignancy (DERM) (Skin Analytics) and Moleanalyzer Pro (FotoFinder systems). The assessment considered existing evidence and identified potential evidence gaps on whether these technologies have the potential to be clinically useful and cost-effective to the NHS.

1.2 Interventions

The Evidence Assessment Group (EAG) evaluated whether two AI technologies, DERM and Moleanalyzer Pro, represent an effective and reliable means of triaging cancer from benign skin lesions, alongside current clinical practice.

1.2.1 Deep Ensemble for Recognition of Malignancy (DERM) (Skin Analytics)

DERM (Skin Analytics) is a UKCA class IIa AI-based skin lesion analysis technology intended for screening, triage and assessment of suspicious skin lesions. It is indicated for use on dermoscopic images of skin lesions where skin cancer is suspected in patients aged 18 years or over.

DERM uses AI-based algorithms to provide a suspected diagnosis of a given lesion and where applicable, a referral recommendation (for example, discharge and give safety netting advice or urgent referral for suspected cancer). DERM can classify lesions as: melanoma, squamous cell carcinoma, basal cell carcinoma, intra-epidermal carcinoma, actinic keratosis, atypical nevus or benign lesions (this includes benign vascular lesion, seborrheic keratosis, dermatofibroma, solar lentigo and melanocytic benign nevus). If a lesion exhibits features of more than one lesion type, DERM uses a risk hierarchy to return the more severe suspected diagnosis. The algorithm was trained on both historical (retrospectively) and prospectively collected images from populations in the UK, US and Italy. DERM uses a fixed algorithm and does not update itself automatically.

The technology has been deployed in the NHS since April 2020, including as a triage tool following a primary care referral. Over 51,000 patients have been assessed following a GP referral on the urgent suspected skin cancer pathway, to identify patients with benign lesions who can be discharged from the pathway without requiring specialist input from secondary care. People with suspicious lesions

after DERM assessment have then been referred to a teledermatology review by a secondary care specialist.

1.2.2 Moleanalyzer Pro (FotoFinder systems)

Moleanalyzer Pro (FotoFinder Systems) is a class IIa CE marked AI-based technology intended to be used by a medical professional for non-invasive visual documentation of skin lesions and aims to help the recognition of melanoma lesions. The technology is not intended to be used to confirm a clinical diagnosis of melanoma, and can be used for any age group. The target population is people with skin lesions, moles or multiple nevus syndrome. Lesions can be between 2 mm to 20 mm and should be on intact skin without additional psoriasis, eczema, acute sunburn or on hair-covered parts of the body.

Moleanalyzer Pro is used with the FotoFinder Universe software platform. The system requires a dermoscopic image for the AI score analysis. The software can only be used with the FotoFinder dermatoscopes: Dermlite Handyscope (this is compatible with any smartphone or tablet) and with Medicam 1000.

FotoFinder provides two options: online AI where the algorithm is updated continuously and offline AI in which the algorithm can be updated annually. This AI score is based on comparisons with images of malignant skin tumours such as: melanoma, basal cell carcinoma, lentigo maligna, squamous cell carcinoma, actinic keratosis, and many others. The score indicates how similar a lesion is to these comparison images, therefore it is only meant to provide a statistical estimate of the similarity to the malignant lesion images. A score between 0-0.0.2 indicates the lesion is inconspicuous, 0.21-0.49 indicates further clarification is necessary, and 0.50-1.0 indicates a conspicuous lesion which should be observed with great attention. Moleanalyzer is already in use in some NHS centres.

1.3 Populations and relevant subgroups

The population of interest was people with skin lesions suspicious of cancer, who have been referred from primary care for further evaluation. The particular setting of interest was patients undergoing is teledermatology assessments, but all settings after primary care referral were considered.

Subgroups relevant to this appraisal were according to skin colour and type, and socioeconomic status.

1.4 Place of the intervention in the treatment pathway

In the UK, dermatology services receive about 1.2 million referrals a year and about 60% of these are suspected skin cancer pathway referrals, but only about 6% are converted to a confirmed case of skin cancer.[1] A significant proportion of people referred by GPs may not require face-to-face appointments

in dermatology departments. The GIRFT report on dermatology highlighted that there are shortages in the workforce leading to delays in the diagnosis and treatment of skin cancer.[2] Furthermore, experts in dermatology mentioned there is a low threshold for referral because GPs don't receive in-depth dermatology training and many do not have access to dermatoscopes, which are essential for confidently identifying both benign skin lesions and skin cancer.

1.4.1 Types of skin cancer

This assessment covers all types of skin cancer. This includes three main types of skin cancer: melanoma, squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), as well as other, rarer, forms of skin cancer.

1.4.1.1 Melanoma

A melanoma is a malignant tumour arising from melanocytes in the skin and is usually seen as a pigmented lesion on the skin. Melanoma is the fifth most common cancer in the UK, accounting for around 4% of all new cancer cases and more cancer deaths than all other skin cancers combined. On average, between 2016 and 2018, 16,744 new cases of melanoma were diagnosed each year in the UK (NICE, 2022).[3] The incidence of melanoma is projected to increase by 7% in the UK between 2014 to 2035.

Prognosis is highly dependent on the stage at diagnosis. For people with stage 1 melanoma (thickness is 2 mm or less, no sign that it has spread) the 5-year survival rate is almost 100%, compared with 30% for people with stage 4 melanoma (spread to distant lymph nodes or other parts of the body).

A weighted 7-point checklist is used to assess pigmented skin lesions and determine the need for referral. A pigmented lesion scoring of 3 or more on the weighted 7-point checklist is referred to the suspected cancer referral pathway.[4]

Weighted 7-point checklist:

- Major features of the lesions (scoring 2 points each):
    - change in size
    - irregular shape
    - irregular colour.
- Minor features of the lesion (scoring 1 point each):
    - largest diameter 7mm or more
    - inflammation
    - oozing
    - change in sensation.

1.4.1.2 Squamous cell carcinoma (SCC)

Squamous cell carcinoma (SCC) is the second most common type of non-melanoma skin cancer. It starts in the cells lining the top of the epidermis (outermost layer of the skin) and accounts for about 20 in every 100 skin cancers (NHS, 2020).[5] Approximately 25,000 squamous cell carcinomas of the skin are diagnosed each year (NICE, 2021).[6]

There is a small risk (up to 5%) of SCC spreading to other parts of the body, such as the lymph nodes (NHS, 2020). The risk of spread with SCCs is greater than with BCCs especially for people who are immunosuppressed. Death from squamous cell carcinoma is rare.

Actinic keratoses are dry, scaly patches of skin caused by damage from sun exposure. There is a small risk that the patches could develop into SCC if untreated (NHS, 2020).[5]

1.4.1.3 Basal cell carcinoma (BCC)

BCC is the most common form of skin cancer and accounts for about 75 in every 100 skin cancers. Approximately 75,000 BCC of the skin are diagnosed each year (NICE, 2021).[6]

BCC does not usually spread to other parts of the body, but if left untreated for a long time, they may get larger and grow deep into the skin and destroy skin, tissue and bone. In rare cases BCC can spread to other parts of the body and sometimes become life-threatening (NHS, 2020).[5] Death from BCC is exceptionally rare.

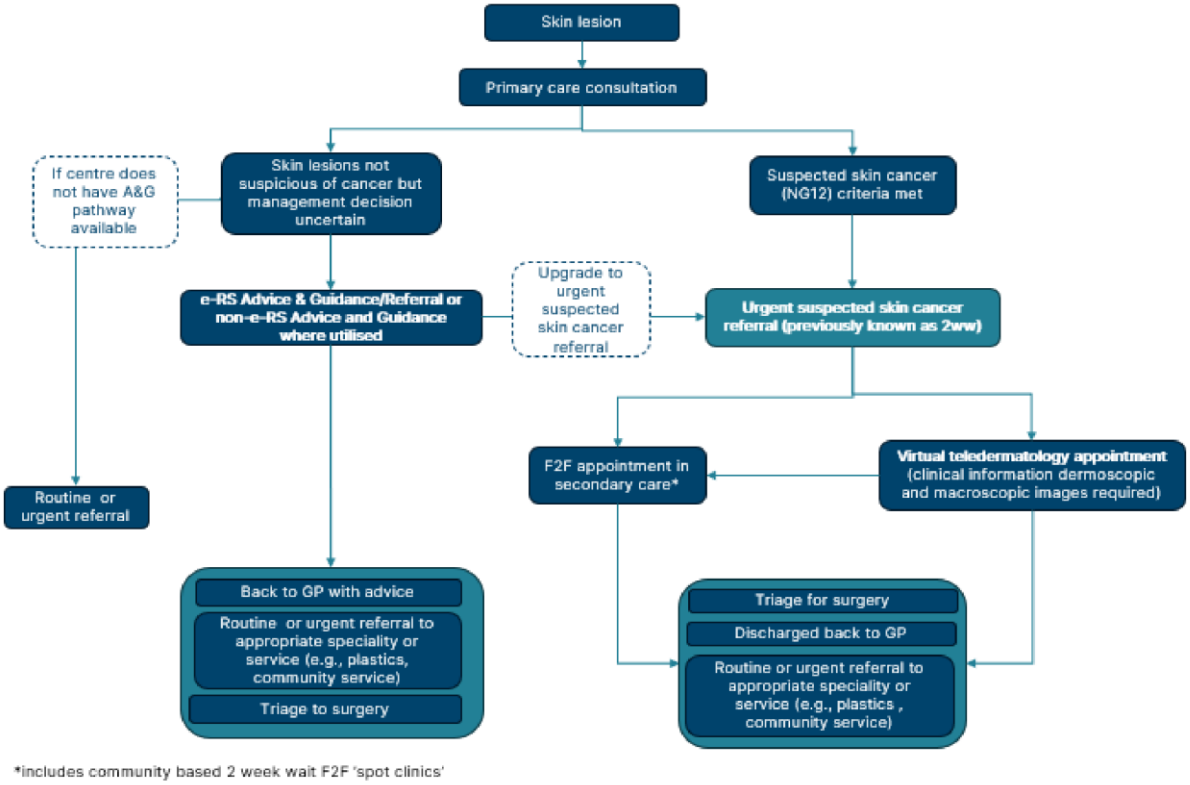
1.4.1.4 Other rare skin cancers

There are 45 other types of non-melanoma skin cancers. Merkel cell carcinoma is rarer and more aggressive than melanoma cancer. It is usually found in the head and neck region. Other types of rare non-melanoma skin cancers can be found in Appendix 1 of the NICE CSG8 guideline.

1.4.2 Current diagnostic pathway

The initial assessment of a person presenting with a skin condition occurs at the primary care level to determine the appropriate referral pathway. Traditionally, GPs directly referred everyone with suspicious skin lesions to secondary care through the urgent suspected skin cancer referral pathway where all referrals required people to attend a secondary care dermatology department for a face-to-face appointment with a consultant dermatologist. This pathway continues to exist where other clinical pathways are unsuitable or unavailable and is particularly well suited for people with multiple suspicious lesions, a history of skin cancer and other risk factors.

Figure 1 Current diagnostic pathway for suspect skin lesions (from NICE scope)



1.4.2.1 Urgent suspected skin cancer referral pathway

A person on the urgent suspected skin cancer referral pathway should receive a diagnosis or ruling out of cancer within 28 days of being referred urgently by their GP. For further details, see NHS England's webpage on faster diagnosis of cancer.[7] Section 1.7 of the NG12 guideline describes the criteria for an urgent referral for skin cancers (melanoma, SCC and BCC) to the urgent suspected skin cancer referral pathway.[8] These are summarised below.

Sections 1.7.1 to 1.7.3 of NICE guideline NG12 recommend that urgent referral using a suspected cancer pathway for melanoma should be arranged for people if:

- they have a suspicious pigmented lesion has a weighted 7-point checklist score of 3 or more,
- dermoscopy suggests melanoma,
- they have a pigmented or non-pigmented skin lesion that suggests nodular melanoma.

Additional criteria [9-11] also recommend urgent referral if:

- any new persistent skin lesion, especially if growing, pigmented, or vascular in appearance and the diagnosis is unclear,

- a new pigmented line in the nail (especially if there is associated damage to the nail), or a lesion growing under the nail,
- there is any doubt about the lesion, or there is a history of recent change,
- a biopsy has confirmed the diagnosis of malignant melanoma. Note: if a lesion is suspected to be melanoma, an urgent referral to a dermatologist or other suitable specialist with experience of melanoma diagnosis should be made, and excision in primary care should be avoided.
- a pigmented or non-pigmented skin lesion suggests nodular melanoma,
- any major features in the 7-point checklist, or any features of the ABCDE system.

Section 1.7.4 of NICE guidelines NG12 recommends a person is referred to an urgent suspected cancer pathway if they present with a skin lesion that raises the suspicion of squamous cell carcinoma. Section 1.7.5 to 1.7.6 of NICE guidelines NG12 recommend a routine referral for people if they have a skin lesion that raises the suspicion of a BCC. An urgent suspected cancer pathway referral should only be considered for a lesion that raises suspicion of BCC if there is a particular concern that a delay may have a significant impact, because of factors such as lesion site or size.

As shown in Figure 1, a referral to the urgent suspected skin cancer pathway results in either an urgent virtual teledermatology review, or an urgent face-to-face appointment in secondary care. If a primary care centre does not have a virtual teledermatology pathway available, the urgent face-to-face appointment pathway is used.

1.4.2.2 Teledermatology service

Teledermatology refers to the use of static digital images and relevant patient information to triage, diagnose, monitor or assess skin conditions remotely. If a person is referred through the urgent suspected skin cancer referral pathway, clinical information along with a high-quality macroscopic image and dermoscopic images of the skin lesion are required. Images should be taken by a healthcare professional trained in medical photography. Images can be taken:

- in a GP surgery
- at a community diagnostic centre (CDC) close to a person's home
- at a teledermatology clinic based at a hospital.

Images are sent to be assessed by a consultant dermatologist using the teledermatology service and stored in the person's record. The person can be either be:

- booked directly for surgery
- discharged back to their GP

- referred for a routine or urgent referral to the appropriate speciality or service.

Virtual teledermatology cannot be used for lesions on difficult sites such as palms, soles, scalp and intimate areas, or for people with multiple lesions. Virtual teledermatology is not used for children.

Teledermatology hubs, also referred to as Community Hubs, have been rolled out in a minority of Trusts in the UK. Patients with a GP referral for a suspicious skin lesions are sent to attend a centre in the community where a clinical photographer or healthcare assistant (CP/HCA) captures standardised photographic images of their lesion(s).

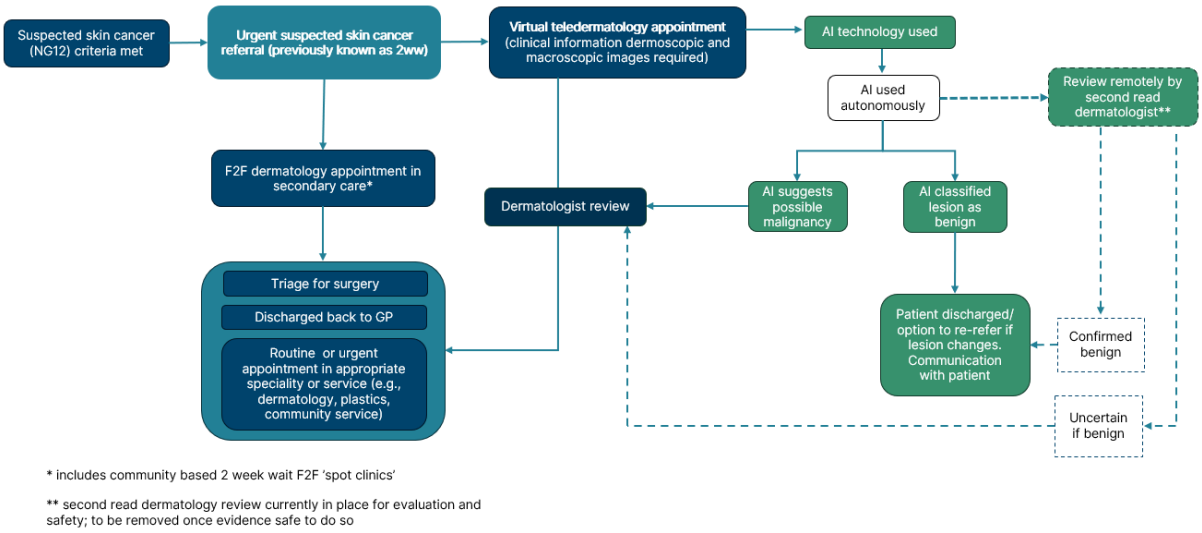
1.4.3 Potential positioning of AI technologies in the pathway

AI technologies to detect skin cancer could be used at various points in the diagnostic pathway:

1. By individuals concerned about suspect lesions, prior to consulting a GP.
2. As an adjunctive diagnostic in primary care settings (e.g. by a GP or nurse), to identify lesions that need referral
3. As an autonomous post referral assessment between primary and secondary care settings.
4. As an adjunctive diagnostic between primary and secondary care settings (e.g. teledermatology triage settings)
5. As an adjunctive diagnostic in a secondary care setting (e.g. by specialist dermatologists at face-to-face consultations)

This report focuses on settings 3 and 4, but considered evidence from other settings where it informed understanding. This aligns with where DERM is currently being used in a pilot to triage suspicious skin lesions after they have been referred by their GP on the urgent suspected skin cancer referral pathway. Figure 2 shows a possible pathway for AI use in post-referral that aligns with the current use of DERM. This post-referral assessment is used to identify those with benign lesions to be discharged from the urgent suspected skin cancer referral pathway. People identified with suspicious lesions (cases that contain at least one atypical, pre-malignant or malignant classification) from an AI assessment go on to a review by a specialist in secondary care.

Figure 2 Proposed positioning of AI technologies in post-referral setting (from NICE scope)



1.4.3.1 Adjunctive use of AI with dermatologist assessment

AI technologies could be used with a dermatologist review. After the AI assessment a dermatologist will review the results. This is done through virtual teledermatology with the aim of minimising false negative results (that is, cancerous lesions missed by the AI technology).

If the lesion is confirmed to be benign by the dermatologist, the patient is discharged from the pathway. The results are communicated to the patient and primary care referring clinician with safety net information to seek further medical advice if the lesion changes. If the dermatologist is uncertain about the diagnosis or if the AI suggests possible malignancy (whether the AI is used autonomously or not), the images are reviewed by a Trust dermatologist and triaged appropriately.

This "second read" dermatology review is currently in place for evaluation and safety, but the long-term plan is to remove this and for AI technologies to work autonomously, maximising efficient use of specialist dermatologists time (see below).

1.4.3.2 Autonomous use of AI

If AI technologies are used autonomously, a lesion classified as benign by the AI technology can be discharged without review by a dermatologist. The patient is discharged from the pathway and the results are communicated to the patient and primary care referring clinician with safety net information to seek further medical advice if the lesion changes. Lesions with suspected malignancy will be transferred to a dermatologist for teledermatology or face-to-face review.

1.4.4 Treatment of confirmed skin cancer

Treatment of skin cancer follows NICE guidance and British Association of Dermatologists guidelines.[11-13] In brief, early stage melanoma is usually treated by surgical excision; later stage

melanoma may also require lymph node resection, radiotherapy or chemotherapy. SCC and BCC are usually treated by surgical excision, but other treatments, including radiotherapy or chemotherapy may occasionally be used.

1.5 Relevant comparators

The comparator for this assessment was clinical assessment and triage of suspicious lesions through the existing diagnostic pathway without use of AI. This can include assessment by specialist dermatologists either remotely or in person.

1.6 Key outcomes addressed as part of the assessment

Outcomes fall into four main areas:

- Diagnostic accuracy
- Implementation, resource use, and practicality
- Clinical impact and patient benefit.
- Costs

1.6.1 Diagnostic accuracy

Diagnostic outcomes are:

- Diagnostic test accuracy (sensitivity and specificity, area under ROC curve)
    - Where available, separately for each type of skin cancer (melanoma, BCC, SCC, rare skin cancers)
- Proportion of cancers missed and detected
- Proportion of benign lesions missed and detected
- Proportion of referrals confirmed to be skin cancer (positive predictive value)

1.6.2 Implementation, resource use, and practicality

Key outcomes relate to resource use and timing:

- Proportion of urgent cancer referrals:
    - needing a face-to-face hospital appointment with a specialist for review of lesion
    - converted to routine referral pathway
    - resulting in a diagnostic biopsy
    - booked for surgical procedure
    - discharged back to GP
- Time to:

    - diagnosis
    - discharge
    - face-to-face consultant appointment
    - treatment (surgery)
- Cancer stage at detection
- Ease of use/acceptability of AI software by healthcare professionals
- Number of people consenting to use the technology
- Test failure rates (with reasons, e.g. image capture issues)
- Proportion of suspicious skin lesions/patients excluded (with reasons, e.g. due to lesion location or scarring)

1.6.3 Clinical impact and patient benefit.

- Clinical morbidities
    - Including distant metastases and adverse outcomes of treatment
- Mortality
- Health-related quality of life
- Non-clinical benefits to patients
    - Reassurance that lesion is not cancerous.
    - Anxiety associated with waiting for a diagnosis.
    - Acceptability of AI technologies or processes

1.6.4 Costs

Costs were considered from an NHS and Personal Social Services perspective. Costs for consideration include:

- Cost of annual subscription for AI software
- Cost of training healthcare professionals to take images and to interpret AI software results
- Cost of consultant dermatologist face-to-face appointments
- Cost of staff time to upload images to AI software platforms and to interpret the results
- Costs related to missed cancers
- Costs of consultant dermatology triage team
- Costs of teledermatology
- Costs of new services required to support AI technologies (such as establishing new teledermatology services and setting up image capture)

1.7 Objectives

The aim of the project was to investigate the clinical and cost-effectiveness of AI technologies as decision aids to triage and diagnose suspicious skin lesions, specifically the two technologies (DERM and Moleanalyzer Pro) described in Section 1.2.

To achieve this, the following objectives were proposed:

Clinical effectiveness

- To perform a rapid systematic review, and if feasible a meta-analysis, of the diagnostic accuracy of the included AI technologies
- To perform a rapid systematic review with a narrative synthesis of the clinical impact and practical implementation of the AI technologies
- Based on the results of the rapid review, to identify evidence gaps and formulate recommendations for future research.

Cost effectiveness

- To perform a rapid systematic review of published cost-effectiveness studies of alternative diagnostic strategies used to aid the diagnosis of skin cancer. This will focus on the included AI technologies but will also include alternative strategies if no evidence is identified for the included technologies.
- To develop a conceptual model that will identify likely drivers of health benefit, harms and cost associated with implementing the included AI technologies in the NHS.
- If evidence and time allows: to develop a budget impact model capturing the direct resource implications of implementing the included AI technologies in the NHS. This may additionally include threshold analysis to explore how health effects or indirect costs may impact cost-effectiveness.

2 METHODS

2.1 Systematic review methods

The systematic reviews were conducted following the general principles recommended in CRD's guidance and reported in accordance with the PRISMA statement.[14, 15]

The review was conducted as a rapid review, aimed at scoping the relevant literature and synthesise studies of key relevance to the UK health setting.

2.1.1 Search strategy

The aim of the literature search was to identify published and unpublished primary studies relating to the use of the proposed AI technologies (DERM and Moleanalyzer Pro) for identifying skin cancer.

An Information Specialist designed the search strategy in Ovid MEDLINE in consultation with the research team. The MEDLINE search strategy was checked by a second information specialist using aspects of the PRESS checklist.[16] This initial search strategy was then divided into two searches so that records highly likely to be about DERM or Moleanalyzer Pro could be identified and screened first. Search 1 contained terms for the two specific technologies and their company names. Search 2 consisted of search terms for skin cancer (in line with those types of skin cancer specified in the NICE scope document) combined with terms for AI and dermoscopy. Both searches were limited to records from 2015 onwards, reflecting the recent development of these technologies.

Bibliographic databases were prioritised for searching, based upon relevance to the topic area of the assessment. The MEDLINE strategies were adapted to run on all the databases and resources specified in the protocol. The searches were run in October 2023 on the following databases and trial registries: MEDLINE ALL (Ovid), Embase (Ovid), Cochrane Central Register of Controlled Trials (CENTRAL, Wiley), the Association for Computing Machinery (ACM) Digital Library, ClinicalTrials.gov and the WHO International Clinical Trials Registry Platform (WHO ICTRP). Records from the searches were imported into EndNote 21 (Clarivate Analytics, US.) for deduplication.

Additionally, company websites were searched to identify relevant publications and other materials relating to the technology. The companies were contacted (via NICE) to provide details of all studies (completed or ongoing) that they have conducted. The search strategies are reported in Appendix 1.

2.1.2 Study selection

Titles and abstracts were screened by one reviewer (NU, AL or MS) and random spot checks were performed by a second reviewer to streamline the screening process. All abstracts of uncertain

inclusion status were checked by a second reviewer. Papers that examined AI technologies but where the technology used was unclear were identified, but did not proceed to full text assessment.

Full papers of any records that were relevant were obtained and independently screened by two reviewers according to the inclusion criteria listed below. Any disagreements were resolved through discussion and, where necessary, consultation with a third reviewer.

A two-phase scoping process identified relevant studies. At the first phase, all relevant studies, (according to the inclusion criteria in Section 2.3) were identified. A scoping process was then used to identify studies of highest quality and most relevance to the decision problem for full data extraction and synthesis (see Section 2.7).

2.1.3 Inclusion criteria

2.1.3.1 Population

People with skin lesions suspicious of cancer, presenting in primary care, local in-person diagnostic clinics, teledermatology, or secondary care settings. The applicability of populations and settings to the NICE scope was assessed and accounted for.

2.1.3.2 Interventions

DERM (Skin Analytics) and Moleanalyzer Pro (FotoFinder systems) used either alone, or in combination with clinical judgment. All versions of the technologies were considered, and their applicability to current NHS practice was assessed and accounted for.

2.1.3.3 Comparators

Clinical judgement and triage of suspect skin lesions as part of the current diagnostic pathway, without AI use. This included, but was not restricted to, urgent teledermatology services and urgent face-to-face secondary care appointments. The applicability of comparators to the NICE scope was assessed and accounted for. Studies without a comparator were also eligible.

2.1.3.4 Reference standard

Histological confirmation or rejection of malignancy from a biopsy of the suspect lesion. For unbiopsied lesions, confirmation of non-malignancy by specialist dermatologists, or ground truth as established by panels of dermatologists, were accepted.

2.1.3.5 Outcomes

See Section 1.6 for a full list of intended outcomes.

2.1.3.6 Study designs

All studies that included adult patients with skin lesions suspicious of cancer, of any design, were eligible for inclusion. Priority was given to studies with prospective recruitment of participants over retrospective reviews. Proof-of-concept, simulations and algorithm training studies were excluded.

2.1.4 Data extraction

A data extraction form was developed and piloted. For the initial scoping process data on intervention, study location and size, setting, type of outcomes reported, design and key quality indicators (randomisation, whether studies are comparative, prospective vs. retrospective design etc.) were extracted by one reviewer and independently checked by a second reviewer.

For studies selected for full data extraction and synthesis, full data on the intervention, patient characteristics and all reported outcomes were extracted by one reviewer and independently checked by a second reviewer. Discrepancies were resolved by discussion, with involvement of a third reviewer where necessary. Where feasible, data were electronically extracted from figures and tables presented in publications.

Data from relevant studies with multiple publications were extracted and reported as a single study. The most recent or most complete publications were used in situations where we could not exclude the possibility of overlapping populations. Where there was evidence that an AI technology has developed or changed over time, only the most recent and complete studies were included. Studies reported as conference abstracts only were excluded.

2.2 Quality assessment

At the scoping phase, all studies were assessed for broad quality using the hierarchy presented in Table 1.

Prioritised diagnostic accuracy studies that reported sensitivity and specificity were assessed using the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2) tool[17] and comparative diagnostic accuracy studies (i.e. with more than one index test) were assessed using the QUADAS-C (Quality Assessment of Diagnostic Accuracy Studies-Comparative) tool, which include items on the quality and applicability of studies.[18] Input from a clinical expert on the applicability of the studies was sought where appropriate. Studies were assessed by a reviewer and checked by a second.

2.2.1 Methods of analysis and synthesis

2.2.1.1 Scoping review

Initially a scoping process was used to classify identified studies for relevance to the decision problem, based on study quality, setting, outcomes reported and relevance to the NHS and population

in the NICE scope (people referred on the urgent suspected skin cancer pathway). The priority hierarchy for the quality of diagnostic accuracy and clinical evidence studies that was used is presented in Table 1.

Table 1 Study priority hierarchy for scoping review

Priority level	Diagnostic accuracy	Clinical and implementation evidence
1 (highest)	Prognostic cohort comparative studies	RCTs
2	Prognostic cohorts of AI technology only	Non-randomised cohort studies
3 (lowest)	Retrospective and case-control studies	Retrospective and case-control studies. Patient or clinician surveys

For each included AI technology only, the studies at highest priority level for that technology were taken forward for full data extraction and narrative synthesis. For example, if there are RCTs of a technology, non-randomised studies were not be considered further. Studies at lower priority levels were taken forward if they were of particular relevance to the NHS and the population in scope or report outcomes were not presented in higher-quality studies.

Studies conducted in teledermatology settings, or equivalent early diagnostic clinics, were preferred for full data extraction and synthesis. However, given variation in diagnostic processes in different countries, other settings, including primary care and specialist dermatology clinics, in studies outside of the UK were considered where no evidence in the preferred settings is available.

2.2.1.2 Narrative synthesis

For studies taken forward from the scoping phase for full synthesis, a narrative synthesis approach was used following appropriate guidance.[19] The results of data extraction for each outcome were presented in structured tables and figures as appropriate, with a text summary. Studies were grouped by population and intervention characteristics. Tabulated results were then compared across studies, interventions and outcomes to identify the broader evidence of effectiveness. Evidence was summarised for specified subgroups (skin colour, skin type and socioeconomic status) where available.

2.2.1.3 Meta-analysis

Where sufficient data on diagnostic accuracy were available, the EAG had planned to pool data relating to sensitivity and specificity by AI technology using bivariate meta-analytic techniques. As data were insufficient for this, separate meta-analyses of sensitivity and specificity were performed instead, using standard random-effects methods. Subgroup analyses were intended for relevant subgroups (skin colour, skin type and socioeconomic status) but no suitable data were available. Heterogeneity was investigated by examining data plots and receiver operating characteristic (ROC) curve plots.

2.3 Methods for synthesising evidence of cost effectiveness

Relevant cost-effectiveness evidence on the use of AI technologies with Class IIa designation (DERM, Moleanalyzer Pro) for early detection of benign skin lesions were identified and narratively summarised. The aim of the review was to examine existing decision-analytic models used to assess the cost-effectiveness of the named AI software options against any comparator, in order to inform parameterisation of a conceptual model to identify key issues, evidence gaps, and areas of uncertainty to help direct future data collection and research.

2.3.1 Identifying and reviewing published cost-effectiveness studies

The searches described in Section 2.1.1 were used identify relevant economic evaluations of name AI technologies in people with suspicious skin lesions in any setting. Study designs included in the review were budget impact models, return on investment analysis, and other cost-only analyses, as well as full economic evaluations considering both costs and consequences (including cost-effectiveness, cost-utility and cost-benefit analyses).

It was anticipated that no relevant studies would be identified for the name technologies therefore additional searches were conducted to identify studies looking at any technology used to aid diagnosis of skin cancer in an NHS setting. The search strategy combined terms for skin cancer with terms for economic evaluations. A search filter was applied to limit retrieval to UK studies, along with date limit of 2013 onwards and a limit to studies published in English. MEDLINE (Ovid) and Embase (Ovid) were searched on 6th November 2023.

Identified economic models were reviewed and discussed in detail, with the aim of informing the design and parameterisation of conceptual model. Material provided by submitting stakeholders pertaining to the value case for their product was also reviewed.

We aimed to answer the following decision questions on the basis of the identified published evidence, and material submitted by the developers of the included technologies:

1. What are the cost and resource use implications of the use of AI technologies following an urgent suspected skin cancer referral to identify benign skin lesions?
2. What would a health economic model to estimate the cost-effectiveness of AI technologies to identify benign skin lesions in this setting look like, and what are the key evidence requirements necessary to populate such a model?

2.3.2 Development of a conceptual cost-effectiveness model

The structure of a conceptual model for AI tools will be necessarily pragmatic and flexible in terms of the number of different diagnostic and care pathways included, and the two potential use cases for AI

technologies in a post-referral setting. The EAG is also clear on the structural limitations of a model of this design, which, whilst based on precedent, may not be able to provide a granular representation of the diagnostic accuracy and outcomes for the many indications included under the skin cancer umbrella, and may not fully represent the impact of these technologies upon consultant capacity and waiting times, amongst other important motivating factors for the present assessment.

The conceptual model described comprises an overview of the structure of a cost-utility model for the assessment of AI technologies for the triage of suspected cancer cases referred on the 2WW pathway. The structure of the conceptual model was designed considering the strengths ad limitations of previously published diagnostic models for skin cancer in an NHS setting, and evidence submitted by stakeholders. The exercise sought to identify key inputs necessary for the linkage of short-term diagnostic accuracy metrics with long-term outcome.

The conceptual model was developed in alignment with the NICE reference case and is described in full in Section 6.

2.4 Handling information from the companies

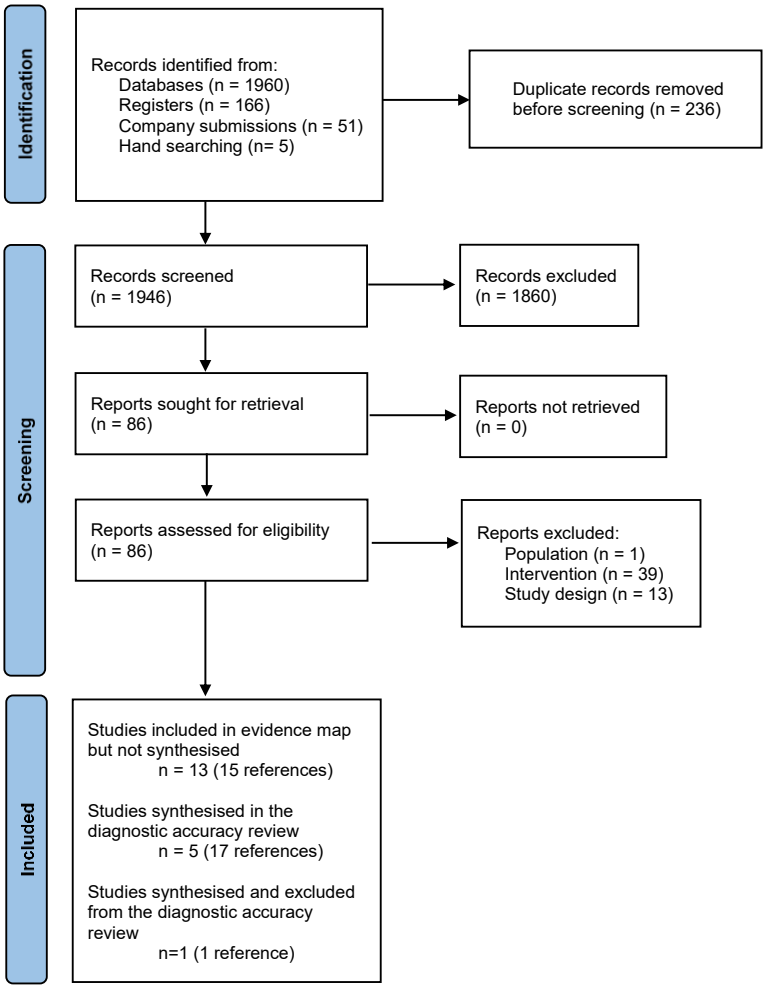
All information submitted by the companies received by the EAG in October 2023 was fully assessed. Information supplied during November 2023 was subject to a more limited assessment. All material supplied was assessed to determine whether it met the inclusion criteria for the reviews. Included studies were data extracted in accordance with the procedures outlined in this protocol.

3 RESULTS: DIAGNOSTIC ACCURACY AND CLINICAL IMPACT

3.1 Search results

Figure 3 presents an overview of the study selection process in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram. A first bibliographic search with named AI technology terms was complemented by a second search with no named technologies, references from company submission and hand searching. A total of 1946 unique records were retrieved and screened. After title and abstract screening, 86 references were retrieved for full text selection. Six studies, including four evaluations of DERM,[20-23] and two studies of Moleanalyzer,[24, 25] were included in the review. In addition, thirteen unique studies, including 11 studies of Moleanalyzer,[26-36] and two studies of DERM[37, 38] that were considered lower priority were included in an evidence map only, and were not fully synthesized. These studies were classed as lower priority because they were either conducted outside of clinical practice (e.g. retrospective design on selected sample of images), or because they evaluated an older or outdated version of an AI technology. A list of studies excluded at full text screening stage is reported in Appendix 2. The publications identified from database searches corresponded with those listed on the company website. The submissions from Skin Analytics and Fotofinder did not include any additional eligible studies, although they provided further details from relevant studies not contained in published material.

Figure 3 Study selection process (PRISMA diagram)



3.2 DERM

3.2.1 Summary of DERM studies

Table 2 summarises the six studies of DERM included in the evidence map. Two studies evaluated an early algorithm version to test the accuracy of DERM for melanoma detection. Phillips (2020)[38] used a retrospective design to train an AI algorithm to detect melanoma from a selected sample of lesions including histologically confirmed melanoma and benign pigmented lesions, along with a meta-analysis of naked-eye examination with or without dermoscopy. Phillips (2019)[37] was a diagnostic cohort where images of suspicious and control skin lesions were collected prospectively in UK hospitals on three different cameras, and retrospectively analysed. Both studies were excluded from the main review as they evaluated an earlier version of DERM for the detection of melanoma only, in a selected sample of histologically diagnosed lesions.

A further four studies were identified, including three prospective diagnostic cohorts,[20-22] and a before-and-after study.[39] All evaluated a more recent version of DERM in a post-referral setting in England. Two studies reported being conducted in a 'teledermatology hub' for triage within the 2WW referral pathway.[22, 23] These studies included patients with a suspicious skin lesions with a GP referral to attend a teledermatology hub where a clinical photographer or healthcare assistant (CP/HCA) captured standardised photographic images of their lesion(s). Following DERM assessment, lesions classed as high-risk were triaged to urgent virtual review by a hospital dermatologist. Lesions classed as low-risk were sent for remote review by a second reader (consultant dermatologist), who would either discharge the patient if in agreement with AI, or overturn the AI risk assessment and proceed with an urgent referral to a hospital dermatologist.

Three studies evaluated sensitivity and specificity of DERM alone against a reference standard that combined histopathology and/or clinical assessment (for non-excised lesions);[20-22] of those, two compared the accuracy of DERM against dermatologist assessment alone concurrently.[20, 21] One unpublished study only reported sensitivity estimates for lesions with histopathological diagnosis;[23] however, the study was included in the review as it also reported clinical output outcomes, and clinician and patient views.

Table 2 Summary of DERM studies identified

Study	Linked material	Design N patients (lesions)	Setting	Period	Diagnostic (index) tests	Outcomes reported
Included in the review						
DERM-003 [NCT04116983] [20]	Marsden[40] Austin[41]	Prosp. DA cohort N= 544 (585)	Hospital	Jun 2020-Feb 2022	DERM [redacted] Dermatol.	DA
DERM-005 Chelsea & Westminster [NCT04123678] [21]	Kawsar 2023 [42] DERM 2023_Q3[43] Skin Analytics 2023[44]	Prosp.DA cohort N= 617 (782)	Hospital	Feb 2020-Aug 2021	DERM [redacted] Dermatol. (TD)	DA Referrals Patient views Economic
UHBFT and WSFT [22]	Andrew[45] DERM 2023_Q3[43] Skin Analytics 2023[46] Jenkins (undated)[47]	Prosp. DA cohort N= NR (8571)	Hospital/'TD hub'	Jul 2021-Oct 2022	DERM version A [redacted] DERM version 'B' [redacted]	DA
UHL [23]	Baker (2023)[39] Skin Analytics 2023[48] Baker (undated)[49]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Included in evidence map only						
Phillips 2019[37]	NA	Algorithm training & prosp/retros. DA cohort N= 501 (551)	Hospital	Sept 2018-Feb 2019	DERM[+] (pre-Aug 2019)	DA
Phillips 2020[38]	NA	Algorithm training +MA N= NR (7102)	NA	NR	NR	DA

Abbreviations: DA: diagnostic accuracy; TD: teledermatology; MA: meta-analysis; UHBFT: University Hospital Birmingham Foundation Trust; WSFT: West Suffolk Foundation Trust; UHL: University Hospital Leicestershire; Prosp.: prospective

*Referrals, procedure duration, waiting time; [#] AI TD introduced in Mar 2022; [+]"earlier version' than DERM v3, only for melanoma, pre-Aug 2019 update.

Based on clinical trial registration, two completed or ongoing studies with no published results were identified.[50] [51] Both are outside the UK, so may be of less relevance to this assessment. These are summarised in Appendix 3, Table 22

A number of evaluations are being carried out across the UK in the post-urgent suspected cancer referral setting, as well as in the pre-referral community setting. The company reported in their November 2023 submission to NICE that outcome data for a number of these evaluations were expected in the fourth quarter of 2024. Further details are presented in Appendix 3, Table 23 and Table 24.

3.2.2 Characteristics of studies

Table 3 summarises the characteristics of participants of DERM studies included in the review. Further participant selection criteria are summarised in Appendix 4, Table 25. Where reported, the large majority of participants were white and very few patients had darker skin (Fitzpatrick types IV-VI). Lesions were most often located on the face and scalp, followed by the chest/back. The proportion of lesions with melanoma was lower in DERM 005, and SCC and BCC rates were higher in DERM 003. No participant characteristic details were reported for the Leicestershire study.[23]

Table 3 Participant characteristics of DERM studies included in the review

	N patients (lesions)	**Age (range)**	**% female**	**Fitzpatrick skin type (%)**	**Ethnicity (%)**	**Lesion location (%)**	**Cancerous lesions (%)[#]**
DERM-003 [20]	544 (585)[+]	Median 73 (18-97)	50	I: 21 II: 57 III: 20 IV: 1 V-VI: 1	White: 94 Black: 0 Asian: 1 Other: 0 Missing/NR: 4	Face/scalp: 46 Posterior chest and back: 15 Arms: 14 Legs: 12 NR/missing: 13	Melanoma: 2.7 SCC: 7.5 BCC: 33.7 Other: 0.3
DERM-005 Chelsea & Westminster [21]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
UHBFT and WSFT [22]	7625 (8571)*	NR (18-93)	NR	I: 8 II: 25 III: 18 IV: 3 V-VI: 1 NR: 44	NR	NR	Melanoma: 2.9 SCC: 3 BCC: 7.2 Other: 0.2

*[#] Expressed as % of all lesions with confirmed diagnosis. [+] Patient/lesions with DERM assessment and confirmed diagnosis. * Participants/lesions received DERM assessment with confirmed diagnosis following referral to trust (and second read for lesions classed by DERM as low-risk).*

3.2.3 Risk of bias

Results of the quality and applicability assessment are reported in Table 4. All studies were at high risk of selection bias due to the exclusion of a significant proportion of participants (15.6% to 27.4% where reported) that would have otherwise been eligible for assessment in clinical practice.[22] The performance of AI is likely to be significantly improved by the exclusion of some of these lesions (e.g. images with body hair, tattoos, subungual lesions).

Two studies (DERM-005 and Thomas 2023)[21, 22] reported separate results for pre-specified thresholds and post-hoc thresholds, therefore the risk of bias was low and high for these respectively. As is standard practice, a significant proportion of lesions did not undergo histology. However, the risk of bias regarding the reference standard was considered to be low in studies that confirmed the absence of cancer using expert consensus and sufficient follow-up. One study (DERM-003)[20] did not report sufficient details on the conduct of the reference standard and was at unclear risk of bias for this domain. There were no significant concerns regarding flow and timing.

All studies raised concerns with regards to the applicability of their populations; the high rate of exclusion of participants with suspected lesions that would normally be seen in practice is a significant limitation. In response to a clarification request, the company noted that the versions of DERM used in all three studies (DERM [redacted] and [redacted]) were older than the current version used in the UK ([redacted]) which, among other elements, includes a different set of thresholds for sensitivity and specificity. Therefore, the applicability of the diagnostic accuracy results for DERM to current practice is uncertain. The teledermoscopy devices used in two studies (Dermlite DL1 Basic (DermLite LLC) system)[20, 21] were considered out of date following clinical advice and therefore raised concern about their applicability to current practice. There were no concerns regarding the applicability of reference standards.

Table 4 Quality assessment of DERM diagnostic accuracy studies

Study	Test	Risk of bias				Applicability concerns		
		P	I	R	FT	P	I	R
DERM 003	DERM	✗	✓	?	✓	✗	?	✓
DERM 005	DERM	✗	✓/✗*	✓	✓	✗	?	✓
Thomas (2023)	DERM	✗	✓/✗#	✓	✓	✗	?	✓
DERM 003	Dermato.	✓	✓	?	✓	✗	✗	✓
DERM 005	TD	✓	✓	✓	✓	✗	✗	✓
DERM 003	DERM vs. dermato.	✓	✓	?	✓	n/a	n/a	n/a
DERM 005	DERM vs. TD	✓	✓/✗*	✓	✓	n/a	n/a	n/a

P = patient selection; I = index test; R = reference standard; FT = flow and timing. Dermato: dermatologist assessment. TD: teledermatologist assessment

✓ indicates low risk; ✗ indicates high risk; ? indicates unclear risk.

* Low risk for the main analyses (pre-specified thresholds for sensitivity and specificity), and high risk for the results of post-hoc analyses where the target sensitivity estimates for melanoma, SCC and BCC were amended to match the DERM algorithm to other settings in 'live development'.
# Low for version A (pre-algorithm change), high for version B (post-algorithm change). The algorithm was changed during the study to improve specificity.

3.2.4 Diagnostic accuracy data from publications

The three fully included studies all reported diagnostic accuracy data for DERM. Studies reported diagnostic accuracy for all melanomas combined and by melanoma type. In all studies the diagnostic accuracy reported was for autonomous use of DERM, without additional assessment by dermatologists.

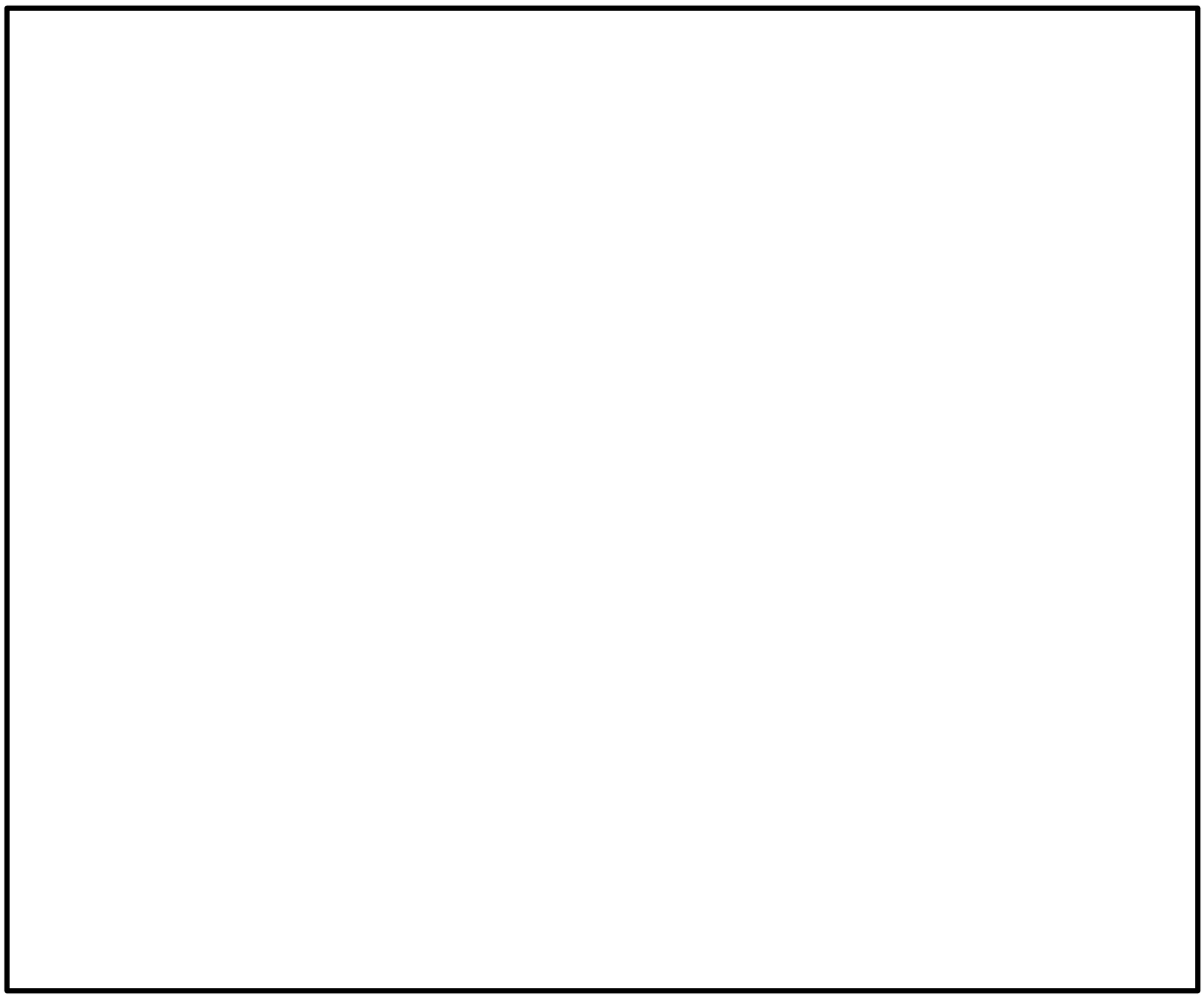
DERM-003[20] reported diagnostic accuracy for three different smartphone cameras when used to take images of lesions (iPhone 11, iPhone 6s, Samsung 10). The EAG have chosen to only report results for the iPhone11, as this was the most recently released phone considered. It should be noted that there were variations in diagnostic accuracy according to phone used. It also reported diagnostic accuracy for dermatologists without AI use.

Thomas (2023)(UHBFT and WSFT)[22] reported results separately for Birmingham and West Suffolk centres. It also reported results for two versions of DERM: DERM-vA (used July 2021 to April 2022) and DERM-vB (used April to October 2022). As DERM-vB appears to have superseded DERM-vA we only report results for the more recent DERM-vB for this study.

Results for DERM-005 were extracted from a preprint manuscript by Marsden et al.[21] This compared DERM to standard of care (dermatologists without AI).

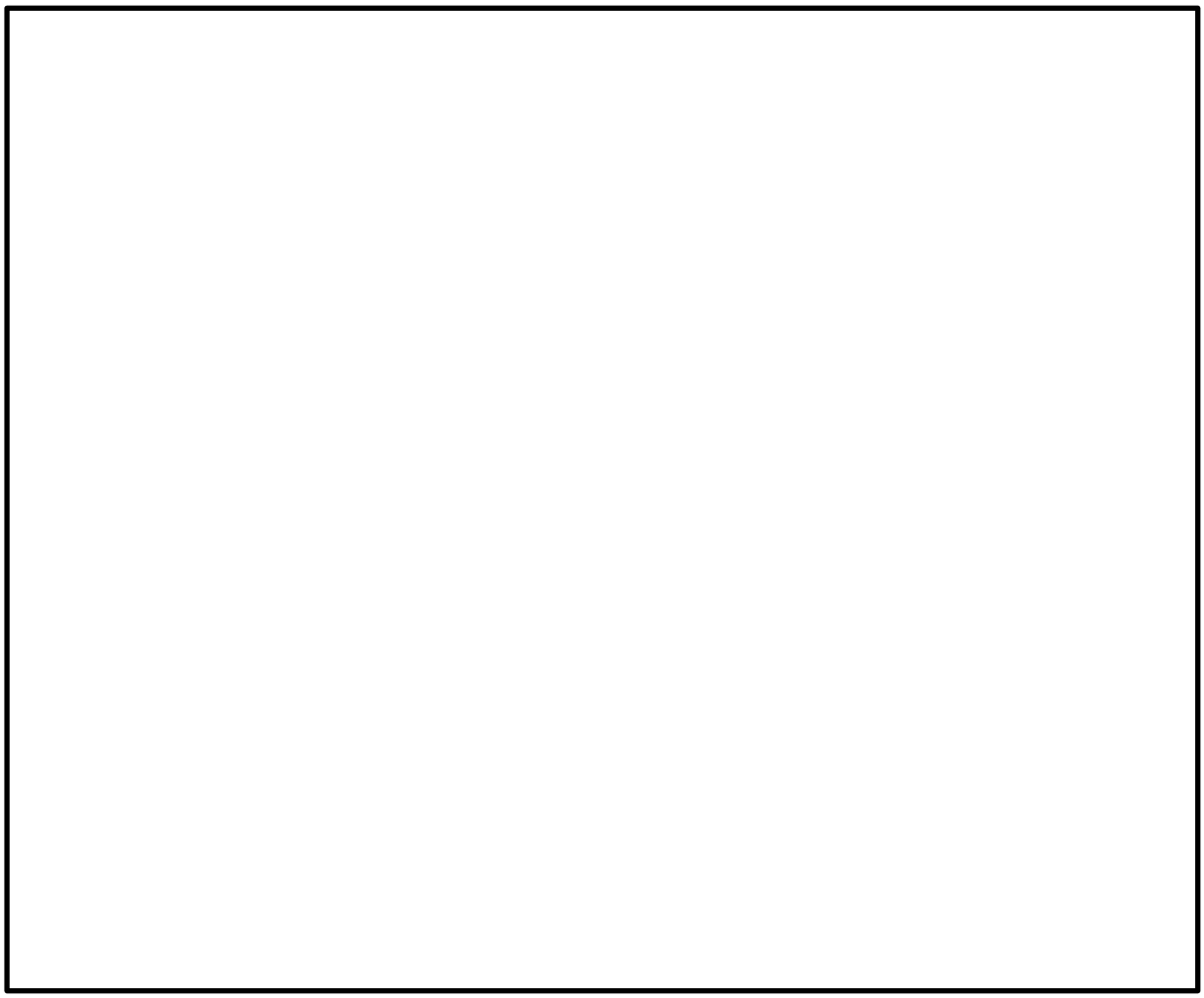
[redacted] Data were extracted from Figure 2 of this preprint, which reported the full categorisation of lesions by true diagnosis and test result, from which sensitivity and specificity estimates were calculated.

Figure 4 Diagnostic accuracy results from DERM publications



summarises the diagnostic accuracy from the three included studies.

Figure 4 Diagnostic accuracy results from DERM publications



Meta-analyses of sensitivity and specificity were performed where two or three studies reported data. These were separate, univariate analyses as data were too limited for bivariate meta-analysis. The meta-analysis results are presented in Table 5. These results suggest a high sensitivity when using DERM autonomously without assessment by a dermatologist is achievable, and may be higher than achievable using a standard diagnostic pathway without DERM. However, some malignant lesions will still be missed. The specificity of DERM is lower than for dermatologists. In particular, specificity was much lower for detecting SCC and BCC, suggesting that DERM has some difficulty in distinguishing these types of cancer from benign lesions.

In DERM-003, for detecting benign lesions, the sensitivity of DERM was significantly lower compared with face-to-face dermatologist assessment (DERM: 43.9% (95% CI 37.4 to 50.6); dermatologist: 73.9% (95% CI 67.6–79.4) although it had comparable specificity (DERM: 93.3% (95% CI 90.0–95.6); dermatologist: 93.7% (95% CI 90.5-95.9)). Hence around 56% of benign lesions

were classified as not benign by DERM, compared with 26% for dermatologists, and approximately 7% and 6% of non-benign (but mostly pre-malignant) lesions were misclassified as benign by DERM and dermatologists respectively.

Further diagnostic accuracy results from studies of DERM are reported in Appendix 4, Table 26.

Table 5 Meta-analysis of diagnostic accuracy from DERM publications

Test	Cancer	Studies	Sensitivity	95% CI		Specificity	95% CI	
DERM	Any malignancy	All	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
DERM	Melanoma	All	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Dermatologists	Any malignancy	DERM-003 & DERM-005	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

3.2.4.1 Subgroup data by skin type

Two studies reported separate diagnostic data for Fitzpatrick skin types V and VI. In Thomas 2023, of 159 lesions assessed, 94 lesions had a final diagnosis, including BCC (n = 1) and IEC (n = 1), and actinic keratosis (n = 1), all correctly referred by DERM (vA or vB).[22] Three atypical nevus were pending face-to face assessment, and the remainder were benign with a benign specificity of 44.3% (39/88). DERM 003 found no Fitzpatrick skin types V and VI.[20]

3.2.5 Diagnostic accuracy from unpublished data

In addition to data in published and unpublished papers, Skin Analytics also provided some original data from the Birmingham and Chelsea & Westminster study centres. This data reported all lesions assessed in those centres from April 2022 up to end of September 2023. These data therefore overlap with the data from publications, but appear more up-to-date. The EAG assumes that all patients were assessed using DERM-vB, given the initiation date. We assume that this data includes all DERM-vB data from the UHBFT and WSFT study up to October 2022, as reported in Thomas 2023. We assume this includes some patients from DERM-005, although the overlap is unclear.

The supplied data also reported detailed numbers of patients by both DERM results and "ground truth" diagnosis, enabling a more thorough analysis of the diagnostic accuracy of autonomous use of DERM than was possible using published data. Diagnostic accuracy was calculated for these data in two ways. The "Exact" analysis considered a DERM result to be a true positive only if it matched exactly the ground truth diagnosis (i.e. a melanoma diagnosed by DERM was a melanoma; an SCC diagnosed by DERM was an SCC). An "All malignant" analysis considered a DERM result to be a true positive if a malignant lesion was diagnosed as malignant even if not in the correct category (i.e. if an SCC was diagnosed by DERM as a melanoma, or vice versa). This "All malignant" analysis approximately matched that performed by the company.

Diagnostic accuracy results from combining the Birmingham and Chelsea & Westminster centres are summarised in Figure 5, and results for the two centres separately are given in Figure 6.

Figure 5 Diagnostic accuracy of DERM from pooled Birmingham and London data

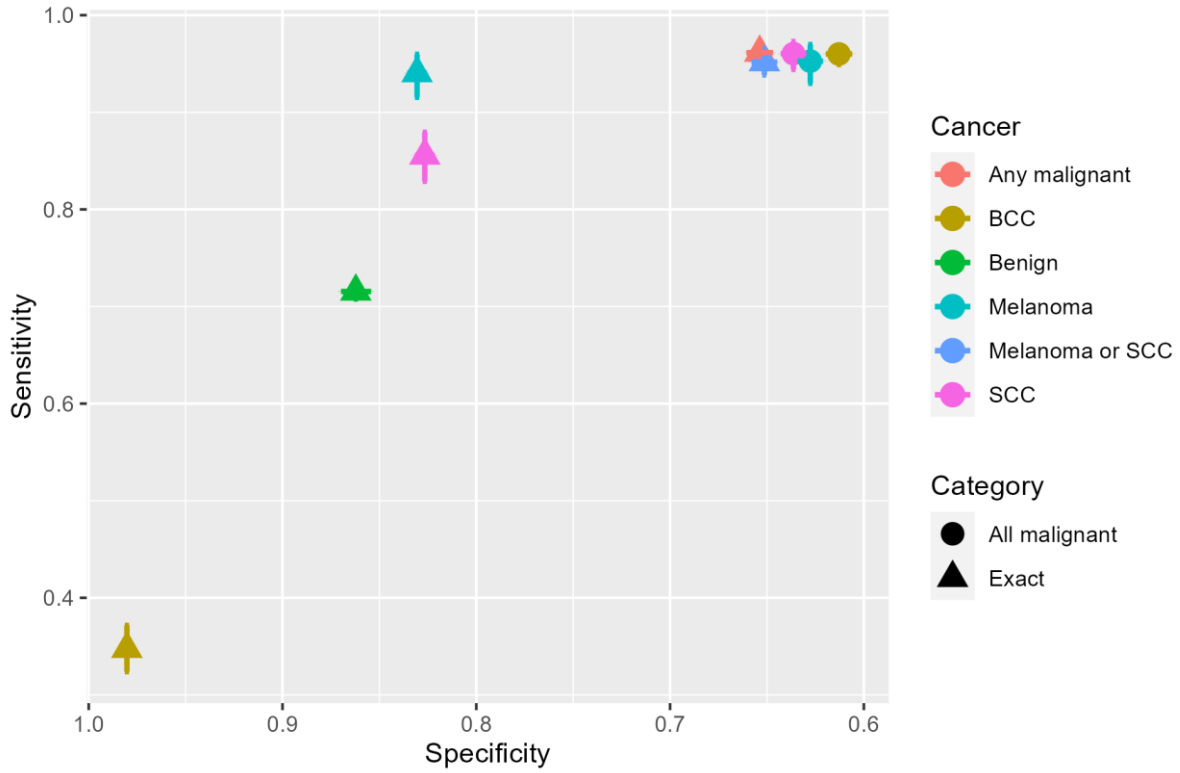
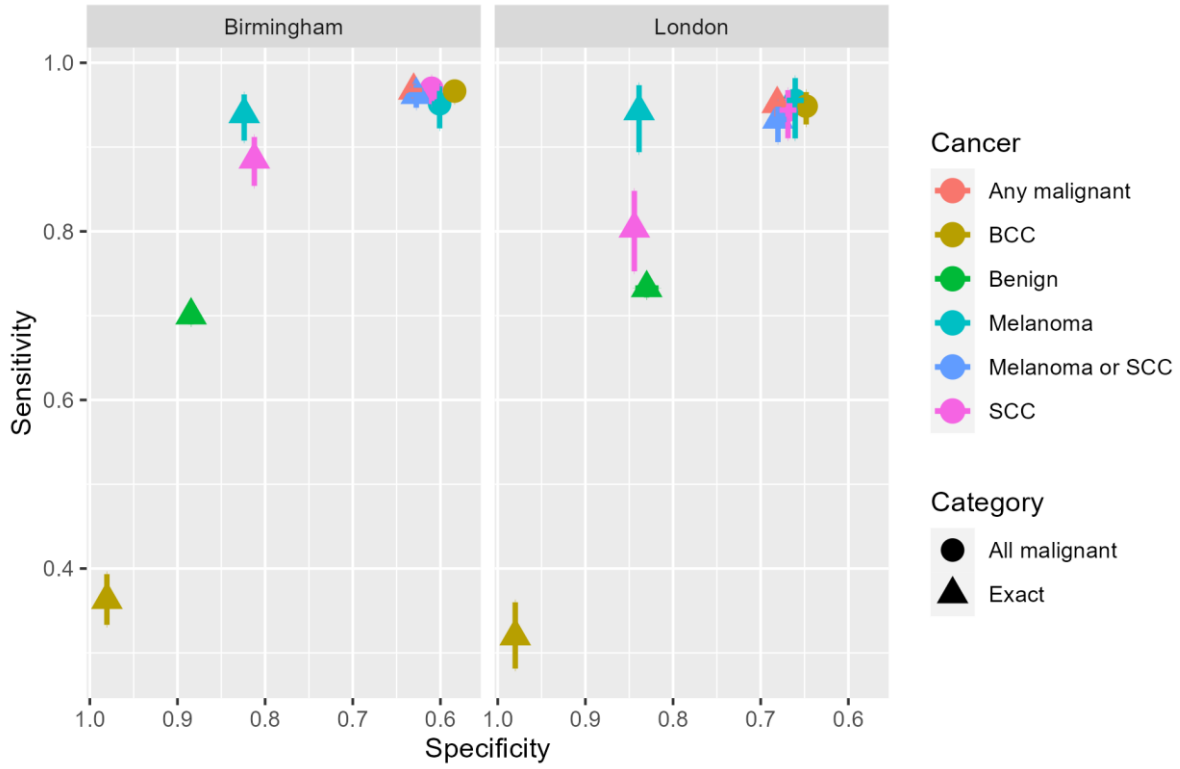


Figure 6 Diagnostic accuracy of DERM from separated Birmingham and London data



These results show a high sensitivity of DERM for detecting malignant lesions when using the "all malignant" classification. For example, detecting any malignant lesion had a sensitivity of 96.1% (95% CI 95.4 to 96.8), and sensitivities were 95% or higher for all types of cancer. Sensitivities were similar in Birmingham and London. The specificity for detecting any malignancy was 65.4% (95% CI 64.7 to 66.1). Specificities varied by type of cancer and were slightly lower in Birmingham than in London, but were generally between 60% and 70%. These results are broadly similar to those extracted from publications.

When using the "Exact" classification there is a decrease in accuracy. For melanoma the sensitivity remains at near 95%, but for SCC and BCC the sensitivity declines substantially. This suggests that both SCC and BCC lesions may be being misclassified as more serious malignancies by DERM (i.e. SCC as melanoma and BCC as SCC or melanoma).

For the detection of explicitly benign lesions the sensitivity was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0). Hence around 28% of benign lesions were classified as not benign by DERM, and 14% of non-benign (but mostly non-malignant) lesions were misclassified as benign.

It should be noted that the reference standard in this analysis was usually a "ground truth" diagnosis made by dermatologists where the lesion was judged to be non-malignant. Therefore, the diagnostic accuracy of DERM may be slightly incorrect as some genuinely malignant lesions may have been incorrectly classified as benign by dermatologists. This also means that estimates of the diagnostic accuracy of dermatologists without DERM may not be reliable.

3.2.5.1 Diagnostic accuracy of full teledermatology pathway

Diagnostic accuracy reported in publications and in the original trial data provided relates only to autonomous use of DERM, and not to the full teledermatology pathway, with or without DERM.

Diagnostic accuracy of the full pathway is largely unknown. Data on assessments by dermatologists after DERM assessment were not reported in publications. In all studies patients who were discharged by a dermatologist were not tested further, so there was no diagnostic reference standard applied.

The unpublished Edge Health report[23] on the Leicestershire study included some data on dermatologist assessment of lesions after the DERM assessment. A summary of these data is given in Table 6. This suggests that a "second read" of lesions classed as benign by DERM [redacted] when compared to using autonomous DERM.

This suggests that using a "second read" for lesions classed as benign by DERM could [redacted] After a final teledermatology assessment this would [redacted] The sensitivity is uncertain because of the lack of a perfect reference standard. However, if the sensitivity of autonomous DERM is 95%, then use of a "second read" could [redacted] based on the Leicestershire data.

Table 6 Results of "second read" assessment in the Leicestershire study

DERM result	After "second read"	After final assessment by Trust dermatologist	Number of lesions	Number of malignant lesions
Benign	Benign	[*Not used*]	[redacted]	[redacted]
Benign	Possibly malignant	Benign	[redacted]	[redacted]
Benign	Possibly malignant	Possibly malignant	[redacted]	[redacted]
Malignant	[*Not used*]	Benign	[redacted]	[redacted]
Malignant	[*Not used*]	Possibly malignant	[redacted]	[redacted]

** No reference standard applied; dermatologist assessment assumed correct*

3.2.5.2 Referral status

As the supplied data included full data on number of malignancies it was possible to estimate how autonomous use of DERM (without a "second read" by a dermatologist) would impact on onward referrals and discharge rates. For this analysis it is assumed that all melanoma, SCC or other-non-BCC malignancy cases should receive an urgent referral; BCC and Bowen's disease should receive a routine referral, and all other case should be discharged or treated locally without referral. We note that this may not be exactly what might happen in practice. The results are summarised in Table 7.

Table 7 Percentages of patients by referral status with autonomous DERM use

Group	**Percentage of total DERM population**	**95% CI**	
Urgent referrals	**39.0**	**38.3**	**39.6**
Correct urgent referrals (Melanoma or SCC, true positive)	5.8	5.5	6.2
Needless urgent referral (False positive)	33.1	32.5	33.8
Missed urgent referral (False negative)	0.3	0.2	0.4
Underdiagnoses (Urgent referral classified as routine)	0.1	0.1	0.2
Routine referrals	**8.7**	**8.3**	**9.1**
Correct routine referrals (true positive)	3.8	3.5	4.0
Needless routine referral (False positive)	4.9	4.6	5.2
Missed routine referral (False negative)	0.6	0.5	0.7
Overdiagnoses (Routine referral classified as urgent)	7.4	7.1	7.8
All referrals (urgent and routine)	**47.7**	**47.0**	**48.4**
Correct referrals (urgent and routine) (True positive)	17.1	16.6	17.6
Needless referral (False positive)	30.6	29.9	31.2
Missed referral (False negative)	0.8	0.7	0.9
Discharged or treated locally	**52.3**	**51.6**	**53.0**
Correct discharge (True negative)	51.5	50.8	52.2
Incorrect discharge (False negative)	0.8	0.7	0.9

The results of this analysis suggest that autonomous use of DERM could approximately halve the number of referrals to a dermatologist (among lesions that can be assessed by DERM). However, a small number of lesions, slightly under 1%, would be both malignant and incorrectly discharged (false negative). Most of those incorrect discharges would be BCC cases and only 0.2% of lesions would be melanomas or SCC and also discharged.

Most referrals would be false positives, with around 64% of all referrals being benign lesions. Among urgent referrals the substantial majority (around 85%) would be false positives. Routine referral would be uncommon (around 9%). This is partly due to a substantial overdiagnosis of BCC cases as being SCC or melanoma.

3.2.6 Implementation, resource use and related outcomes

One study reported data on referral and exclusion rates.[22]

Two studies of DERM reported data that related to implementation outcomes (as listed in Section 1.6.2).[22, 23] Data on these outcomes were mostly taken from the unpublished Edge Health report of patients in Leicestershire.[23] Two studies of DERM reported data on cancer stages as diagnosed by a reference standard. Most melanoma were had superficial spreading and had Breslow thickness <1.0mm. Most SCC identified were Stage 1. Further details are reported in Appendix 4, Table 27 and Table 28.

No evidence, published or unpublished, was identified for numbers of patients transferred to surgery, or test failure rates.

3.2.6.1 Referral and exclusion rates

Thomas 2023[22] reported data on the diagnostic pathway for patients assessed with DERM-vB. This is summarised in Table 8.

There were some differences between the two locations in terms of rate of use of DERM and referral rates, suggesting that use of DERM may vary by location. A notable issue was the substantial number of lesions that could not be assessed using DERM.

With DERM vB, between 64% and 76% of lesions eligible for AI assessment and judged non-malignant by DERM were subsequently discharged after second read or referral: none of these lesions that were subsequently biopsied were malignant.

Table 8 Diagnostic pathway for patients in Thomas 2023 when using DERM-vB

		Birmingham	West Suffolk
Not assessed using DERM		25%	17%
Referred to dermatologist by DERM	**Total**	44%	62%
	Malignant lesions	7.5%	9.7%
Judged non-malignant by DERM	**Total**	31%	21.6%
	Discharged at second read	18.7%	10.7%
	Discharged after referral	4.8%	2.7%
	Malignant lesions	0	0

** All % are out of total n of cases/patients, including those not assessed by DERM*

In the Leicestershire study [redacted] of lesions could not be assessed by DERM. The main reasons for this were

[redacted] In that study, after DERM assessment and teledermatology, [redacted].

The rates for DERM alone without combining with a second read were not reported.

[redacted]

3.2.6.2 Impact on resource use

In the Leicestershire study, each virtual dermoscopy review with DERM required on average [redacted] per lesion for a Trust dermatologist to review on the Skin Analytics platform. In contrast, a UHL consultant dermatologist estimated an average of [redacted] per patient for a face-to-face or telephone review. Following the introduction of the AI-teledermatology pilot pathway, the total time required to review cases through virtual, phone and face-to-face appointments was [redacted] Compared with the time estimated to review all [redacted] patients face-to-face or through telephone review prior to the introduction of the pilot, the Leicestershire study estimated that the AI pathway resulted in a [redacted]

3.2.6.3 Timings

In the Leicestershire study, the introduction of an AI-teledermatology pathway led to [redacted] following a referral for 2WW suspected skin cancer referral. Median time from referral to face-to-face outpatient appointments was [redacted] for patients who were on the traditional pathway. During the period of the pilot, an [redacted] were observed compared with the period preceding its introduction.

[redacted] No other data on waiting times, including time to discharge and time to treatment were reported for any of the DERM studies.

3.2.6.4 Cancer stage

No evidence on cancer stages at times of diagnosis was identified.

3.2.6.5 Acceptability to healthcare professionals

One study of DERM (versions not reported) collected feedback from healthcare professionals on benefits and limitations of the tool [23].

In the study conducted across Leicestershire community hubs, [redacted] shared their views on their confidence with DERM, its impact on the trust and on patients. Response rates were not reported. [redacted]

[redacted]

3.2.7 Clinical impact and patient benefit

3.2.7.1 Clinical morbidities and mortality

The included studies did not report medium or long-term data on clinical morbidities such as metastases or adverse outcomes of cancer treatment, nor were data on mortality reported.

3.2.7.2 Health-related quality of life

No data identified in the included studies.

3.2.7.3 Non-clinical benefits to patients

Two studies of DERM (versions not reported) collected feedback from patients on benefits and limitations of the tool.[23, 42] A total of 266 respondents (38.2% response rate) completed questions on their experience with DERM as part of the DERM-005 study. In the Leicestershire study, [redacted] with a suspicious lesion across three community hubs reported on confidence with DERM after having their lesion assessed with DERM (response rate not reported).

Reassurance that lesion is not cancerous

In the DERM-005 study, patients expressed confidence in DERM being used on a visual analogue scale (VAS) from 0 to 100, with higher scores indicating a higher level of agreement. Participants generally responded positively when considering AI as a tool to help doctors, but more cautiously when considering the use of AI to replace a dermatologist. Median levels of agreement with interquartile range (IQR) are illustrated in Figure 7.

Figure 7 Summary of responses DERM-005 relating to confidence in AI diagnosis

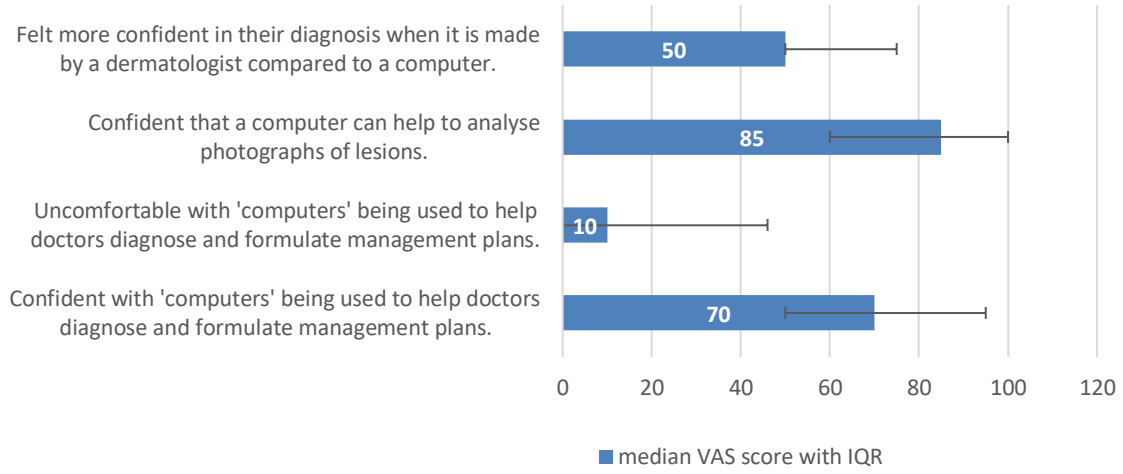


Figure created by EAG based on Kawsar 2023 Table 3.

Waiting for a diagnosis and associated anxiety

In the DERM-005 study, the photography service was generally considered an efficient use of patients' time, and respondents agreed that a computer assessing the photographs saves time compared to face-to-face consultation (Figure 8). No respondents felt the time needed to take photographs was too long (median score 0), and most would rather have their lesion assessed by a computer than waiting weeks to see an in-person dermatologist.[42]

Figure 8 Summary of responses DERM-005 relating to waiting for a diagnosis

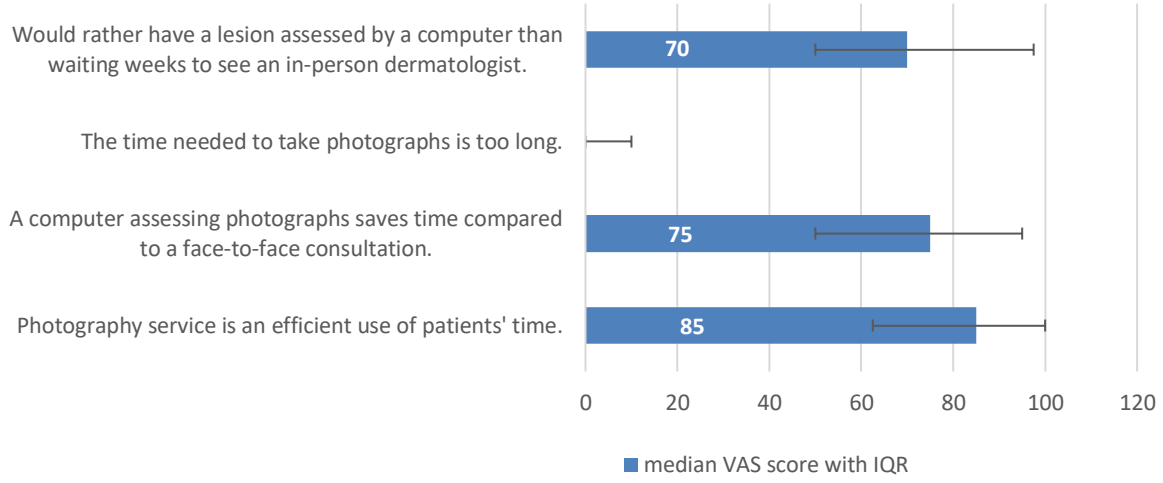


Figure created by EAG based on Kawsar 2023 Table 3.

Of the patient responses in the Leicestershire study,

[redacted]

Acceptability of AI technologies or processes

In the DERM-005 study, patients generally indicated they felt comfortable with the use of AI and the dermoscopic images required, but there was a mixed response to a statement on preference for a face-to-face dermatologist appointment. No participants found it embarrassing to have photos taken (median score 0, IQR 0-5) (Figure 9).

Figure 9 Summary of responses DERM-005 relating to acceptability

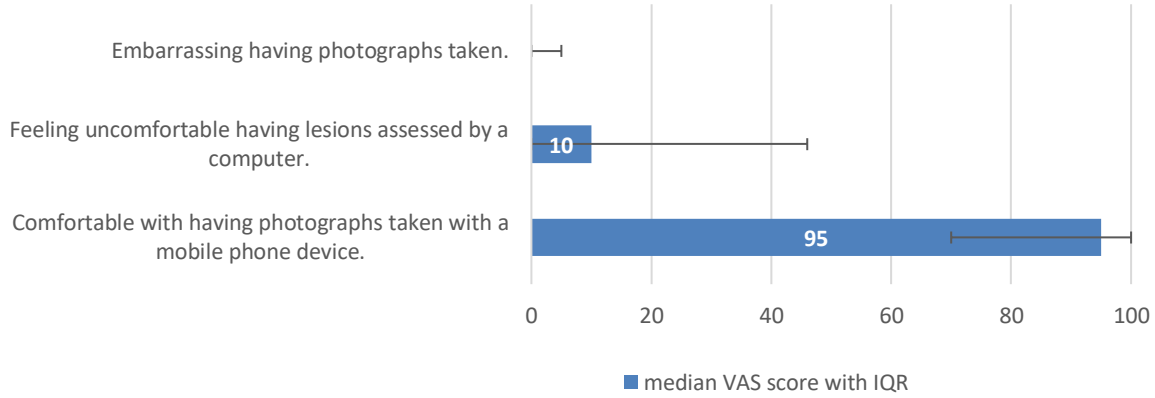


Figure created by EAG based on Kawsar 2023 Table 3.

In the Leicestershire study,

[redacted]

3.3 Moleanalyzer Pro

3.3.1 Summary of Moleanalyzer Pro studies

A total of 13 distinct studies of Moleanalyzer Pro were identified, in two prospective, cross-sectional diagnostic accuracy cohort studies,[24, 25] and 11 retrospective reviews of image datasets.[26-36]

The two prospective cohort were multi-centre studies conducted in a post-referral, secondary care setting; Winkler 2023[25] was conducted in Germany, and McLellan 2021[24] in Canada. Winkler 2023 evaluated the accuracy of Moleanalyzer for detecting melanoma against dermatologist assessment with and without Moleanalyzer in patients with suspected melanocytic lesions; final diagnosis was confirmed by biopsy (in 55% of patients) or clinical follow-up and/or expert consensus. In addition to diagnostic accuracy, the study reported the number of unnecessary excisions, and acceptability of the AI-tool from dermatologists and patients. MacLellan 2021 evaluated the diagnostic accuracy Moleanalyzer Pro for detecting malignancies against: a dermatologist face-to-face assessment, remote dermatologist assessment, and other non-invasive technologies beyond the scope of this assessment.

Clinical management decisions were recorded, and all suspected lesions were excised regardless of the clinical decision or AI output.

Eleven studies performed a retrospective review of existing image datasets to test the diagnostic accuracy of an AI-algorithm against a reference standard test. They are summarised in Table 9. Where reported, the Moleanalyzer algorithm in these studies was based on a modified version of Google's Inception v4 CNN architecture. The reference standard in these studies included histopathology, dermatologist consensus and/or clinical follow-up; four studies analysed only or nearly only excised lesions.[26, 28, 29, 34] Five studies compared the accuracy of AI-algorithm against dermatologists' assessment .[26-29, 36] and three were compared against other AI tools.[24, 30, 36] Due to the lack of prospective evaluation in a clinical setting, eleven studies were excluded from the main review,[26-36] and two studies were retained for full data extraction, quality assessment and synthesis.[24, 25]

Table 9 Summary of Moleanalyzer (Fotofinder) studies included in the evidence map

Study	**Design**	**N participants (lesions)**	**Diagnostic (index) tests**	**Outcomes**
Fink 2019[26]	Retrospective review	72 (72)	Moleanalyzer Pro Dermatologists	Diagnostic accuracy
Haenssle 2018[27]	Retrospective review	NR (300)	Moleanalyzer Pro Dermatologists	Diagnostic accuracy
Haenssle 2020[28]	Retrospective review	100 (100)	Moleanalyzer Pro Dermatologists	Diagnostic accuracy
Kommoss 2023[29, 52]	Retrospective review	100 (100)	Moleanalyzer Pro Dermatologists	Diagnostic accuracy Clinical management decisions
MacLellan 2021 [24]	Prospective cohort	184 (209)	Moleanalyzer Pro Dermatologist (with/without dermatoscope) Teledermatologists Various TD-AI tools[#]	Diagnostic accuracy Clinical management decisions
Sies 2020[30]	Retrospective review	435 (1981)	Moleanalyzer Pro Moleanalyzer Dynamole	Diagnostic accuracy
Sies 2021[31, 53]	Retrospective review	108 (233)	Moleanalyzer Pro	Diagnostic accuracy
Sies 2022[32]	Retrospective review	465 (1549)	Moleanalyzer Pro	Diagnostic accuracy
Winkler 2020[33]	Retrospective review	180 (780)	Moleanalyzer Pro	Diagnostic accuracy
Winkler 2021a[34]	Retrospective review	30 (30)	Moleanalyzer Pro Dermatologists	Diagnostic accuracy
Winkler 2021b[35]	Retrospective review	NR (130)	Moleanalyzer Pro	Diagnostic accuracy

Winkler 2022[36, 54]	Retrospective review	59 (236)	Moleanalyzer Pro Dermatologists Other AI-tool*	Diagnostic accuracy
Winkler 2023[25]	Prospective cohort	188 (228)	Moleanalyzer Pro Dermatologists Both combined	Diagnostic accuracy Unnecessary excisions Dermatologist and patient acceptability

[#] Teledermoscopy DermEngine, MetaOptima), MelaFind, Verisante Aura. *Based on resnet34 architecture trained with images from the HAM10000 database

3.3.2 Characteristics of Moleanalyzer Pro studies

Table 10 summarises the characteristics of participants included in MacLellan[24] and Winkler 2023.[25] Further participant selection criteria are summarised in Appendix 4, Table 25. The large majority of patients had lighter skin colours (Fitzpatrick types II-III). Where reported, lesions were most often located on the trunk, followed by extremities. The prevalence of melanoma (respectively 28.2% and 16.7%) was high in both studies compared with an urgent referral population in the UK.

Table 10 Participant characteristics of the Moleanalyzer Pro studies included in the review

	Mean age (range)	**% female**	**Fitzpatrick skin type (%)**	**Ethnicity (%)**	**Lesion location (%)**	**Melanoma lesions (%)**
MacLellan[24]	52 (31-86)	46	I: 3 II: 60 III: 36 IV-VI: <1	NR	Head/neck: 24* Trunk: 42* Extremities: 31* Acral: 3*	28.2%
Winkler 2023[25]	53 (19-91)	48	I: 3 II: 34 III: 56 IV: 6 V-VI: 1	NR	Head/neck: 8 Trunk: 65 Upper extremities: 10 Lower extremities: 15 Acral: 1 Nail: 1	16.7%

**Only reported for lesions confirmed as melanoma*

3.3.3 Risk of bias

Results of the quality and applicability assessment are reported in Table 11.

Both studies were at high risk of selection bias due to the exclusion of participants that would have otherwise been eligible for assessment in clinical practice, including non-melanocytic lesions, and for MacLellan (2023), Fitzpatrick skin types higher than III. The threshold for a positive diagnosis with AI was not reported in MacLellan 2023, therefore the index test domain was at unclear risk of bias.

The reference standard tests were at low risk of bias. As with DERM studies, excision and histology was not performed in all participants in Winkler (2023). However, the risk of bias regarding the reference standard was low, due to the use of clinical follow-up data and expert consensus for non-excised lesions. There was insufficient information from the study to assess the flow of study participants and exclusions from analysis, therefore this domain was at unclear risk of bias.

Study exclusions in both studies (notably non-melanocytic lesions), the high prevalence of melanoma, and the inclusion of lesions deemed "challenging" by a dermatologist in Maclellan (2023) limits the applicability of both studies to an urgent referral population. The model of dermatoscope used was not reported in Winkler 2023, and was out of date in MacLellan 2023, which limited the applicability of dermatologist assessments. There were no concerns regarding the applicability of reference standards.

Table 11 Quality assessment of Moleanalyzer Pro diagnostic accuracy studies

Study	Test	Risk of bias				Applicability concerns		
		P	I	R	FT	P	I	R
MacLellan 2021	Moleanalyzer	✗	?	✓	?	✗	✓	✓
Winkler 2023	Moleanalyzer	✗	✓	✓	?	✗	✓	✓
MacLellan 2021	Dermato.	✗	✓	✓	?	✗	✗	✓
Winkler 2023	Dermato.	✗	✓	✓	?	✗	?	✓
Maclellan	Teledermato.	✗	✓	✓	?	✗	✗	✓
Maclellan	Moleanalyzer vs. dermato & teledermato.	✗	?	✓	?	n/a	n/a	n/a
Winkler	Moleanalyzer vs. dermato.	✗	✓	✓	?	n/a	n/a	n/a

P = patient selection; I = index test; R = reference standard; FT = flow and timing.

✓ indicates low risk; ✗ indicates high risk; ? indicates unclear risk. Dermato: Face-to-face dermatologist assessment. Teledermato.: Remote dermatologist assessment on images

3.3.4 Diagnostic accuracy

Winkler 2023 reported the diagnostic accuracy of using Moleanalyzer Pro both with and without clinical input; MacLellan reported results for Moleanalyzer Pro, face-to-face dermatology and remote dermatologist diagnosis alone. The results presented were for the diagnosis of melanoma only; no data were reported for other types of skin cancer or for premalignant and benign lesions. Results for Winkler 2023 and MacLellan 2021 are presented in Table 12 and Table 13 respectively. Winkler

(2023) also reported ROC curves of diagnostic performance. These are reproduced, in a simplified form, in Figure 10. PPV and NPV were not reported.

Table 12 Diagnostic accuracy in Winkler 2023

	Sensitivity*	**Specificity***	**Accuracy***
Dermatologist alone	84.2 (69.9-92.6)	72.1 (65.3-78.0)	74.1 (68.1-79.4)
Moleanalyzer Pro alone	81.6 (66.6-90.8)	88.9 (83.7-92.7)	87.7 (82.8-91.4)
Dermatologist with Moleanalyzer Pro	100.0 (90.8-100.0)	83.7 (77.8-88.3)	86.4 (81.3-90.3)

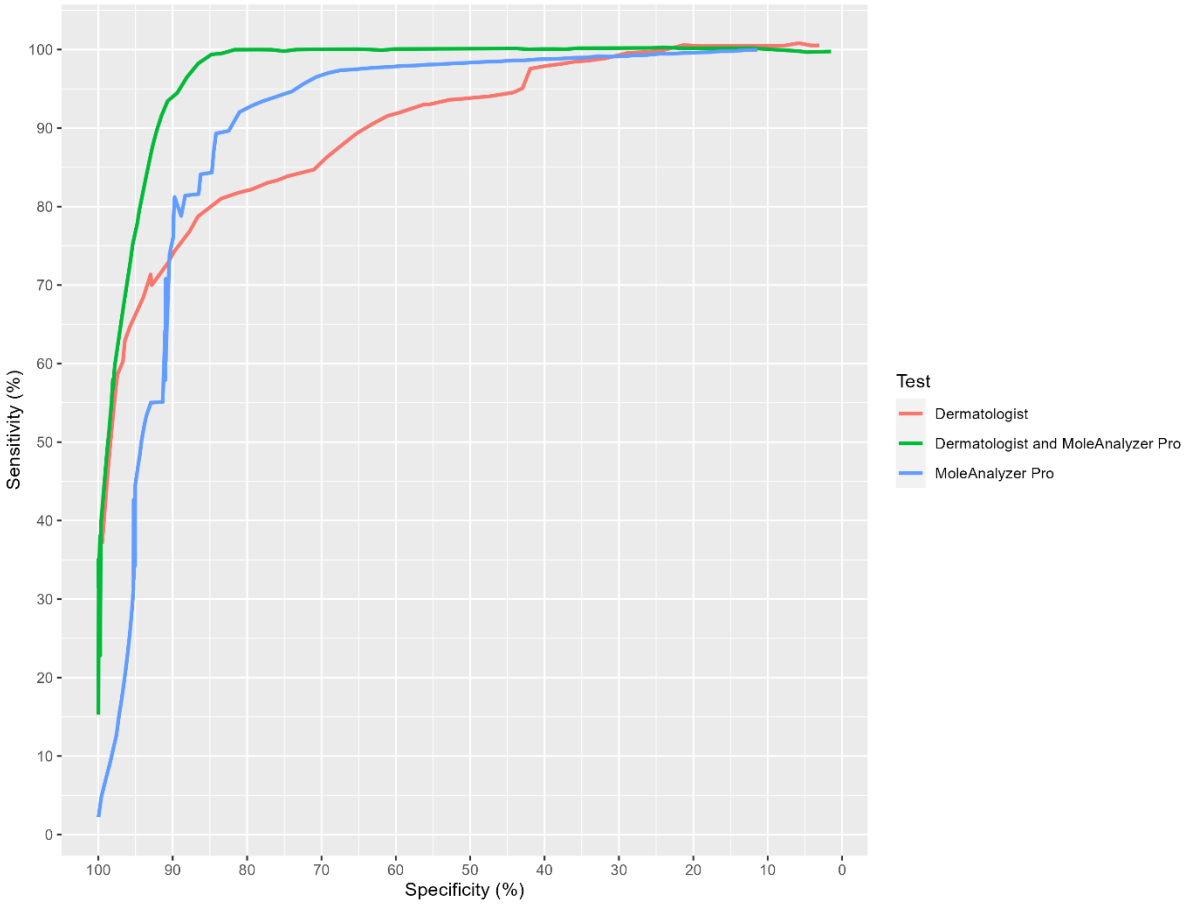
* Results expressed as % and 95% CI

*Table 13 Diagnostic accuracy in MacLellan 2021**

	Sensitivity*	**Specificity***
Moleanalyzer Pro alone	88.1 (79.4-96.9)	78.8 (71.5-86.2)
Dermatologist alone	96.6 (91.9-100)	32.2 (18.4-46.0)
Teledermatologist alone	89.8 (79.6-96.2)	66.0 (57.8-73.5)

* Results expressed as % and 95% CI

Figure 10 ROC curves for Moleanalyzer Pro for melanoma diagnosis (adapted from Winkler 2023)



A meta-analysis of the two studies found that Moleanalyzer Pro alone had a sensitivity of 84.4% (95% CI 73.9 to 91.0) and a specificity of 84.5% (95% CI 72.0 to 92.1) to detect melanoma. In both studies, Moleanalyzer Pro had somewhat poorer sensitivity, but higher specificity for detecting melanoma than face-to-face dermatologists. Compared with teledermatology, Moleanalyzer Pro had slightly lower sensitivity and higher specificity in MacLellan (2021). Combining Moleanalyzer Pro with dermatologist assessment had higher sensitivity and specificity than assessment by dermatologists alone. The estimated sensitivity is lower than observed for DERM, but the ROC curve in Figure 10 suggests than Moleanalyzer Pro could achieve a specificity of around 60 to 75% at a sensitivity of over 95%, which is similar to that observe for DERM (see for example Figure 5).

The EAG did not identify any evidence on the diagnostic accuracy of Moleanalyzer Pro for the detection of SCC, BCC or malignant lesions in general.

3.3.5 Implementation, resource use and related outcomes

3.3.5.1 Referral rates

One study of Moleanalyzer Pro evaluated referral decisions with face-to-face dermatology alone and following the integration of AI into decision making. In Winkler (2023),[25] dermatologists originally

recommended the excision of 104 of 190 (54.7%) benign nevi. After reviewing and integrating Moleanalyzer Pro results into decision-making, the estimated rate of unnecessary excisions was reduced by 19.2% from 104 to 84 nevi ($p < .001$), whilst the rate of excision of malignant lesions was not significantly changed ($p > 0.99$). The percentage of nevi managed by follow-up examinations was increased with the integration of Moleanalyzer Pro results into decision-making (from 37.9% to 44.7%, $p=0.053$).

The EAG did not identify any other evidence from Moleanalyzer Pro studies included in the synthesis on implementation, resource use, and related outcomes.

3.3.5.2 Acceptability to healthcare professionals

Winkler 2023 reported on feedback from dermatologists.[25] Dermatologists were asked after every assessment of a lesion whether or not they judged the CNN scores to be helpful and/ or reassuring. For 205 out of 228 lesions, dermatologists completed the evaluation. Out of 205 replies, 159 indicated CNN scores were reassuring (77.6%) and 173 CNN scores were perceived to be helpful (84.4%).

3.3.6 Clinical impact and patient benefit

The EAG did not identify any evidence on Moleanalyzer Pro, published or unpublished, on any clinical outcomes.

3.3.6.1 Non-clinical benefits to patients

Patients were provided with a questionnaire including ten statements, based on the 'trust in medical technology' instrument. For each item, response categories indicated the level of agreement with a statement, from very high to none, and undecided. Results are summarised below, pragmatically grouped by categories referring to the outcomes of interest for this report (reassurance, waiting for diagnosis, acceptability), although several items could be considered to contribute to multiple outcomes.

Reassurance that lesion is not cancerous

Responses indicated that patients generally trusted the CNN results (76% very high/high agreement) (Figure 11). CNN results were considered trustworthy by 81.5% of respondents (very high/high agreement) and the CNN exam provided a feeling of increased safety for 88.5% of respondents (very high/ high agreement). The same level of reassurance was not found when considering autonomous use of Moleanalyzer. When asked whether the AI tool may offer a higher diagnostic quality than a physician, 41.1% of respondents indicated low or no agreement. The overwhelming majority of respondents indicated that they would like the opinion of an expert physician besides an AI-assisted diagnosis (97.8% very high/ high agreement).

Figure 11 Reassurance offered by Moleanalyzer results, percentage agreement

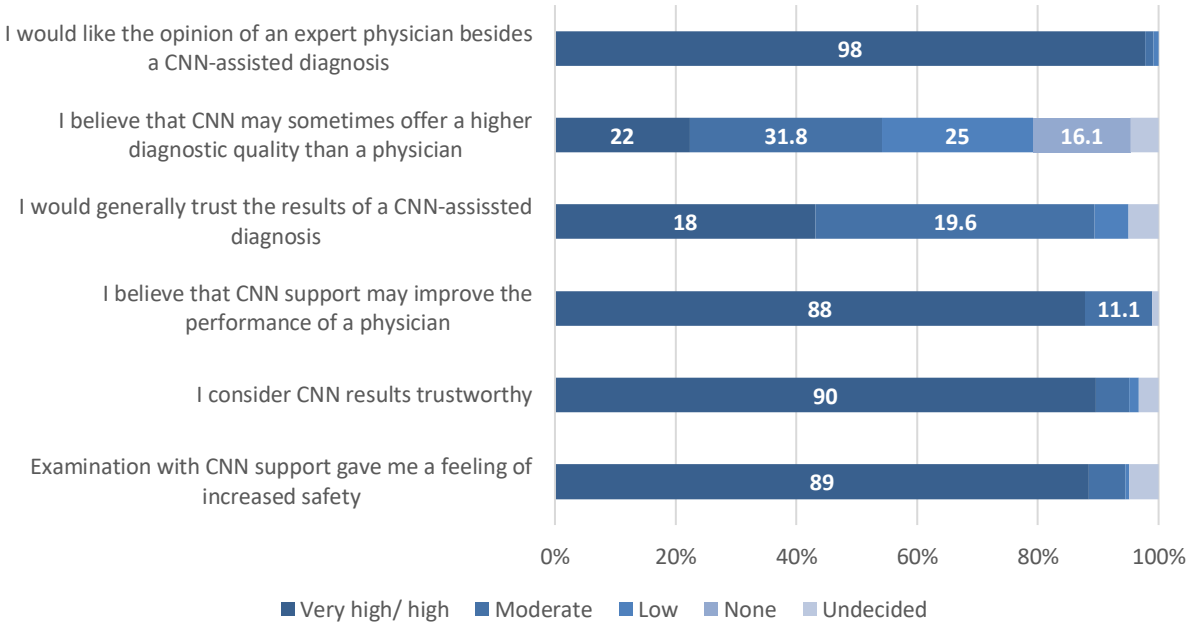


Figure was created by the EAG based on data from Figure S2.[36]

Waiting for a diagnosis and associated anxiety

Patients were asked whether they would accept longer examination times for an additional CNN-assisted diagnosis, and 33% expressed very high or high agreement with this statement.

Acceptability of AI technologies or processes

Three questionnaire items related to acceptability of using Moleanalyzer in the diagnostic process (Figure 12). Respondents generally did not believe that a CNN may completely replace the examination by a physician (26% moderate agreement, 23% low agreement, 28% no agreement). However, responses relating to the use of AI to assist the diagnosis made by a clinician were more favourable, with patients generally indicating they accepted the use of the tool by clinicians (85% no agreement with statement that CNN should not be used).

Figure 12 Acceptability of Moleanalyzer in diagnostic process, percentage agreement

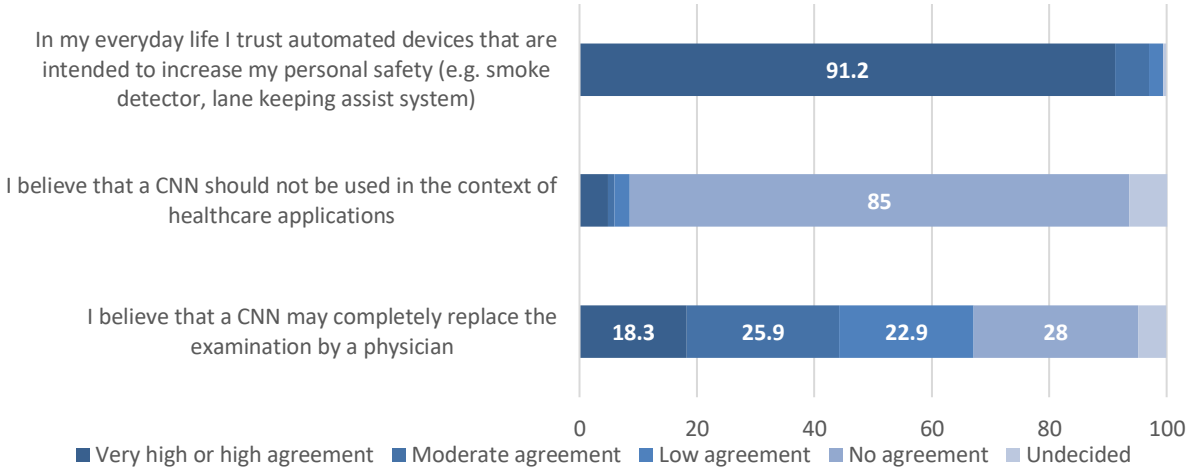


Figure was created by the EAG based on data from Figure S2 [36]

3.4 Conclusions

3.4.1 DERM

The review identified three recent studies of DERM that were suitable for assessment, and one currently unpublished study that was also considered. All were performed in the UK and embedded DERM within a post-referral setting.

3.4.1.1 Diagnostic accuracy

Both published and unpublished data sources for DERM suggested it has a high diagnostic accuracy for detection of malignant lesion when used autonomously: with a sensitivity of around 96.1% (95% CI 95.4 to 96.8) for a specificity of around 65.4% (95% CI 64.7 to 66.1). Diagnostic accuracies for detecting specific types of cancer (melanoma or SCC) were similar to this. There was some evidence that DERM might tend to misdiagnose BCC, with many BCC cases being classified as SCC or melanoma. The sensitivity when detecting benign lesions was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0).

The diagnostic accuracy for autonomous use of DERM was broadly similar to the diagnostic accuracy of dermatologists. In the DERM studies dermatologists had a

[redacted]

[redacted] This is similar to a previous systematic review of dermatology which found a summary sensitivity of 94.9% and specificity of 84.3%. [55]

The diagnostic accuracy of the whole teledermatology pathway including DERM could not be assessed because of a lack of any independent reference standard of diagnosis, but there was some evidence that a "second read" of lesions classified as benign by DERM [redacted] This is a key area of uncertainty in assessing the actual clinical value of using DERM.

3.4.1.2 Clinical outcomes

The EAG identified very limited published evidence on any clinical outcomes. Unpublished data suggested that autonomous use of DERM could approximately halve the number of referrals to a dermatologist (among lesions that can be assessed by DERM). However, a small number of lesions, slightly under 1%, would be both malignant and incorrectly discharged (false negative).

Between [redacted] of lesions could not be assessed by DERM in the studies that reported this data. The main reasons for this were because [redacted]

3.4.1.3 Patient and clinician perspectives

Some evidence was found for patient and clinician opinions of the use of DERM. Consultants overwhelmingly thought that AI should not be used autonomously, and there was a concern that AI used as a decision-aid was increasing patient' time on the diagnostic pathway. However, the evidence is limited to very small samples of responders.

Patients were perhaps more positive than clinicians about the use of DERM alongside a face-to-face diagnostic appointment with a clinician. Patients with experience of having a lesion assessed with DERM were generally accepting of the use of DERM as a tool aiding clinical diagnosis, but up to 50% of patients indicated they preferred a face-to-face dermatology appointment.

3.4.2 Moleanalyzer Pro

Fourteen studies of Moleanalyzer were identified, but only two prospectively evaluated patients in practice and so only they were considered for full synthesis. Neither was performed in the UK. No relevant unpublished material was identified.

3.4.2.1 Diagnostic accuracy

A meta-analysis of two studies found that Moleanalyzer Pro alone had a sensitivity of 84.4% (95% CI 73.9 to 91.0) and a specificity of 84.5% (95% CI 72.0 to 92.1) to detect melanoma. In both studies, Moleanalyzer Pro had somewhat poorer sensitivity, but higher specificity for detecting melanoma than face-to-face dermatologists. The diagnostic accuracy of Moleanalyzer Pro for the detection of SCC, BCC or other malignant lesions is unknown.

3.4.2.2 Clinical outcomes

The EAG did not identify any evidence for Moleanalyzer Pro for any clinical outcome.

3.4.2.3 Patient and clinician perspectives

The use of Moleanalyzer Pro was generally supported by both clinicians and patients, and its results were trusted. However, the overwhelming majority of patients indicated that they would like the opinion of an expert physician besides an AI-assisted diagnosis.

4 RESULTS: COST-EFFECTIVENESS REVIEW

4.1 Results of literature searches

Two sets of database searches were conducted to identify any cost-effectiveness evidence on the named studies and to inform the development of a conceptual decision-analytic model. The first of these searches was strictly confined to economic studies relating to the use of the named technologies. The second comprised a targeted literature search to identify economic evaluations of any approach to skin cancer diagnosis in an NHS setting. Conference abstracts were excluded from this search. Search strategies can be found in full in Appendix 1. Identified studies were summarised narratively. No formal data extraction or quality appraisal was undertaken.

4.1.1.1 Economic studies relating to the named technologies

479 records were identified through database searches related to the named technologies. Only one of these records related to health economics review - a clinical trial registration for an economic evaluation of DERM (Skin Analytics), for which there were no corresponding publications or abstracts. As a result, no studies from this search were considered in the literature review.

4.1.1.2 Economic studies related to diagnostics in skin cancer

The broader search for economic studies relating to the diagnosis of skin cancer in a UK setting returned 999 unique records (date limit of 2013 onwards). Three cost-effectiveness studies were identified following full-text screening, namely Wilson *et al*. 2013, Edwards *et al*., 2016, and Wilson *et al*. 2018.[56-58] These studies were considered relevant to the development and parameterisation of the conceptual model, although none related specifically to adjunctive or autonomous use of AI technologies for diagnosis of lesions suspicious of skin cancer.

4.1.1.3 Other identified studies

A submission from Skin Analytics provided two unpublished reports relevant to the cost-effectiveness of DERM, with some relevance for the development of the conceptual model. The first of these comprised an evaluation of a pilot of DERM implemented across the University Hospitals of Leicester NHS Trust in the 2WW pathway.[23] The second study comprised a preliminary report describing a *de novo* cost-utility model produced by the University of Exeter and Skin Analytics. No executable model was made available to the EAG. As these two studies are directly relevant to the decision problem, they are discussed separately in Section 5.

A report commissioned by the NHSE AI Award group also included economic analyses. This report is only subject to a brief overview in Section 5.3 below as it was made available to the EAG only shortly before the end of the project. The documentation provided was also incomplete and did not include an executable model.

4.2 Summary of identified evidence

The characteristics of the identified studies are summarised in Table 14. All three identified studies were decision analytic models. In line with inclusion criteria for the broad review of any approach to skin cancer diagnosis, all were from a UK perspective.

Table 14 Characteristics of identified studies

Study details	Intervention and comparator	Study population, study design, data sources	Costs (perspective, description and values) and outcomes (description and values)	Results: cost-effectiveness
[56] English primary care setting Cost-effectiveness analysis	MoleMate diagnostic aid plus best practice vs best practice alone	Patients aged 18 or over who have at least one suspicious pigmented lesion, that could not immediately be diagnosed as benign in a primary care setting. Study design: modelling study, decision tree with Markov extension design, lifetime time horizon, 3.5% discount rate applied to costs and benefits Source of clinical data: MoleMate RCT n=1293 patients in 15 general practices in the East of England Source of resource-use data: MoleMate trial, published literature Unit costs: NHS reference costs, published literature Utility data: Published literature	National Health Service (NHS) perspective Costs included: intervention costs including MoleMate device and annual maintenance costs, GP staff time; referral costs and follow-up tests and procedures. Treatment costs associated with true positives were based on 2010 UK guidelines for the management of cutaneous melanoma comprising biopsy excision, staging, and definitive surgery Outcome measure: quality-adjusted life-year	MoleMate strategy is estimated to cost an extra £18 compared to best practice alone, and yield 0.01 QALYs per patient. Corresponding ICER is £1,896/QALY
[57] United Kingdom Cost-effectiveness analysis	Alternative risk-stratified surveillance policies (based on Williams score) vs current practice (ad hoc presentation)	UK population Study design: modelling study, patient-level simulation design, 30-year time horizon, 3.5% discount rate applied to costs and benefits Source of clinical data: published literature, expert opinion	National Health Service (NHS) perspective Costs included: primary care costs, referral, diagnosis, treatment, follow-up and end-of-life costs Outcome measure: quality-adjusted life-year	The most cost-effective surveillance strategy (highest net benefit) was for those with a Williams score of 15 to 21 to be offered a one-off full-body skin examination, and for those with a score of 22 or more to be enrolled into a quinquennial monitoring program, rising to annual recall for those with a risk score greater than 43. For implementation of the overall surveillance program, the ICER was

Study details	Intervention and comparator	Study population, study design, data sources	Costs (perspective, description and values) and outcomes (description and values)	Results: cost-effectiveness
		Source of resource-use data: published literature, guidelines Unit costs: NHS reference costs Utility data: published literature		£10,199. Per patient QALYs are improved by 0.016 and costs are increased by £165.
[58] United Kingdom Systematic review and cost-effectiveness analysis	VivaScope 1500 and 3000 imaging systems versus routine management and monitoring	Three study populations were considered: 1. People with suspected melanoma who have equivocal lesions following dermoscopy 2. People with suspected basal cell carcinoma (BCC) whose lesions have an equivocal or positive result on dermoscopy, to make or confirm diagnosis, as an alternative to diagnostic biopsy 3. Patients with lentigo maligna (LM) prior to surgical management Study design: modelling study, decision tree with Markov extension design, lifetime time horizon, 3.5% discount rate applied to costs and benefits Source of clinical data: SLR of available evidence of VivaScope Unit costs: company data, NHS tariff and reference costs Utility data: published literature	NHS and Personal Social Services (PSS) perspective Costs included: intervention costs (including equipment, maintenance, consumables, staff training, staff time), comparator costs (biopsies, histological examination, monitoring, clinician time), costs associated with management of positive and negative results, and future health events (e.g., recurrence, progression) Outcome measure: quality-adjusted life-year	Where VivaScope is used exclusively in the melanoma population, the ICER was between £8,877 - £19,095. Incremental health was improved by 0.009 – 0.016 QALYs, incremental costs were between £138 - £178 (ranges indicate use of different diagnostic accuracy data). When also used for other indications, VivaScope becomes the dominant strategy in this population. For use exclusively in the BCC population, results show a dominant strategy – average per patient costs are reduced by £52 and QALYs are increased by 0.011. In the LM population, the model indicates a cost-effective strategy (ICER: £10,241) where VivaScope is used only for LM mapping where average per-patient costs are increased by £70.75 - £ and QALYs are increased by 0.007. Where VivaScope is used across indications, per-patient average costs are reduced by £74.12 and QALYs increased by 0.007 (indicating a dominant strategy).

4.2.1 Wilson et al. (2013)

Wilson and colleagues developed a decision analytic model to assess the cost-effectiveness of the MoleMate handheld SIAscopy scanner and proprietary algorithm as a diagnostic aid for primary care clinicians to direct more appropriate referral of pigmented lesions to specialists, compared to current practice. The economic model drew on data generated in the MoleMate UK trial, which enrolled 1,293 participants across 15 English general practices.

4.2.1.1 Model structure

The authors adopted a decision tree model structure to capture the initial decision to *refer* or *not refer* patients to specialist care. Three Markov models were used to estimate the long-term costs and health outcomes of patients based on their diagnosis at the terminal nodes of the decision tree (i.e. true positive, false negative, and true negative/false positive). The reference standard (i.e., the definition of an appropriate referral) was whether secondary care clinicians decided to biopsy or monitor a lesion – matching the reference standard used in the trial. The model did not structurally distinguish between melanoma and other types of skin cancer, and only accounted for disease stage at diagnosis.

Patients correctly identified (true positives) were assumed to be appropriately treated at the point of diagnosis, and thus remained within the same Markov state according to stage at diagnosis until death i.e., treated patients cannot experience progression. Patients with a false-negative diagnosis similarly entered the Markov model according to stage at diagnosis, but could experience disease progression, could be diagnosed and treated (entering a corresponding Markov state according to their post-treatment prognosis by stage at diagnosis), or could die. Patients without cancer (correctly identified or not) simply followed a normal life expectancy with zero cost or health consequences.

4.2.1.2 Mechanism of cost consequences

For patients in both arms of the model, two potential outcomes at primary care were possible – referral or non-referral to secondary care. Improved specificity reduced the number of (inappropriate) referrals and therefore reduced costs associated with follow-up investigations in secondary care. In the model, the specificity of MoleMate was lower than that of best practice (82.1% vs 89.2%), suggesting increased costs versus current care. Improved sensitivity had the effect of increasing immediate costs associated with follow-up investigations and treatment but lowered the cost associated with treatment of later-stage disease from initially unidentified melanomas. As treatment costs differed by disease stage at diagnosis (i.e., higher treatment costs for more severe disease), the net cost impact of improved sensitivity depends on the scale of the cost difference between treating early and late-stage disease and the effect of discounting. The MoleMate system was more sensitive than current practice (98.4% vs 95.6%). Use of the MoleMate system itself was associated with a small additional cost

(about £14) and the MoleMate strategy increased average patient costs by £18, suggesting little cost impact beyond the cost of the device itself.

4.2.1.3 Mechanism of health consequences

Staging of disease at diagnosis was based on the American Joint Committee on Cancer (AJCC) Melanoma Staging Database report from 2009.[59] The distribution of disease stage at diagnosis was implicitly assumed not to be affected by underlying diagnosis. Incidence of malignancy in patients referred to specialist care was 5%, the majority of whom had Stage 1a/b disease.[59]

Undiagnosed disease was associated with a 70-80% annual probability of remaining at the current stage, a 10% probability of being detected in a given year, and a 10-20% probability of progressing one or more stages. These transition probabilities were based on an earlier cost-effectiveness model for screening of melanoma.[60] Health state utilities were derived from a 2004 conference abstract - Bendeck et al.[61] Stage 4 disease was associated with the most significant quality of life impact (no cancer: 1.00 vs Stage 4: 0.52) where other stages were associated with more modest impacts on quality of life. Disease prognosis worsened commensurately with disease stage, with risk of death calculated using a log-odds ratio vs 1a melanoma where patients were at greater risk of mortality in later stages. Patients with Stage 4 disease had a log-odds ratio of death of 5.743, based on the AJCC report. Given the opportunities in the model for disease progression, and the worse outcomes associated with later disease stages in terms of mortality and quality of life, missing a case of cancer at the point of diagnosis has a negative health consequence in the model.

As MoleMate was associated with increased sensitivity compared to current practice (98.4% vs 95.6%), more patients with skin cancer were correctly referred to specialist care and were subsequently treated, generating a small QALY benefit of 0.093 versus current practice. This improvement in patient health offsets cost increases associated with lower specificity of the MoleMate system (as described in Section 4.2.1.2). MoleMate was associated with an ICER of £1,896 per QALY gained in the base-case analysis compared to current practice.

4.2.2 Edwards et al. (2016)

Edwards and colleagues performed a systematic review and economic evaluation to evaluate the clinical and cost-effectiveness of VivaScope 1500 and 3000 for the diagnosis of equivocal skin lesions and in lesion margin delineation prior to surgical excision. VivaScope is a technology designed to be used in conjunction with dermoscopic examination to aid diagnosis of suspicious lesions.

For the evaluation of cost-effectiveness, three ‘part’ models were built covering three populations 1) people with suspected melanoma who have equivocal lesions following dermoscopy 2) people with

suspected basal cell carcinoma (BCC) whose lesions have an equivocal or positive result on dermoscopy, to make or confirm diagnosis, as an alternative to diagnostic biopsy, and 3) patients with lentigo maligna (LM) prior to surgical management. Only the first two of these models related to the diagnosis of suspicious lesions and therefore are the models relevant to this review; both are discussed in detail below.

4.2.2.1 Model structure – diagnosis of melanoma

The authors employed a decision tree model structure to calculate the short-term outcomes of patients with suspected melanoma with equivocal lesions following dermoscopic assessment and used Markov models to represent the long-term outcomes of patients.

The current practice arm of the model details the current patient pathway – some patients directly undergo biopsy and excision (where melanoma status is confirmed) and some undergo monitoring. Monitoring can result in referral for biopsy and excision if melanoma is suspected or discharged if not. Given that biopsy is considered the gold standard test, the melanoma status of all patients who have undergone biopsy and excision is ultimately known. Some patients without melanoma are biopsied unnecessarily (and suffer associated health losses and procedure costs). In the VivaScope arm of the model, all patients undergo an examination with VivaScope, where positive cases are excised and biopsied, while negative cases are discharged without further investigation or treatment. Those patients with a positive VivaScope result (or patients undergoing biopsy in standard care) without melanoma will have unnecessarily undergone biopsy (and its associated harms) and those with melanoma who tested negative at VivaScope (or discharged at monitoring) will have been discharged inappropriately. Patients then enter one of three Markov models based on their diagnostic outcome and true disease status.

The first Markov model represents the outcomes of patients who are correctly identified as having melanoma (VivaScope TP or identified using biopsy). Identified melanomas were assumed to be identified as either in situ (60% of lesions) or Stage 1 (1a or 1b) (40% of lesions). A number of key assumptions were made:

- Following identification and treatment, melanomas were assumed not to progress;
- Patients with identified melanomas had a reduction in their HRQoL applied as a one-off disutility at treatment which then returned to that of the general population;
- Patients with lesions on their head and neck experienced an additional permanent reduction in HRQoL due to scarring following excision and biopsy;
- Patients with an identified melanoma 1b were at increased risk of mortality for 10 years, returning to the general population thereafter.

The second Markov model represents false negatives with melanoma. Key model assumptions were as follows:

- All melanomas were assumed to be in situ or Stage 1 (1a or 1b) at the time of assessment;
- Melanoma could progress by only a single stage;
- All unidentified melanomas are identified when they reach Stage 2 (2a, 2b, 2c), or within 5 years of the first assessment;
- People with an unidentified melanoma had a HRQoL equal to that of the age-adjusted general population until their melanoma was identified (when a decrement is applied);
- People with a lesion on their head and neck experienced an additional permanent reduction in their HRQoL due to scarring following excision and biopsy;
- Unidentified melanomas did not incur any costs;
- Melanoma was assumed to be successfully treated upon diagnosis (no further progression);
- People with an unidentified or identified melanoma at Stage 1b or 2 were at increased risk of mortality, from the outset of the model until 10 years after diagnosis, after which point mortality risk was equal to that of the general population.

The third Markov model represented people without melanoma (VivaScope false positive or true negative, or negative following biopsy/monitoring). Key assumptions were as follows:

- People with a lesion on their head and neck experienced an additional permanent reduction in their HRQoL due to scarring following excision and biopsy;
- Otherwise, HRQoL was equal to that of the general population of the same age.

Transition probabilities in the model were informed by assumptions regarding the progression of patients. A progression probability of 15.3% was used, calculated based on the assumption that the mean duration of Stage 1 melanoma is 50 months, and 50% melanomas progress. This transition probability was applied to progression from both in situ to Stage 1, and Stage 1 to Stage 2. An annual probability of opportunistic diagnosis (given initial non-identification) of 35% was applied based on the assumption that all unidentified melanomas would be diagnosed by the time they reach Stage 2 at the latest, and it was structurally imposed that yet unidentified melanomas at 5-years were diagnosed.

4.2.2.2 Mechanism of cost consequences – diagnosis of melanoma

The costs associated with the VivaScope pathway in the model depend on a) whether the VivaScope technology is used for other potential indications (and thus the fixed costs of VivaScope are spread over a larger population) and b) which diagnostic accuracy figures are used in the model (Alarcon et al.[62] vs Pellacani et al.[63]) where the former has the greatest impact on pathway costs. If VivaScope has higher sensitivity, more cases will be identified correctly, and treatment costs will be higher.

In the model, there is little difference in treatment costs across disease stages (except for the increased use of SLNB in later stages) and therefore identification of melanoma at an earlier stage is unlikely to drive value in terms of reducing modelled treatment costs. This contrasts to Wilson et al. 2013,[56] where there was a steeper gradient in costs. Higher specificity for VivaScope results in fewer non-melanoma patients undergoing unnecessary excision and biopsy (thus costs saved – excision and biopsy is £151). Given that monitoring is not available on the VivaScope pathway, monitoring costs are also saved (£93). The incremental cost results are shown in Table 15 and are shown according to different diagnostic accuracy inputs and whether VivaScope is used for melanoma only, for the two diagnostic indications or for all three indications. The cost of the device itself appears to be a large driver of pathway costs.

Table 15 Incremental costs associated with VivaScope pathway

	Alarcon diagnostic accuracy data	**Pellacani diagnostic accuracy data**
VivaScope for melanoma diagnosis only	£137.99	£177.03
VivaScope for melanoma and BCC	-£52.71	-£13.67
VivaScope for all indications	-£56.95	-£17.91

4.2.2.3 Mechanism of health consequences – diagnosis of melanoma

A more specific VivaScope test will reduce the number of patients undergoing excision and biopsy – this reduces the number of patients experiencing anxiety while waiting for biopsy results, and the number of patients experiencing permanent disutility from scarring on their head and neck.

Unlike other studies in this review, the key driver of value in the model appears to be the reduction in health harms associated with biopsy and excision used for the detection of melanoma. In the model, under routine management, 67% of lesions were excised despite a prevalence of melanoma of only 15%. Given the large health decrement applied in the model following biopsy and excision, the main value case of VivaScope appears to be the reduction of the use of biopsy and excision and its associated harms. It is unlikely that earlier diagnosis of melanoma would be a key driver of patient health in the model, as it assumes that patients with unidentified melanomas have HRQoL equal to that of the general population until diagnosis, when a one-off decrement applies at the point of treatment (in addition to a permanent decrement for some patients with scarring). This contrasts to other studies such as Wilson et al. 2013[56] which assumed a persistent impact on HRQoL for treated patients (thus making the health consequence for a missed case higher). The fact that structural limitations are placed on progression in Edwards et al. 2016[58] also means that patients with unidentified cancers are unable to progress to later stages and incur greater health decrements, again meaning that the health consequence is lower for a missed case of melanoma.

The incremental QALY results associated with each pathway are shown in Table 16.

Table 16 Incremental QALYs associated with VivaScope pathway

	Alarcon diagnostic accuracy data	**Pellacani diagnostic accuracy data**
VivaScope arm	0.016	0.009

4.2.2.4 Model structure – diagnosis of BCC

A decision tree model structure was used to estimate the short-term outcomes of patients with suspected BCC lesions and positive or equivocal findings in dermoscopy. According to the model structure, patients with lesions suspicious of BCC are examined either according to current practice (who all receive diagnostic biopsy), or with VivaScope. Given that diagnostic biopsy is the gold standard for diagnosis, all patients in the current practice arm have their treatment status correctly determined and are treated or discharged accordingly. Diagnostic biopsy incurs a one-off disutility related to the procedure (-0.02), a 6-week disutility related to anxiety while waiting for biopsy results (-0.008), and a permanent disutility for 5% of patients with scarring on their head or neck (-0.016). Although all patients are appropriately treated or discharged in the current practice arm, some patients undergo unnecessary diagnostic biopsy and experience a utility decrement. In the VivaScope arm, patients testing positive at VivaScope progressed to treatment (without the need for diagnostic biopsy). Patients for whom VivaScope indicated a negative result received diagnostic biopsy (because the original dermoscopic outcome suggested malignancy) and are discharged or treated as appropriate. All patients with the BCC in the VivaScope arm were correctly treated. A proportion of patients who tested positive at VivaScope will be false positives and therefore will have been inappropriately treated.

Treatment in the model comprised both surgical and non-surgical therapies. Patients undergoing surgical treatment (75% of patients) experienced a utility decrement (-0.004) from the procedure itself, and a proportion experienced a permanent disutility associated with scarring (-0.019 for surgical excision, -0.021 for Mohs surgery). Given that all patients in both arms of the model with BCC are correctly identified as having the condition, only one Markov model is required for those patients who have experienced scarring from unnecessary biopsy as the net difference in long-term treatment outcomes between arms is because of scarring.

4.2.2.5 Mechanism of cost consequences – diagnosis of BCC

The immediate mechanism by which VivaScope impacts costs in the model is by reducing diagnosis costs – the cost of biopsy in the model is £134 whereas the cost of VivaScope is £70 (exclusive use on BCCs). This cost benefit in favour of VivaScope will be somewhat reduced by the unnecessary treatment costs incurred through treatment of FP patients at VivaScope.

4.2.2.6 Mechanism of health consequences – diagnosis of BCC

It appears that the mechanism by which health is impacted in the model is through the avoidance of health harms associated with diagnostic biopsy, which carries a large health decrement in the model associated with anxiety while waiting for results, the procedure itself, and scarring. The driver of the value of VivaScope appears to be as a result of the fact that it is non-invasive (unlike diagnostic biopsy) and so not associated with scarring and not associated with a long wait for results, thus no anxiety-related decrement. VivaScope allows some patients (who test positive) to proceed directly to treatment without biopsy, appearing to generate value in the model through a reduction in the proportion of patients receiving diagnostic biopsy.

Given that all patients with BCC in both arms are treated, there is no mechanism for health gains associated with improving identification of disease in the model.

The model results showed that where VivaScope is used exclusively for suspected melanomas with equivocal dermoscopy, the ICER is £8,877 - £19,095 depending on clinical data used. When also used for other indications, VivaScope becomes the dominant strategy in these patients. For use exclusively in the BCC population, results show a dominant strategy.

4.2.3 Wilson et al. (2018)

The authors adapted a previously developed decision analytic model (Wilson et al. 2013[56]) to evaluate the potential cost-effectiveness of a risk-stratified population surveillance programme. The authors estimated the costs and outcomes associated with surveillance strategies of different risk groups. The population was segmented by Williams score, a clinical tool for identifying the risk of melanoma. The main purpose of this study was to identify the risk score cut-off at which it is most cost-effective to enrol patients into a surveillance programme consisting of 1) a one-off visit to the patient's primary care practitioner, 2) an ongoing primary care-based monitoring programme (and the optimal frequency of visits). The authors estimated outcomes over 30-year time horizon.

4.2.3.1 Model structure

The authors employed a patient-level simulation model based on the structure of Wilson. The model is comprised of two 'modules' – patients enter the model in the natural history module according to the distribution of prevalent melanomas and their disease stages. When contact is made with the health system, the patient enters the clinical module which has a decision tree-like structure where referral, treatment, and discharge decisions occur. The clinical module allows patients to present in primary care: both of their own initiative and if they are told to do so following a risk assessment. Following presentation, any suspicious moles are inspected at primary care and a decision is made to either refer to secondary care or discharge the patient. The model categorises melanoma into four main types: superficial spreading, lentigo maligna (LM), acral lentiginous, and nodular, each with

nine stages of invasion (1a – 4) plus an in-situ stage (except for nodular melanoma). The authors assumed that invasive disease would progress at the same rate irrespective of the primary melanoma subtype, but the model allowed different progression probabilities from in situ disease. Patients with melanoma correctly identified as such (true positives) receive appropriate treatment according to their disease stage – they are then flagged by the model as having a history of melanoma and are at risk of stage-specific mortality. False-negative patients are discharged and returned to the natural history module in which they are at risk of disease progression and mortality. False positives incur the cost of referral and are discharged into the community. The authors assume that patients who are unaware they have melanoma suffer no impairment in quality of life. At the point of diagnosis, a disutility is assigned.

4.2.3.2 Mechanism of cost consequences

All optimal surveillance strategies were associated with incremental costs, which included the cost of the surveillance strategy itself and increased costs associated with the treatment of identified cases. The benefits of surveillance were primarily driven by health consequences not cost savings.

4.2.3.3 Mechanism of health consequences

In the model, early disease detection of disease prevents progression to later stages which are associated with greater health decrements and higher rates of mortality. Early detection via surveillance therefore generates health benefits by avoiding cases of late-stage diagnosis compared to when the disease is identified opportunistically.

The most cost-effective surveillance strategy (highest net benefit) was for those with a Williams score of 15 to 21 to be offered a one-off full-body skin examination, and for those with a score of 22 or more to be enrolled into a quinquennial monitoring program, rising to annual recall for those with a risk score greater than 43. The overall ICER associated with the implementation of the surveillance strategies was £10,199.

4.3 Discussion

All three studies identified in the cost-effectiveness review employed similar model structures – a decision tree structure to represent the short-term outcomes associated with different diagnostic pathways, and Markov models to estimate long-term outcomes.

All three studies incorporated multiple indications, but none were so broad as the scope of the present assessment, and the extent to which different diagnoses were distinguished between prognostically and diagnostically varied. In Wilson 2013, the model tracked outcomes of malignant skin disease

which comprised basal cell carcinoma, squamous cell carcinoma, and malignant melanoma. The model did not distinguish between melanoma or squamous cell carcinoma and estimated outcomes based only on disease stage at diagnosis. Wilson 2018 included melanoma only but distinguished between the following sub-types: superficial spreading, lentigo maligna, acral lentiginous, and nodular. The model assumed that invasive disease would progress at the same rate irrespective of type but allowed the rate of progression from in situ disease to vary by subtype. The Edwards 2016 model accounted for melanoma, basal cell carcinoma, and lentigo maligna which were considered individually within three separate 'part' models.

The approach of the authors to the progression of undiagnosed cancers incorporated a range of data sources and fixed assumptions. Wilson 2013 and 2018 used data from Losina 2007[60] and Wilson 2017[64], respectively as sources for expert-elicited progression probabilities. The authors assumed a 10% annual probability of opportunistic detection of previous false negatives. Wilson 2013 and Wilson 2018 did not place limits on the progression of patients with false negative test results, whereas the Edwards 2016 model assumed all undetected cancers would be opportunistically detected by the time they progress to Stage 2. For the BCC model presented by Edwards et al., no cancers remained undetected and so there was no progression possible in the Markov model component. In the Edwards melanoma model, an annual progression probability of 15.3% was applied, regardless of the current disease stage. A 35% annual probability of identification (if initially undetected) was applied based on the assumption that all unidentified melanomas should be identified by the time they reach Stage 2, or 5 years after initial assessment. All studies assumed differential mortality rates according to disease stage for both identified and unidentified melanomas. BCC was assumed not to be associated with elevated mortality rates.

The mechanism by which costs and health outcomes are impacted in the three publications differ substantively. Value in the Edwards model was driven through the reduction of inappropriate procedures (most notably biopsy and excision and diagnostic biopsy) on health outcomes and costs. Biopsy and excision for melanoma were associated with a permanent disutility from scarring (for those with a lesion on their head or neck), temporary disutilities from anxiety while waiting for test results, and a disutility from the procedure itself. This model placed less value on diagnostic sensitivity, assuming that unidentified cancers have a utility equal to that of the general population until later diagnosis. In the BCC model, implicit in the structure is that 100% of BCC cases are always correctly identified as such and so there is no cost or health consequence from improved diagnosis.

This approach contrasts with that of the other two publications, whereby increased sensitivity drove value. In Wilson 2013, MoleMate increased average per patient testing costs, but improved patient health because of increased detection of cancers which were associated with improvements in long-term health outcomes. This is likely due to two structural differences: firstly, the model assumed a

differential utility decrement by cancer stage independent of diagnosis, and secondly, any health decrement associated with procedures undertaken in secondary care (e.g., scarring from biopsy and excision) was not captured. However, Wilson 2018 differs from Wilson 2013 in that they assume undiagnosed melanomas only impacts HRQoL after diagnosis (in line with Edwards) but did allow disease progression beyond that assumed in Edwards. Neither Wilson 2013 nor Wilson 2018 captured the health impact of scarring on patient health or other harms of procedures such as diagnostic biopsy e.g. anxiety while waiting for results.

5 ECONOMIC MODELS SUBMITTED BY SKIN ANALYTICS

5.1 Cost-utility model (Exeter Test Group/Skin Analytics)

Skin Analytics provided a preliminary report on a cost-utility model developed with Exeter Test Group during the latter part of the EVA process. The executable model itself was not made available to the EAG for review. Due to the late provision of the company cost-effectiveness report, and the incomplete description of the analysis in the submitted documents, the EAG is unable to provide the usual level of scrutiny of a company cost-effectiveness model and does not accord with the template used in the assessment of company model used within the single technology process.

The decision problem considered in this analysis aligned with the scope of the EVA, i.e. triage of patients referred from primary care via the dermatology urgent skin cancer referral pathway. The model assesses two models of implementation of DERM in this setting: DERM with a second read, in which the images from DERM-negative patients are assessed by a consultant prior to discharge; and DERM without a second read, where DERM-negative patients are discharged without a further assessment. The model considered two comparators: face-to-face assessment, and teledermatology.

5.1.1 Modelled population

The characteristics of the modelled population was based on NHS sources. It was assumed that 87.2% of patients screened had precancerous or benign lesions. 5.9% of patients were assumed to have melanoma, SCC, and rare skin cancers, and 6.9% had BCC. The model assumed that disease stage of melanoma at the point of diagnosis would also apply to SCC and other rare cancers. Evidence supporting this assumption was not in the presented the provided report.

5.1.2 Model structure

The model structure was adapted from Wilson et al. 2013[56] (described in Section 4.2.1), comprising a decision tree with Markov models at each terminal node to link specific diagnostic outcomes with long-term costs and outcomes. The model described differs from that presented in Wilson et al. 2013 in that it explicitly models BCC as a separate diagnostic category to the high-risk cancers (i.e. melanoma, SCC, and rare cancers), reflecting the different prognosis and treatment of these indications. The model applies three diagnostic categories, each with a distinct diagnostic accuracy profile for each strategy, and associated treatment costs. These are:

- ‘High-risk cancers’, including melanoma, squamous cell carcinoma, and ‘other high-risk cancers’;
- basal cell carcinoma;
- low-risk lesions (benign and precancerous).

The model adopted a lifetime time horizon (up to 100 years of age), with a 1-year cycle length in the Markov phase of the model. A half-cycle correction was applied. The model adopted an NHS and Personal Social Services perspective. Costs and benefits are discounted at 3.5% per annum.

There are possible four diagnostic pathways represented by decision trees. Whilst the report does not contain a complete model schematic, it can be inferred from the provided description.

DERM without a second read

- DERM-positive patients are referred to a face-to-face dermatologist, and can then be diagnosed malignant (TP & FP), or benign (TN & FN);
- DERM-negative patients are discharged and enter the false-negative or true-negative Markov model;
- Patients who are ineligible for DERM assessment or whose DERM assessment is unsuccessful are referred directly to a face-to-face assessment.

DERM with a second read

- DERM-positive patients are referred to a face-to-face dermatologist, and can then be diagnosed malignant (TP & FP), or benign (TN & FN);
- DERM-negative patients undergo a virtual triage by a consultant dermatologist from which they can be discharged or referred to a face-to-face dermatologist who can diagnose malignant (TP & FP), or benign (FN & TN);
- Patients who are ineligible for DERM assessment or whose images are unsuccessful are referred directly to face-to-face assessment.

Face-to-face assessment

- Patients are assessed by a dermatologist and are either discharged (FN & TN) or referred for histological assessment (TP & FP).

Teledermatology

- Patients who are ineligible for teledermatology assessment are referred directly to face-to-face assessment;
- Images are assessed remotely by a consultant, and patients are either discharged (FN & TN) or referred for a face-to-face assessment;

There are five Markov models used to represent the differing prognoses of patients by diagnostic outcome and indication beyond the terminal nodes of the decision tree:

High-risk cancer (melanoma, SCC, rare cancers)

- True positives: Patients enter a health state corresponding to the stage of their disease at the point of diagnosis and treatment. The prognosis of patients with in-situ or Stage 1a cancer is equal to that of the general population for the remainder of the modelled time horizon. Later cancer stages are essentially modelled using a series of three tunnel states, wherein a patient is subject to an elevated mortality risk for the first five years which declines for the following five years, and returns to that of the general population thereafter.
- False negatives: Patients enter a health state corresponding to the stage of their disease. Every year the patient can remain undetected and remain at the same stage, progress to a more advanced stage, or be opportunistically diagnosed and treated. The outcomes of these patients upon diagnosis are modelled in the same way as true positives.

Basal cell carcinoma

- True positives: Patients correctly diagnosed with BCC are treated and experience general population mortality risk. A small disutility is applied to some patients reflecting the impact of scarring on the head or neck upon HRQoL. This Markov model comprises two health states – alive and dead.
- False negatives: Patients with undetected BCC have a 20% annual probability of being opportunistically diagnosed and treated. There is no risk of progression associated with having undetected BCC, nor is there any impact on HRQoL. A proportion of patients whose BCC is detected and treated experience a small utility decrement as above. A four-state model is described – undiagnosed BCC, opportunistic detection and treatment, treated, and dead. It is unclear what purpose the separate health state representing detection and treatment serves.

5.1.3 Mechanism of cost consequences

Costs relating to diagnosis include face-to-face assessment, biopsy/excision, and multidisciplinary team meetings (MDT). Costs associated with each of the diagnostic processes are replicated in Table 17 for comparison.

Table 17 Exeter model diagnostic costs

Parameter	Cost	Source
Photo clinic appointment (medical photography for DERM, teledermatology)	£14.30	Skin Analytics – 45 minutes of Band 3 time.
Teledermatology review	£25.00	Skin analytics – 10 minute slot (2020 PSSRU cost – hospital based consultant, medical) plus 15 mins Band 3 administration time.
Teledermatology system price per image	£7.00	Skin Analytics – list price of Cinapsis, Dermicus.
DERM second read	£17.00	Skin Analytics consultant time*
DERM assessment price per image	£38.20	Skin Analytics list price*
Face-to-face dermatologist appointment	£142.00	WF01B, 2023-25 NHS Payment Scheme. NHS England

** Unit prices provided to the EAG differed from those applied in the Exeter model*

The costs of further follow-up and treatment following a referral to a face-to-face appointment differ by the modelled indication, with further costs associated with more advanced stage at diagnosis. On the melanoma pathway, initial biopsy/excision and sentinel lymph node biopsy had a unit cost of £507, and was applied in addition to an MDT (£123) to all patients who were not discharged following their face-to-face assessment, as was a vitamin D test at a cost of £178. The source of the £178 cost of a vitamin D test was unclear, and appeared to be substantially higher than other literature sources,[65] which tend to inflate from a figure of £16.50 based on previous NICE guidance.[66] The costs associated with biopsy and treatment appear high relative to the studies discussed in Section 4, and were not consistently based on NHS Reference Costs/PSSRU costs. This punishes diagnostic strategies with lower specificity and may inflate the potential cost savings associated with a higher-specificity strategies.

Frequency of clinical follow-up was determined by disease stage at diagnosis, with a unit cost of £77 for each visit. Patients with Stage 1B or higher disease were assumed to require frequent ongoing follow-up imaging (e.g. MRI, CT, ultrasound). Terminal care costs of £15,531 were applied to patients who died with Stage 1b or higher disease. Costs of further investigations were applied to melanoma patients with Stage 2 or higher disease at diagnosis, including histology testing, and further medical imaging. Further surgical and systemic treatment was included for patients with Stage 3 or 4 disease.

Treatment costs associated with BCC were calculated using a weighted cost of £556.82 per patient, comprising various alternative treatment strategies from McFerran et al.[67], with costs inflated to 2024 values using the EPPI-Centre cost converter. It was noted that phototherapy, which contributed £38.84 to the weighted cost, is not used for treatment of BCC on the NHS. As there are no health consequences of a missed BCC diagnosis, the only meaningful outcome of a correct BCC diagnosis is

incurring this cost. This counterintuitively means technologies with poorer sensitivity generate value by having a lower sensitivity.

The primary mechanism of cost-savings in the model was the avoidance of face-to-face assessments and biopsy. The specificity of a face-to-face assessment with a consultant was 79.7%, resulting in a proportion of patients receiving costly biopsy unnecessarily if they are not discharged using teledermatology or DERM. A diagnostic pathway with higher sensitivity also avoids of missed cases which have the potential to develop into advanced disease, with substantially increased treatment costs. An important assumption with the model that the sensitivity of face-to-face assessments is increased to 99% following triage with either DERM or teledermatology. This means that fewer cancers are missed in model pathways including an additional triage step. The plausibility of this assumption is not clear and may not reflect real-world practice given the low assumed specificity of DERM and teledermatological assessment.

The assumption that a relatively high number of unnecessary biopsies resulting from face-to-face assessments, and the improvement in sensitivity of face-to-face assessment following triage, are likely key drivers of benefit in the model and as such the associated diagnostic parameters are central to the value proposition.

5.1.3.1 DERM (with or without second read) versus teledermatology

Results of the company's model suggest DERM either with or without second read generates costs savings relative to teledermatology, see Table 18. A simple comparison of first-line assessment costs inclusive of DERM however suggests that both DERM strategies are more costly than teledermatology (£72 vs £57 average cost per patient). These higher costs associated with both DERM strategies are driven fewer by patients being eligible for assessment by DERM than teledermatology (81% vs 90%). This results in more patients receiving more expensive face-to-face assessments.

The first line incremental costs associated with both DERM strategies are, however, offset by improved specificity relative to teledermatology which results in higher effective discharge rates. Effective discharge rates are 36.9% for DERM without second read, 15.7% for DERM with a second read and 30.9% for teledermatology. The higher discharge rates associated with DERM without a second read generates cost savings as fewer face-to-face appointments are required and fewer biopsies conducted, whilst DERM with a second read generates cost savings through the avoidance of missed diagnoses. The specificity of teledermatology was assumed to be 35% based on an average observed across UK DERM pilot pathways and other real-world data sources, this compares with a specificity of 42% assumed for DERM without second read based on performance across secondary care pilot sites. The specificity of DERM with a second read can be estimated using the DERM specificity of 42% and the specificity of the second read of 60%. The assumed specificity of teledermatology

however appears low compared with published sources. Teledermatology specificity was reported as 84.3% in the Cochrane review referenced in the preliminary Exeter report which may indicate the assumed specificity is lower than in practice.[55]

The total average costs of the peri-referral pathway (i.e. between referral and initial secondary care consultation) are approximately £146 for teledermatology, £118 for DERM without a second read and £172 for DERM with a second read (assuming there is no additional step which can overrule Skin Analytics dermatologists). That is, the reduction in face-to-face dermatology referrals achieved by DERM used autonomously generates cost savings per patient referred from primary care. DERM with a second read may be the most costly approach, but may be associated with non-cash releasing benefits related to outsourcing of teledermatology review to Skin Analytics consultants. Note that using the modelled assumptions, the inclusion of teledermatology in this pathway is more costly than simply referring all patients to a face-to-face assessment. However, if the Cochrane diagnostic accuracy values are applied for teledermatology, DERM strategies become more costly than teledermatology. Teledermatology also becomes cost saving versus the traditional pathway.

5.1.3.2 DERM without second read versus face-to-face assessment

Compared to face-to-face assessment, results of the company's model suggest both DERM strategies incur lower costs, see Table 18. As above, this driven by lower costs associated with unnecessary referrals and inappropriate biopsies. In the BCC population additional cost savings are also generated dues to the lower sensitivity for DERM compared to face-to-face assessment (90% vs 95%). This occurs because of the assumption that missed cases of BCC have no consequences in terms of costs.

5.1.3.3 DERM with a second read versus DERM without a second read

DERM with a second read is associated with incremental costs compared with DERM without a second read, see Table 18. The cost difference between the two strategies is in part driven by the addition of the second read which increases costs in the DERM with a second read strategy. However, the incremental costs associated with DERM with a second read are partially offset by the lower rate of missed diagnoses.

Table 18 Results of cost-effectiveness of DERM

			Inc. (vs usual care)		**Inc. (vs teledermatology)**		
Strategy	**Cost (£)**	**QALY**	**Cost (£)**	**QALY**	**Cost (£)**	**QALY**	**ICER**
DERM + second read	£465.84	11.1925	-£31.14	+0.0077	-£6.27	+0.0039	£24,655.23
DERM	£445.09	11.1917	-£51.89	+0.0069	-£27.02	+0.0031	-
Teledermatology	£472.11	11.1886	-£24.87	+0.0038	-	-	Strictly dominated
Usual care – baseline	£496.98	11.1848	N/A				Strictly dominated

Inc: incremental; discrepancies due to rounding.

5.1.4 Mechanism of health consequences

The annual risk of progression with melanoma, SCC, and other rare cancers is derived from Wilson 2013 as described in Section 4.2.1. As in Wilson 2013, the distribution of disease stage at diagnosis was implicitly assumed not to be affected by the underlying diagnosis. Health outcomes in the model were a consequence of both treatment and underlying disease which were applied as either a utility decrement or mortality modifier. Specific assumptions were applied for the BCC population, and melanoma, SCC and other rare cancer populations.

The model assumes BCC does not progress if not diagnosed, and undiagnosed BCC has no further health consequences. There is a 20% annual probability of opportunistic detection of undiagnosed BCC in the Markov phase of the model, with all patients assumed to achieve general population health outcomes following treatment, with no risk of recurrence. The treatment of BCC is associated with costs and causes a permanent disutility in 15% of the 58.9% of patients with scarring on their head or neck. There are therefore negative outcomes in terms of both costs and QALYs associated with correctly diagnosing a case of BCC, meaning that in the model it appears that more benefit is yielded by missing a given case of melanoma than by detecting it. The assumptions underpinning the modelling of BCC may not be clinically plausible. This means that a diagnostic strategy with a higher sensitivity for BCC is likely to be less cost-effective than one that misses BCC more often and postpones diagnosis. The sensitivity of DERM for BCC is 90%, lower than the 95% assumed for face-to-face assessment and teledermatology. This is likely to lead to increased costs and reduced QALYs for the latter two strategies, despite achieving a better diagnostic outcome. The clinical plausibility of this is unclear and runs counter to expectations that improving diagnostic outcomes improve health outcomes.

Melanoma, SCC, and other rare cancers were assumed be associated with lower quality of life dependent on disease stage. Utilities were based on a 2014 study by Tromme and colleagues,[68] which used the EQ-5D-5L questionnaire in a population with melanoma to derive utility weights according to disease stage and whether patients were actively undergoing treatment or were in remission. These utilities were adjusted to the mean age of the modelled population using Sullivan et al.[69] Utilities reflecting the impact of treatment were applied as a one-off disutility in the first year of treatment. Patients with Stage 1b or 2 disease were assumed to return to an age-adjusted general population-equivalent utility two months after treatment. Those with Stage 3 or 4 cancer at diagnosis have a reduced quality of life for the remainder of their lifetime. These utilities were based on small samples and are not necessarily logically consistent – e.g. a patient who has recovered from Stage 3 cancer has a utility of 0.701, but 0.797 for a patient who has recovered from Stage 4 disease. A single utility representing recovered patients with Stage 2 or above cancer may have been more appropriate.

Melanoma, SCC, and rare cancers were assumed to occur on the head or neck in 40.2% of patients, 15% of whom experienced scarring following treatment, and a permanent disutility, the magnitude of which was not reported. It was also assumed that patients would experience a disutility of -0.505 for the period over which they are waiting for a result after GP referral to capture the impact of anxiety and psychological distress.

Mortality rates for high-risk cancers were taken from Edwards et al. 2016, ultimately based on Balch et al. 2009.[59] As described above, patients with high-risk cancers at Stage 1b or higher had an increased risk of cancer-related mortality for the first 10 years following diagnosis and treatment, after which time they have the same mortality risk as a healthy member of the general population (using ONS data). An annual probability of death was calculated from 5-year survival data. The mortality risk, and thus potential QALY loss associated with undiagnosed high-risk cancers is a potentially significant driver of benefit generated by more sensitive treatment strategies.

The most effective strategy was DERM with a second read, generating 11.1925 QALYs at a cost of £465.84. The ICER for DERM with a second read compared to DERM alone was £24,655 per additional QALY gained. Both DERM strategies were predicted to be less costly and more effective than teledermatology and usual care. Teledermatology was less costly and more effective than usual care. The observed shortfall in QALYs accrued on usual care is likely to be driven by the assumption that the sensitivity of a face-to-face assessment is significantly improved in patients who have undergone previous DERM triage. This structurally confers health benefits onto strategies employing an intermediary step between primary and secondary care and may not be reflective of real patient outcomes.

5.1.5 Summary of critique

The submitted model represents the most recent and complete attempt to represent the NHS urgent skin cancer referral pathway, but is subject to a number of weaknesses which may mean it does not appropriately characterise the main drivers of value in this pathway.

As the most common form of skin cancer, the consequences of diagnosis and treatment of BCC is an influential driver of cost accrual in the model. The model essentially punishes correct BCC diagnoses, as excision is associated with accrual of costs and a QALY decrement. This introduces a disincentive to improve diagnostic sensitivity, and indeed DERM is less sensitive for BCC than teledermatology or face-to-face assessment. This may reduce QALYs and increase costs on the two comparator pathways, and this is somewhat concealed in the cohort structure.

The model also structurally imposes a 99% sensitivity for face-to-face assessment following triage, without evidential support. This means that the simple introduction of a triage step (i.e. DERM,

teledermatology) prior to consultation with a dermatologist reduces missed diagnoses and avoids the associated cost and health implications. This assumption may not reflect the respective real-world holistic sensitivity of these pathways but would invariably result in better cost-effectiveness estimates for DERM and teledermatology.

The costs associated with biopsy and treatment of high-risk cancers drives cost accrual in triage strategies with lower specificity, as more patients will undergo unnecessary and expensive diagnostic biopsy, in addition to the costs of a face-to-face consultation. Significant value is therefore generated by triage strategies with higher specificity. The magnitude of uncertainty surrounding the specificity of teledermatology is vital to understanding the potential for cost-effective use of DERM in this model structure. Given higher rate of ineligibility for assessment with DERM vs AI, the true discharge effective rates are closer than implied by simple comparison of the respective diagnostic accuracy statistics of each technology, as a higher proportion of patients on the DERM pathway proceed immediately to face-to-face assessment. The specificity of teledermatology reported in published sources is substantially higher than that observed in the pilot sites (which were largely not set up for teledermatology services). It is therefore highly plausible that in the presented model structure, teledermatology would be more cost-effective than a pathway incorporating DERM.

5.2 East Midlands Academic Health Science Network (2023)

The authors report an evaluation of a pilot of a Skin Analytics AI-powered teledermatology (i.e. DERM with a second read) for the skin cancer 2WW pathway at University Hospitals of Leicester (UHL) sites in March 2022. The evaluation uses a mixed-methods framework, combining patient and staff feedback surveys with quantitative data collected as part of the pilot. The existing pre-intervention pathway prior to the implementation of the pilot involved patients referred from primary care on the urgent skin cancer referral pathway (NG12).

All patients referred in this way progressed to a face-to-face appointment with a consultant dermatologist, at which point a decision would be made to discharge or investigate further. Patients deemed eligible for DERM attended a clinical hub where a healthcare professional captured an image of the patient's skin lesion. This image is then reviewed by the AI technology, and patients are provided with an instant diagnosis. Patients who receive a diagnosis of concern, or whose images could not be read by the AI, are reviewed by a hospital consultant first remotely and then in person if required. Lesions identified as benign by AI were further reviewed by a Skin Analytics dermatologist who validated the result.

[redacted]

[redacted]

Other indicative scenarios were performed, exploring the potential benefit of mitigating the cost of a second read which indicated cost-benefit ratios from 1.27 to 1.88.

[redacted]

5.3 NHSE AI in Health and Care Economic Evaluations

The contents of this report remain confidential at the time of submission.

[redacted]

[redacted]

6 MODEL CONCEPTUALISATION AND IDENTIFICATION OF EVIDENCE GAPS

6.1 Model conceptualisation

The following sections describe a conceptual model based primarily on a synthesis of the economic analyses identified in the economic review, and evidence submitted by skin analytics. While cost-utility models have recently been built to address the present decision problem (see Sections # and #), the EAG considers currently available evidence insufficient to answer the issue of the potential cost-effectiveness of AI technologies for detecting benign lesions following referral from primary care. This section expands upon the EAG reasoning for this conclusion and details key data necessary to fully address the decision problem.

6.1.1 Decision problem

The outlined conceptual model considers both use cases for AI technologies proposed for this evaluation, i.e., autonomous AI triage following referral from primary care, and AI triage with a second read 'safety net' prior to discharge following referral from primary care. The use case of AI technologies to be assessed is the identification of benign lesions and the direction of discharge prior to contact with secondary care. These decisions could be made autonomously by AI or following dermatologist review (second read). A holistic modelling approach to the diagnostic accuracy of these technologies is necessary in order to assess the potential value to the NHS.

The modelled population should include all patients referred on the urgent referral skin cancer pathway from primary care. The prevalence of cancer subtypes should be sourced from appropriate and recent UK national sources. Staging of disease at the point of entry into the model should be based on UK data if available. If there are differences by stage at presentation according to indication, this should also be reflected. See Section 6.2.1 for further discussion.

To reflect current service provision, two alternative comparator diagnostic pathways are considered - the teledermatology model, and the conventional model of referral to face-to-face assessment model. Current provision varies across the English NHS, with no nationally standardised alternative model to the usual referral pathway. Fully reflecting regional variations may therefore require additional comparator pathways to be modelled. Modelled outcomes should include diagnostic outcomes, i.e. TP, FP, FN, TNs; costs; and QALYs. Disaggregation of outcomes by indication should be possible.

The proposed model should be built in full alignment with the NICE Reference Case and should adopt an NHS and Personal Social Services perspective. Costs and benefits should be discounted at 3.5%

per annum. A lifetime time horizon should be applied on the basis of the age of the modelled population.

6.1.2 Proposed model structure

In line with previous economic analyses, the EAG proposes a cohort model in which all patients enter a common decision tree structure, regardless of underlying indication. Different Markov models would then be used to reflect differences in long-term costs and outcomes as a result of diagnostic outcome. Differences in model inputs relating to costs and health outcomes, would allow the model to be parameterised to address specific indications e.g., melanoma, SCC, and BCC.

The level of granularity possible in the model will be data dependent. However, as in previous models, it will be important to differentiate melanoma and other high-risk cancers from BCC, as the costs and consequences of diagnosis and misdiagnosis can be radically different.

The proposed model applies three broad diagnostic categories which each have distinct long-term consequences, which are represented by different Markov models. The capacity of the conceptual model to account for specific diagnoses within these categories is dependent upon the availability of data to inform specific diagnostic accuracy and natural history parameters. The diagnostic categories are as follows, based on the groupings proposed in the Exeter Test Group/Skin Analytics model described in Section 5.1:

- 'High-risk cancers', including melanoma, squamous cell carcinoma, and other rare high-risk cancers;
- basal cell carcinoma;
- low-risk lesions (benign and precancerous).

A key concern regarding the use of AI technologies for the diagnosis of skin cancers is the identification of rarer indications. Given that these technologies may have limited experience of rare cancers, there remains uncertainty as to whether their high sensitivity to melanoma and SCC is maintained across these rarer indications. Treating them as a single diagnostic category in terms of diagnostic accuracy, stage at diagnosis, rate of progression, and impact upon mortality may therefore be subject to uncertainty. Where possible, sensitivity analysis should be undertaken in which rare cancers are categorised separately, and alternative sources of diagnostic and prognostic data are used to parameterise this sub-population in the model.

Decision trees

Patients enter the decision tree following a urgent referral from primary care, according to the chosen approach to AI implementation (i.e. with or without a second read), and to each comparator (face-to-face & teledermatology). The decision tree directs patients through a series of tests and clinical

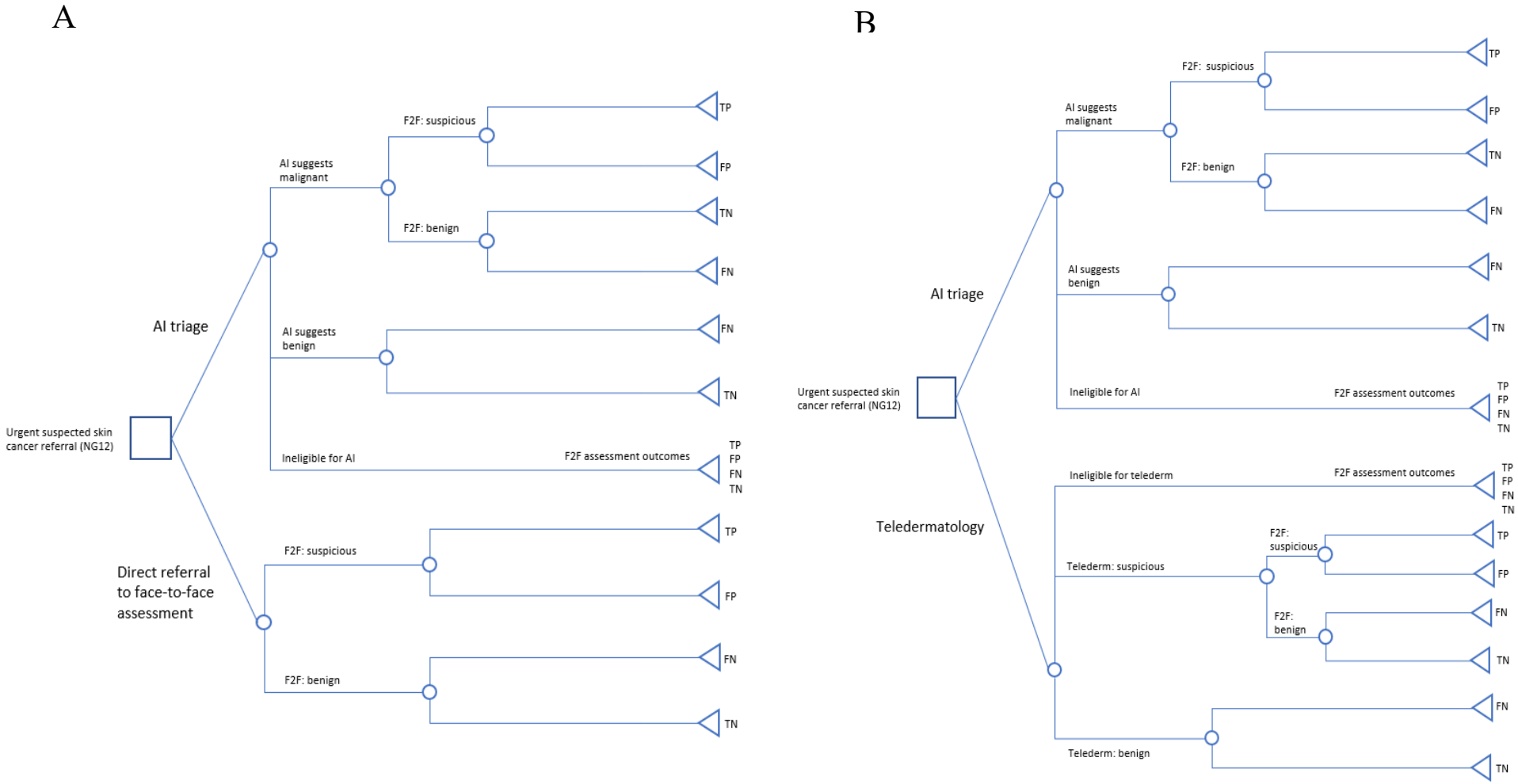
decision points, determining their accumulation of any costs associated with testing and appointments, and their ultimate diagnostic classification, i.e. TP, FP, FN, TN.

The comparator combinations of AI with and without second read with teledermatology and the direct referral pathways generate four diagnostic pathways, illustrated in the simplified decision tree schematics in Figure 13. Whilst the schematic depicts individual head-to-head comparisons, the proposed model would generate results in a fully incremental format.

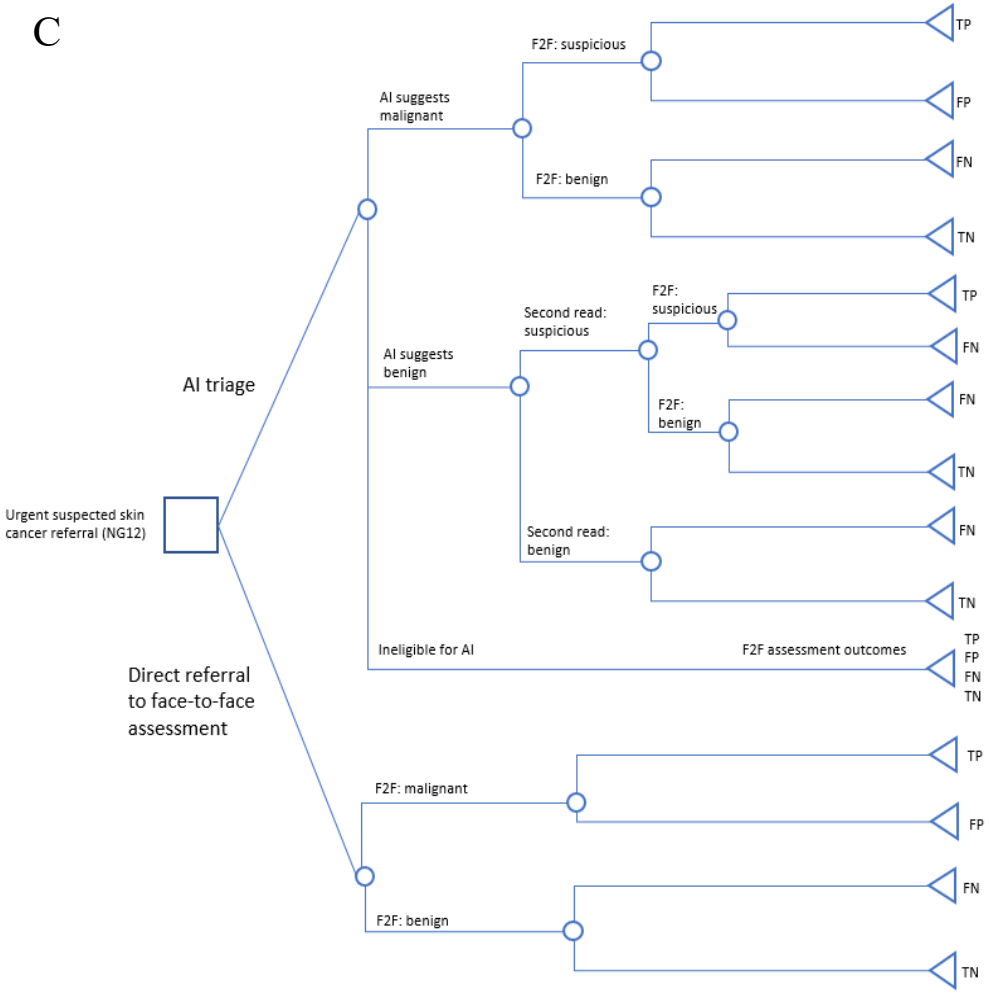
Only a proportion of patients are eligible AI and teledermatology assessment this is represented in the decision trees by a third initial branch. Different proportions of patients are eligible for each of these technologies, with current eligibility criteria more restrictive for the use of AI triage technologies than for teledermatology. This may have a significant impact on the costs and outcomes achieved on each pathway. Patients ineligible for AI/teledermatology are routed straight to face-to-face assessment, with a proportion whose ineligibility was not assessed prospectively, and were thus subject to additional costs associated with unsuccessful photography/an indeterminate AI result. The diagnostic accuracy of a consultant dermatologist may also differ for patients whose lesions are ineligible for each technology. Where possible this should be accounted for in the economic analysis or otherwise explored in relevant sensitivity analysis.

All decision trees determine the proportion of patients with TP, FP, TN, and FN under each diagnostic strategy, with long-term outcomes for each determined by each of the respective Markov models depicted in Figure 14. At the terminal nodes representing TP and FP, patients are assumed to undergo biopsy and/or treatment appropriate to their stage at diagnosis.

Figure 13 Proposed model structure: Decision tree schematic (A) AI without second read vs referral to face-to-face assessment; (B) AI without second read vs teledermatology; (C) AI with second read vs referral to face-to-face assessment; (D) AI with second read



C


Urgent suspected skin cancer referral (NG12)
AI triage
Direct referral to face-to-face assessment
AI suggests malignant
F2F: suspicious
TP
FP
F2F: benign
FN
TN
AI suggests benign
Second read: suspicious
F2F: suspicious
TP
FN
F2F: benign
FN
TN
Second read: benign
FN
TN
Ineligible for AI
F2F assessment outcomes
TP
FP
FN
TN
F2F: malignant
TP
FP
F2F: benign
FN
TN

D

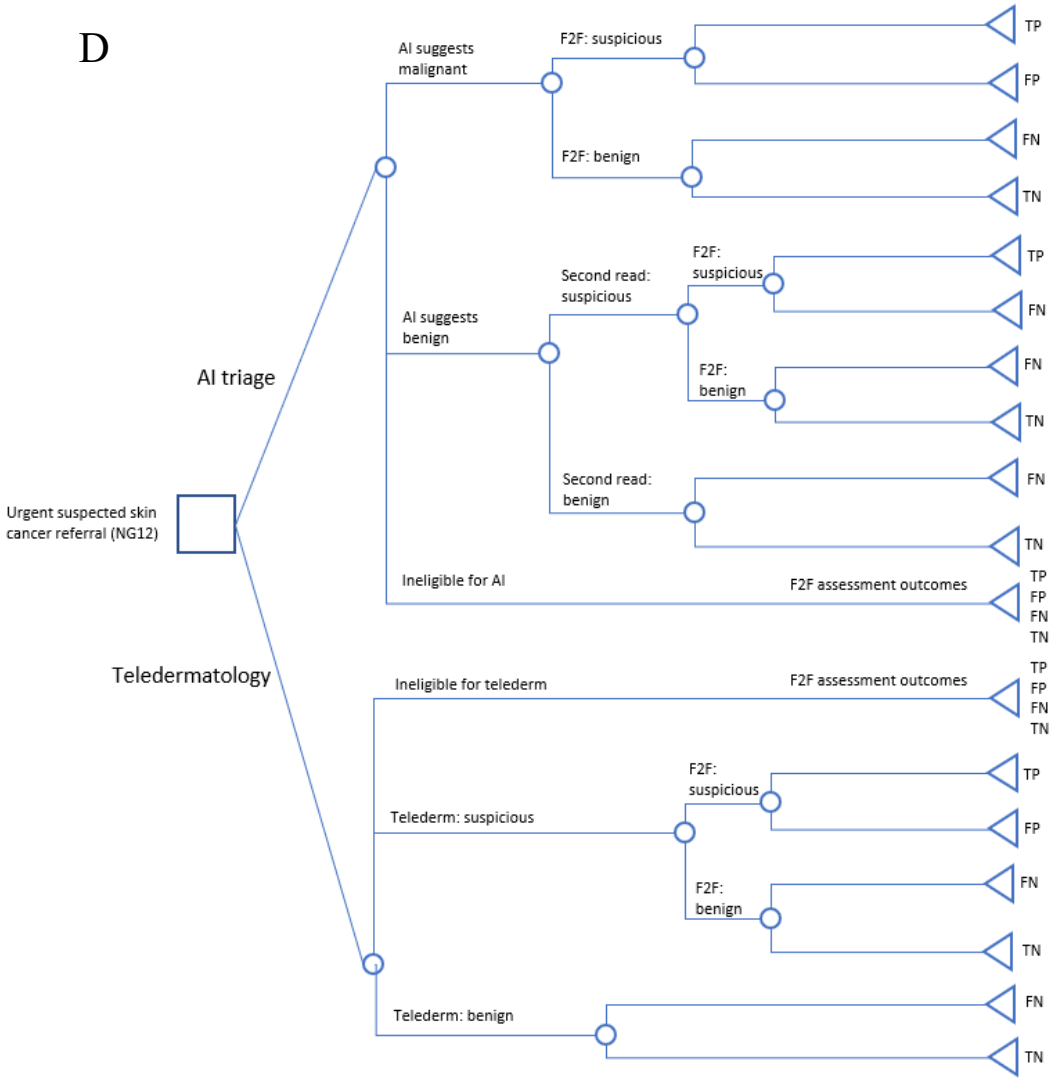
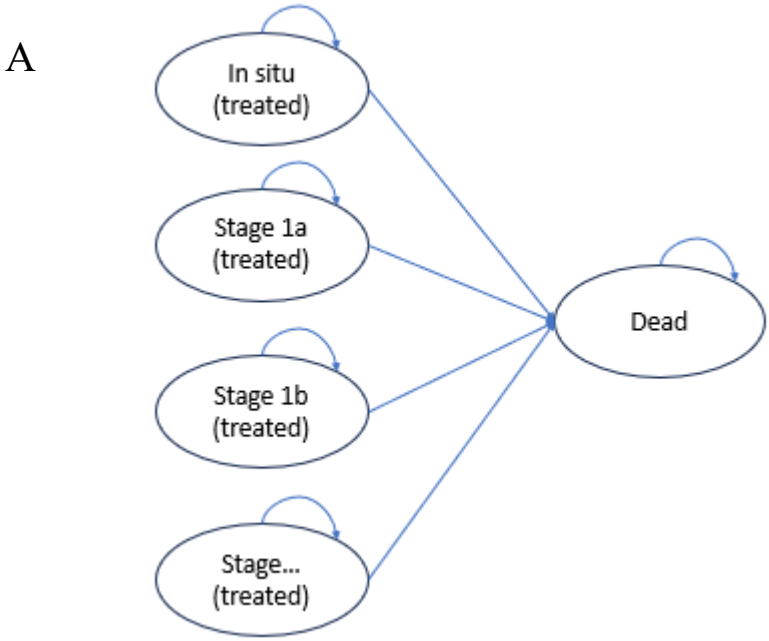
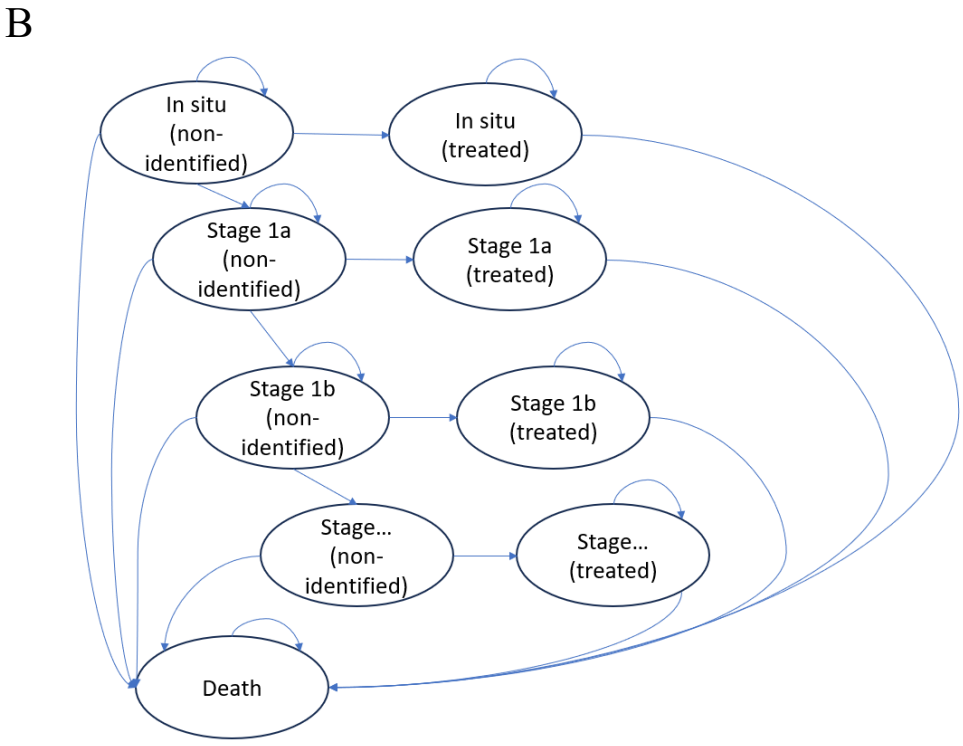

Urgent suspected skin cancer referral (NG12)
AI triage
Teledermatology
AI suggests malignant
F2F: suspicious
TP
FP
F2F: benign
FN
TN
AI suggests benign
Second read: suspicious
F2F: suspicious
TP
FN
F2F: benign
FN
TN
Second read: benign
FN
TN
Ineligible for AI
F2F assessment outcomes
TP
FP
FN
TN
Ineligible for telederm
F2F assessment outcomes
TP
FP
FN
TN
Telederm: suspicious
F2F: suspicious
TP
FP
F2F: benign
FN
TN
Telederm: benign
FN
TN

Figure 14 Markov model components: (A) True positives; (B) False negatives; (C) True negatives and false positives





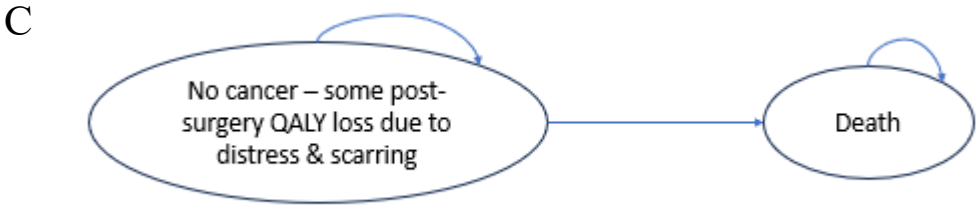
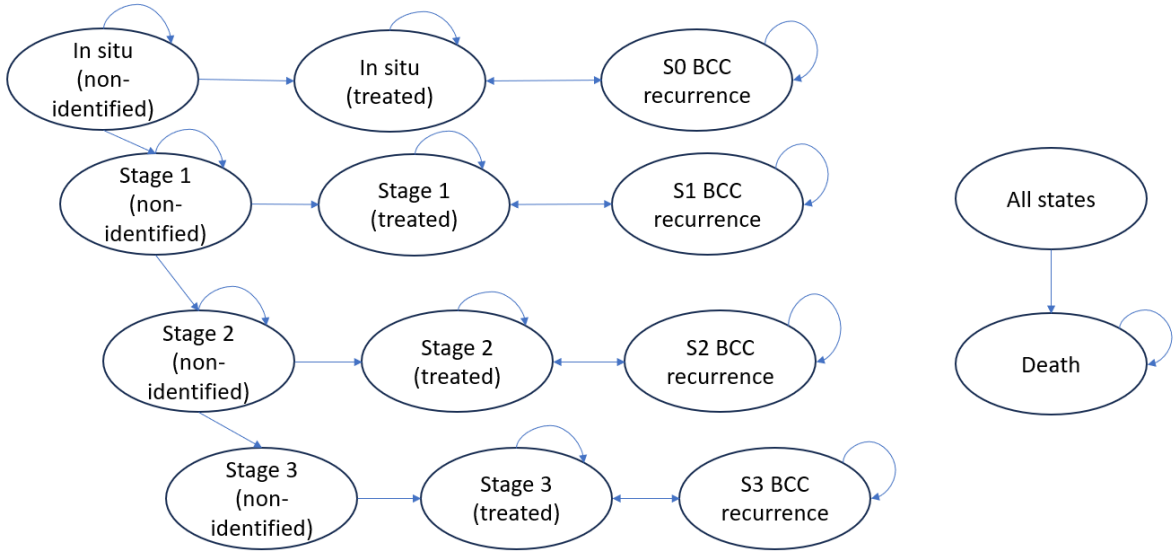


Figure 15 Markov model component capturing outcomes of basal cell carcinoma false negatives



Markov models

Patients correctly identified at the terminal nodes of the decision tree enter Markov model A (Figure 14), these patients have ongoing mortality and HRQoL implications following treatment depending on the disease stage at the point of diagnosis. This Markov model comprises a Markov state (not depicted) for every possible disease stage at the point of diagnosis and treatment, and a series of tunnel states reflecting mortality risks post-treatment. The use of tunnel states permits declining risks of post-treatment mortality to be modelled (per Edwards *et al.*) and may be applied as long as clinically appropriate, at which point they return to a general population risk of mortality (see Section 6.2).

Patients who reach a false negative terminal node enter Markov model B (Figure 14), and have a stage-specific risk of progression, mortality, and opportunistic detection. Patients with a true negative or false positive diagnosis enter Markov model C (Figure 14). These patients have general population mortality and HRQoL outcomes. Utility decrements may be applied to account for the long-term impact of scarring due to inappropriate biopsy on the head and neck in false-positive patients.

The EAG note that existing modelling approaches assume no adverse implications of a missed BCC in terms of cost or health outcomes. The clinical plausibility of this approach is unclear and in the context of a cost-utility model, this essentially rewards strategies with lower sensitivity, as the costs associated with BCC are avoided or postponed (and are subject to more discounting). The EAG therefore proposes an alternative approach to capturing the long-term impact of missed diagnoses of BCC. Under this approach, true positives, true negatives, and false positives follow the same Markov model structures as the high-risk cancers, but in the proposed model, false negatives for BCC follow the structure presented in Figure 15. Whilst BCC is associated with a low risk of spread and

progression to metastatic forms of the disease, if left undiagnosed, some subtypes can be invasive and can cause local destruction of deeper tissues such as muscle and bone,[70] which can be particularly impactful for lesions located on the head and neck. Untreated BCC may become more advanced over time and can be prone to higher rates of recurrence.[71-73] Whilst recurrence of BCC remains manageable, it is associated with additional treatment costs[11]. The Markov structure in Figure 15 therefore intends to capture the slow development of non-identified BCC and its opportunistic detection. Following detection and treatment, patients are then subject to a stage-specific recurrence rate. It is assumed that recurrent BCC is immediately detected and treated (with an associated cost), and patients return to the stage-specific 'treated' health state, with an ongoing risk of further recurrence. This assumes that recurrence of BCC does not have a modifying effect on the probability of future recurrences. Mortality is possible from any health state. Ideally, stage-specific treatment costs and disutilities would be modelled to allow differences in treatment costs arising from differences in the complexity of surgical intervention and reconstruction, and the potential impact upon HRQoL to be accounted for in the model.[74] The modelling of BCC in this way will be dependent on the availability of data to inform progression and stage-specific recurrence rates.

This model structure is in broad alignment with the analyses described in Section 4, including that built by the Exeter Test Group, which in itself was adapted from Wilson et al. 2013[56] and Edwards et al. 2016,[58] with the addition of a Markov component to capture the long-term outcomes of a missed case of BCC. A model of this design captures the differential in core costs and consequences of alternative diagnostic tests which impact the routing of patients through a diagnostic pathway. This includes the financial consequences of appropriately discharging patients with benign lesions, and avoiding unnecessary resource intensive face-to-face consultations, but also the impact of missed diagnoses on cost and health outcomes.

An important omission from the proposed model structure is the ability to capture non-cash resource benefits. Doing so would require a more complex approach which estimates the effect of the technologies upon downstream dermatologist capacity, and the impact of its deployment in this and other populations upon health outcomes.

6.2 Clinical input parameters

6.2.1 Prevalence of disease and distribution by disease stage

Estimates of prevalence are required for the proposed model and should be based on the population described in the decision problem i.e., patients referred on the urgent referral pathway from primary care. A systematic review should be conducted to identify the prevalence of each disease type considered in the model for the UK urgent referral population or identify sources of NHS data to inform this parameter. All three studies identified in the cost-effectiveness review in Section 4.2 relate

to different patient populations, and are therefore not relevant to the current decision problem. The Exeter model used post-market surveillance data from Skin Analytics to obtain estimates of prevalence but noted this was a placeholder with a preference for acquiring national data in the future. Sensitivity analysis should be conducted on plausible estimates of prevalence to represent uncertainty or regional variation in prevalence estimates.

Also required for the proposed model is the distribution of each disease by stage at identification. Data should be obtained based on stage at presentation for lesions examined on the urgent referral pathway across each disease type considered in this model. Distribution of disease by stage will impact cost and health outcomes estimated by the model. If presentation is typically at later stages of disease, there will be reduced scope to generate benefits via early detection and vice versa.

6.2.2 Diagnostic accuracy data

To inform a future cost-effectiveness model, data on diagnostic accuracy are required for all relevant diagnostic strategies This data should ideally be obtained for each indication considered in the model, as diagnostic accuracy may differ by condition.

The use case for AI technologies in the proposed model involves the identification of benign lesions to allow patients to be discharged following referral, but prior to F2F assessment. The key statistic to estimate the capacity of a test to correctly identify true negative cases is the specificity associated with this pathway. Discharge of patients with benign lesions reduces the cost and health implications associated with unnecessary investigations. Whilst diagnostic accuracy may be framed with regards to its sensitivity to benign lesions at this point in the diagnostic pathway, it is helpful to refer to sensitivity and specificity for detection of malignancy, for consistency with the intent of subsequent/comparative F2F assessment.

The value implications of differing diagnostic performance across the comparators under consideration will depend on the following assumptions: follow-up costs for patients after a positive test (e.g., cost of biopsy and excision), the health consequence of treatment itself (e.g., scarring due to excision, anxiety while waiting for biopsy results), Markov state stage-dependent treatment costs/health decrements, and assumptions regarding the progression of patients. The net effect of this (along with the impact of discounting) will determine how diagnostic accuracy drives costs and outcomes.

In the case of the present pathway, this relationship is somewhat complicated by the application of sequential tests, for example, the use of a second read following AI assessment. For two sequential tests with imperfect sensitivity, independence between tests would imply that overall sensitivity of a pathway would decrease. This may not be reflective of actual practice and would punish pathways

with more steps. It is unlikely that test sensitivity is fully independent between steps, a lesion deemed malignant by AI may be more likely to be deemed malignant by a read by a dermatologist and so it may be inappropriate for subsequent sensitivity values to be applied to one other. Equally, assuming interdependence of two sequential tests may also not be completely appropriate, especially where testing is subjective. Similar issues also apply to test specificity. For two tests with imperfect specificity, the overall specificity would increase assuming tests are independent with specificity generally increasing with the number of steps in the pathway. This may not hold true in practice, false-negative cases detected by AI might reflect unusual or difficult cases that might also be identified as positive by a dermatologist upon second read.

Given these complexities and the number of diagnostic decision points in the decision trees described in Section 6.1.2, care must be taken that diagnostic accuracy values are not simply pieced together from different sources to estimate whole-pathway sensitivity and specificity. In order to understand the resource use implications of post-referral use of AI or teledermatology, data on the sensitivity and specificity of both the whole pathway and its constituent components must be collected. This data should ideally be generated comparatively on the same clinical population (i.e. having undergone the same pre-screening) in the same conditions. In the case of teledermatology, it is important to ensure the intention is the same as AI, i.e. with the express intention of identifying and ruling out benign lesions (as opposed to triage/prioritisation of all lesions), otherwise estimates of specificity are not comparable with the use case of AI technologies in this space.

The clinical evidence supplied in support of the DERM and MoleMate technologies is described and synthesised in Section 3. This evidence is largely derived from pilot studies. The EAG consider that further development of the evidence base is required to inform a future cost-effectiveness model. Available diagnostic accuracy evidence for the technologies across alternative pathways and settings. To inform a future cost-effectiveness model, data should be based on studies with the following characteristics:

- **Setting:** UK post-referral (before secondary care investigations);
- **Intervention:** AI technologies (with and without human confirmatory read);
- **Comparators**: Face-to-face and teledermatology with intent to exclude benign lesions;
- **Outcome**: Diagnostic accuracy (sensitivity and specificity) of individual component tests and the overall pathway.

6.2.3 Progression and opportunistic detection parameters

Parameters describing the ongoing probability of undiagnosed progressing or being opportunistically detected are necessary to inform the transition between states in the long-term Markov components of the model, representing the natural history of skin cancer in patients with a false negative diagnostic

outcome. These parameters are likely to be influential in determining the mechanism of benefit in a future cost-effectiveness model. A model which applies more rapid progression or a lower chance of subsequent detection will impose greater value on improved sensitivity of a diagnostic pathway.

Data on the progression of unidentified skin cancers to inform progression probabilities appear limited. The approach taken by cost-utility models reported in the cost-effectiveness review and the Exeter Test Group model relied on expert-elicited progression probabilities, including Losina et al. 2007[60] and Wilson et al. 2017.[64] In the absence of more recently published alternative data sources, the proposed model may need to adopt transition probabilities based upon these studies.

Two contrasting approaches were taken by identified studies to structural assumptions regarding the opportunistic detection of false negatives of patients – Edwards et al. assumed that false negatives must be identified 5 years following initial assessment or upon progression Stage 2. However, this latter restriction may be reflective of the typically earlier staging at presentation of the population considered in the Edwards study. The other identified studies placed no structural limitations on patient progression. However, it may be appropriate to impose a time-based limit on the period over which a malignant lesion remains undetected, to avoid implausibly long durations of patients living with progressed disease.

Consideration should also be paid to whether it is appropriate to use common progression and identification parameters across multiple diseases in a future cost-effectiveness model or whether separate values should be used if disease processes are sufficiently different.

In the conceptual model, Markov state-specific mortality rates are likely appropriate, i.e. the mortality rate for a modelled patient is dependent upon their disease stage at presentation (if correctly identified and treated), or a patient's current disease stage (where undetected). Within previous models, mortality risks have increased with disease stage, with mortality risks converging with that of the general population following successful treatment. A permanent increase to mortality rates may also be appropriate in patients who experienced more aggressive treatment at later stages of disease. Mortality in patients with benign lesions can be reflected by general population rates.

A consensus across the models considered in this report is that people with 1a melanoma have a risk of mortality close to that of the general population and so no additional risk was assumed. Regarding patients with disease initially identified (or subsequently identified following initial non-identification), an assumption should be made regarding the duration of elevated mortality following treatment, reflecting the residual risk associated with the disease. Edwards assumed that following identification, patients would experience elevated mortality for 10-years after which their risk of mortality would return to that of the general population – 5 years at a higher rate and 5 years at a

lower rate. An alternative approach applied by Wilson 2013[56] (parameters obtained from Balch[59]) and Wilson 2018[57] (parameters obtained from a previous NICE appraisal [75]) calculated log-odds ratios for each stage and applied them to general population mortality rates.

All models identified in this report applied differential rates of mortality according to disease stage, reflecting differences in prognoses. Given the large sample size of the Balch et al. reference (n=30,946)[59], the EAG considers this a suitable source for populating a future cost-effectiveness analysis but may require re-analysis reflecting more recent techniques. Given the age of this study, further searches should be undertaken to identify more recent estimates of mortality in this population (or other secondary analyses of Balch *et al*), although this is unlikely to be an important driver of model outcomes.

6.3 Health-related quality of life

Health-related quality of life is represented in the model through the application of health utilities. Previous models have applied utilities to represent dimensions including:

- Utility decrements representing the disutility associated with diagnostic and treatment procedures (e.g., anxiety associated with the wait for biopsy results, scarring as a result of biopsy/excision);
- Health-state utilities representing diagnostic status and disease stage (or presence of disease).

The utilities reported in Tromme *et al*. 2014 (adapted for use in Edwards *et al.* 2016) to represent health-state utilities specific to disease stage and treatment status may be adequate to represent the impact of skin cancer and its treatment upon HRQoL. However, this data should be re-analysed – perhaps by pooling EQ-5D scores for patients with Stage 3 and 4 melanoma to avoid logical inconsistencies arising from the small sample size of patients with Stage 4 disease in remission. EQ-5D-5L summary scores should be crosswalked to EQ-5D-3L using the Hernández Alava mapping algorithm,[76] and should be adjusted for age and sex balance using the EEPRU value set established by the NICE Decision Support Unit.[77]Given the age of the Tromme *et al.* dataset and its aforementioned limitations, a systematic review of HRQoL studies should be undertaken to identify any more recently published data sources. Where alternative values are identified these should be mapped to EQ-5D-3L for consistency with the NICE Reference Case. Utilities should also be adjusted for age and sex balance using the EEPRU value set.[77]

An anxiety-related disutility in line with that used in Edwards *et al.* 2016 could also be applied for the period over which patients await a final diagnostic result following GP referral. The impact of AI

technologies on this interval should be identified from existing evidence sources (such as the DERM pilot studies), and its effects explored in sensitivity analysis, reflecting the potential for lengthened waiting times as seen in the UHL pilot. Previous models have also applied utility decrements with scarring on the head or neck following treatment. A disutility of an appropriate magnitude should be identified from literature sources.

A systematic review of HRQoL values should also seek to identify disutilities associated with BCC treatment. In the absence of disease-specific HRQoL data, it may also be appropriate to apply a one-off disutility equivalent to that applied for the treatment of melanoma in situ, which is typically managed using excision in a similar manner to BCC.

6.4 Cost and resource use parameters

Relevant costs in the proposed cost-effectiveness model include those related to diagnosis (e.g., the cost of the technologies, comparators, and clinical appointments), treatment and investigation-related costs (e.g., biopsy, excision, imaging), and long-term state-dependent management costs based on treatment and disease-stage. Those related to the technologies themselves should be based on information provided by the companies and any implementation costs likely to be incurred should be considered in the model (e.g., staff training, establishing new medical photography infrastructure).

The costs and resource use assumptions applied in a cost-effectiveness model are likely to be a key driver of the value of technologies in this space. There is a degree of control over the valuation of each diagnostic accuracy parameter in models of diagnostic technologies, that is, greater value can be ascribed to improving sensitivity by emphasising the costs of a missed diagnosis on the cost of delayed treatment. Equally, a technology which prioritises specificity may be made to generate more apparent value through increasing specificity and thus avoid unnecessary further investigations. The proposed model should aim for consistency in sources of cost data with precedent in NICE appraisals to ensure costs to the NHS and PSS are represented as accurately as possible.

Any costs associated with NHS procedures should be based on the latest national sources in alignment with the NICE methods guide for consistency with previous (and future) NICE decisions. These sources include the Unit Costs of Health and Social Care[78], NHS Reference Costs to and the NHS Drug Tariff[79]. Any costs without appropriate NHS reference costs (e.g., long-term state-dependent costs) should be based on a synthesis of the available evidence with costs inflated to the current cost year. The application of unit costs in the model should be made based on treatment guidelines provided by NICE and authoritative clinical guidelines.

6.4.1 Technology costs

Costs of the relevant technologies were provided by Skin Analytics and Moleanalyzer as part of the assessment process. Available information for each company is described below.

6.4.1.1 Skin Analytics DERM

Skin Analytics provided information regarding pricing for DERM. Pricing information is provided according to two options on a per-year basis a) per 10,000 catchment population covered; and b) per 2WW referral. It is unclear whether both pricing models are available to trusts, or if the cost per 2WW referral is for indicative purposes only, as annual payments are stated to be made up-front.

The pricing options are presented in Table 19. The total cost per 2WW image processed is £30.00, with an additional optional unit cost of £8.20 per referral to store images in order to allow remote review by trust clinical staff.

The company state that pricing is inclusive of training and data storage costs. The proposed model should identify relevant costs of establishing the infrastructure necessary to take and process photographs, administer patients through the DERM process, and any further steps further to the implementation of the technology in settings with and without existing teledermatology services.

Table 19 Skin Analytics DERM pricing

Component	**£ per 10k**	**£ per 2ww**	**Description**
Base platform with DERM review	3,300	30.00	Image and medical history capture platform, DERM assessment, PDF report with suspected diagnosis and recommended next steps.
Teledermatology functionality add on (optional)	900	8.20	Specialist teledermatology functionality within Skin Analytics' system to allow clinical staff to virtually review patient's cases and decide on the most appropriate outcome.
Discount if contributing outcome data (optional)	(250)	(2.30)	Discount provided if > 50% of biopsy results for patients through the pathway are shared with Skin Analytics.
Total cost per year (ex VAT) – *with outcomes discount*	**3,950**	**35.90**	
Total cost per year (ex VAT) – *without outcomes discount*	**4,200**	**38.20**	
Second read (Skin Analytics dermatologist)	**£17 per case**		

6.4.1.2 FotoFinder Moleanalyzer

FotoFinder provided details of the costs associated with Moleanalyzer. The company provided the costs in Table 20 for the technology. It was unclear from the company's submission how these pricings applied, for example, whether on a per user basis or otherwise. The company stated that there was no cost for training and indicated that there was a discount for multi-user access. Full pricing details should be incorporated into a future cost-effectiveness model.

Table 20 Moleanalyzer pricing

Pricing option	**Cost**
FotoFinder Moleanalyzer AIMEE scoring (flat per year)	£1,210
FotoFinder Moleanalyzer Pro includes AIMME offline package (per year)	£1,750

AIMME: artificial intelligence mole examination and evaluation

6.4.2 Diagnosis, treatment and follow-up costs

As discussed, any future cost-utility model should be parameterised using NHS Reference Costs and costs provided by the Personal Social Services Research Unit (PSSRU) for consistency with other models considered by NICE. Unit costs should be applied to resource use assumptions informed by NICE guidelines.

The EAG have outlined a non-exhaustive list of unit costs in Table 21 below that could be adapted to implement into a future cost-effectiveness model, alongside a comparison with the values used by the Exeter/Skin Analytics model. For implementation into a future cost-effectiveness model, unit costs should be updated based on the most recent published reference costs.

Table 21 Cost items required for the proposed model

Parameter	**Exeter Model Value**	**EAG Identified Costs**	**EAG comment**
Dermatological appointment (outpatient)	£142 NHS reference costs (WF01B 330 - first attendance)	**WF01A** – non-admitted, follow-up: Non-consultant led: £129.26 Consultant-led: £163.41 **WF01B** – non-admitted, first visit: Non-consultant led: £143.81 Consultant-led: £163.39	Clarification should be sought as to the appropriate reference cost.
Teledermatology	£25 10 minutes of 'hospital-based consultant' time, with additional 15 minutes band 3 administration time – unit costs from PSSRU	**WF01C** – non-admitted, non-face to face, follow-up: Non-consultant-led: £121.20 Consultant-led: £115.44 **WF01D** – non-admitted, non-face to face, first visit: Non-consultant-led: £284.09 Consultant-led: £114.52	As NHS reference costs appear considerably higher than the values applied in the Exeter model, clarification should be sought as to appropriate unit costs.

Biopsy + excision	£507 – inclusive of biopsy, sentinel lymph node biopsy, and surgical treatment in a single sitting, NIHR costing	**JC42C** – outpatient, intermediate skin procedures, 19 years and over: £257.43	
Sentinel lymph node biopsy (SLNB)	See above	**WH54A** – admitted patient care, day case, CC Score 1+: £1,584.52 **WH54B –** admitted patient care, day case, CC Score 0: £1,510.75	
BRAF testing & reporting	Testing: £374 Reporting: £113	£37 – Olaparib STA[80]	
Ultrasound	£248	**RD43Z**: ultrasound scan duration 20 mins+: By department code: IMAGDA: £155.34 IMAGOP: £293.54 SI: £160.26	
CT scan	£108	**RD26Z**: computerised tomography scan, three areas, with contrast: By department code: IMAGDA: £139.49 IMAGOP: £146.34 IMAGOTH: £88.74 SI: £164.08	

6.5 Strengths and limitations of the proposed modelling approach

The conceptual model described by the EAG is based primarily on a synthesis of the economic evidence identified in the economic review, as well as evidence submitted by Skin Analytics. The presented model considers the currently available evidence and identifies areas where further research is required.

Strengths of the EAG's approach to the conceptual model include that it draws on precedent within the indication and other analyses considered by NICE to inform the structure, key assumptions, and parameterisation. The conceptual model better aligns with the NICE reference case, through the use of more consistent cost and utility data sources and methods of analysis. The alternative structure proposed by the EAG for patients with BCC better represents the long-term consequence of BCC in terms of recurrence and therefore better captures the consequence of a false-negative case.

Limitations of the model proposed by the EAG include that the model cannot capture one of the primary benefits of the system, namely non-cash-releasing benefits (in common with other identified models). The hybrid structure proposed (a decision tree and Markov extension) cannot meaningfully

quantify the impact of reducing demand on services in terms of reducing waiting times (and potential improvements in quality of care) for a specialist consultation across all dermatological indications. A more complex modelling approach would be required to capture demand, capacity, and temporal dynamics.

6.6 Summary of evidence requirements

To inform a future cost-effectiveness model, future research should focus on addressing the limitations of the clinical evidence that would allow greater certainty in comparative diagnostic accuracy of AI technologies against comparators. As discussed above, the clinical evidence identified in Section 3, was based on heterogenous pathways and settings and may not provide appropriate diagnostic accuracy inputs for the pathway described in this model. The EAG consider that studies reporting the diagnostic accuracy of should have the following characteristics:

- **Setting:** UK peri-referral (following referral from primary care, before secondary care investigations);
- **Intervention:** AI technologies (with and without human confirmatory read);
- **Comparators**: Face-to-face and teledermatology;
- **Outcome**: Diagnostic accuracy (sensitivity and specificity) of individual tests and the overall pathway.

7 DISCUSSION

7.1 Statement of principal findings

7.1.1 Diagnostic accuracy and clinical impact

7.1.1.1 DERM

Three studies of DERM were examined to assess diagnostic accuracy. Autonomous use of DERM appears to have a high diagnostic accuracy for detection of malignant lesions: with a sensitivity of around 96.1% (95% CI 95.4 to 96.8) for a specificity of around 65.4% (95% CI 64.7 to 66.1). Similar diagnostic accuracies were found for detecting specific types of cancer (melanoma or SCC). There was some evidence that DERM might misdiagnose BCC cases as SCC or melanoma. Results for malignancy were similar across published and unpublished data. The sensitivity when detecting benign lesions was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0).

The diagnostic accuracy of autonomous use of DERM appears to be similar the diagnostic accuracy of dermatologists without DERM. In the DERM studies dermatologists [redacted] The diagnostic accuracy of the whole teledermatology pathway including DERM could not be reliably assessed because of a lack of any independent reference standard of diagnosis.

The EAG found very limited evidence on the broader clinical impact of DERM, most of it unpublished. The evidence suggested that if DERM were used on its own around half of all patients would be discharged, and half referred for further assessment (either in person or through teledermatology). About 0.8% of patients would be discharged with a malignant lesion, mostly with BCC. Between [redacted] of patients have lesions that are unsuitable for DERM assessment. [redacted]

Patient opinion was broadly supportive of using DERM in some form as part of their diagnosis, but patients were divided on whether they preferred teledermatology to face-to-face appointments. Clinicians were generally very resistant to using DERM in isolation without human assessment of lesions.

7.1.1.2 Moleanalyzer Pro

Two prospective studies of Moleanalyzer Pro were identified; neither were performed in the UK. Moleanalyzer Pro alone had a sensitivity of 84.4% (95% CI 73.9 to 91.0) and a specificity of 84.5%

(95% CI 72.0 to 92.1) to detect melanoma from a meta-analysis of the studies. This appeared similar to the accuracy of dermatologists alone. No eligible evidence was found for the diagnosis of SCC, BCC or other cancers.

The EAG did not identify any relevant evidence on the clinical impact of using Moleanalyzer Pro.

Patient and clinician opinion was generally supportive of using Moleanalyzer Pro in some way to aid diagnosis. However, the overwhelming majority of patients indicated that they would like the opinion of an expert physician besides an AI-assisted diagnosis.

7.1.2 Cost-effectiveness review and stakeholder submissions

No published assessments of the cost-effectiveness of the named AI technologies in an NHS setting were identified. Three published cost-effectiveness studies were identified evaluating any diagnostic technology for skin cancer in an NHS setting. All three studies focused on melanoma but also consider other skin cancers (e.g. BCC). While all identified studies adopted similar model structures, the mechanisms by which diagnostic accuracy generated value (in terms of either cost savings or QALY gain) differed substantively across studies. In particular, diagnostic sensitivity had less value in some models with value instead generated by the avoidance of unnecessary referral and diagnostic procedures. This is exemplified in one identified model of BCC in which it was assumed that all cases were correctly identified, and as such there were no cost or health consequences from improving diagnostic accuracy. Conversely, in other models improved sensitivity and reduced frequency of missed diagnoses were the main driver of benefits. In these models, greater emphasis was placed on the consequences of missed diagnoses, with more granular modelling of the consequences of disease progression and mortality.

The EAG received several submissions that included relevant economic analysis. This included a preliminary report describing a cost-utility model developed by Exeter Test Group and Skin Analytics, a pilot evaluation of DERM for the skin cancer 2WW pathway at University Hospitals of Leicester, and several economic analyses commissioned by NHSE (Unity Insights and University of Surrey). All three submission assessed the value of using DERM in an NHS setting. No economic evidence in support of Moleanalyzer Pro was submitted.

The most comprehensively reported and relevant of these was the cost-utility model developed by Exeter Test Group. This model built upon the three previous skin cancer models identified in the EAG's review. Aligning with the proposed use case, this model represents an assessment of DERM in a post-referral setting, with and without a second read, compared with teledermatology and the conventional urgent referral model (face-to-face). It considered three diagnostic categories – high-risk cancer, BCC, and non-/pre-cancer.

The EAG considered the model structure largely appropriate to assess core aspects of the potential value of AI technologies for identifying benign lesions in a post-referral setting, but noted several issues which may mean that the main value drivers may not be appropriately characterised. Namely, the model imposed disincentives for the correct diagnosis and treatment of BCC, which rewarded the comparatively lower sensitivity of DERM; assumptions around post-triage diagnostic accuracy of face-to-face assessment which structurally assumed benefits for any strategy incorporating a triage step; costs associated with diagnostic investigations and treatment may be inconsistent with sources generally used in NICE appraisals, and may overvalue specificity in terms of generating cost-savings; and the derivation of the HRQoL value set is not aligned with the NICE Reference Case. It remains highly uncertain whether currently available diagnostic accuracy evidence is sufficient to reliably populate a cost-utility model, particularly with regards to the comparative specificity of AI technologies to an effectively implemented teledermatology service. Therefore, whilst this analysis predicted that DERM with or without a second read would dominate all other options, this was highly dependent on the relative specificity of teledermatology.

7.1.3 Conceptual model

The EAG outlines a conceptual model which aims to provide an alternative to that described in the Skin Analytics submission. The proposed conceptual model seeks to address methodological issues identified in the reviewed literature and to explore the necessary structure and evidence required for future model development. For patients with high-risk cancers, the model structure described in the Skin Analytics model would be preserved. An alternative structure is, however, proposed to capture the natural history of BCC in false negatives, to better reflect the long-term health- and cost-consequences of BCC.

While cost-utility models have recently been built in support of the present decision problem, the EAG considers the available evidence inadequate to characterise the potential value of these technologies in an NHS setting. In particular, the EAG highlights limitations in comparative diagnostic accuracy evidence for the named technologies. Current evidence for both DERM and Moleanalyzer is lacking with regard to the diagnostic accuracy of the whole diagnostic pathway (i.e. inclusive of subsequent steps). Availability of this data is essential to understanding the likelihood of missed cases which cannot be inferred from the partial data currently available. Similarly, comparable diagnostic accuracy data describing current service provision is lacking, particularly for the teledermatology pathway. Without comparative evidence on the diagnostic accuracy of AI technologies and teledermatology, their relative value for safe and cost-effective identification of benign lesions will remain unclear.

The EAG also note a lack of robust data available to inform progression probabilities in undiagnosed disease, and a focus on expert-elicited parameters in previous cost-utility models. Establishing rates of progression and ultimately the consequences of missed diagnosis is important to characterising trade-offs in sensitivity and other potential cost-savings. While adjunctive AI technologies have principally been positioned as means of more efficiently identifying benign lesions, the introduction of further triage steps may also impact pathway sensitivity and are likely to represent part of the value case for AI technologies.

The EAG propose that a future cost-effectiveness model should use unit costs obtained based on the NICE Reference Case from national sources, namely the latest NHS Reference Costs and Unit Costs of Health and Social Care (PSSRU) where available, with costs supplemented with those identified by a systematic review of the literature. The EAG also notes that costs of establishing the necessary services to implement the technology in trusts without existing teledermatology infrastructure have not been characterised. It may be appropriate to also include these start-up costs within any economic analysis.

7.2 Strengths and limitations of the assessment

7.2.1 Strengths

This report presents an extensive systematic review of all published and unpublished evidence on DERM and Moleanalyzer Pro. The consistency between evidence identified through database searches, and that supplied by the companies suggests that this report covers all the relevant evidence on the two technologies.

Skin Analytics supplied a large quantity of evidence on DERM, including raw study data and unpublished study reports and economic analyses. This enabled a more thorough investigation of the clinical value and cost-effectiveness of DERM than would have been possible if using only published studies.

The outlined conceptual model addresses limitations with currently proposed models to more comprehensively evaluate both the short-term costs and consequences associated with alternative diagnostic strategies.

7.2.2 Limitations

Given the short time frame for this project a rapid review approach was used. Database searches were more limited than for a full review and were focused on publications explicitly naming DERM or Moleanalyzer Pro. We acknowledge that some relevant material may have been missed, although the consistency of our findings with material supplied by the companies reduces this risk.

The use of a rapid review approach also meant that we restricted full data extraction and synthesis to studies with prospective inclusion of patients, and to the most recent versions of the two technologies. This may mean that some useful evidence has not been considered. However, we consider that our approach has focused on the highest quality evidence of most relevance to practice.

The rapid review approach and limitations in the evidence base meant that the capacity to synthesise evidence was limited. Meta-analysis was not feasible for most outcomes, and many key outcomes were only reported in one publication or source.

The EAG consider that while the proposed conceptual model improves upon the approaches taken by existing studies, the proposed model (as with all other identified studies) fails to capture non-cash benefits associated with demand on dermatologist time. To capture these benefits, a more complex simulation approach would be required, capturing demand, capacity, and temporal dimensions.

The EAG were unable to provide an assessment of the likely budget impact and resource use which was a stated objective of the project. This in part reflect the compressed timelines and late provision of materials by Skin Analytics. However, uncertainties in the applicable unit costing and underlying diagnostic accuracy associated with each technology would likely limit the strength of conclusions that could be drawn from such analysis.

7.3 Key limitations of the evidence base

7.3.1 Diagnostic accuracy

Only three studies of DERM and two studies of Moleanalyzer Pro that prospectively evaluated the diagnostic accuracy of AI in clinical practice were identified. Hence the evidence base for the technologies is modest. The prospective studies of Moleanalyzer Pro were conducted outside of the UK, were not explicitly in a teledermatology setting, and did not evaluate the accuracy of AI for detecting non-melanoma cancer.

The DERM versions (in particular, the set sensitivity/specificity thresholds) and the dermatoscopes used for clinical assessments were out of date; therefore, the applicability of the diagnostic accuracy results to current practice is uncertain.

Most patients included in diagnostic accuracy had lighter skin colours (Fitzpatrick types II-III). The restricted eligibility to DERM and Moleanalyzer Pro and the systematic exclusion of a significant proportion of participants who would normally be assessed in practice meant that the evidence base for both devices was considered to be at high risk of bias and raises concerns about its applicability to practice.

In all except one diagnostic accuracy study, only a subset of participants (those with suspected malignancy) had a reference standard test that included histopathology. Although this is reflective of practice, the risk of reference standard test misclassification in these studies cannot be excluded.

7.3.2 Clinical impact and benefit

There is no evidence on the impact of DERM or Moleanalyzer Pro on clinical morbidities, mortality, health-related quality of life. In particular, the EAG notes that there is no substantive evidence on the benefits or harms AI use might have for patients.

Evidence from healthcare practitioners on their confidence in DERM and its clinical and broader impact on the pathway and patient management is limited, although initial evidence from limited samples suggested that patients and clinicians do not support the autonomous use of AI tools.

7.3.3 Resource use

Evidence on resource use for DERM was mostly limited to some unpublished results. Much of this evidence compared DERM as part of the teledermatology pathway to face-to-face dermatology. Consequently, the impact on resource use attributable specifically to DERM is uncertain. In particular, how autonomous use of DERM might compare to DERM combined with dermatologist assessment is unclear.

The EAG found no evidence on the impact of Moleanalyzer Pro on resource use.

7.3.4 Cost-effectiveness

No evidence on the cost-effectiveness of either DERM or Moleanalyzer Pro was identified in the EAG's review of published evidence. Evidence on cost-effectiveness of DERM submitted by Skin Analytics and NHSE (Unity Insights and University of Surrey) was both preliminary and incomplete. Uncertainties in the main value drivers including diagnostic accuracy of both DERM and comparator technologies limit the conclusions that can be drawn from this evidence. A more complete understanding of the economic analysis commissioned by NHSE may address some of these uncertainties.

7.4 Patient and public inclusion

The short time frame of this assessment meant the EAG did not seek any independent public or patient involvement. Patient representatives were included on the scoping committee for this assessment and will be involved in the decision-making process based on this report.

At scoping, patient representatives identified several key issues for consideration:

- The need to ensure that use of AI does not lead to malignant lesions being missed.
- Concerns around equality due to difficulty in assessing lesions covered by tattooing, hair or scarring, or in hard to assess areas.
- Equality issues around diagnosis of skin cancer in people with darker skin or non-white ethnicity.
- The need to reduce anxiety created by the diagnostic process (e.g. due to long waits for diagnoses, or incorrect initial diagnoses).

The EAG notes that this report was largely unable to resolve these issues; see discussion in Sections 7.3 and 8.2.

7.5 Equality, diversity and inclusion

As this was a rapid review of existing evidence the EAG could not consider equality issues beyond what was available in publications or supplied material.

The EAG notes several equality concerns arising from our review:

The evidence base for both technologies included few patients with non-white ethnicity or darker skin tones. Since skin cancer may be harder to detect in these people this is of concern. It is unclear whether the AI tools have been properly validated in people with darker skin tome, and what is the resulting diagnostic accuracy. Differences in diagnostic accuracy could lead to inequalities due to different diagnostic pathways, such as if some people have to wait for a face-to-face appointment because an AI assessment was inconclusive.

DERM could not be used for a substantial number of patients, due to lesions being too large to assess; lesions being in areas with tattoos, scarring or hair covering; or lesions being on parts of the body unsuited to assessment with a dermatoscope. This could potentially cause inequalities due to resulting differences in diagnostic pathways and access to diagnostic services.

Use of AI could improve access to skin cancer diagnosis as it may reduce the need for face-to-face appointments, so reducing patient time commitment and need to travel to appointments.

8 CONCLUSIONS

8.1 Implications for service provision

The high diagnostic accuracy of DERM suggests that it has potential for use as a triage and diagnostic tool for skin cancer in a post-referral setting. This could be either as part of a teledermatology pathway alongside assessment by dermatologists, or as an autonomous diagnostic tool where it replaces some of the need for consultant-led teledermatology.

Although evidence on the clinical impact of DERM was limited, it did suggest that, in eligible lesions, autonomous use of DERM could reduce the need for human dermatology assessment, without substantially adversely affecting accuracy. The practical impact and clinical benefit of using DERM in combination with dermatologist assessment is currently unclear, particularly when compared to teledermatology without using DERM. Current economic evidence to support the cost-effectiveness of DERM is also limited and it is unclear whether the plausible advantages of DERM represent value for money. On the basis of early modelling exercises, there is a reasonably high certainty that DERM has the potential to be used cost-effectively in the post-referral setting, compared to the traditional urgent skin cancer referral pathway. It is less clear whether DERM has potential to be cost-effective compared to teledermatology without DERM. DERM with a second read is less likely to generate cost-savings versus conventional teledermatology, but may have non-cash releasing benefits (e.g. reduced waiting times, quality of care improvements) associated with outsourcing of consultant review to Skin Analytics.

The EAG considers that the evidence on Moleanalyzer Pro is too limited to judge how it might be used in practice. Currently, prospective studies in clinical practice have only assessed its accuracy in diagnosing melanoma. It is unclear whether it could be adapted to detect all forms of skin cancer, or if not, how a melanoma-only AI tool would be used in practice. As Moleanalyzer Pro has not been tested in the UK as part of a teledermatology programme, it is currently unclear what clinical benefits it could have within NHS practice. There is, similarly, no economic evidence to support the use of Moleanalyzer Pro in an NHS setting. Assuming a similar use case to DERM and appropriate data collection, the value of Moleanalyzer Pro could be assessed using the conceptual model outlined by the EAG.

The substantial resistance from both patients and clinicians to using AI without any human dermatological assessment means that if AI is to be used autonomously in some way, more robust evidence that is applicable to current practice is needed to demonstrate that it has clear benefits to patients, without sacrificing accuracy.

8.2 Suggested research priorities

8.2.1 Diagnostic accuracy

There is a need for further research on the diagnostic accuracy of AI compared to standard teledermatology in specific patient subgroups:

- In individuals with darker skin types (Fitzpatrick IV-VI) and a broad range of ethnicity groups;
- For lesion types and lesions located in body sites and not currently covered by DERM and Moleanalyzer Pro evidence;
- To identify rare skin cancers.

Future diagnostic studies should, where possible, examine and compare the diagnostic accuracy of:

- AI as a standalone device;
- AI in combination with human teledermatology (e.g. with a "second read" for all AI assessed lesions;
- Teledermatology without AI;
- Face-to-face assessment without teledermatology.

Studies should use adequate blinding between AI and dermatologists, and use an appropriate and robust reference standard. Particularly, an independent and blinded "ground truth" diagnosis from dermatologists not involved in the teledermatology process, and with appropriate follow-up, is needed for all lesions that are not assessed with histology. Future studies should use up-to-date dermatoscopes to address the limited applicability of existing studies.

Future studies should also follow relevant reporting guidance.[81] This includes clarity on the pathway and positioning of AI within it, clear documentation of reasons for test failures and exclusions (including eligibility assessment and exclusions from analysis), diagnostic accuracy cut-offs (and timing at which these are specified) and reference standard definitions. Versions of AI devices (including algorithms versions, whether used offline or online) and dermatoscopes where applicable should be reported clearly to inform applicability to practice. Diagnostic accuracy should preferably be reported with sufficient granularity (such as with detailed matrices) so as to evaluate sensitivity and specificity by type of cancer. For patients with multiple lesions, studies should specify whether and how the risk of within-patient correlation was addressed.

All DERM and most Moleanalyzer Pro studies were co-authored by staff affiliated to their respective device manufacturer. There is a need for independent evaluations of DERM and Moleanalyzer Pro in

clinical practice, using commonly agreed, standardised interoperable systems and agreed standards of data collection. Evaluations of Moleanalyzer Pro (ideally, in head-to-head designs against DERM) within a UK post-referral context are required to assess whether Moleanalyzer Pro is a suitable, autonomous alternative to DERM, including for the detection of non-melanoma cancer.

8.2.2 Clinical impact

The overall impact of AI requires evaluation, including clinical output (such as referral types and waiting times) throughout the clinical pathway, and longer-term morbidity and mortality outcomes. There is a lack of prospective data available to inform progression of disease in patients with missed diagnosis which is required to appropriate populate an economic model. Larger, independent prospective studies are needed that examine all clinical impact outcomes where evidence is currently absent. These studies should also examine the perceptions of DERM from healthcare professionals and patients, and the impact of DERM on the diagnostic pathway and patient care, to further understand potential barriers to implementation.

Further evaluations of DERM are ongoing across a range of centres in the UK, including the post-two-week-wait referral pathway and in the pre-referral setting. Although few details were reported, it is hoped that this future evidence will address whether DERM can provide clear clinical benefit, perform consistently and be received positively across a range of local services with differing case mix and pathways.

Evidence on the clinical value of Moleanalyzer Pro is required. This should ideally be through prospective observational cohort studies where Moleanalyzer Pro is used within an NHS two-week-wait referral pathway setting, along similar lines to the existing studies of DERM.

8.2.3 Cost-effectiveness and resource use

The use of AI technologies to direct the discharge of patients with benign lesions following referral from primary care has a range of cost and resource consequences which have not been adequately characterised in existing models. Company-sponsored analyses suggested that DERM used autonomously and with a second read could be highly cost-effective compared to current urgent skin cancer referral models. However, much of this value is generated through potentially optimistic assumptions around the diagnostic accuracy of comparators, and of the surrounding pathway. The parameterisation of these analyses is not aligned with NICE precedent, which may overvalue the cost and health implications of DERM.

[redacted] Notably, the magnitude of uncaptured non-cash releasing benefits remains unquantified.

Whilst a conventional cost-utility analytical approach is able to capture important direct cost and health implications of alternative diagnostic strategies, a lack of key comparative data means the relative clinical and cost-effectiveness of pathways incorporating AI technologies and teledermatology remains highly uncertain. Directly comparable evidence on the diagnostic accuracy of AI technologies and teledermatology in a post-referral setting compared with unassisted teledermatology is required to assess the potential value of AI technologies.

A better understanding of the resource implications associated with the implementation of AI technologies will also require further research to establish the costs to the NHS associated with current pathways. Evidence submitted to the EAG demonstrated that the costs of both teledermatology and face-to-face assessments are key value drivers.

Where possible future studies should seek to address these uncertainties by collecting appropriate data on resource implications including impacts on health care professionals' time, set-up and operational costs associated with both teledermatology and AI technologies in trusts without existing infrastructure, as well as the proportion of patients eligible (and the effect of characteristics determining ineligibility) for AI/teledermatology assessment. Further research must also be undertaken to quantify the benefits to population health within skin cancer and other dermatology indications associated with any release of NHS consultant dermatologist resource, and understand the effects of these technologies on waiting times for final diagnosis.

9 REFERENCES

1. *The two-week wait skin cancer pathway: innovative approaches to support early diagnosis of skin cancer as part of the NHS COVID-19 recovery plan. 4 April 2022 Version 1*: NHS and British Association of Dermatologists; 2022.

2. Levell N. *Dermatology. GIRFT Programme National Specialty Report*: GIRFT and NHS; 2021.

3. *Clinical Knowledge Summary - Melanoma*. National Institute of Health and Care Excllence (NICE); 2022. URL: https://cks.nice.org.uk/topics/melanoma/ (accessed 26th October, 2023).

4. *Clinical Knowledge Summary - Melanoma. How should I assess a lesion?* National Institute of Health and Care Excellence (NICE); 2022. URL: https://cks.nice.org.uk/topics/melanoma/diagnosis/assessment/ (accessed 26th October, 2023).

5. *Non-melanoma skin cancer*. NHS. URL: https://www.nhs.uk/conditions/non-melanoma-skin-cancer/ (accessed 26th October, 2023).

6. *Clinical Knowledge Summary - Skin cancers - recognition and referral*. National Institute of Health and Care Excellence (NICE); 2021. URL: https://cks.nice.org.uk/topics/skin-cancers-recognition-referral/ (accessed 26th October, 2023).

7. *Faster diagnosis*. NHS England. URL: https://www.england.nhs.uk/cancer/faster-diagnosis/ (accessed 26th October, 2023).

8. National Institute for Health and Care Excellence. *Suspected cancer: recognition and referral (NG12)*. London: NICE; 2023.

9. Bhave P, Wu Y, Fernandez-Peñas P. *Melanoma*. BMJ Best Practice; 2023. URL: https://bestpractice.bmj.com/topics/en-gb/268 (accessed 26th October, 2023).

10. Marsden JR, Newton-Bishop JA, Burrows L, Cook M, Corrie PG, Cox NH*, et al.* Revised UK guidelines for the management of cutaneous melanoma 2010. *J Plast Reconstr Aesthet Surg* 2010;**63**:1401-19. https://doi.org/https://doi.org/10.1016/j.bjps.2010.07.006

11. National Institute of Health and Care Excellence. *Improving outcomes for people with skin tumours including melanoma. Cancer service guideline [CSG8]*. NICE; 2010. URL: https://www.nice.org.uk/guidance/csg8 (accessed 26th October, 2023).

12. Keohane SG, Botting J, Budny PG, Dolan OM, Fife K, Harwood CA*, et al.* British Association of Dermatologists guidelines for the management of people with cutaneous squamous cell carcinoma 2020. *Br J Dermatol* 2021;**184**:401-14. https://doi.org/10.1111/bjd.19621

13. National Institute of Health and Care Excellence. *Melanoma: assessment and management. NICE guideline [NG14]*. NICE; 2022. URL: https://www.nice.org.uk/guidance/ng14/chapter/Recommendations#assessing-melanoma (accessed 26th October, 2023).

14. Centre for Reviews and Dissemination. *Systematic Reviews: CRD's guidance for undertaking reviews in health care.* York: CRD, University of York; 2009.

15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD*, et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev* 2021;**10**:89. https://doi.org/10.1186/s13643-021-01626-4

16. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. *J Clin Epidemiol* 2016;**75**:40-6. https://doi.org/10.1016/j.jclinepi.2016.01.021

17. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB*, et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529-36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

18. Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, *et al.* QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. *Ann Intern Med* 2021;**174**:1592-9. https://doi.org/10.7326/m21-2234

19. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, *et al.* Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* 2020;**368**:l6890. https://doi.org/10.1136/bmj.l6890

20. Marsden H, Morgan C, Austin S, DeGiovanni C, Venzi M, Kemos P, *et al.* Effectiveness of an image analyzing AI-based digital health technology to identify non-melanoma skin cancer and other skin lesions: results of the DERM-003 study. *Front Med (Lausanne)* 2023;**10**:1288521. https://doi.org/https://dx.doi.org/10.3389/fmed.2023.1288521

21. Marsden H, Kemos C, Venzi M, Noy M, Maheswaran S, Francis N, *et al.* Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service (unpublished). In: Frontiers in Medicine; 2023: 36.

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

23. *An evaluation of AI powered teledermatology for skin cancer 2WW pathway*: East Midlands Academic Health Science Network; 2023.

24. MacLellan AN, Price EL, Publicover-Brouwer P, Matheson K, Ly TY, Pasternak S, *et al.* The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *J Am Acad Dermatol* 2021;**85**:353-9. https://doi.org/https://dx.doi.org/10.1016/j.jaad.2020.04.019

25. Winkler JK, Blum A, Kommoss K, Enk A, Toberer F, Rosenberger A, Haenssle HA. Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: human with machine. *JAMA Dermatol* 2023;**159**:621-7. https://doi.org/https://dx.doi.org/10.1001/jamadermatol.2023.0905

26. Fink C, Blum A, Buhl T, Mitteldorf C, Hofmann-Wellenhof R, Deinlein T, *et al.* Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *J Eur Acad Dermatol Venereol* 2020;**34**:1355-61. https://doi.org/https://dx.doi.org/10.1111/jdv.16165

27. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, *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. https://doi.org/https://dx.doi.org/10.1093/annonc/mdy166

28. Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, *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. https://doi.org/https://dx.doi.org/10.1016/j.annonc.2019.10.013

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

30. Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Buhl T, *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. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2020.04.043

31. Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Kommoss FKF, *et al.* Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification. *J Dtsch Dermatol Ges* 2021;**19**:842-50. https://doi.org/https://dx.doi.org/10.1111/ddg.14384

32. Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Buhl T, *et al.* Does sex matter? Analysis of sex-related differences in the diagnostic performance of a market-approved convolutional neural network for skin cancer detection. *Eur J Cancer* 2022;**164**:88-94. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2021.12.034

33. Winkler JK, Sies K, Fink C, Toberer F, Enk A, Deinlein T, *et al.* Melanoma recognition by a deep learning convolutional neural network - performance in different melanoma subtypes and localisations. *Eur J Cancer* 2020;**127**:21-9. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2019.11.020

34. Winkler JK, Sies K, Fink C, Toberer F, Enk A, Abassi MS, *et al.* Collective human intelligence outperforms artificial intelligence in a skin lesion classification task. *J Dtsch Dermatol Ges* 2021;**19**:1178-84. https://doi.org/https://dx.doi.org/10.1111/ddg.14510

35. Winkler JK, Sies K, Fink C, Toberer F, Enk A, Abassi MS, *et al.* Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *Eur J Cancer* 2021;**145**:146-54. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2020.12.010

36. Winkler JK, Tschandl P, Toberer F, Sies K, Fink C, Enk A, *et al.* Monitoring patients at risk for melanoma: may convolutional neural networks replace the strategy of sequential digital dermoscopy? *Eur J Cancer* 2022;**160**:180-8. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2021.10.030

37. Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, *et al.* Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Network Open* 2019;**2**:e1913436. https://doi.org/https://dx.doi.org/10.1001/jamanetworkopen.2019.13436

38. Phillips M, Greenhalgh J, Marsden H, Palamaras I. Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatol Pract Concept* 2020;**10**:e2020011. https://doi.org/https://dx.doi.org/10.5826/dpc.1001a11

39. Baker KA, Roberts E, Harman K, Mullarkey D, Kalsi D. Using artificial intelligence to triage skin cancer referrals: outcomes from a pilot study. *Br J Dermatol* 2023;**188**:iv173. https://doi.org/https://dx.doi.org/10.1093/bjd/ljad113.372

40. Marsden H, Palamaras I, Kemos P, Greenhalgh J. Effectiveness of an image-analysing artificial intelligence-based digital health technology to diagnose nonmelanoma skin cancer and benign skin lesions. *Br J Dermatol* 2023;**188**:iv43. https://doi.org/https://dx.doi.org/10.1093/bjd/ljad113.091

41. Austin S, Palamaras I, Greenhalgh J, Kemos P, Marsden H. *Effectiveness of an image analysing AI-based digital health technology to diagnose non-melanoma skin cancer (NMSC) and benign skin lesions: results of the DERM-003 study (unpublished)*.

42. Kawsar A, Hussain A, Kalsi D, Kemos P, Marsden H, Thomas L. Patient perspectives of artificial intelligence as a medical device in a skin cancer pathway *Front Med (Lausanne)* 2023;**10**:1259595. https://doi.org/10.3389/fmed.2023.1259595

43. DERM. *DERM Clinical Performance Report Q3 2023 (unpublished)*: DERM; 2023.

44. Skin Analytics. *Skin Analytics case study: Chelsea & Westminster Hospital NHS Foundation Hospital [unpublished]*. London: Skin Analytics; 2023.

45. Andrew K, Price T, Barlow N, Morris E, Ungureanu S, Zaki I. Continued improvement of artificial intelligence in identifying skin cancer. Paper presented at: EADV Congress, 11-14 October 2023; Berlin.

46. Skin Analytics. *Skin Analytics case study: University Hospitals Birmingham NHS Foundation Trust [unpublished]*. London: Skin Analytics; 2023.

47. Jenkins R, Brewer F, Kalsi D, Mullarkey D. *Clinical performance of an artificial intelligence as a medical device deployed within an urgent suspected skin cancer pathway [unpublished]*. London: Skin Analytics; undated.

48. *Case study: artificial intelligence helping to speed up skin cancer diagnosis in Leicester, Leicestershire, and Rutland integrated care system [unpublished]*: NHS England; 2023.

49. Baker KA, Roberts E, Kalsi D, Mullarkey D, Harman K. *Using artificial intelligence to triage skin cancer referrals: outcomes from a pilot study [unpublished]*; undated.

50. Skin Analytics Limited. *DERM US and EU validation study*. ClinicalTrials.gov. Bethesda (MD): National Library of Medicine (US). URL: https://classic.clinicaltrials.gov/show/NCT05126173 (accessed 1st November, 2023).

51. Skin Analytics. *A study to assess whether an Artificial Intelligence (AI) technology can be used to assess images of suspicious skin lesions to determine the likelihood of skin cancer; and reduce the number, and cost assocated with, biopsies of skin lesions in primary care*. Australia New Zealand Clinical Trials Registry; 2019. URL: https://anzctr.org.au/ACTRN12619000398101.aspx (accessed 1st November, 2023).

52. Haenssle HA, Winkler JK, Fink C, Toberer F, Enk A, Stolz W, *et al.* Skin lesions of face and scalp - classification by a market-approved convolutional neural network in comparison with 64 dermatologists. *Eur J Cancer* 2021;**144**:192-9. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2020.11.034

53. Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Kommoss FKF, *et al.* [Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification]. *J Dtsch Dermatol Ges* 2021;**19**:842-51. https://doi.org/https://dx.doi.org/10.1111/ddg.14384_g

54. Tschandl P, Hofmann L, Fink C, Kittler H, Haenssle HA. Melanomas vs. nevi in high-risk patients under long-term monitoring with digital dermatoscopy: do melanomas and nevi already differ at baseline? *J Eur Acad Dermatol Venereol* 2017;**31**:972-7. https://doi.org/10.1111/jdv.14065

55. Chuchu N, Dinnes J, Takwoingi Y, Matin RN, Bayliss SE, Davenport C, *et al.* Teledermatology for diagnosing skin cancer in adults. *Cochrane Database Syst Rev* 2018;**12**:CD013193. https://doi.org/10.1002/14651858.CD013193

56. Wilson EC, Emery JD, Kinmonth AL, Prevost AT, Morris HC, Humphrys E, *et al.* The cost-effectiveness of a novel SIAscopic diagnostic aid for the management of pigmented skin lesions in primary care: a decision-analytic model. *Value Health* 2013;**16**:356-66. https://doi.org/https://dx.doi.org/10.1016/j.jval.2012.12.008

57. Wilson ECF, Usher-Smith JA, Emery J, Corrie P, Walter FM. A modeling study of the cost-effectiveness of a risk-stratified surveillance program for melanoma in the United Kingdom. *Value Health* 2018;**21**:658-68. https://doi.org/https://dx.doi.org/10.1016/j.jval.2017.11.009

58. Edwards SJ, Mavranezouli I, Osei-Assibey G, Marceniuk G, Wakefield V, Karner C. VivaScope R 1500 and 3000 systems for detecting and monitoring skin lesions: a systematic review and economic evaluation. *Health Technol Assess* 2016;**20**:1-260. https://doi.org/https://dx.doi.org/10.3310/hta20580

59. Balch CM, Gershenwald JE, Soong SJ, Thompson JF, Atkins MB, Byrd DR, *et al.* Final version of 2009 AJCC melanoma staging and classification. *J Clin Oncol* 2009;**27**:6199-206. https://doi.org/10.1200/jco.2009.23.4799

60. Losina E, Walensky RP, Geller A, Beddingfield FC, 3rd, Wolf LL, Gilchrest BA, Freedberg KA. Visual screening for malignant melanoma: a cost-effectiveness analysis. *Arch Dermatol* 2007;**143**:21-8. https://doi.org/10.1001/archderm.143.1.21

61. Hadley JC, Bendeck S, Brown K, Chen S. Quality of life impact by melanoma as measured by utilities. *J Am Acad Dermatol* 2004;**50**:P119. https://doi.org/10.1016/j.jaad.2003.10.402

62. Alarcon I, Carrera C, Palou J, Alos L, Malvehy J, Puig S. Impact of in vivo reflectance confocal microscopy on the number needed to treat melanoma in doubtful lesions. *Br J Dermatol* 2014;**170**:802-8. https://doi.org/10.1111/bjd.12678

63. Pellacani G, Pepe P, Casari A, Longo C. Reflectance confocal microscopy as a second-level examination in skin oncology improves diagnostic accuracy and saves unnecessary excisions: a longitudinal prospective study. *Br J Dermatol* 2014;**171**:1044-51. https://doi.org/10.1111/bjd.13148

64. Wilson ECF, Usher-Smith JA, Emery J, Corrie PG, Walter FM. Expert elicitation of multinomial probabilities for decision-analytic modeling: an application to rates of disease progression in undiagnosed and untreated melanoma. *Value Health* 2018;**21**:669-76. https://doi.org/https://doi.org/10.1016/j.jval.2017.10.009

65. Veena P, Clare G, Prashanth P, Timothy D, Sajeda H, Virginia L*, et al.* Reducing vitamin D requests in a primary care cohort: a quality improvement study. *BJGP Open* 2020;**4**:bjgpopen20X101090. https://doi.org/10.3399/bjgpopen20X101090

66. Filby A, Lewis L, Taylor M. *An economic evaluation of interventions to improve the uptake of vitamin D supplements in England and Wales*: National Institute for Health and Care Excellence; 2014.

67. Ethna M, Sarah D, Mark L. *Skin in the game: the cost consequences of skin cancer diagnosis, treatment and care in Northern Ireland.* https://doi.org/10.1101/2023.05.09.23289720

68. Tromme I, Devleesschauwer B, Beutels P, Richez P, Leroy A, Baurain JF*, et al.* Health-related quality of life in patients with melanoma expressed as utilities and disability weights. *Br J Dermatol* 2014;**171**:1443-50. https://doi.org/https://doi.org/10.1111/bjd.13262

69. Sullivan PW, Slejko JF, Sculpher MJ, Ghushchyan V. Catalogue of EQ-5D Scores for the United Kingdom. *Med Decis Making* 2011;**31**:800-4. https://doi.org/10.1177/0272989X11401031

70. McDaniel B, Badri T, Steele RB. Basal cell carcinoma. In: *StatPearls*Treasure Island (FL): StatPearls Publishing; 2023.

71. Thomson J, Hogan S, Leonardi-Bee J, Williams HC, Bath-Hextall FJ. Interventions for basal cell carcinoma of the skin. *Cochrane Database of Systematic Reviews* 2020;**11**:CD003412. https://doi.org/10.1002/14651858.CD003412.pub3

72. *Basal cell carcinoma - an overview*. The Primary Care Dermatology Society; 2023. URL: https://www.pcds.org.uk/clinical-guidance/basal-cell-carcinoma-an-overview (accessed 20th December, 2023).

73. *Basal cell carcinoma PIL*. British Association of Dermatologists; 2022. URL: https://www.bad.org.uk/pils/basal-cell-carcinoma/ (accessed 20th December, 2023).

74. Fijałkowska M, Bonczar M, Jastrzębski I, Ostrowski P, Antoszewski B, Koziej M. Growth rate of basal cell carcinoma: a meta-analysis and systematic review. *Postepy Dermatol Alergol* 2023;**40**:220-4. https://doi.org/10.5114/ada.2023.124795

75. National Institute for Health and Care Excellence. *Nivolumab for treating advanced (unresectable or metastatic) melanoma [TA384]*. NICE; 2016. URL: https://www.nice.org.uk/guidance/ta384/ (accessed 20th December, 2023).

76. Hernández Alava M, Pudney S, Wailoo A. Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from a UK population study. *Pharmacoeconomics* 2023;**41**:199-207. https://doi.org/10.1007/s40273-022-01218-7

77. Hernández Alava M, Pudney S, Wailoo A. *Estimating EQ-5D by age and sex for the UK*. London: NICE: Decision Support Unit; 2022.

78. Jones KC, Weatherly H, Birch S, Castelli A, Chalkley M, Dargan A*, et al. Unit costs of health and social care 2022 manual*. Kent: Personal Social Services Research Unit (University of Kent) & Centre for Health Economics (University of York); 2023.

79. NHS Prescription Services. *December 2023 drug tariff*. NHS Business Services Authority; 2023. URL: https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/drug-tariff (accessed 20th December, 2023).

80. National Institute for Health and Care Excellence. *Olaparib for maintenance treatment of BRCA mutation-positive advanced ovarian, fallopian tube or peritoneal cancer after response to first-line platinum-based chemotherapy (review of TA598) [ID6191]*. URL: https://www.nice.org.uk/guidance/indevelopment/gid-ta11315 (accessed).

81. Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M*, et al.* Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR Derm Consensus Guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group. *JAMA Dermatol* 2022;**158**:90-6. https://doi.org/https://dx.doi.org/10.1001/jamadermatol.2021.4915

APPENDICES

Appendix 1: Literature search strategies

Named technology searches

MEDLINE(R) ALL
via Ovid http://ovidsp.ovid.com/
1946 to October 26, 2023
Date searched: 27th October 2023
Records retrieved: 65

1 "Deep Ensemble for the Recognition of Malignancy".af. (1)
2 (DERM and (Algorithm$ or Artificial Intelligen$ or AI)).tw. (12)
3 "Melanoma Image Analysis Algorithm".af. (0)
4 (Skin Analytics$ or SkinAnalytics$).af. (6)
5 (Moleanalyzer$ or Mole analyzer$ or Moleanalyser$ or Mole analyser$ or FotoFinder$).af. (63)
6 1 or 2 or 3 or 4 or 5 (79)
7 exp animals/ not humans.sh. (5163374)
8 6 not 7 (77)
9 limit 8 to yr="2015 -Current" (65)

Embase
via Ovid http://ovidsp.ovid.com/
1974 to 2023 October 26
Date searched: 27th October 2023
Records retrieved: 398 (NB - date limit 2015 onwards applied in EndNote)

1 "Deep Ensemble for the Recognition of Malignancy".af. (1)
2 DERM.dv. (114)
3 (DERM and (Algorithm$ or Artificial Intelligen$ or AI)).mp. (22)
4 "Melanoma Image Analysis Algorithm".af. (0)
5 (Skin Analytics$ or SkinAnalytics$).af. (8)
6 (Moleanalyzer$ or Mole analyzer$ or Moleanalyser$ or Mole analyser$ or FotoFinder$).af. (273)
7 or/1-6 (415)
8 Nonhuman/ not Human/ (5308649)
9 7 not 8 (398)

Cochrane Central Register of Controlled Trials (CENTRAL)
via Wiley http://onlinelibrary.wiley.com/
Issue 10 of 12, October 2023
Date searched: 27th October 20203
Records retrieved: 19

#1 "Deep Ensemble for the Recognition of Malignancy" 0
#2 DERM 653
#3 (Algorithm* or Artificial Intelligen* or AI) 29320
#4 #2 and #321
#5 "Melanoma Image Analysis Algorithm" 0
#6 (Skin next Analytics* or SkinAnalytics*) 1
#7 (Moleanalyzer* or Mole next analyzer* or Moleanalyser* or Mole next analyser* or FotoFinder*) 20
#8 #1 or #4 or #5 or #6 or #7 with Publication Year from 2015 to 2023, in Trials 19

ACM Digital Library
https://dl.acm.org/
Date searched: 27th October 2023
Records retrieved: 128 records

Search of the The ACM Guide to Computing Literature

1. 35 Results for: [All: "deep ensemble for the recognition of malignancy"] OR [All: "melanoma image analysis algorithm"] OR [All: "skin analytics"] OR [All: "skin-analytics"] OR [All: "skinanalytics"] OR [All: moleanalyzer*] OR [All: "mole-

analyzer"] OR [All: "mole analyzer"] OR [All: moleanalyser*] OR [All: "mole analyser"] OR [All: "mole-analyser"] OR [All: fotofinder*] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]

2. 93 Results for: [All: derm] AND [[All: algorithm* or] OR [All: "artificial intelligence"] OR [All: or ai]] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]

ClinicalTrials.gov
https://classic.clinicaltrials.gov/ct2/home
Date searched: 27th October 2023
Records retrieved: 50

Basic search screen used unless otherwise stated with terms entered into the "other terms" search box.

1. 3 Studies found for: "Deep Ensemble for the Recognition of Malignancy"
2. 1 Study found for: "Melanoma Image Analysis Algorithm" OR "Melanoma Image Analysis Algorithms"
3. 21 Studies found for: "DERM" AND (Algorithm OR algorithms OR "Artificial Intelligence" OR AI)
4. 4 Studies found for: "Skin Analytics" OR "Skin-Analytics" OR "SkinAnalytics"
5. 4 Studies found for: "Skin Analytics" OR "Skin-Analytics" OR "SkinAnalytics" in sponsor field in advanced search screen
6. 2 Studies found for: Moleanalyzer OR "Mole-analyzer" OR "Mole analyzer" OR Moleanalyser OR "Mole analyser" OR "Mole-analyser"
7. 14 Studies found for: FotoFinder
8. 1 Study found for: FotoFinder in sponsor field in advanced search screen

WHO International Clinical Trials Registry Platform (WHO ICTRP)
https://trialsearch.who.int/
Date searched: 27th October 2023
Records retrieved: 37

1. Basic search screen:
11 records for 11 trials found for: "Deep Ensemble for the Recognition of Malignancy" OR "Melanoma Image Analysis Algorithm" OR "Melanoma Image Analysis Algorithm" OR "Skin Analytics" OR "Skin-Analytics" OR "SkinAnalytics" OR Moleanalyzer* OR "Mole-analyzer" OR "Mole analyzer" OR Moleanalyser* OR "Mole analyser" OR "Mole-analyser" OR FotoFinder*

2. Basic search screen:
4 records for 4 trials found for: "DERM" AND (Algorithm OR algorithms OR "Artificial Intelligence" OR AI)

3. Advanced search screen, recruitment status set to all:
2 records for 2 trials found: Tile field - "Deep Ensemble for the Recognition of Malignancy" OR "Melanoma Image Analysis Algorithm" OR "Melanoma Image Analysis Algorithm" OR "Skin Analytics" OR "Skin-Analytics" OR "SkinAnalytics" OR Moleanalyzer* OR "Mole-analyzer" OR "Mole analyzer" OR Moleanalyser* OR "Mole analyser" OR "Mole-analyser" OR FotoFinder*

4. Advanced search screen, recruitment status set to all:
10 records for 10 trials found: Intervention field - "Deep Ensemble for the Recognition of Malignancy" OR "Melanoma Image Analysis Algorithm" OR "Melanoma Image Analysis Algorithm" OR "Skin Analytics" OR "Skin-Analytics" OR "SkinAnalytics" OR Moleanalyzer* OR "Mole-analyzer" OR "Mole analyzer" OR Moleanalyser* OR "Mole analyser" OR "Mole-analyser" OR FotoFinder*

5. Advanced search screen, recruitment status set to all:
5 records for 5 trials found: Primary Sponsor field - "Skin Analytics" OR "Skin-Analytics" OR SkinAnalytics* OR FotoFinder*

6. Advanced search screen, recruitment status set to all:
4 records for 4 trials found : Title field -"DERM" AND (Algorithm OR algorithms OR "Artificial Intelligence" OR AI)

7. Advanced search screen, recruitment status set to all:
1 records for 1 trials found: Intervention field - "DERM" AND (Algorithm OR algorithms OR "Artificial Intelligence" OR AI)

AI and dermoscopy search strategies

MEDLINE(R) ALL
via Ovid http://ovidsp.ovid.com/
1946 to October 30, 2023
Date searched: 31st October 2023
Records retrieved: 676

1 exp Skin Neoplasms/ (144404)
2 melanoma/ or hutchinson's melanotic freckle/ or melanoma, amelanotic/ (99075)
3 exp Carcinoma, Basal Cell/ (19781)
4 Carcinoma, Squamous Cell/ (141659)
5 Bowen's Disease/ (2003)
6 Carcinoma, Merkel Cell/ (3172)
7 Carcinoma, Neuroendocrine/ (5888)
8 exp Nevus/ (17238)
9 (skin adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (51808)
10 (cutaneous adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (18524)
11 melanoma$.ti,ab. (138894)
12 (nonmelanoma$ or non-melanoma$ or NMSC).ti,ab. (7225)
13 (melanocyt$ or non-melanocyt$ or nonmelanocyt$ or keratinocyt$).ti,ab. (75689)
14 ((melanotic or malignan$ or Hutchinson$) adj2 freckle$).ti,ab. (66)
15 (lentigo$ adj2 maligna$).ti,ab. (1363)
16 ((basal adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)) or basalioma$ or BCC).ti,ab. (30190)
17 ((squamous cell adj2 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)) or SCC or cSCC).ti,ab. (137121)
18 (Bowen$ adj3 (disease$ or lesion$ or cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (2466)
19 (Merkel cell adj2 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (4168)
20 ((intra-epiderm$ or intraepiderm$ or intra-derm$ or intraderm$) adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (864)
21 ((neuroendocrine or neuro-endocrine) adj2 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (28719)
22 ((skin or cutaneous or pigmented or nonpigmented) adj3 (lesion$ or nodul$ or macule$)).ti,ab. (59252)
23 (mole$1 or nevus or nevi or naevus or naevi).ti,ab. (43265)
24 ((acitinic or solar or senile) adj2 kerato$).ti,ab. (535)
25 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 (617501)
26 Artificial Intelligence/ (40908)
27 algorithms/ (306306)
28 exp Machine Learning/ (61112)
29 exp neural networks, computer/ (61671)
30 ((artificial$ or machine$ or computational$) adj2 intelligen$).ti,ab. (35799)
31 computer vision.ti,ab. (7427)
32 (AI or AIDHT or AIaMD).ti,ab. (47517)
33 (augment$ adj2 intelligen$).ti,ab. (209)
34 algorithm$.ti,ab. (366699)
35 deep learning.ti,ab. (46831)
36 machine learning.ti,ab. (85994)
37 ((supervised or unsupervised or semi-supervised) adj2 learning).ti,ab. (11807)
38 ((neural or convolutional) adj2 network$).ti,ab. (100150)
39 (CNN or CNNs or DCNN or DCNNs).ti,ab. (18024)
40 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 (727630)
41 25 and 40 (8447)
42 Dermoscopy/ (5910)
43 (dermoscop$ or dermascop$ or dermatoscop$).ti,ab. (7658)
44 ((skin or cutaneous or epidermis) adj3 (microscopy or microscopies)).ti,ab. (1062)
45 (epiluminescen$ adj3 (microscopy or microscopies)).ti,ab. (229)
46 (teledermoscop$ or teledermascop$ or teledermatoscop$).ti,ab. (150)
47 (videodermoscop$ or videodermascop$ or videodermatoscop$).ti,ab. (188)
48 (Dermlite Handyscope$ or "Medicam 1000").ti,ab. (4)
49 (teledermatolog$ or tele-dermatolog$).ti,ab. (1283)
50 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 (11821)

51 41 and 50 (987)
52 exp animals/ not humans.sh. (5164446)
53 51 not 52 (984)
54 limit 53 to yr="2015 -Current" (676)

Embase
via Ovid http://ovidsp.ovid.com/
1974 to 2023 October 30
Date searched: 31st October 2023
Records retrieved: 1035

1 exp skin tumor/ (242335)
2 exp "nevi and melanomas"/ (217266)
3 (skin adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (71175)
4 (cutaneous adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (24674)
5 melanoma$.ti,ab. (195684)
6 (nonmelanoma$ or non-melanoma$ or NMSC).ti,ab. (11666)
7 (melanocyt$ or non-melanocyt$ or nonmelanocyt$ or keratinocyt$).ti,ab. (102721)
8 ((melanotic or malignan$ or Hutchinson$) adj2 freckle$).ti,ab. (73)
9 (lentigo$ adj2 maligna$).ti,ab. (1951)
10 ((basal adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)) or basalioma$ or BCC).ti,ab. (39985)
11 ((squamous cell adj2 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)) or SCC or cSCC).ti,ab. (190640)
12 (Merkel cell adj2 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (6086)
13 (Bowen$ adj2 (disease$ or lesion$ or cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (3026)
14 ((intra-epiderm$ or intraepiderm$ or intra-derm$ or intraderm$) adj3 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (1092)
15 neuroendocrine carcinoma/ (4182)
16 ((neuroendocrine or neuro-endocrine) adj2 (cancer$ or carcinoma$ or tumour$ or tumor$ or neoplas$ or oncolog$ or adenoma$ or adenocarcinoma$ or epithel$ or malignan$ or premalignan$ or precancer$)).ti,ab. (48955)
17 skin defect/ (66760)
18 ((skin or cutaneous or pigmented or nonpigmented) adj3 (lesion$ or nodul$ or macule$)).ti,ab. (87454)
19 (mole$1 or nevus or nevi or naevus or naevi).ti,ab. (50346)
20 ((acitinic or solar or senile) adj2 kerato$).ti,ab. (623)
21 or/1-20 (870026)
22 exp artificial intelligence/ (88413)
23 exp algorithm/ (594577)
24 exp machine learning/ (425179)
25 convolutional neural network/ (26698)
26 ((artificial$ or machine$ or computational$) adj2 intelligen$).ti,ab. (42771)
27 computer vision.ti,ab. (7980)
28 (AI or AIDHT or AIaMD).ti,ab. (63751)
29 (augment$ adj2 intelligen$).ti,ab. (292)
30 algorithm$.ti,ab. (463852)
31 deep learning.ti,ab. (54317)
32 machine learning.ti,ab. (101443)
33 ((supervised or unsupervised or semi-supervised) adj2 learning).ti,ab. (13757)
34 ((neural or convolutional) adj2 network$).ti,ab. (117758)
35 (CNN or CNNs or DCNN or DCNNs).ti,ab. (21381)
36 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 (1085732)
37 21 and 36 (18333)
38 exp epiluminescence microscopy/ (14216)
39 (dermoscop$ or dermascop$ or dermatoscop$).ti,ab,mv,my. (10281)
40 ((skin or cutaneous or epidermis) adj3 (microscopy or microscopies)).ti,ab,mv,my. (1437)
41 (epiluminescen$ adj3 (microscopy or microscopies)).ti,ab,mv,my. (282)
42 (teledermoscop$ or teledermascop$ or teledermatoscop$).ti,ab,mv,my. (203)
43 (videodermoscop$ or videodermascop$ or videodermatoscop$).ti,ab,mv,my. (256)
44 (Dermlite Handyscope$ or "Medicam 1000").ti,ab. (2)
45 teledermatology/ (1803)
46 (teledermatolog$ or tele-dermatolog$).ti,ab. (1887)
47 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 (19385)

48 37 and 47 (1393)
49 limit 48 to yr="2015 -Current" (1035)

Cochrane Central Register of Controlled Trials (CENTRAL)
via Wiley http://onlinelibrary.wiley.com/
Issue 10 of 12, October 2023
Date searched: 31st October 2023
Records retrieved: 10

ID Search Hits
#1 MeSH descriptor: [Skin Neoplasms] explode all trees 2152
#2 MeSH descriptor: [Melanoma] this term only 2742
#3 MeSH descriptor: [Hutchinson's Melanotic Freckle] this term only 14
#4 MeSH descriptor: [Melanoma, Amelanotic] this term only 2
#5 MeSH descriptor: [Carcinoma, Basal Cell] explode all trees 451
#6 MeSH descriptor: [Carcinoma, Squamous Cell] this term only 3422
#7 MeSH descriptor: [Bowen's Disease] this term only 41
#8 MeSH descriptor: [Carcinoma, Merkel Cell] this term only 34
#9 MeSH descriptor: [Carcinoma, Neuroendocrine] this term only 80
#10 MeSH descriptor: [Nevus] explode all trees 104
#11 (skin near/3 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)):ti,ab,kw 4751
#12 (cutaneous near/3 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)):ti,ab,kw 454
#13 melanoma*:ti,ab,kw 6573
#14 (nonmelanoma* or non-melanoma* or NMSC):ti,ab,kw 844
#15 (melanocyt* or non-melanocyt* or nonmelanocyt* or keratinocyt*):ti,ab,kw 1465
#16 ((melanotic or malignan* or Hutchinson*) near/2 freckle*):ti,ab,kw 17
#17 (lentigo* near/2 maligna*):ti,ab,kw 49
#18 ((basal near/3 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)) or basalioma* or BCC):ti,ab,kw 1585
#19 (("squamous cell" near/2 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)) or SCC or cSCC):ti,ab,kw 10125
#20 (Bowen* near/3 (disease* or lesion* or cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)):ti,ab,kw 115
#21 ("Merkel cell" near/2 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)):ti,ab,kw 108
#22 ((intra-epiderm* or intraepiderm* or intra-derm* or intraderm*) near/3 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)):ti,ab,kw 46
#23 ((neuroendocrine or neuro-endocrine) near/2 (cancer* or carcinoma* or tumour* or tumor* or neoplas* or oncolog* or adenoma* or adenocarcinoma* or epithel* or malignan* or premalignan* or precancer*)):ti,ab,kw 1248
#24 ((skin or cutaneous or pigmented or nonpigmented) near/3 (lesion* or nodul* or macule*)):ti,ab,kw 2762
#25 (mole or moles or nevus or nevi or naevus or naevi):ti,ab,kw 684
#26 ((acitinic or solar or senile) near/2 kerato*):ti,ab,kw 38
#27 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24 or #25 or #26 24686
#28 MeSH descriptor: [Artificial Intelligence] this term only 554
#29 MeSH descriptor: [Algorithms] this term only 4515
#30 MeSH descriptor: [Machine Learning] explode all trees 931
#31 MeSH descriptor: [Neural Networks, Computer] explode all trees 540
#32 ((artificial* or machine* or computational*) near/2 intelligen*):ti,ab,kw 1756
#33 "computer vision":ti,ab,kw 140
#34 (AI or AIDHT or AIaMD):ti,ab,kw 5476
#35 (augment* near/2 intelligen*):ti,ab,kw 13
#36 algorithm*:ti,ab,kw 17728
#37 "deep learning":ti,ab,kw 1037
#38 "machine learning":ti,ab,kw 2501
#39 ((supervised or unsupervised or semi-supervised) near/2 learning):ti,ab,kw 207
#40 ((neural or convolutional) near/2 network*):ti,ab,kw 1888
#41 (CNN or CNNs or DCNN or DCNNs):ti,ab,kw 320
#42 #28 or #29 or #30 or #31 or #32 or #33 or #34 or #35 or #36 or #37 or #38 or #39 or #40 or #41 26379
#43 #27 and #42 356
#44 MeSH descriptor: [Dermoscopy] this term only 103
#45 (dermoscop* or dermascop* or dermatoscop*):ti,ab,kw 473
#46 ((skin or cutaneous or epidermis) near/3 (microscopy or microscopies)):ti,ab,kw 78
#47 (epiluminescen* near/3 (microscopy or microscopies)):ti,ab,kw 161

#48	(teledermoscop* or teledermascop* or teledermatoscop*):ti,ab,kw	26
#49	(videodermoscop* or videodermascop* or videodermatoscop*):ti,ab,kw	14
#50	(Dermlite next Handyscope* or "Medicam 1000"):ti,ab,kw	0
#51	(teledermatolog* or tele-dermatolog*):ti,ab,kw	110
#52	#44 or #45 or #46 or #47 or #48 or #49 or #50 or #51	660
#53	#43 and #52 with Publication Year from 2015 to 2023, in Trials	10

ACM Digital Library
https://dl.acm.org/
Date searched: 31st October 2023
Records retrieved: 424 records

Search of the The ACM Guide to Computing Literature using advanced search interface.

1. 20 Results for: [[Title: skin] OR [Title: cutaneous] OR [Title: pigmented] OR [Title: nonpigmented] OR [Title: freckle*] OR [Title: lentigo*] OR [Title: basal] OR [Title: "squamous cell"] OR [Title: bowen*] OR [Title: "merkel cell"] OR [Title: intra-epiderm*] OR [Title: intraepiderm*] OR [Title: intra-derm*] OR [Title: intraderm*] OR [Title: neuroendocrine] OR [Title: neuro-endocrine]] AND [[Title: cancer*] OR [Title: carcinoma*] OR [Title: tumour*] OR [Title: tumor*] OR [Title: neoplas*] OR [Title: oncolog*] OR [Title: adenoma*] OR [Title: adenocarcinoma*] OR [Title: epithel*] OR [Title: maligna*] OR [Title: melanotic] OR [Title: premalignan*] OR [Title: precancer*] OR [Title: lesion*] OR [Title: nodul*] OR [Title: macule*]] AND [[Title: "artificial intelligence"] OR [Title: "machine intelligence"] OR [Title: "computational intelligence"] OR [Title: "computer vision"] OR [Title: ai] OR [Title: ai-dht] OR [Title: aidht] OR [Title: aiamd] OR [Title: "augmented intelligence"] OR [Title: algorithm*] OR [Title: "deep learning"] OR [Title: "machine learning"] OR [Title: "supervised learning"] OR [Title: "unsupervised learning"] OR [Title: "semi-supervised learning"] OR [Title: "neural network"] OR [Title: "neural networks"] OR [Title: "neural networking"] OR [Title: convolutional] OR [Title: cnn] OR [Title: cnns] OR [Title: dcnn] OR [Title: dcnns]] AND [[Title: dermoscop*] OR [Title: dermascop*] OR [Title: dermatoscop*] OR [Title: teledermoscop*] OR [Title: teledermascop*] OR [Title: teledermatoscop*] OR [Title: videodermoscop* or videodermascop* or videodermatoscop*] OR [Title: teledermatolog*] OR [Title: tele-dermatolog*] OR [Title: microscopy] OR [Title: microscopies] OR [Title: epiluminescen*] OR [Title: handyscope*] OR [Title: "medicam 1000"]] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]

2. 218 Results for: [[Abstract: skin] OR [Abstract: cutaneous] OR [Abstract: pigmented] OR [Abstract: nonpigmented] OR [Abstract: freckle*] OR [Abstract: lentigo*] OR [Abstract: basal] OR [Abstract: "squamous cell"] OR [Abstract: bowen*] OR [Abstract: "merkel cell"] OR [Abstract: intra-epiderm*] OR [Abstract: intraepiderm*] OR [Abstract: intra-derm*] OR [Abstract: intraderm*] OR [Abstract: neuroendocrine] OR [Abstract: neuro-endocrine]] AND [[Abstract: cancer*] OR [Abstract: carcinoma*] OR [Abstract: tumour*] OR [Abstract: tumor*] OR [Abstract: neoplas*] OR [Abstract: oncolog*] OR [Abstract: adenoma*] OR [Abstract: adenocarcinoma*] OR [Abstract: epithel*] OR [Abstract: maligna*] OR [Abstract: melanotic] OR [Abstract: premalignan*] OR [Abstract: precancer*] OR [Abstract: lesion*] OR [Abstract: nodul*] OR [Abstract: macule*]] AND [[Abstract: "artificial intelligence"] OR [Abstract: "machine intelligence"] OR [Abstract: "computational intelligence"] OR [Abstract: "computer vision"] OR [Abstract: ai] OR [Abstract: ai-dht] OR [Abstract: aidht] OR [Abstract: aiamd] OR [Abstract: "augmented intelligence"] OR [Abstract: algorithm*] OR [Abstract: "deep learning"] OR [Abstract: "machine learning"] OR [Abstract: "supervised learning"] OR [Abstract: "unsupervised learning"] OR [Abstract: "semi-supervised learning"] OR [Abstract: "neural network"] OR [Abstract: "neural networks"] OR [Abstract: "neural networking"] OR [Abstract: convolutional] OR [Abstract: cnn] OR [Abstract: cnns] OR [Abstract: dcnn] OR [Abstract: dcnns]] AND [[Abstract: dermoscop*] OR [Abstract: dermascop*] OR [Abstract: dermatoscop*] OR [Abstract: teledermoscop*] OR [Abstract: teledermascop*] OR [Abstract: teledermatoscop*] OR [Abstract: videodermoscop*] OR [Abstract: videodermascop*] OR [Abstract: videodermatoscop*] OR [Abstract: teledermatolog*] OR [Abstract: tele-dermatolog*] OR [Abstract: "epiluminescence microscopy"] OR [Abstract: "epiluminescence microscopies"] OR [Abstract: "dermlite handyscope"] OR [Abstract: "medicam 1000"]] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]

3. 11 Results for: [[Title: melanoma*] OR [Title: nonmelanoma*] OR [Title: non-melanoma*] OR [Title: nmsc] OR [Title: melanocyt*] OR [Title: non-melanocyt*] OR [Title: nonmelanocyt*] OR [Title: keratinocyt*] OR [Title: mole] OR [Title: moles] OR [Title: nevus] OR [Title: nevi] OR [Title: naevus] OR [Title: naevi] OR [Title: basalioma*] OR [Title: bcc or scc or cscc] OR [Title: "hutchinson freckle"] OR [Title: "hutchinson's freckle"] OR [Title: "solar keratosis"] OR [Title: "solar keratoses"] OR [Title: "acitinic keratosis"] OR [Title: "acitinic keratoses"] OR [Title: "senile keratosis"] OR [Title: "senile keratoses"]] AND [[Title: "artificial intelligence"] OR [Title: "machine intelligence"] OR [Title: "computational intelligence"] OR [Title: "computer vision"] OR [Title: ai] OR [Title: ai-dht] OR [Title: aidht] OR [Title: aiamd] OR [Title: "augmented intelligence"] OR [Title: algorithm*] OR [Title: "deep learning"] OR [Title: "machine learning"] OR [Title: "supervised learning"] OR [Title: "unsupervised learning"] OR [Title: "semi-supervised learning"] OR [Title: "neural network"] OR [Title: "neural networks"] OR [Title: "neural networking"] OR [Title: convolutional] OR [Title: cnn] OR [Title: cnns] OR [Title: dcnn] OR [Title: dcnns]] AND [[Title: dermoscop*] OR [Title: dermascop*] OR [Title: dermatoscop*] OR [Title: teledermoscop*] OR [Title: teledermascop*] OR [Title: teledermatoscop*] OR [Title: videodermoscop* or videodermascop* or videodermatoscop*] OR [Title: teledermatolog*] OR [Title: tele-dermatolog*] OR [Title: "epiluminescence microscopy"] OR [Title: "epiluminescence microscopies"] OR [Title: "dermlite handyscope"] OR [Title: "medicam 1000"]] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]

4. 175 Results for: [[Abstract: melanoma*] OR [Abstract: nonmelanoma*] OR [Abstract: non-melanoma*] OR [Abstract: nmsc] OR [Abstract: melanocyt*] OR [Abstract: non-melanocyt*] OR [Abstract: nonmelanocyt*] OR [Abstract: keratinocyt*] OR [Abstract: mole] OR [Abstract: moles] OR [Abstract: nevus] OR [Abstract: nevi] OR [Abstract: naevus] OR [Abstract: naevi] OR [Abstract: basalioma*] OR [Abstract: bcc or scc or cscc] OR [Abstract: "hutchinson freckle"] OR [Abstract: "hutchinson's freckle"] OR [Abstract: "solar keratosis"] OR [Abstract: "solar keratoses"] OR [Abstract: "acitinic keratosis"] OR [Abstract: "acitinic keratoses"] OR [Abstract: "senile keratosis"] OR [Abstract: "senile keratoses"]] AND [[Abstract: "artificial intelligence"] OR [Abstract: "machine intelligence"] OR [Abstract: "computational intelligence"] OR [Abstract: "computer vision"] OR [Abstract: ai] OR [Abstract: ai-dht] OR [Abstract: aidht] OR [Abstract: aiamd] OR [Abstract: "augmented intelligence"] OR [Abstract: algorithm*] OR [Abstract: "deep learning"] OR [Abstract: "machine learning"] OR [Abstract: "supervised learning"] OR [Abstract: "unsupervised learning"] OR [Abstract: "semi-supervised learning"] OR [Abstract: "neural network"] OR [Abstract: "neural networks"] OR [Abstract: "neural networking"] OR [Abstract: convolutional] OR [Abstract: cnn] OR [Abstract: cnns] OR [Abstract: dcnn] OR [Abstract: dcnns]] AND [[Abstract: dermoscop*] OR [Abstract: dermascop*] OR [Abstract: dermatoscop*] OR [Abstract: teledermoscop*] OR [Abstract: teledermascop*] OR [Abstract: teledermatoscop*] OR [Abstract: videodermoscop* or videodermascop* or videodermatoscop*] OR [Abstract: teledermatolog*] OR [Abstract: tele-dermatolog*] OR [Abstract: "epiluminescence microscopy"] OR [Abstract: "epiluminescence microscopies"] OR [Abstract: "dermlite handyscope"] OR [Abstract: "medicam 1000"]] AND [E-Publication Date: (01/01/2015 TO 12/31/2023)]

ClinicalTrials.gov
https://classic.clinicaltrials.gov/ct2/
Date searched: 2nd November 2023
Records retrieved: 270

1. 30 Studies found for: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm | "skin cancer" OR "skin neoplasm" OR "skin tumor" OR "skin tumour" OR "skin carcinoma"

2. 7 Studies found for: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs | "skin cancer" OR "skin neoplasm" OR "skin tumor" OR "skin tumour" OR "skin carcinoma"

3. 29 Studies found for: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm | "cutaneous cancer" OR "cutaneous neoplasm" OR "cutaneous tumor" OR "cutaneous tumour" OR "cutaneous carcinoma"

4. 6 Studies found for: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs | "cutaneous cancer" OR "cutaneous neoplasm" OR "cutaneous tumor" OR "cutaneous tumour" OR "cutaneous carcinoma"

5. 55 Studies found for: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm | melanoma OR nonmelanoma OR non-melanoma OR melanocytic OR non-melanocytic OR nonmelanocytic OR keratinocytic OR melanocyte OR non-melanocyte OR nonmelanocyte OR keratinocyte OR melanotic OR "lentigo maligna"

6. 11 Studies found for: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs | melanoma OR nonmelanoma OR non-melanoma OR melanocytic OR non-melanocytic OR nonmelanocytic OR keratinocytic OR melanocyte OR non-melanocyte OR nonmelanocyte OR keratinocyte OR melanotic OR "lentigo maligna"

7. 38 Studies found for: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm | "Basal cell cancer" OR "Basal cell neoplasm" OR "Basal cell tumor" OR "Basal cell tumour" OR "Basal cell carcinoma" OR "Squamous cell cancer" OR "Squamous cell neoplasm" OR "Squamous cell tumor" OR "Squamous cell tumour" OR "Squamous cell carcinoma"

8. 13 Studies found for: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs | "Basal cell cancer" OR "Basal cell neoplasm" OR "Basal cell tumor" OR "Basal cell tumour" OR "Basal cell carcinoma" OR "Squamous cell cancer" OR "Squamous cell neoplasm" OR "Squamous cell tumor" OR "Squamous cell tumour" OR "Squamous cell carcinoma"

9. 23 Studies found for: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm | (skin OR cutaneous OR dermatology OR dermal OR dermis OR epidermal OR epidermis) AND ("Neuroendocrine cancer" OR "Neuroendocrine neoplasm" OR "Neuroendocrine tumor" OR "Neuroendocrine tumour" OR "Neuroendocrine carcinoma")

10. 5 Studies found for: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs | (skin OR cutaneous OR dermatology OR dermal OR dermis OR epidermal OR epidermis) AND ("Neuroendocrine cancer" OR "Neuroendocrine neoplasm" OR "Neuroendocrine tumor" OR "Neuroendocrine tumour" OR "Neuroendocrine carcinoma")

11. 45 Studies found for: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm | Bowen OR Bowens OR Bowen's OR "Merkel cell" OR mole OR moles OR nevus OR nevi OR naevus OR naevi OR "senile keratosis" OR "acitinic keratosis" OR “solar keratosis”

12. 8 Studies found for: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs | Bowen OR Bowens OR Bowen's OR "Merkel cell" OR mole OR moles OR nevus OR nevi OR naevus OR naevi OR "senile keratosis" OR "acitinic keratosis" OR "solar keratosis"

WHO International Clinical Trials Registry Platform (WHO ICTRP)
https://trialsearch.who.int/
Date searched: 2nd November 2023
Records retrieved: 177

1. Advanced search screen, recruitment status set to all:
Condition field: "skin cancer" OR "skin neoplasm" OR "skin tumor" OR "skin tumour" OR "skin carcinoma" Intervention field: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*
8 records for 8 trials found

2. Advanced search screen, recruitment status set to all:
Condition field: "skin cancer" OR "skin neoplasm" OR "skin tumor" OR "skin tumour" OR "skin carcinoma" Intervention field: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs
0 records for 0 trials found

3. Advanced search screen, recruitment status set to all:
Condition field: "cutaneous cancer" OR "cutaneous neoplasm" OR "cutaneous tumor" OR "cutaneous tumour" OR "cutaneous carcinoma" Intervention field: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*
0 records for 0 trials found

4. Advanced search screen, recruitment status set to all:
Condition field: "cutaneous cancer" OR "cutaneous neoplasm" OR "cutaneous tumor" OR "cutaneous tumour" OR "cutaneous carcinoma" Intervention field: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs
0 records for 0 trials found

5. Advanced search screen, recruitment status set to all:
Condition field: melanoma* OR nonmelanoma* OR non-melanoma* OR melanocyt* OR non-melanocyt* OR nonmelanocyt* OR keratinocyt* OR "lentigo maligna" Intervention field: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*
10 records for 10 trials found

6. Advanced search screen, recruitment status set to all:
Condition field: melanoma* OR nonmelanoma* OR non-melanoma* OR melanocyt* OR non-melanocyt* OR nonmelanocyt* OR keratinocyt* OR "lentigo maligna" Intervention field: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs
0 records for 0 trials found

7. Advanced search screen, recruitment status set to all:
Condition field: "Basal cell cancer" OR "Basal cell neoplasm" OR "Basal cell tumor" OR "Basal cell tumour" OR "Basal cell carcinoma" OR "Squamous cell cancer" OR "Squamous cell neoplasm" OR "Squamous cell tumor" OR "Squamous cell tumour" OR "Squamous cell carcinoma" Intervention field: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*
6 records for 6 trials found

8. Advanced search screen, recruitment status set to all:

Condition field: "Basal cell cancer" OR "Basal cell neoplasm" OR "Basal cell tumor" OR "Basal cell tumour" OR "Basal cell carcinoma" OR "Squamous cell cancer" OR "Squamous cell neoplasm" OR "Squamous cell tumor" OR "Squamous cell tumour" OR "Squamous cell carcinoma" Intervention field: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs
0 records for 0 trials found

9. Advanced search screen, recruitment status set to all:
Condition field: (skin OR cutaneous OR dermatology OR dermal OR dermis OR epidermal OR epidermis) AND ("Neuroendocrine cancer" OR "Neuroendocrine neoplasm" OR "Neuroendocrine tumor" OR "Neuroendocrine tumour" OR "Neuroendocrine carcinoma") Intervention field: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*
0 records for 0 trials found

10. Advanced search screen, recruitment status set to all:
Condition field: (skin OR cutaneous OR dermatology OR dermal OR dermis OR epidermal OR epidermis) AND ("Neuroendocrine cancer" OR "Neuroendocrine neoplasm" OR "Neuroendocrine tumor" OR "Neuroendocrine tumour" OR "Neuroendocrine carcinoma") " Intervention field: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs
0 records for 0 trials found

11. Advanced search screen, recruitment status set to all:
Condition field: Bowen OR Bowens OR Bowen's OR "Merkel cell" OR mole OR moles OR nevus OR nevi OR naevus OR naevi OR "senile keratosis" OR "acitinic keratosis" OR “solar keratosis” Intervention field: "artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*
0 records for 0 trials found

12. Advanced search screen, recruitment status set to all:
Condition field: Bowen OR Bowens OR Bowen's OR "Merkel cell" OR mole OR moles OR nevus OR nevi OR naevus OR naevi OR "senile keratosis" OR "acitinic keratosis" OR “solar keratosis” Intervention field: "deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs
0 records for 0 trials found

13. Basic search screen:
17 records for 17 trials found for: ("skin cancer" OR "skin neoplasm" OR "skin tumor" OR "skin tumour" OR "skin carcinoma") AND ("artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*)

14 1 trial found for: ("skin cancer" OR "skin neoplasm" OR "skin tumor" OR "skin tumour" OR "skin carcinoma") AND ("deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs)

15. 44 records for 31 trials found for: (melanoma* OR nonmelanoma* OR non-melanoma* OR melanocyt* OR non-melanocyt* OR nonmelanocyt* OR keratinocyt* OR "lentigo maligna") AND ("artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*)

16. 1 trial found for: (melanoma* OR nonmelanoma* OR non-melanoma* OR melanocyt* OR non-melanocyt* OR nonmelanocyt* OR keratinocyt* OR "lentigo maligna") AND ("deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs)

17. 25 records for 25 trials found for: ("Basal cell cancer" OR "Basal cell neoplasm" OR "Basal cell tumor" OR "Basal cell tumour" OR "Basal cell carcinoma" OR "Squamous cell cancer" OR "Squamous cell neoplasm" OR "Squamous cell tumor" OR "Squamous cell tumour" OR "Squamous cell carcinoma") AND ("artificial intelligence" OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*)

18. 8 records for 8 trials found for: ("Basal cell cancer" OR "Basal cell neoplasm" OR "Basal cell tumor" OR "Basal cell tumour" OR "Basal cell carcinoma" OR "Squamous cell cancer" OR "Squamous cell neoplasm" OR "Squamous cell tumor" OR "Squamous cell tumour" OR "Squamous cell carcinoma") AND ("deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs)

19 9 records for 9 trials found for: (Bowen OR Bowens OR Bowen's OR "Merkel cell" OR mole OR moles OR nevus OR nevi OR naevus OR naevi OR "senile keratosis" OR "acitinic keratosis" OR "solar keratosis") AND ("artificial intelligence"

OR "machine intelligence" OR "computational intelligence" OR "computer vision" OR AI OR "augmented intelligence" OR algorithm*)

20. 1 trial found for: (Bowen OR Bowens OR Bowen's OR "Merkel cell" OR mole OR moles OR nevus OR nevi OR naevus OR naevi OR "senile keratosis" OR "acitinic keratosis" OR "solar keratosis") AND ("deep learning" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "semi-supervised learning" OR "neural network" OR "neural networks" OR convolutional OR CNN OR CNNs OR DCNN OR DCNNs)

Appendix 2: excluded studies at full text screening stage

Exclude population (n=1)

Corbin A, Marques O. Exploring Strategies to Generate Fitzpatrick Skin Type Metadata for Dermoscopic Images Using Individual Typology Angle Techniques. *Multimedia Tools Appl.* 2022;**82**:23771–95

Exclude intervention (n=39)

Abbes W, Sellami D. Deep Neural Networks for Melanoma Detection from Optical Standard Images Using Transfer Learning. *Procedia Comput. Sci.* 2021;**192**:1304–12

Abhishek K, Kawahara J, Hamarneh G. Predicting the clinical management of skin lesions using deep learning. *Scientific Reports* 2021;**11**:7769.

Abouche H, Jimi A, Zrira N, Benmiloud I. Segmentation and Classification of Dermoscopic Skin Cancer on Green Channel. *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 2023:347–54

Arlette J, Wong A, Khodadad I, Kazemzadeh F. Deep Tissue Sequencing Using Augmented Intelligence to Probe Melanocytic Lesions. *Journal of Cutaneous Medicine & Surgery* 2017;**21**:572.

Assuta H, Systems. *Artificial Intelligence-assisted Evaluation of Pigmented Skin Lesions*. NCT03362138

Babino G, Lallas A, Agozzino M, Alfano R, Apalla Z, Brancaccio G, et al. Melanoma diagnosed on digital dermoscopy monitoring: A side-by-side image comparison is needed to improve early detection. Journal of the American Academy of Dermatology 2021;85(3):619-25.

Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer* 2019;**119**:11-7.

Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer* 2019;**111**:148-54.

Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer* 2019;**113**:47-54.

Camacho-Gutierrez Jos A, Solorza-Calderon S, lvarez-Borrego J. Multi-Class Skin Lesion Classification Using Prism- and Segmentation-Based Fractal Signatures. Expert Syst. Appl. 2022;197.

Cerminara SE, Cheng P, Kostner L, Huber S, Kunz M, Maul JT, et al. Diagnostic performance of augmented intelligence with 2D and 3D total body photography and convolutional neural networks in a high-risk population for melanoma under real-world conditions: A new era of skin cancer screening? *European Journal of Cancer* 2023;**190**:112954.

Cinotti E, Haouas M, Grivet D, Perrot JL. In vivo and ex vivo confocal microscopy for the management of a melanoma of the eyelid margin. *Dermatologic Surgery* 2015;**41(12)**:1437-40.

Cinotti E, Santi F, Perrot JL, Habougit C, Tognetti L, Rubegni P. Squamous cell carcinoma arising on acrodermatitis continua of Hallopeau: clinical and noninvasive skin imaging features. *International Journal of Dermatology* 2021;**60(6)**:763-5.

Del Rosario F, Farahi JM, Drendel J, Buntinx-Krieg T, Caravaglio J, Domozych R, et al. Performance of a computer-aided digital dermoscopic image analyzer for melanoma detection in 1,076 pigmented skin lesion biopsies. *Journal of the American Academy of Dermatology* 2018;**78**:927-34.e6.

Eduardo Pr, Reyes s. Performing Melanoma Diagnosis by an Effective Multi-View Convolutional Network Architecture. *Int. J. Comput. Vision* 2023;**131**:3094–117

Eduardo Pr, Ventura Sebasti n. An Ensemble-Based Convolutional Neural Network Model Powered by a Genetic Algorithm for Melanoma Diagnosis. *Neural Comput. Appl.* 2022;**34**:10429–48

Francese R, Frasca M, Risi M, Tortora G. A Mobile Augmented Reality Application for Supporting Real-Time Skin Lesion Analysis Based on Deep Learning. *J. Real-Time Image Process.* 2021;**18**:1247–59

Gu R, Wang L, Zhang L. DE-Net: A Deep Edge Network with Boundary Information for Automatic Skin Lesion Segmentation. *Neurocomput.* 2022;**468**:71–84

Gupta A, Tiwari D, Agarwal S, Jain M. Fuzzy Based Support System for Melanoma Diagnosis. *Proceedings of the Third International Conference on Mining Intelligence and Knowledge Exploration - Volume 9468* 2015:235–46

Haugsten ER, Vestergaard T, Trettin B. Experiences Regarding Use and Implementation of Artificial Intelligence-Supported Follow-Up of Atypical Moles at a Dermatological Outpatient Clinic: Qualitative Study. *MIR Dermatology* 2023;**6**:e44913.

He X, Wang Y, Zhao S, Chen X. Co-Attention Fusion Network for Multimodal Skin Cancer Diagnosis. *Pattern Recogn.* 2023;**133**.

Huang K, Jiang Z, Li Y, Wu Z, Wu X, Zhu W, et al. The Classification of Six Common Skin Diseases Based on Xiangya-Derm: Development of a Chinese Database for Artificial Intelligence. *Journal of Medical Internet Research* 2021;**23**:e26025.

Inthiyaz S, Altahan Baraa R, Ahammad Sk H, Rajesh V, Kalangi Ruth R, Smirani Lassaad K, et al. Skin Disease Detection Using Deep Learning. *Adv. Eng. Softw.* 2023;**175**.

Jahn AS, Navarini AA, Cerminara SE, Kostner L, Huber SM, Kunz M, et al. Over-Detection of Melanoma-Suspect Lesions by a CE-Certified Smartphone App: Performance in Comparison to Dermatologists, 2D and 3D Convolutional Neural Networks in a Prospective Data Set of 1204 Pigmented Skin Lesions Involving Patients' Perception. *Cancers* 2022;**14**:07.

Jamil U, Khalid S, Akram MU, Ahmad A, Jabbar S. Melanocytic and Nevus Lesion Detection from Diseased Dermoscopic Images Using Fuzzy and Wavelet Techniques. *Soft Comput.* 2018;**22**:1577–93

Janda M, Horsham C, Vagenas D, Loescher LJ, Gillespie N, Koh U, et al. Accuracy of mobile digital teledermoscopy for skin self-examinations in adults at high risk of skin cancer: an open-label, randomised controlled trial. *The Lancet Digital Health* 2020;**2(3)**:e129-e37.

Jiji GW, DuraiRaj PJ. Content-Based Image Retrieval Techniques for the Analysis of Dermatological Lesions Using Particle Swarm Optimization Technique. *Appl. Soft Comput.* 2015;**30**:650–62

Karolinska University H, Karolinska I, Medical University of V, Stockholm School of E. *AI-Augmented Skin Cancer Diagnosis in Teledermatoscopy*. NCT06080711

Majumder S, Ullah Muhammad A. A Computational Approach to Pertinent Feature Extraction for Diagnosis of Melanoma Skin Lesion. *Pattern Recognit. Image Anal.* 2019;**29**:503–14

Menzies SW, Sinz C, Menzies M, Lo SN, Yolland W, Lingohr J, et al. Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. *The Lancet. Digital Health* 2023;**5**:e679-e91.

Mery D, Romero P, Garib G, Pedro A, Salinas Maria P, Sepulveda J, et al. On Skin Lesion Recognition Using Deep Learning: 50 Ways To Choose Your Model. *Image and Video Technology: 10th Pacific-Rim Symposium, PSIVT 2022, Virtual Event, November 12–14, 2022, Proceedings* 2023:103–16

Oliveira Roberta B, Marranghello N, Pereira Aledir S, Tavares Jo o Manuel RS. A Computational Approach for Detecting Pigmented Skin Lesions in Macroscopic Images. *Expert Syst. Appl.* 2016;**61**:53–63

Salido Julie Ann A, Ruiz C. Using Morphological Operators and Inpainting for Hair Removal in Dermoscopic Images. *Proceedings of the Computer Graphics International Conference*

Surowka G, Ogorzallek M. Segmentation of the Melanoma Lesion and Its Border. *Int. J. Appl. Math. Comput. Sci.* 2022;**32**:683–99

University Hospital Basel S. *Melanoma Detection in Switzerland With VECTRA*. NCT04605822

University Hospital G. *ARTIficial Intelligence-based Smartphone Application for Skin Cancer Detection*. NCT06080711

Yang Y. Privacy-Preserving With Sonification For Training of Convolutional Deep Neural Networks for Melanoma Diagnosis. *Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing* 2021:106–15

Zhang Y, Xie F, Chen J. TFormer: A throughout Fusion Transformer for Multi-Modal Skin Lesion Diagnosis. *Comput. Biol. Med.* 2023;**157**.

Zhao X-Y, Wu X, Li F-F, Li Y, Huang W-H, Huang K, et al. The Application of Deep Learning in the Risk Grading of Skin Tumors for Patients Using Clinical Images. *J. Med. Syst.* 2019;**43**:1–7

Exclude design (n=13)

Beltrami EJ, Brown AC, Salmon PJM, Leffell DJ, Ko JM, Grant-Kels JM. Artificial intelligence in the detection of skin cancer. *Journal of the American Academy of Dermatology* 2022;**87**:1336-42.

Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M, et al. Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology: CLEAR Derm Consensus Guidelines From the International Skin Imaging Collaboration Artificial Intelligence Working Group. *JAMA Dermatology* 2022;**158**:90-6.

Ferrante di Ruffano L, Takwoingi Y, Dinnes J, Chuchu N, Bayliss SE, Davenport C, et al. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. *Cochrane Database of Systematic Reviews* 2018;**12**:CD013186.

Filho M, Ma Z, Tavares Jo o M. A Review of the Quantification and Classification of Pigmented Skin Lesions: From Dedicated to Hand-Held Devices. *J. Med. Syst.* 2015;**39**:1–12

Groh M, Harris C, Daneshjou R, Badri O, Koochek A. Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm. *Proc. ACM Hum.-Comput. Interact.* 2022;**6**.

Haggenmuller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill RL, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer* 2021;**156**:202-16.

Marchetti MA, Liopyris K, Navarrete-Dechent C. Net benefit and decision curve analysis of competing diagnostic strategies for cutaneous melanoma. *Journal of the American Academy of Dermatology* 2021;**85(2)**:e87-e8.

Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, et al. Attitudes towards Trusting Artificial Intelligence Insights and Factors to Prevent the Passive Adherence of GPs: A Pilot Study. *Journal of Clinical Medicine* 2021;**10**:14.

Skin Analytics (undated) The Evidence behind Skin Analytics [unpublished]. London, Skin Analytics

Skin Analytics (2023) Patient sentiment towards AI use in dermatology pathways [unpublished]. London, Skin Analytics

Sydney Melanoma Diagnostic Centre. *The assessment of experimental artificial intelligence (AI) algorithms for the diagnosis of skin tumours against human performance*. ACTRN12620000695909

Winkler JK, Haenssle HA. [Artificial intelligence-based classification for the diagnostics of skin cancer]. *Dermatologie (Heidelberg, Germany)* 2022;**73**:838-44.

Zhelev et al 2023 Cost-effectiveness of DERM: A modelling study. Report version 1.0 16th November 2023. [unpublished]. London, Skin Analytics

Appendix 3: Ongoing studies

Table 22 DERM registered, ongoing studies

Study	Start-completion	Location	Population	Design	Main outcomes
DERM-006 (NCT05126173)	Mar 2022 – Sept 2022	USA, Italy	Skin biopsy N=1111 (actual)	Prosp. cohort	DA
ACTRN12619000398101	Mar 2022- Dec 2022 (anticipated)	Australia	Primary care (GP) N=750 (target)	Prosp. cohort	DA Biopsy referral QoL

Table 23 Ongoing evaluations of DERM in the UK – post-referral

[redacted]

Table 24 Ongoing evaluations of DERM in the UK – pre-referral

[redacted]

Appendix 4: Data extraction tables

Table 25 Selection criteria of studies included in the synthesis

Study	Reported selection criteria
DERM-003 [20]	Dermatology clinic patients with >=1 suspicious skin lesion suitable for photographing. Include lesions <15 mm diameter, on a site suitable for photographing, not in area of visible scarring or tattooing, and not previously biopsied, excised or otherwise traumatised.
DERM-005 Chelsea & Westminster [21]	Adult patients with at least one suspicious lesion being photographed as part of standard care teledermatology; lesions less than 15 mm, in a location suitable for photography, no previous trauma including biopsy or excision, no visible scarring or tattooing.
UHBFT and WSFT [22]	Adults with 1 to 3 suspicious lesions not larger than 15 mm. Exclusions are lesions that are not potentially malignant, those requiring monitoring for treatment response or staging of disease, non-dermascopic images of lesions, open ulcerated lesions, obscured by hair, tattoos or scars, subungual or on mucosal, genital or palmoplantar surfaces, previously biopsied lesions.
UHL [23]	Exclude <18 years, 2+ lesions, genital lesions. No further details.
MacLellan 2021	Exclude: recurrent lesions or metastases; previously biopsied or excised; lesions less than 2 mm or greater than 2 cm in diameter; lesions not accessible to the devices; lesions located on scars, crusts, psoriasis, eczema, sunburn, or other skin condition; lesions covered by thick hair; inaccessible genital, mucosal, obscured by foreign matter, ulcerated, sole, palm, close to eye; Fitzpatrick skin >III.
Winkler 2023	Melanocytic lesions. No further details.

Table 26 Full diagnostic accuracy results (DERM studies)

Study	Index test	Outcome	Sensitivity*	Specificity*	AUROC*	PPV*	NPV*
DERM-003	DERM [redacted] (iPhone 11)	Melanoma	93.3 (66.0–99.7)	73.6 (69.6–77.1)	92.6 (84.3–100)	8.7 (5.0–14.4)	99.8 (98.4–100)
		SCC	93.2 (80.3–98.2)	45.7 (41.3–50.1)	90.1 (86.1–94.0)	12.8 (9.5–17.1)	98.7 (96–99.7)
		BCC	95.8 (91.7–98)	45 (39.5–50.6)	92.0 (89.7–94.3)	51.1 (45.8–56.4)	94.7 (89.5–97.5)
		Malignant	96.0 (92.6–98)	45 (39.5–50.6)	NR	58 (53.1–62.7)	93.5 (88.1–96.7)
		IEC	100 (67.9–100)	46.6 (41–52.3)	89.0 (84.2–93.8)	6.2 (3.3–11.2)	100 (96.8–100)
		AK	84.8 (72.5–92.4)	47.2 (40.9–53.6)	81.1 (75.0–87.2)	27.5 (21.3–34.7)	92.9 (86.6–96.5)
		Atypical	59.1 (36.7-78.5)	43.9 (37.4–50.6)	89.4 (82.7–96.2)	9.2 (5.2–15.6)	91.7 (84.5–95.9)
		Benign	43.9 (37.4–50.6)	93.3 (90.0–95.6)	80.9 (77.3–84.5)	81.3 (73.1–87.5)	71.4 (67.0–75.5)

	Clinicians	Melanoma	81.2 (53.7–95.0)	98.9 (97.6–99.6)	90.3 (80.4–100)	68.4 (43.5–86.4)	99.5 (98.3–99.9)
		SCC	63.6 (47.7–77.2)	89.1 (86–91.5)	76.9 (69.6–84.3)	32.9 (23.4–44.1)	96.7 (94.5–98.0)
		BCC	97.5 (93.9–99.1)	77.4 (72.4–81.8)	90.0 (87.3–92.7)	72.6 (66.7–77.7)	98 (95.2–99.3)
		Malignant	93.8 (90–96.3)	77.4 (72.4–81.8)	NR	77 (71.9–81.4)	94.3 (90.6–96.7)
		IEC	90.9 (57.1–99.5)	78.8 (73.8–83.2)	63.6 (49.8–77.4)	13.2 (6.8–23.3)	99.6 (97.4–100)
		AK	96.7 (87.6–99.4)	79.3 (73.6–84)	85.0 (79.2–90.8)	53.1 (43.5–62.6)	99 (96.1–99.8)
		Atypical	76.2 (52.5–90.9)	73.9 (67.6-79.4)	85.1 (75.1–95)	21 (12.9–32.2)	97.1 (93.1–98.9)
		Benign	73.9 (67.6–79.4)	93.7 (90.5-95.9)	82.1 (78.8–85.5)	88.5 (83.0–92.5)	84.6 (80.5–87.9)
DERM-005	DERM [redacted]	Malignant	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	Teledermatologist		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	DERM (post-hoc analysis)[#]		[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Thomas (2023)	Derm vA (UHB)	Melanoma	95.0 (90–97.6)	58.80 (57.4–60.2)	NR	6.7 (5.7–7.9)	99.7 (99.5–99.9)
	Derm vA (WSFT)		97.0 (84.7–99.5)	63.20 (59.5–66.7)	NR	11.4 (8.2–15.6)	99.8 (98.7–100)
	Derm vB (UHB)		100.0 (93.8–100)	80.90 (79.3–82.4)	NR	10.7 (8.4–13.6)	100.0 (99.8–100.0)
	Derm vB (WSFT)		100.0 (82.4–100)	80.40 (77.2–83.4)	NR	12.9 (8.3–19.4)	100.0 (99.2–100)
	Derm vA (UHB)	Malignant	96.0 (94.4–97.2)	45.00 (43.4–46.6)	NR	25.3 (23.7–26.9)	98.3 (97.6–98.8)
	Derm vA (WSFT)		99.3 (96.3–99.9)	33.1 (29.3–37.1)	NR	28.5 (24.8–32.5)	99.5 (97–99.9)
	Derm vB (UHB)		98.9 (96–99.7)	64.8 (62.9–66.7)	NR	17.4 (15.2–19.8)	99.9 (99.5–100.0)
	Derm vB (WSFT)		100.0 (94.7–100)	60.6 (56.6–64.5)	NR	23.1 (18.7–28.3)	100.0 (98.9–100.0)

[*] All results expressed as % (95% confidence interval).

[#] Target sensitivity changed to >95% for melanoma and SCC and >90% for BCC

CI = Confidence interval, PPV = Positive predictive value, NPV = Negative predictive value; SCC = Squamous cell carcinoma; BCC = Basal cell carcinoma; IEC = Intraepidermal squamous cell carcinoma; AK = Actinitic keratoses; UHB = University Hospital Birmingham; WSFT = West Suffolk Foundation Trust[redacted]

Table 27 Included studies reporting subtype, Breslow thickness, and stage of melanoma

Type of cancer	Lesion characteristics	DERM-003	DERM-005	MacLellan 2021	Winkler 2023
Subtype of melanoma					
Superficial spreading		9	[redacted]	NR	NR
Lentigo melanoma		1	[redacted]	NR	NR
Other/ not available/ ambiguous		6	[redacted]	NR	NR
Breslow thickness					
in situ		2	[redacted]	27	12
< 1.0 mm		7	[redacted]	NR Mean (SD) 0.72 (0.56) Median (range) 0.57 (0.19-2.9)	NR Invasive: 26
1.01-2.0 mm		2	[redacted]		
> 2.0 mm		4	[redacted]		
> 4 mm		0	[redacted]		
Not available		1	[redacted]	NR	NR
TOTAL		**16**	[redacted]	**59**	**38**

Table 28 Included studies reporting subtype and stage of SCC, BCC, and other malignancies

Lesion characteristics	DERM-003	DERM-005
SCC		
Subtype		
Poorly differentiated	4	[redacted]
Moderately differentiated	15	[redacted]
Well differentiated	16	[redacted]
Other/ unknown	8	[redacted]
Stage		
Tis	1	[redacted]
T1	38	[redacted]
T2	0	[redacted]
T3	NR	[redacted]
T4	3	[redacted]
Not available/ other/ unknown	2	[redacted]
TOTAL	*44*	[redacted]
BCC		
Subtype		
Superficial	13	[redacted]
Nodular	94	[redacted]
Infiltrative	17	[redacted]
Morphoeic	0	[redacted]

Micronodular	2	[redacted]
Basosquamous	1	[redacted]
Not available/ other/ unknown	70	[redacted]
Stage		
Tis	3	[redacted]
T1	141	[redacted]
T2	2	[redacted]
T3	NR	[redacted]
T4	0	[redacted]
Not available/ unknown	51	[redacted]
TOTAL	*197*	[redacted]
Other malignancies		
TOTAL	*2*	[redacted]

Table 29 Diagnostic pathway outcomes for patients in Thomas 2023

		DERM vA		DERM vB	
Total number of cases (patients)		**Birmingham N=7,171**	**West Suffolk N=1,119**	**Birmingham N=4,800**	**West Suffolk N=1,410**
Not assessed with DERM*		27.4%	15.6%	25%	17%
Assessed with DERM*		72.6%	84.5%	75%	83%
Referred to dermatologist by DERM*	**Total**	53.2%	69.4%	44%	62%
	Malignant lesions	48.8%	67.0%	7.5%	9.7%
Judged non-malignant by DERM*	**Total**	19.4%	15.0%	31%	21.6%
	Discharged at second read	12.4%	7.8%	18.7%	10.7%
	Discharged after referral	2.8%	0.8%	4.8%	2.7%
	Malignant lesions	4.3%	6.4%	0	0

* All % are out of total n of cases/patients, including those not assessed by DERM

NICE National Institute for Health and Care Excellence

Health Technologies Programme

Artificial Intelligence technologies for assessing skin lesions selected for referral on the urgent suspected cancer pathway to detect benign lesions and reduce secondary care specialist appointments: early value assessment

Professional organisation submission

Thank you for agreeing to give us your organisation's views on the technology and the possible use in the NHS.

You can provide a unique perspective on the technology in the context of current clinical practice that is not typically available from the published literature.

To help you give your views, please use this questionnaire. You do not have to answer every question – they are prompts to guide you. The text boxes will expand as you type.

Information on completing this submission

- Please do not embed documents (such as a PDF) in a submission because this may lead to the information being mislaid or make the submission unreadable
- We are committed to meeting the requirements of copyright legislation. If you intend to include **journal articles** in your submission you must have copyright clearance for these articles. We can accept journal articles in NICE Docs.
- Your response should not be longer than 13 pages.

NICE National Institute for Health and Care Excellence

About you

1. Your name	Tania von Hospenthal
2. Name of organisation	The British Association of Dermatologists
3. Job title or position	Director of Transformation, Quality, & Improvement Unit (TQIU)
4. Are you (please select Yes or No):	An employee or representative of a healthcare professional organisation that represents clinicians? Yes or ~~No~~ A specialist in the treatment of people with this condition? ~~Yes or~~ No A specialist in the clinical evidence base for this condition or technology? Yes or No Other (please specify):
5a. Brief description of the organisation (including who funds it).	The British Association of Dermatologists (BAD) is the professional membership body for dermatologists in the UK. The BAD is a registered charity whose charitable objectives are the practice, teaching, training, and research of dermatology. We work with NHS England, the Department of Health, NICE, ICS, NHS Trusts, local authorities, and patient bodies advising on best practice for dermatology service provision across all UK service settings.
5b. Has the organisation received any funding from any company which provides technologies for these topics in the last 12 months? **If so, please state the name of company, amount, and purpose of funding.**	No
5c. Do you have any direct or indirect links with, or funding from, the tobacco industry?	No

The aim of artificial intelligence technologies for assessing skin lesions selected for referral on the urgent suspected cancer pathway to detect benign lesions and reduce secondary care specialist appointments

6. What is the main aim of these technologies? (For example, initial diagnosis, clinical monitoring, treatment triage, assessing stages of disease progression or risk stratification.)	Initial diagnosis of selected types of malignant skin cancers and some benign lesions in secondary care populations.
7. In your view, how could these technologies meet an unmet need in the indicated population?	The unmet need for accurately diagnosing benign skin lesions is in primary care. This would reduce the demand burden on secondary care services.

What is the expected place of the technology in current practice?

8. What are the current interventions offered to the indicated population?	Consultant-led triage
9a. Are there any relevant clinical guidelines we should be aware of, and if so, which?	Skin cancer guidelines Basel Cell Carcinoma Squamous Cell Carcinoma Melanoma clinical guidelines
9b. Is the pathway of care well defined? Does it vary or are there differences of opinion between professionals across the	The pathway of care is not clearly defined and a clear "use case" definition is crucial for applications used in clinical pathways. There is a potential difference of clinical opinion as to whether these apps should sit in primary or secondary care to have the most impact. However, without evidence-based research and analysis of data in the general population particularly in all skin colours this is not an accurate interpretation of any available feedback data.

NHS? (Please state if your experience is from outside England.)	NICE IOG / Faster Diagnostic pathway guidance
9c. What impact would the technology have on the current pathway of care (i.e., where do they fit into the care pathway, and would they be used as an alternative or in addition to standard care)?	This would be used in the primary secondary care interface- the question of alternative or additional is important as there will be health economic consequences.
10a. How does healthcare resource use differ between the technology and current care?	Current use of technology relies on back up reading from consultants- when this is removed the key question will be how the settings are dialled...a high sensitivity for melanoma needs to be matched by a high specificity or there is a risk that there are increased referrals rerouted to secondary care, thus increasing workload rather than the opposite.
10b. In what clinical setting should the technology be used? (For example, primary or secondary care, specialist clinics.)	This technology sits in the post referral space.
10c. What investment is needed to introduce the technology? (For example, for facilities, equipment, or training.)	Community Diagnostic Centres (CDC), primary care various models, IMI training for Healthcare Assistants (HCA) to be able to take images - rigorous costings/health economic modelling/evidence of total effect on workload required.
11a. Do you expect the technology to provide additional benefits compared with current care?	No, because a diagnostic tool should be entirely accurate as a standalone device driven by AI. Without the correct case mix and proper evaluation, there's a risk of misdiagnosis. Each local service has differing case mix and pathways, and there may be considerable variation across services. The technology is based on a very early model dermoscope, which is regarded as inferior to current models. This should be considered. Are there any issues re equality, diversity, inclusion, and access to this model?

11b. Do you expect the technology to increase health-related quality of life more than current care?	Some producers of skin smartphone apps claim that they can classify (diagnose), monitor and treat a range of skin disorders. Many of these are flawed in their design due to the lack of availability of clinical images representing the full breadth and depth of skin disorders on which these algorithms are developed. This leads to many apps focusing on diagnosis of a limited number of skin cancers and neglecting potentially more serious and rare diseases. If it's a standalone tool lacking adequate evidence, its implementation could result in harm.
12. Are there any groups of people for whom the technology would be more or less effective (or appropriate) than the general population?	The accuracy of AI algorithms intended to support skin cancer diagnoses may be overestimated where studies are conducted in settings which do not reflect clinical practice. For example, a study using a retrospective image database without supplementary clinical information, excluding atypical presentations (e.g., body sites such as palms/soles), using highly selected patient groups (e.g., excluding skin of colour) or limiting study cases to those already selected for excision introduces significant bias, risks missing serious but rare diagnoses which can lead to patient harm and is unlikely to provide strong evidence for widespread use. Currently, AI targeting specific populations is not designed to ensure equity for all individuals with skin cancer, as it should encompass people of all skin of colour.

The use of the technology

13. Do you consider that the use of the technology will result in any substantial health-related benefits that are unlikely to be included in the quality-adjusted life year (QALY) calculation?	This required health economists to assess. Health economic considerations in dermatology AI
14. Do you consider the technology to be innovative in its potential to make a significant and substantial impact on health-related benefits and how might it improve the way that current need	This technology is first to market scaled CE marked, and thus innovative. The question of the successes of such technologies will require attention to development of a healthy competitive market for skin AI in the UK. Any such market should include a more rigorous assessment before regulatory approval (compare with drugs approval), open competitions to compare, and agreement to use commonly agreed, standardised interoperable systems and data collection using internationally agreed standards- only in this way will there be prevention of

is met? If so, please state the innovative aspects.	monopolies, room for new improving technologies, and competitive evolution of technologies to allow the UK to compete internationally.

Sources of evidence

15. Are you aware of any relevant evidence that might not be found by a systematic review of the evidence?	Much evidence in this field is held commercially, by private companies.
16. How do data on real-world experience compare with the available data?	Most skin image-based Artificial Intelligence (AI) systems are trained on publicly available datasets that are high-resolution, well-centred and low-noise. These do not represent most real-world community-captured photographs. Key issues facing innovators have been a lack of real-world datasets for training, poor comprehension of the specific use cases in dermatology, lack of solutions derived from NHS pathways, restriction of AI to purely images without metadata, lack of diversity in skin types in datasets, and the absence of UK standards for image capture, transmission, storage, and use for AI. An AI technology developed using a referred (e.g., secondary care) population is not generalisable to a primary care (unreferred) population. Other examples of selected populations include patients within limited age groups, restricted ethnicity groups or specific skin lesions at certain body sites or rarer skin cancers. An AI technology is likely to be more cost-effective in certain subgroups of a population, therefore screening every individual attending primary care could result in a large use of resources to detect few cases. The prevalence of the disease and specific clinical characteristics (case-mix) in a population will impact the accuracy of the AI tool, highlighting the importance of having peer-reviewed published effectiveness evidence from a comparable population.

Equality

17. Are there any potential equality issues that should be taken into account when considering this technology?	Yes, need to consider the specific risks that the use of AI brings, such as differential performance of AI health technologies, and the risk of 'AI bias' against marginalised groups including ethnic minorities. Common reasons for this bias include under-representation of that group within the training and testing datasets ('health data poverty') and the encoding of human biases into the data. We must ensure that any new AI regulatory framework ensures 'safety for all' and not just 'safety on average'.

Topic-specific questions

18. What criteria and/or personal and clinical requirements would need to be met for people to be offered AI technologies for assessing skin lesions?	AI technology should undergo rigorous clinical validation through studies involving diverse populations to confirm its effectiveness in various skin types, ethnicities, and lesion types. Regulatory capacity and capability in AIaMD should be reinforced for both pre-market and post-market stages by providing specialised training to regulators, gatekeepers, and key contributors involved in the entire product life cycle.
19. Are these technologies currently used in the NHS? How would they be used?	Yes, various adoption and test scenarios have been implemented; however, there isn't substantial data to support the overall continuous development. The data held does not adequately represent the entire UK population.
20. What features and technical and clinical requirements would AI technologies need in order to be considered for assessing skin lesions selected for referral on the urgent suspected cancer pathway to detect benign lesions and reduce secondary care specialist	

appointments?	
21. Is there an unmet need in the urgent suspected cancer referral pathway that AI technologies could address? Should the scope of the current assessment be expanded beyond those selected for referral to the urgent suspected cancer pathway (i.e., expanded beyond the post-referral setting)?	
22. What do you consider the most relevant clinical and patient outcomes for evaluating AI technologies assessing skin lesions in people selected for referral on the urgent suspected cancer pathway?	**Clinical:** Sensitivity and specificity eg % of patients with skin cancer misdiagnosed as benign using AI (false negative rate) % of patients with benign lesions misdiagnosed as malignant using AI (false positive rate) Cost of medical photography / AI including job plan resourcing compared to traditional face to face appointment e.g., health economics of pathway. % of patients meeting Faster Diagnosis Standard through AI pathway eg clinical diagnosis of cancer or no cancer communicated within 28 days Average time to reach FDS target compared to traditional face to face urgent skin cancer pathway. % of patients eligible for AI pathway (varies across populations) All of the above across the whole range of skin types including type 5 and 6 **Patient:** % of patients requiring additional face to face appointment to make diagnosis following medical photography / AI % of patients representing to GP team with uncertainty about diagnosis following AI diagnosis Patient satisfaction with photography / AI appointment Road miles travelled through AI pathway versus traditional face to face pathway (relevant to climate too). Time between GP referral and AI outcome communication to patient

	Patient satisfaction with quality of information provided if diagnosed as benign lesion with no further follow up (these patients will not have any direct communication with a dermatologist so may not be given a definite diagnosis but will be told that the lesion is benign).
23. What are the limitations of adopting AI technologies for assessing skin lesions in people selected for referral on the urgent suspected cancer pathway in the NHS, if any?	It is time-consuming, involves additional steps in the pathway, and unless deployed as a direct diagnostic tool, it doesn't contribute value to the pathways.
24. What level and type of healthcare professional support is needed?	Consultants, GPs

Key messages

25. In up to 5 bullet points, please summarise the key messages of your submission.	• an intended use that addresses a clinical unmet need and which integrates in patient pathways to enhance the service provided by health care professionals, improving patient care. • a clearly defined use case (population) which can be safely and demonstrably applied to the largest dataset including where possible all ages, skin types, and body sites, and all presentations of skin disease. • a robust, transparent evidence base to support the proposed AI diagnostic/monitoring functions in the intended patient population for a clearly specified set of skin diseases.

Thank you for your time.

Please log in to your NICE Docs account to upload your completed submission.

Your privacy

NICE National Institute for Health and Care Excellence

The information that you provide on this form will be used to contact you about the topic above.

Please select YES if you would like to receive information about other NICE topics - ~~YES~~ or NO

For more information about how we process your personal data please see our privacy notice.

NICE National Institute for Health and Care Excellence

Early Value Assessment (EVA): Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments

Section A: External Assessment Report - Comments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
NHSE Dermatology Outpatient Recovery and Transformation Programme	1	3 and through out	Objectives	The objectives of the evidence review and external assessment report (EAR) to support the NICE EVA are stated by the authors at the outset as follows: *To investigate the clinical and cost-effectiveness of two AI technologies: DERM (Skin Analytics) and Moleanalyzer Pro (Fotofinder), as decision aids to triage and diagnose suspicious skin lesions following a primary care referral.* This objective is not aligned to the title of the EVA which is '*AI technologies for assessing skin lesions for referral on the urgent suspected skin cancer pathway to detect benign lesions and reduce secondary care specialist appointments.*' The NICE EVA was requested to consider the use of AI in the urgent skin cancer pathway to exclude benign lesions because this approach is already being piloted in several NHS sites around the country funded by the AI in diagnostics awards and funding for further sites has been agreed. The medium term view of NHSE, subject to safety considerations, is to roll out this model and for it to become business as usual. The NHSE ask of the NICE EVA was to assess the evidence in respect of the assessment of this real world skin cancer pathway. The EAR provided covers a different and much wider scope than that agreed in the title and much of it is unhelpful in answering the real-world question that relates to the NHSE funded AI pilots. There is an emphasis on the '*diagnosis of suspicious lesions'* throughout, particularly skin cancers. The EVA question relates to the use of the AI tools to detect benign lesions.	The EAG is concerned that some stakeholders have fundamentally misunderstood how the EAG produces an EVA report. We remind the committee and stakeholders that we are tasked with assessing the evidence in line with the topic scope (which was agreed by the scoping committee). We note in particular that the scope required us to examine the diagnostic accuracy to detect malignant lesions (See scope page 21, Table 1) and to assess the evidence on Moleanalyzer Pro (See scope page 20, Table 1) It is the EAG's opinion that many of the comments in this document are disagreements with the EVA scope and process, and are not reasonable criticisms of the EAG report. We therefore do not respond to such comments here. The EAG questions the idea that AI can be used to diagnose benign lesions, somehow without using it to diagnose malignant lesions. By definition, most lesions NOT diagnosed as benign by AI will be treated as potentially

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				In summary NHSE are seeking to use autonomous AI for the diagnosis of benign skin lesions in the urgent suspected skin cancer pathway NOT to use the tool for the diagnosis of malignant skin lesions. This is the fundamental difference between the EVA question asked and the information provided in this EAR. *Inclusion of Molanalyzer Pro (Fotofinder)* The inclusion of the MoleAnalyzer Pro in the EAR is unhelpful and we believe that this product should be removed for the following reasons: • Data are from Canada and Germany only which are different health care systems and no experience in an NHS setting was provided • The data reviewed consider the use of the tool as a decision aid for melanoma and there is not evidence provided of its use in an urgent suspected skin cancer pathway to diagnose benign skin lesions (the question for the EVA) The evidence evaluation considers the diagnostic accuracy of the tool for melanoma only; this is outwith the scope of the EVA which relates to detecting benign lesions and reducing secondary care specialist appointments	malignant, with implications for clinical practice and cost-effectiveness. We note also that a “suspicious lesion” is any lesion that a person or GP is concerned about: it includes benign lesions. The EAG considers that, to understand whether AI can be used to diagnose benign lesions requires a full understanding of the performance of AI tools. As patient representatives made clear at the scoping workshop, using AI to diagnose benign lesions would be unacceptable if that led to malignant lesions being misdiagnosed: so it is essential to first establish that AI can accurately identify malignant lesions. The stakeholder comments frequently conflate the term ‘model’ with reference to economic modelling with model, referring to the design of the NHS diagnostic pathway. This appears to be the source of much misdirected criticism, particularly in the context of discussion of the Skin Analytics/Exeter Test Group cost-utility model. The stakeholders also confuse the description of identified and submitted economic evidence as novel analysis presented by the EAG, which is criticised as being beyond the remit of the EVA. Economic evaluation represents a core part of NICE’s decision making processes. NICE takes

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
					an NHS and Personal Social Services perspective in their assessments of diagnostic technologies. When considering the cost-effectiveness of AI technologies in this context, a cost-utility model must be able to characterise the consequences of all diagnostic outcomes. Correct identification of benign lesions leading to discharge is just one dimension of the diagnostic accuracy data used to populate a typical diagnostic model, which must capture the trade-offs between cost-savings arising from discharge of patients with benign lesions, and the health and resource implications of missed malignancies. Whilst the outcome of interest to clinicians in this EVA differs to that in other diagnostic assessments (i.e. TNs rather than TPs), the economic model must take a holistic view of the value of the technology from an NHS perspective. To do this we still need to know the sensitivity for malignancies, and to understand the consequences of missed diagnoses. On the issue of including Moleanalyzer Pro, the EAG notes that it was necessary to conduct a systematic review of Moleanalyzer in order to demonstrate that there was no UK evidence, and no evidence relevant to diagnose benign lesions. While the EAG could have reported nothing on Moleanalyzer, we consider it more useful to report what evidence does exist.

Early Value Assessment (EVA): Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
NHSE Dermatology Outpatient Recovery and Transformation Programme	2	3	Results	*AI recruited highly selected populations and raised concerns about bias and applicability* The populations recruited in the NHSE pilots uses real-world clinical situations and the patients excluded from the AI tool are much the same as those excluded using a teledermatology pathway. The following patients are not suitable for TD or AI: lesions on scalp, nails, intimate areas and people with multiple skin cancers and people who do not want to have a TD/AI interaction.	We consider our interpretation to be valid. We agree that such similar exclusions would apply to teledermatology, but such lesions still need to be assessed, and it is important to make clear that their exclusion from AI studies raises concerns about the applicability of the evidence. The risk of bias and applicability assessment followed the Cochrane QUADAS-2 guidance and needs to be understood within the context of this appraisal and the NICE scope. The applicability of DERM needs to be assessed for all people with suspicious skin lesions who are referred on the urgent suspected skin cancer pathway for it to be meaningful to patients and clinicians; comparisons between AI and face-to-face clinical assessment need to account for the potential bias associated with the exclusion of harder to diagnose lesions. A clinical adviser who sits on the advisory committee indicated that such lesions are harder for AI to diagnose, therefore the exclusion of these lesions from AI studies is likely to introduce in favour of AI. See also comment 73 from the BAD.
NHSE Dermatology Outpatient Recovery and Transformation Programme	3	3	Results and throughout document	*MoleAnalyzer Pro had lower sensitivity, but higher specificity for detecting melanoma than face-to-face dermatologists* See comment 1 above, the data relating to the diagnosis of melanoma by this tool is outwith the scope of the EAR for the EVA as the question relates to the diagnosis of benign lesions in the urgent skin cancer pathway	See response to comment 1

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
NHSE Dermatology Outpatient Recovery and Transformation Programme	4	3	Results	*Patient and clinical opinions showed substantial resistance to using AI without any assessment of lesions by a dermatologist* This statement is made several times in the document and is not well-evidenced in the report. The assumption is that this is a view from the redacted text of the Leicester data and/or reflects the view of experts on the group who have no experience of the use of the tool in NHS clinical practice. The evidence presented suggests some reticence from patients to replace face-to-face (F2F) interactions with an autonomous AI interaction. This is particularly described in the context of the MoleAnalyzer Pro but this tool has not been utilised in an NHS care pathway and we do not believe it appropriate to include the limited published information relating to this tool in the evidence review with regards to patient and clinician views as they do not relate to an NHS setting and an urgent skin cancer pathway.	These conclusions are clearly based on data, both published and unpublished. See EAG report Sections 3.2.6.5, 3.2.7.3, 3.3.5.2, 3.3.6.1, 3.4.1.3 and 3.4.2.3. We also recognise that the data is limited and recommend further research on patient and healthcare professional opinions in Section 8.2.2.
NHSE Dermatology Outpatient Recovery and Transformation Programme	5	3	Conclusions	*DERM shows promising diagnostic accuracy for triage and diagnosis of suspicious cancer lesions in selected patients referred from primary care. Its impact on the diagnostic pathway and patient care is, however, uncertain. Moleanalyzer Pro shows promising accuracy for diagnosing melanoma, but its evidence base is limited*. Whilst this information is helpful, the conclusion does not answer the question that is asked by the EVA about its ability to diagnose benign lesions and reduce secondary care appointments; it refers to diagnostic accuracy for suspicious lesions rather than benign lesions	See response to comment 1
NHSE Dermatology Outpatient Recovery and	6	3	Conclusions	*'While AI has the potential to be cost-effective for the identification of benign lesions, further research addressing the limitations in the diagnostic accuracy evidence is necessary*.'	See response to comment 1

Early Value Assessment (EVA): Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Transformation Programme				The question of the EVA was the diagnosis of benign lesions not overall diagnostic accuracy so this statement suggests that the data suggests that AI has the potential to be cost-effective in the pathway in which it is being used. The limitations in diagnostic accuracy would appear to relate to malignant lesions which is not the remit of this review.	
NHSE Dermatology Outpatient Recovery and Transformation Programme	7	5	Background	*'As benign skin lesions and skin cancer are so common, this places a very high burden on dermatology clinics, which may lead to long waiting times for a diagnosis.'* This statement is not strictly accurate. Skin cancer diagnosis is prioritised and as a result delays in diagnosis of skin cancer through the urgent skin cancer pathway are unusual (although there are downstream delays in surgical treatment of skin cancer). This prioritisation of skin cancer activity leads to a reduction in clinical capacity for other people with inflammatory skin conditions who often require F2F assessment. The NHS Plan requires us to reduce unnecessary F2F activity by the use of technology including teledermatology. Replacing F2F urgent skin cancer activity with teledermatology (TD) and in due course AI has the potential to free up F2F capacity and benefit those patients that need to be seen for a F2F appointment.	We can amend this statement at a later date.
NHSE Dermatology Outpatient Recovery and Transformation Programme	8	6	Data sources	The NHSE meta-evaluation considers the current NHS pilots and information from this is essential to inform the EVA, it should be available in time to be considered by the group as part of the EAR	The EAG can only consider evidence supplied in reasonable time during the evaluation period. Section 3.2.1 and Appendix 3 refer to ongoing NHS evaluations known to the EAG.
NHSE Dermatology Outpatient Recovery and Transformation Programme	9	7	Modelling	*Use of the term '2 week wait pathway'* Please correct this terminology to urgent skin cancer referral pathway where it appears (except in research studies where the 2ww pathway was in place at the time of the work)	We can amend this where appropriate at a later date.

Early Value Assessment (EVA): Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
NHSE Dermatology Outpatient Recovery and Transformation Programme	10	8	Economic evidence	The three models considered in great detail consider the impact of missed diagnoses of malignant skin lesions particularly BCC (with later statements indicating that BCCs are over-diagnosed as melanoma or SCC rather than under-diagnosed). Whilst this is of interest, this is not relevant to the question being asked about use of AI to exclude benign lesions. The models take no account of the real-world safety netting that occurs after a F2F, TD or AI interaction to ensure that patients seek further advice if their lesion continues to cause concern. This lack of pragmatism is also reflected in the lack of reference to/discussion of missed diagnoses inexperienced clinicians.	See response to comment 1
NHSE Dermatology Outpatient Recovery and Transformation Programme	11	8	Economic evidence	*'namely, the model imposed disincentives for the correct diagnosis and treatment of BCC'* This is a bold statement to make based on the evidence reported and, provided safety netting is provided for all patients (F2F/AI and TD), there is a similar outcome potentially. Later comments suggest that BCCs are over-diagnosed as SCC or melanoma which means that from a clinical perspective there is less/little cause for concern.	This comment appears to misinterpret the use of the word 'model'. This is a factual statement relating to the structure of the Skin Analytics/Exeter economic model.
NHSE Dermatology Outpatient Recovery and Transformation Programme	12	9	Conceptual model	*'the EAG propose an alternative structure for patients with BCC, aimed at better capturing the cost- and health-consequences of BCC, particularly with reference to disease recurrence.'* Whilst this is of interest, this is out of scope of the EVA which asks a specific question about the role of AI to diagnose benign lesions in the urgent skin cancer pathway and reduce secondary care activity	See response to comment 1
NHSE Dermatology Outpatient Recovery and Transformation Programme	13	10	Further research needs	*'Directly comparable evidence on the diagnostic accuracy of AI technologies and teledermatology in a post-referral setting compared with unassisted teledermatology is required to assess the potential value of AI technologies. This would require studies comparing AI with dermatologists assessments, recruiting a representative population and case-mix, use up-to-date versions of AI and dermoscopy, and with a robust independent reference standard for all patients'*	We agree that such data may be available in existing studies.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				This information is very important. However, our understanding is that these data are available from the current NHSE pilot sites as the second read provides this comparator data.	
NHSE Dermatology Outpatient Recovery and Transformation Programme	14	10	Further research needs	*'Further research must also be undertaken to quantify the benefits to population health within skin cancer and other dermatological indications associated with any release of NHS consultant dermatologist resource, and understand the effects of these technologies on waiting times for final diagnosis*.' To streamline and improve timely access to care for people with skin conditions, it is essential to reduce unnecessary F2F interactions and this requires a pan-system approach to outpatient transformation of dermatology services. The NHSE OPRT are delivering on this and key to this is the use of TD across dermatology pathways including the urgent skin cancer pathway. In systems that have implemented these approaches there is evidence of patient benefits and timely access to care. However, the unintended consequence of implementing successful TD and AI pathways is that whilst early diagnosis is possible there are downstream challenges in treating skin cancer which can skew the skin cancer target data (as occurred with the East Midlands Leicester AI pilot). The release of clinician time by the use of autonomous AI to exclude benign lesions provides an opportunity to transfer resource appropriately to these other areas of the skin cancer pathway. Data suggest a potential productivity gain of around 20% if autonomous AI is utilised to diagnose benign skin lesions in the post-referral urgent suspected skin cancer pathway.	No response needed
NHSE Dermatology Outpatient Recovery and Transformation Programme	15	23	1.4.2.	*'Traditionally, GPs directly referred everyone with suspicious skin lesions to secondary care through the urgent suspected skin cancer referral pathway where all referrals required people to attend a secondary care dermatology department for a face-to-face appointment with a consultant dermatologist.'*	Not an error.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				It is no longer the case that all patients referred on an urgent suspected skin cancer pathway have to be seen for a F2F appointment. The new virtual pathway was introduced in 2021 and this is shown and referenced in the scoping pathway charts on page 27. It is within this virtual pathway that the AI tool is used to exclude benign lesions in a post-referral urgent suspected skin cancer pathway.	
NHSE Dermatology Outpatient Recovery and Transformation Programme	16	23	1.4.2	*'The NICE guidelines (NG12) on recognition and referral of skin cancer describes the use of the NHS e-RS advice and guidance (A&G) service as the main pathway in primary care for access to specialist'* This is incorrect, NG12 does not recommend the use of A&G. This is recommended in the NHSE OPRT referral optimisation guidance and this is a separate pathway from the virtual urgent skin cancer pathway.	We can amend this at a later date.
NHSE Dermatology Outpatient Recovery and Transformation Programme	17	25	1.4.2.2 teledermatology service	Please make clear that the pathway referred to is the urgent suspected skin cancer pathway	We can amend this at a later date.
NHSE Dermatology Outpatient Recovery and Transformation Programme	18	26	1.4.2.2	*'Virtual teledermatology cannot be used for lesions on difficult sites such as palms, soles, scalp and intimate areas, or for people with multiple lesions. Virtual teledermatology is not used for children.'* These patients are excluded from both TD and AI pathways	Not an error.
NHSE Dermatology Outpatient Recovery and Transformation Programme	19	26	1.4.2.2	*'Following DERM assessment, lesions classed as high-risk are triaged to urgent virtual review by a hospital dermatologist, whilst lesions classed as low-risk are sent for remote review by a second reader (consultant dermatologist), who will either discharge the patient if in agreement with AI, or overturn the AI risk assessment and proceed with an urgent referral to a hospital dermatologist.'*	Not an error.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				This is now describing AI not TD so needs a separate section. *Teledermatology hubs* TD hubs are used for TD alone and TD/AI and not all TD/AI is done in hubs	
NHSE Dermatology Outpatient Recovery and Transformation Programme	20	27	1.44	Radiotherapy and chemotherapy not used for melanoma, immunotherapy needs a mention if you are to discuss melanoma treatments; for any economic models (which we believe sit outside the scope of this evidence review) the authors would need to become familiar with the up-to-date investigations and treatments for the different stages of disease	Not an error. This section was intended only as a brief comment on treatment, as it is beyond the scope of the assessment.
NHSE Dermatology Outpatient Recovery and Transformation Programme	21	28	1.6 key outcomes	Diagnostic accuracy is given as a key measure; this is not the question for the EVA which relates to the detection of benign lesions rather than diagnostic accuracy of all skin lesions	See response to comment 1.
NHSE Dermatology Outpatient Recovery and Transformation Programme	22	28	1.6.1	• *Diagnostic test accuracy (sensitivity and specificity, area under ROC curve)* *Where available, separately for each type of skin cancer (melanoma, BCC, SCC, rare skin cancers)* • *Proportion of cancers missed and detected* • *Proportion of benign lesions missed and detected* • *Proportion of referrals confirmed to be skin cancer (positive predictive value)* The diagnostic accuracy outcome measures for the skin cancer types do not sit with the question asked of the EVA. The specific question for the EVA relates to the proportion of benign lesions missed and detected, the outcome measures relevant to benign lesions are pivotal to the evidence review.	See response to comment 1. We note that these outcomes are taken directly form the NICE scope.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
NHSE Dermatology Outpatient Recovery and Transformation Programme	23	32	2.1.3.1	• What is meant by a rapid diagnostic clinic?	This was meant as an in-person local diagnostic clinic. This can be corrected.
NHSE Dermatology Outpatient Recovery and Transformation Programme	24	42	3.2.3	*'Results of the quality and applicability assessment are reported in Table 4. All studies were at high risk of selection bias due to the exclusion of a significant proportion of participants (15.6% to 27.4% where reported) that would have otherwise been eligible for assessment in clinical practice. The performance of AI is likely to be significantly improved by the exclusion of some of these lesions (e.g. images with body hair, tattoos, subungual lesions).'* • These are exclusions for all TD models so its not clear why there is mention of bias as these patients are unsuitable for TD and AI. If these patients had a TD or AI interaction it would be double activity as they would inevitably require a F2F interaction so their exclusion at the outset is to be recommended	We think it reasonable to highlight that these patients will not be assessed by AI (therefore affecting estimate accuracy), even if they would not go to teledermatology. Please see our response to comment 2 for further details.
NHSE Dermatology Outpatient Recovery and Transformation Programme	25	42	3.2.3	*'The teledermoscopy devices used in two studies (Dermlite DL1 Basic (DermLite LLC) system)20, 21 were considered out of date following clinical advice and therefore raised concern about their applicability to current practice.'* Whose clinical advice was taken in order to justify the statement without supporting evidence? These devices are still widely used in dermatology departments in the UK and one is listed as a recommended piece of equipment on the BAD website. The DERM AI tool has learnt using this kit so it's the correct kit to use for the algorithm to work.	Clinical advice was drawn from clinical experts on the scoping committee.
NHSE Dermatology Outpatient	26	43	3.2.4	*'Thomas (2023)(UHBFT and WSFT) reported results separately for Birmingham and West Sussex centres. It also reported results for two versions of DERM: DERM-vA (used July 2021 to April*	The reference to assessing DERM vB here relates only to this section and the Thomas study, and is simply to avoid confusion from

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Recovery and Transformation Programme				*2022) and DERM-vB (used April to October 2022). As DERM-vB appears to have superseded DERM-vA we only report results for the more recent DERM-vB.'* Some confusion here as later you report on DERM-vA and actually shouldn't we consider both? Or at least be clear about which you are reporting.	presenting multiple results. Reporting on DERM vA may be removed on editing.
NHSE Dermatology Outpatient Recovery and Transformation Programme	27	47	3.2.5	*'We assume that this data includes all DERM-vB data from the UHBFT and WSFT study up to October 2022, as reported in Thomas 2023. We assume this includes some patients from DERM-005, although the overlap is unclear'* Would be helpful to check?	Not feasible in timeframe of EVA.
NHSE Dermatology Outpatient Recovery and Transformation Programme	28	47	3.2.5	*'When using the "Exact' classification there is a decrease in accuracy. For melanoma the sensitivity remains at near 95%, but for SCC and BCC the sensitivity declines substantially. This suggests that both SCC and BCC lesions may be being misclassified as more serious malignancies by DERM (i.e. SCC as melanoma and BCC as SCC or melanoma).'* Exact diagnostic classification for malignant tumours sits outwith the remit of the EAR (which considers benign lesions). Even so, this statement means that there is less to worry about if the tool is overdiagnosing in terms of the lesions as the tumours will not be inappropriately discharged from care without treatment?	See response to comment 1.
NHSE Dermatology Outpatient Recovery and Transformation Programme	29	48	3.2.5	*'It should be noted that the reference standard in this analysis was usually a "ground-truth" diagnosis made by dermatologists where the lesion was judged to be non-malignant. Therefore, the diagnostic accuracy of DERM may be slightly incorrect as some genuinely malignant lesions may have been incorrectly classified as benign by dermatologists. This also means that estimates of the diagnostic accuracy of dermatologists without DERM may not be reliable.'* There is little mention of the diagnostic accuracy of AI vs clinicians throughout and this is a really important point.. Both AI	A reasonable point, we consider that we present as much on this as we can, giving evidence limitations.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				and clinicians make mistakes in diagnosing benign lesions but provided safety netting of a similar type is in place for both clinical interactions surely there can be less concern?	
NHSE Dermatology Outpatient Recovery and Transformation Programme	30	*55*		*'Acceptability of AI technologies or processes* *'but there was a mixed response to a statement on preference for a face-to-face dermatologist appointment (Figure 9).'* These data are not shown in Figure 9	We can amend this at a later date.
NHSE Dermatology Outpatient Recovery and Transformation Programme	31	*55*	*3.2.7.3*	*Reassurance that lesion is not cancerous* *'In the DERM-005 study, patients expressed confidence in DERM being used on a visual analogue scale (VAS) from 0 to 100, with higher scores indicating a higher level of agreement. Participants generally responded positively when considering AI as a tool to help doctors, but more cautiously when considering the use of AI to replace a dermatologist. Median levels of agreement with interquartile range (IQR) are illustrated in Figure 7.'* These data are incorrectly interpreted in the conclusions and summary where you state: '*Patient and clinical opinions **showed substantial resistance to using AI** without any assessment of lesions by a dermatologist'*	We disagree that this is an incorrect interpretation of the overall evidence (not just that described here)
NHSE Dermatology Outpatient Recovery and Transformation Programme	32	*63*		The inclusion of the MoleAnalyzer Pro in the EAR is unhelpful and we believe that this product should be removed from the review for the following reasons: • Data are from Canada and Germany only which are different health care systems and no experience in the assessment of benign skin lesions in an NHS post-referral urgent suspected skin cancer pathway was provided • The data reviewed consider the use of the tool as a decision aid not in a suspected skin cancer pathway to diagnose benign skin lesions	See response to comment 1.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				The evidence evaluation considers only the diagnostic accuracy of the tool for melanoma; this is outwith the scope of the EVA which relates to detecting benign lesions and reducing secondary care specialist appointments	
NHSE Dermatology Outpatient Recovery and Transformation Programme	33	*66*	*4*	*Cost-effectiveness review* The studies referenced are unhelpful as they do not relate to the models being used in NHS clinical practice for which an evidence review is being support for the EVA	See response to comment 1.
NHSE Dermatology Outpatient Recovery and Transformation Programme	34	*80*	*5.1.1*	*Modelled population* *.. based on NHS sources. It was assumed that 87.2% of patients screened had precancerous or benign lesions. 5.9% of patients were assumed to have melanoma, SCC, and rare skin cancers, and 6.9% had BCC. The model assumed that disease stage of melanoma at the point of diagnosis would also apply to SCC and other rare cancers.* We cannot relate to this data, around 20% of most USC referral is BCC, where are these data from? Also not clear why modelling is included relating to the malignant lesions. This is outwith the remit of the EVA which relates to the use of AI to exclude benign skin lesions in a post-referral urgent suspected skin cancer pathway	This section describes the cost-utility model developed by Skin Analytics and Exeter Test Group submitted in support of NICE's decision making. Economic evaluation represents a core part of NICE's decision making processes. NICE takes an NHS and Personal Social Services perspective in their assessments of diagnostic technologies. Cost-utility analysis must therefore capture the trade-offs between cost-savings arising from discharge of patients with benign lesions, and the health and resource implications of missed malignancies.
NHSE Dermatology Outpatient Recovery and Transformation Programme	35	*80*	*5.1.2*	Models: All models need to include book direct to surgery after TD/AI interaction and similar booked for surgery after F2F interaction; this modelling then highlights/identifies the reduced need for F2F activity for the TD/AI pathway	This section describes the cost-utility model developed by Skin Analytics and Exeter Test Group submitted in support of NICE's decision making. Whilst we can include points relating to the modelling into our critique and own conceptual

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
					model, these comments do not relate to the EAG's work.
NHSE Dermatology Outpatient Recovery and Transformation Programme	36	*82*	*5.1.3*	The information included in these costings is inaccurate in terms of staging investigations (for example Stage 1B melanoma does not have any staging investigations, SLNB has different indications now than it did a few years ago). Please update your knowledge of the investigations and management of these patients if the data are to be included.	As per comment 35.
NHSE Dermatology Outpatient Recovery and Transformation Programme	37	*84*	*5.1.3*	*'The assumption that a relatively high number of unnecessary biopsies resulting from face-to-face assessments, and the improvement in sensitivity of face-to-face assessment following triage, are likely key drivers of benefit in the model and as such the associated diagnostic parameters are central to the value proposition.'* Evidence of use of AI/TD to date suggests: reduced number of F2F interactions (TD), TD clinician interactions (autonomous AI), fewer biopsies, more direct booking to surgery (TD/AI). All of these can release capacity and create productivity gains within the system.	This sentence describes mechanisms of benefit in the SA/Exeter model.
NHSE Dermatology Outpatient Recovery and Transformation Programme	38	*84*	*5.1.3.1*	*'These higher costs associated with both DERM strategies are driven fewer by patients being eligible for assessment by DERM than teledermatology (81% vs 90%).'* These data are surprising as the same exclusions would mostly apply for TD as for AI with around 20% of patients being unsuitable for either a TD or an AI pathway	This appears to disagree with the company's position and evidence provided from the pilots.
NHSE Dermatology Outpatient Recovery and Transformation Programme	39	*84*	*5.1.3.1*	*Teledermatology specificity was reported as 84.3% in the Cochrane review* Which Cochrane review is this referring to please and how many patients, what cohort etc?	This is Chuchu et al 2018 [our ref 55]. We can add the missing citation at a later date.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
NHSE Dermatology Outpatient Recovery and Transformation Programme	40	*85*	*5.1.3.2*	*'Compared to face-to-face assessment, results of the company's model suggest both Derm strategies incur lower costs, see Table 18. As above, this driven by lower costs associated with unnecessary referrals and inappropriate biopsies.'* If safety data supported the use of autonomous AI to exclude benign lesions in the urgent suspected skin cancer pathway, there is the potential to release specialist capacity to do what only specialists can do.	This is a key point made by the EAG throughout the report.
NHSE Dermatology Outpatient Recovery and Transformation Programme	41	*88*	*5.1.5*	*'Given higher rate of ineligibility for assessment with DERM vs AI…'* Again surprising as the exclusions are largely similar	See comment 38
NHSE Dermatology Outpatient Recovery and Transformation Programme	42	*91*	*6.1.1*	*'To reflect current service provision, two alternative comparator diagnostic pathways are considered - the teledermatology model, and the conventional model of referral to face-to-face assessment model. Current provision varies across the English NHS, with no nationally standardised alternative model to the usual referral pathway.'* This is inaccurate, there is now a virtual urgent skin cancer pathway (see earlier) which is being rolled out across the country with a KPI of 50%	If this represents a legitimate comparator it is a scoping issue which should have been highlighted at an earlier stage of the EVA process.
NHSE Dermatology Outpatient Recovery and Transformation Programme	43	*92*	*6.1.2*	Conceptual model: whilst this is an interesting project this is way out of scope of this EVA. NHSE is piloting an AI tool to reduce the number of benign lesions seen on the urgent skin cancer pathway. The aim is for this to become an autonomous AI tool without a second read and the benefit will be to free up clinician capacity and reduce unnecessary clinic appointments for patients. The purpose of this EVA is to review the safety of this model. The conceptual model proposed is not relevant to this question but instead focusses on the diagnostic accuracy of AI for skin lesions which is a different question.	See response to comment 1.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				The statement below highlights the difference between the approach of the different models being considered which are outwith the scope of the EVA, the pragmatic model in use currently in the NHS that seeks to exclude the benign and an idealistic model that considers diagnostic accuracy across all skin lesions. *'A key concern regarding the use of AI technologies for the diagnosis of skin cancers is the identification of rarer indications. Given that these technologies may have limited experience of rare cancers, there remains uncertainty as to whether their high sensitivity to melanoma and SCC is maintained across these rarer indications. Treating them as a single diagnostic category in terms of diagnostic accuracy, stage at diagnosis, rate of progression, and impact upon mortality may therefore be subject to uncertainty. Where possible, sensitivity analysis should be undertaken in which rare cancers are categorised separately, and alternative sources of diagnostic and prognostic data are used to parameterise this sub-population in the model.'*	
NHSE Dermatology Outpatient Recovery and Transformation Programme	44	*94*	*Figure 13*	Decision tree models: figures B and D with all skin lesions passing through AI/TD will never be possible because of the 20% of patients that are unsuitable for images to support a diagnosis	The decision trees in Figure 13 include an 'ineligible for AI' and 'ineligible for telederm' branch, to capture the diagnostic outcomes of these patients.
NHSE Dermatology Outpatient Recovery and Transformation Programme	45	*99*	*6.2.2*	*'The use case for AI technologies in the proposed model involves the identification of benign lesions to allow patients to be discharged following referral, but prior to F2F assessment. The key statistic to estimate the capacity of a test to correctly identify true negative cases is the specificity associated with*	See comment 1.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				this pathway. Discharge of patients with benign lesions reduces the cost and health implications associated with unnecessary investigations.' This is the aim of the current pragmatic approach to the use of AI in NHS urgent suspected skin cancer pathways and this is the question that we seek to answer (exclusion of the benign not diagnosis of the malignant). The AI technology also needs to be shown to be 'safe'/low risk, with appropriate safety-netting to allow removal of the second read and maximise the productivity gain.	
NHSE Dermatology Outpatient Recovery and Transformation Programme	46	*100*	*6.2.2*	*To inform a future cost-effectiveness model, data should be based on studies with the following characteristics:* • ***Setting:*** *UK post-referral (before secondary care investigations);* • ***Intervention:*** *AI technologies (with and without human confirmatory read);* • ***Comparators****: Face-to-face and teledermatology with intent to exclude benign lesions;* • ***Outcome****: Diagnostic accuracy (sensitivity and specificity) of individual component tests and the overall pathway*. We have been sighted on much of this data from our pilot studies and the NHSE commissioned meta-evaluation will provide much of this information	We agree that existing studies can potentially provide this data.
NHSE Dermatology Outpatient Recovery and Transformation Programme	47	*108*	*7.1.1.1*	*Three studies of DERM were examined to assess diagnostic accuracy. Autonomous use of DERM appears to have a high diagnostic accuracy for detection of malignant lesions: with a sensitivity of around 96.1% (95% CI 95.4 to 96.8) for a specificity of around 65.4% (95% CI 64.7 to 66.1). Similar*	Not an error.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				diagnostic accuracies were found for detecting specific types of cancer (melanoma or SCC). There was some evidence that DERM might misdiagnose BCC cases as SCC or melanoma. This is very important. The over diagnosis of BCC is good not bad. It is much preferable to over-diagnose skin cancers	
NHSE Dermatology Outpatient Recovery and Transformation Programme	48	*108*	*7.1.1.1*	*'Clinicians were generally very resistant to using DERM in isolation without human assessment of lesions.'* We cannot see the evidence in the document to support this statement and it is not our experience from the pilots that we work with.	See response to point 4.
NHSE Dermatology Outpatient Recovery and Transformation Programme	49	*112*	*7.3.1*	*The DERM versions (in particular, the set sensitivity/specificity thresholds) and the dermatoscopes used for clinical assessments were out of date; therefore, the applicability of the diagnostic accuracy results to current practice is uncertain.'* This is a strong assertion that does not sit with clinical practice. Newer equipment exists but the equipment used currently works with the AI algorithm and is doing very well in terms of excluding benign lesions.	We note that uncertainty is not a negative. We mean only that as newer versions of DERM/newer dermatoscopes are used the accuracy may change.
NHSE Dermatology Outpatient Recovery and Transformation Programme	50	*113*	*7.3.2*	*Evidence from healthcare practitioners on their confidence in DERM and its clinical and broader impact on the pathway and patient management is limited, although initial evidence from limited samples suggested that patients and clinicians do not support the autonomous use of AI tools.* There is insufficient evidence provided in this report to make this statement	We disagree. See response to comment 4.
NHSE Dermatology Outpatient	51	*114*	*7.4*	*The need to ensure that use of AI does not lead to malignant lesions being missed.*	Not an error.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Recovery and Transformation Programme				*Concerns around equality due to difficulty in assessing lesions covered by tattooing, hair or scarring, or in hard to assess areas. Equality issues around diagnosis of skin cancer in people with darker skin or non-white ethnicity.* *The need to reduce anxiety created by the diagnostic process (e.g. due to long waits for diagnoses, or incorrect initial diagnoses).* These concerns from patients are very important and we have addressed the issues around ensuring equity of access to dermatology care and health inequalities and described potentail mitigations, particularly for teledermatology in our guidance document. There will always be around 20% of patients unsuitable for a TD/AI interaction for the reasons outlined above and we are aware that more works needs to be done in people with rich skin tones	
NHSE Dermatology Outpatient Recovery and Transformation Programme	52	*115*	*8.1*	*The substantial resistance from both patients and clinicians to using AI without any human dermatological assessment means that if AI is to be used autonomously in some way, more robust evidence that is applicable to current practice is needed to demonstrate that it has clear benefits to patients, without sacrificing accuracy.* The team have not provided the evidence in the document to support this statement	We disagree. See response to comment 4.
NHS England	53	3 and through out	Throughou t	The originally scope of this Assessment report which was commissioned is reflected in the title. 'Early Value Assessment EVA: Artificial Intelligence Technologies for Assessing skin Lesions for Referral on the Urgent Suspected Cancer Pathway to detect Benign Lesion and Reduce Secondary Care Specialist Appointment." NHSE requested this EVA specifically for the purpose of evaluating AI's ability to exclude benign lesions without a second read in the urgent skin cancer pathway because this approach is already being piloted in several NHS sites and further funding	See response to comment 1.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				has been agreed to roll out further to other sites. With the view, subject to safety considerations, for it to become business as usual. The evidence review has fundamentally strayed from its intended purpose. 1. What should have been a synthesis centred around the AI's ability to autonomously exclude benign lesions from the pathway (i.e. do the technologies have sufficient specificity?) has instead largely morphed into a test of the technologies' ability to detect and diagnose malignancies with suitable accuracy (i.e. do the technologies have sufficient sensitivity?). 2.For reasons in line with the above, MoleAnalyzer's product should not have formed part of the evidence review as it is not intended to autonomously exclude. It is instead a diagnostic tool for use in a physician appointment as part of physician examination and is therefore out of agreed scope. Our view is that it should be excluded from further work on this EVA. 3.The review makes several strong references to "substantial clinical resistance" in the summaries and conclusions (i.e. the only sections which a lot of people will read!) but does little to substantiate this claim in the main body of the review. It would appear that these references are at least partly the result of informal feedback from clinicians (known to us) who have been present throughout this process, and that this is why there is almost no reference from the body of evidence. 4.Linked with the above point and independent to the level of robustness of evidence substantiating this claim, it is our view that the opinions of clinicians should not hold weight in an evidence review which focuses on the empirical evidence re: the economic viability and clinical effectiveness of an intervention. In summary NHSE are seeking to use autonomous AI for the diagnosis of benign skin lesions in the urgent suspected skin cancer pathway NOT to use the tool for the diagnosis of malignant skin lesions. This is the fundamental difference	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				between the EVA question asked and the information provided in this EAR.	
NHS England	54	3 and through out	Throughou t	The basic aims of the review as it was conducted appear to be flawed and drift away from the agreed scope which is reflected in the title. For example, on page 3 the objective is cited as: "*To investigate the clinical and cost-effectiveness of two AI technologies: DERM (Skin Analytics) and Moleanalyzer Pro (Fotofinder), as decision aids to triage and diagnose suspicious skin lesions following a primary care referral."* This is not in keeping with the agreed scope, which was to assess the effectiveness of selected AI technologies to autonomously exclude benign lesions from the urgent suspected skin cancer pathway and reduce the need for secondary care appointments. In other words, the aim of the review is to determine the specificities of the technologies as opposed to their sensitivities. This same error of assessing the technologies' ability to identify and diagnose all (or specific types of) malignant lesions are carried forward in the wording and working throughout the review. For example, in the plain English summary, the review findings state: "*The evidence we reviewed suggests DERM could potentially identify 95% of all skin cancers but would require about half of all patients to be referred to dermatologists. Moleanalyzer Pro could identify about 85% of malignant melanomas. This appears to be a similar accuracy to that achieved by dermatologists alone. How their use would impact diagnosis and treatment for patients in practice, and the burden on clinicians, is currently unclear."* These findings are irrelevant to the agreed scope of the review, since the agreed intention was never to assess diagnostic accuracy of malignant lesions or what proportion of malignant lesions these technologies could detect. The relevant KPI should instead be the percentage of appropriate lesions which are determined by the AI to be suitable for immediate discharge without further review, but which are later	See response to comment 1.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				determined to be malignant when subjected to secondary assessment.	
NHS England	55	3 and through out	Throughou t	For reasons stated in Comment 1, the Inclusion of MoleAnalyzer Pro is scope creep as it is not designed to autonomously exclude benign legions. It merely uses AI to provide a risk score which helps to inform physician assessment. MoleAnalyzer Pro should be removed from this review.	See response to comment 1.
NHS England	56	50	3.2.6	Linked with our previous point (in comment 1) regarding the most relevant KPI being a low % of malignant lesions which were recommended for immediate discharge, there has been a fundamental misunderstanding regarding the rates of missed malignant lesions as a result of the use of DERM. At the bottom of page 50, the review states that "*between 4.3% and 6.4% of lesions were judged as eligible for discharge by DERM vA and, following a second read and referral to the trust, were subsequently diagnosed as malignant."* This is incorrect. The figure of 4.3-6.4% relates to the proportion of cases recommended for immediate discharge by DERM which were not, upon review, recommended for immediate discharge by a human dermatologist. Of this cohort which were referred to a Trust and seen face to face as a result of a human disagreeing with DERM's recommendation, only a small proportion were found to be malignant. My understanding is that this figure was in the 0.1-0.2% range as a proportion of the whole, and that missed malignant lesions were therefore 21-64x less common than the review states. This error, when corrected, has significant implications for the evidence base for DERM.	We will remove this incorrect statement on editing.
NHS England	57	3, 8, 10, 53, 108, 115	Abstract, scientific summary, 3.2.6.5 7.1.1 8.1	The weight given to the resistance towards AI of some clinicians should not be a consideration in an evidence review of this nature. The review should instead focus on the clinical and economic evidence which clinical opinion should be guided by, and should aim to influence said medical opinion through clear synthesis and presentation of that evidence.	Assessment of clinician and patient opinion is standard practice in reviews like this. The EAG considers it important that the committee are aware of this body of opinion, even if those opinions are not based on evidence.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				Strong language is used to state this claim several times, especially in the abstract & scientific summary which are the only sections some readers will attempt: "*clinical opinions showed substantial resistance*" (page 3); "*very substantial resistance, particularly among clinicians*" (page 8); "*substantial resistance from both patients and clinicians*" (page 10). There are other similar references in the closing sections of the review; "*Clinicians were generally very resistant to using DERM in isolation…"* (page 108); "*substantial resistance from both patients and clinicians…"* (page 115). However, upon a full reading, the review does little to attempt to substantiate this core claim. A study of unknown size in Leicestershire community hubs is included on page 53 but redacted, however the findings of such a study could hardly be extrapolated across the wider clinical workforce, especially since the dermatology clinical workforce size in this setting and geography would be small and response rates for this study were not reported. The clinical leads from secondary care services who have already deployed DERM have been incredibly positive about the technology with several keen to move away from the current level of safety netting at the earliest opportunity due to the increasing belief that it is unnecessary, based on their experience of the outcomes the technology provides. These positive clinical opinions were a driver for the NHSE Outpatient Recovery and Transformation Programme requesting this EVA in the first place.	
NHS England	58	3, 42, 50, 58, 112, 116	Abstract 3.2.6.1 3.3.3 7.3.1 7.5 8.2.1	A criticism of the technologies at various points in the review is that they are inappropriate for too many people, citing lesions in the hair and within tattoos as examples amongst others. The review also states that EDI concerns arise as a result of the fact that those for whom it is appropriate are disproportionately white. The low specificity figures cited on page 3 and elsewhere appear to be determined using a denominator of all patients as opposed to just those patients for whom the technology is appropriate.	We note that these EDI concerns were explicitly raised by patient representatives during scoping. We therefore think it reasonable to raise those issues in our report.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				This is an inappropriate assessment since, whilst such limitations are inconvenient and unfortunate, this should not impact an evidence review which looks at the outcomes of patients on whom the intervention was used. An intervention should not be assessed in scenarios where it is neither appropriate nor indeed used in practice. We acknowledge that there are EDI implications, however increased service efficiency as a result of deployment of AI for patients within its established (if indeed limited) capabilities would positively impact the availability of face-to-face services for all patients on waiting lists; including those for whom the use of AI is not currently appropriate.	
NHS England	59	3, 10, 53, 108, 115	Abstract, Scientific summary, 7.1.1 8.1	We do not have access to the figures previously presented to us, but our clear understanding is that patient attitudes towards autonomous AI are significantly more favourable than presented in this review.	No evidence supplied. The EAG cannot comment.
The British Association of Dermatologists	60	3-4	Abstract	The results are reported like a systematic review evaluating diagnostic accuracy for malignant lesions – the scope of this report was to assess skin lesions to detect benign lesions and to reduce secondary care referrals – neither of these have been adequately included. For MoleAnalyzer it needs to be clear that there is no data for benign lesions. Bottom of page 3 – suggests that unpublished data suggests that it reduces around half of patients – this is a misleading statement as it is based on one study which has flaws. This needs to be clarified – the number of patients in that study need to be provided. There is no mention in the abstract about the exclusion of lesions from studies and more importantly there is no mention that DERM excludes skin of colour lesions and most studies have selected fair skinned individuals. This is of major concern for generalisability to a population where all patients should be offered access to all technologies in the NHS unless there is a specific reason why they are not. This potential risk of exacerbating racial bias and the risk of perpetuating inequities has	See response to comment 1.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				been glossed over and should be included in the abstract and in the summary. Page 4 – the conclusion states that there is promising accuracy for triage and diagnosis – triage is not the same as diagnosis – the studies were not evaluated to assess triage impact – it was clear that the studies have not reviewed clinical impact, so this statement is inaccurate and false. Furthermore, stating that the diagnosis of suspicious skin lesions is overstating the results when the specificity for melanoma was 65.4% suggesting a high rate of false positives which would lead to an increase in referrals to secondary care.	
The British Association of Dermatologists	61	4	Abstract	It is unclear why the conceptual model characterises the long-term consequences of BCC, as this is not within the scope of the document?	See response to comment 1 regarding the relevance of economic evaluations to NICE decision making. Basal cell carcinoma is explicitly within the scope of this appraisal.
The British Association of Dermatologists	62	5	Scientific summary	Line 2- Melanoma is not the deadliest cancer, Merkel cell cancer is, therefore this statement is factually incorrect. Second paragraph – not sure what system the AI is being used in when 'in combination with a Dermatologist' as this is not a decision assist tool for a dermatologist – this needs to be clarified. The objectives listed under the aim states 'as decision aids to triage and diagnose suspicious lesions', but the scope of the project was to detect benign lesions and reduce referrals to secondary care. The review has been performed like systematic reviews and reported the evidence, but the role of the evaluation team was to look deeper into the evidence to answer the scope, as was requested by NHS England.	We can delete this at a later date. We consider this to be clear as is. See response to comment 1.

Early Value Assessment (EVA): Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
The British Association of Dermatologists	63	7	Scientific summary	2nd paragraph results- The reports detection of benign lesions sensitivity was 71.5% and then states this is comparable to dermatologists, what comparative study has been used and is this the same? If not, it needs to be made clear that the comparative data are lacking and the quality of the evidence available is weak.	This refers to the same studies as for DERM.
The British Association of Dermatologists	64	9	Scientific summary	Page 9 conclusions – line 3; states 'within a subset of patients.' This needs to be very clear who these are and that darker skin types have been excluded; this is a major oversight of the team who wrote the report not to recognise this potential racial bias that would be perpetuated for patients attending to see their GP in the community.	This has been included in the amended summary.
The British Association of Dermatologists	65	11	Plain English summary	Incorrectly states in second paragraph that 'the evidence was reviewed to investigate ...can accurately identify skin cancer.' – as per the title of this EVA that was not the scope – it was not to see if they improve the diagnosis process, instead it was to see if the AI tools can accurately detect benign lesions and aid GPs and secondary care specialists to reduce the number of referrals to secondary care	See response to comment 1.
The British Association of Dermatologists	66	21	1.2.2	Second paragraph states which dermatoscopes used – same needs to be reported for DERM	This was taken from NICE scope, we may not be able to provide this detail.
The British Association of Dermatologists	67	26	1.4.2	Second paragraph – not all teledermatology hubs have DERM assessment – this paragraph is misleading suggesting that the only option would be to have DERM as part of a teledermatology community hub. Most of Scotland has this set up and none of the hubs use DERM	Note: from "Following DERM assessment…" onwards is a formatting error: this text will be removed on editing.
The British Association of Dermatologists	68	41	3.2.2	DERM-003 and DERM-005 have notable differences mixes in terms of average age, % female, lesion location and consequently in prevalences of the different types of skin cancer. Any similar disparities in the case mix of benign lesions are likely to have contributed to the observed differences in specificities	While this may be the case, we consider that the switch to DERM vB is a more likely explanation of the differences.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				between these two studies (specificity was 45% for detection of any malignant lesion in DERM-003 and 80.4% in DERM-005).	
The British Association of Dermatologists	69	41-42, and 45-46	3.2.2 and 3.24	Large proportions of lesions are excluded from the DERM studies, especially the Thomas one. For this study, the authors and subsequently the DAP team have only included data for lesions with a confirmed diagnosis (on histology or clinical diagnosis) (n=8571 lesions), however Figure 3 in the paper shows that DERM was used for 10,925 lesions, with 2028 of those with a DERM 'discharge from pathway' recommendation (so presumably a low risk DERM rating) subsequently discharged following remote review by 'Second reader' (i.e. image-based clinical diagnosis of benign). Removing these 'benign' (but presumably with no specific lesion diagnosis recorded) from the dataset leaves 8897 lesions rather than 8571 so it is not at all clear how the 8571 included lesions map back to the original lesion set that was assessed by DERM.	We agree there is some uncertainty here that we could not fully resolve in the EVA timeframe.
The British Association of Dermatologists	70	44,46	Figure 4 and Figure 5	The scales on these plots should ideally range from 0 to 100 to give a clearer picture of where these datasets lie in ROC space.	This suggestion would make differences hard to perceive.
The British Association of Dermatologists	71	56	3.3.1	Both MoleAnalyzer Pro studies were conducted in secondary care, one including only melanocytic lesions and the other including only excised lesions. The participant characteristics demonstrate that both studies focused on Fitzpatrick skin type III or lower even though only one study explicitly restricted eligible skin types. Winkler 2023 further restricted inclusion to lesions already identified as melanocytic, therefore the prevalence of melanoma was high and spectrum of included lesions relatively narrow (likely impacting on specificity). In the McLellan study, all lesions were excised thereby indicating some degree of clinical suspicion (the paper states lesions were included if they warranted further investigation and were 'clinically challenging'), which is likely to have a considerable effect on both accuracy and applicability.	Not an error. We accept this point.

Early Value Assessment (EVA): Artificial Intelligence Technologies for Assessing Skin Lesions for Referral on the Urgent Suspected Cancer Pathway to Detect Benign Lesions and Reduce Secondary Care Specialist Appointments

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
The British Association of Dermatologists	72	56	3.3.4	The apparently outlying results for the Dermatologist alone in the MacLellan study are likely because the accuracy is based on the in-person triage decision on initial presentation – excise, no excise or monitor - as opposed to a clinical diagnosis of melanoma or not. This gives a misleading impression of the difference in accuracy between the dermatologist and the AI tool which recorded results in terms of suspicion for melanoma. Given the differences in the populations, pooling these studies may not have been appropriate as the resulting data is not indicative of clinical performance in any particular setting.	We accept this point
The British Association of Dermatologists	73	32, 43-45, 50	*2.1.3, 3.2.4, 3.2.6,*	The intended use of the test is clearly set out and the eligibility criteria largely reflect this (p32, section 2.1.3). In principle, the focus on studies using more current algorithms will increase the applicability of the evidence and prospective collection of data at least partly reduces selection bias. More care could have been taken when considering studies of patients in a secondary care setting however, as studies conducted in dermatology clinics may not fully document prior testing undergone by participants prior to study inclusion (e.g. both MoleAnalyzer studies and at least one of four DERM studies). The quality assessment was mostly appropriate and highlights key flaws in the evidence base, in terms of participant selection and (lack of) applicability of the evidence to the review question. The EAG recorded no problems with flow and timing in any study but the exclusion of eligible lesions after recruitment occurred would usually be considered a flow and timing issue (Thomas 2023 particularly affected due to apparent exclusion of lesions with no specific lesion diagnosis assigned). Lack of data on test failure rates (reported in 3.2.6) should also be considered an important flow and timing issue – a test with excellent accuracy will have limited impact if it has high failure rates.	Thank you. We agree with several of these points (including risk of bias and further research) and will consider incorporating these at a later date.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				Differences in study settings are reflected in the reported participant characteristics (e.g. DERM-003 is notably different to the others) and this is likely to have translated through to some of the observed differences in sensitivities and specificities between studies. Meta-analysis of DERM data was conducted to provide an overall indication of average performance (pg 43, section 3.2.4) however care is needed to ensure that important and true differences between settings are not masked by the focus on average estimates. The individual studies are large and likely to be adequately powered so although the analysis gives an overall average accuracy, real differences in AI performance in different clinical populations should not be ignored. Similarly, the rationale for omitting data for DERM-vA from the Thomas study was to focus on more recent version of DERM however if versions similar to DERM-vA were used in other included studies then it would have been more sensible to include all data to demonstrate how test deployment in new settings might affect accuracy (i.e. DERRM-A required algorithm updating to obtain optimal performance however the lower performing vA was used for a number of months prior to the update). Care is also needed when comparing the accuracy of the AI tools to dermatologist assessments (pg 44-45, section 3.2.4). For example a straight comparison of detection of malignancy suggests lower sensitivity for dermatologists however detection of malignancy does not necessarily equate to the clinical action taken following a dermatologist assessment (e.g. a referral or excision may be recommended for lesions not considered to be malignant if clinical suspicion is sufficiently high), such that the clinical consequences of 'missing' a malignant diagnosis may be less for dermatologists than for a DERM assessment as low risk. There is a lack of evidence for the clinical impact of introducing these types of tools (pg 50, section 3.2.6). Ideally a prospective comparison of health outcomes and health service outcomes with and without the use of the AI tool in a TD hub would be needed in	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				order to fully evaluate impact. Time to diagnosis for cases missed by the tool, lesions not imaged due to test failure or ineligibility for imaging, and possible effects on GP referral patterns could all impact on the ability of the tool to improve health outcomes.	
The British Association of Dermatologists	74	63-65	3.4	The conclusions for both tools summarise key findings but could do more to point to the study limitations and applicability of the evidence to the review question which are clearly highlighted by the results of the quality assessment. Differences in study settings, the focus on particular skin types, the potential for lesion exclusions and unknown test failure rates are likely to considerably impact on observed accuracy and the clinical utility of the tools.	We agree with this point. We can amend conclusions at a later date.
The British Association of Dermatologists	75	115	8	Following on from the comment above, the overall conclusions potentially overstate the evidence for DERM at the current time, although some qualification in terms of lack of evidence for its practical impact and clinical benefit are given and the Research priorities section clearly outlines relevant evidence gaps. The Discussion also does an excellent job of highlighting inadequacies in the evidence and lack of diversity in the data. Generally, I have some concerns that the headline accuracy results could be taken at face value despite real concerns about selection bias and applicability. The fact that the majority of evidence is based on people with lighter skin types is of further concern and gives no insight into how the AI tool would perform in communities with more mixed ethnic groups.	Thank you. As per above, the report places a further emphasis on the lack of evidence for darker skin colours.
The British Association of Dermatologists	76	*Throughout*	*Throughout*	The redacted data implies a clinician cannot be trusted with confidential information, when clinicians are managing confidential patient information on a daily basis.	This is a NICE requirement.
The British Association of Dermatologists	77	*8*	Scientific summary	The EAG also noted several issues which may mean that the main value drivers were not appropriately characterised' - it is important to confirm that any inaccuracies in previous evaluations relating to phototherapy usage and costings in the management	This statement refers to the Skin Analytics and Exeter Test Group model. The point is noted and can be added to our critique at a later date.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				of basal cell carcinoma have been fully recognised and taken into account.	
The British Association of Dermatologists	78	*23*	*1.4.2*	The pathway isn't entirely accurate. 'Traditionally, GPs directly referred everyone with suspicious skin lesions to secondary care through the urgent suspected skin cancer referral pathway where all referrals required people to attend a secondary care dermatology department for a face-to-face appointment with a consultant dermatologist.' This should read 'GP teams' as referring, and 'Consultant-led teams' reviewing the patients.	We think our description is reasonable, but can consider editing at a later date.
The British Association of Dermatologists	79	29	1.6.4	Costs need to include administrative costs. One of the highest costs for units piloting AI can be employing new admin staff to book patients into photography clinics and set up new systems for patient tracking lists.	This is a scoping issue but 'employing new admin staff' appears to be covered under 'Costs of new services required to support AI technologies (such as establishing new teledermatology services and setting up image capture)'
The British Association of Dermatologists	80	71	4.2.2	Large focus on Vivascope and Confocal microscopy which is not standard care	A review of published cost-effectiveness studies is a key stage in the development of a new economic model.
The British Association of Dermatologists	81	80	5.1	Exeter report based on 2013 modelling. Please ensure inaccuracies are highlighted eg phototherapy is not used for BCC treatment yet used in costings £94 per session x 46 sessions, inflated to 2024.	Noted – this can be edited at a later date.
Skin Analytics	82	*20, 30*	*1, 1.7*	Drift from EVA purpose of, and inconsistent focus on, scope of benign lesions and reduction in secondary care specialist appointments - The EVA process exists to assess new technologies that are most needed and in demand[1]. The need was discussed at length in the scoping call and included	See response to comment 1. We also note that there was no published evidence and the company supplied no robust clinical evidence on "current challenges faced by NHS dermatology departments ", "the severity of workforce challenges" or "delays already

[1] Early Value Assessment (EVA) for medtech [Internet]. National Institute for Health and Care Excellence. [cited 2024 Jan 11]. Available from: https://www.nice.org.uk/about/what-we-do/eva-for-medtech

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				within the title of the EVA, namely to 'detect benign lesions and reduce secondary care specialist appointments'. We are therefore surprised and disappointed to see a lack of any detail relating to the current challenges faced by NHS dermatology departments. Whilst referenced in passing, the severity of workforce challenges are unquantified, and there is seemingly no reference to delays already experienced within both urgent suspected cancer and routine dermatology pathways: a. 24% of consultant Dermatology posts in the UK are unfilled and there are >140 locums[2]; b. Urgent suspected skin cancer referrals have increased 5x faster than skin cancer detection over the past 10 years[3]; c. Increasing routine referral backlog where a third of melanoma and SCC are found[4] due to prioritisation of skin cancer referrals, which disproportionately affects Black and Asian patients[5]. Without this context, there may be a risk of biassing interpretation of the evaluation, considering the new technologies against an idealised standard of care pathway which does not reflect reality.	experienced within both urgent suspected cancer and routine dermatology pathways".

[2] Levell N. Dermatology GIRFT Programme National Specialty Report [Internet]. Getting It Right First Time | NHS England & NHS Improvement. 2021 Aug. Available from: https://www.gettingitrightfirsttime.co.uk/wp-content/uploads/2022/07/DermatologyReport-Sept21o.pdf

[3] CancerData [Internet]. www.cancerdata.nhs.uk. NHS; Available from: https://www.cancerdata.nhs.uk/cwt_conversion_and_detection

[4] A teledermatology roadmap: implementing safe and effective teledermatology triage pathways and processes [Internet]. www.england.nhs.uk. NHS England; 2023 [cited 2023 Nov 20]. Available from: https://www.england.nhs.uk/long-read/a-teledermatology-roadmap-implementing-safe-and-effective-teledermatology-triage-pathways-and-processes/

[5] Public Health England. Routes to diagnosis 2015 update: malignant melanoma | National Cancer Intelligence Network Short Report [Internet]. National Cancer Registration and Analysis Service; 2015 [cited 2023 Nov 14]. Available from: http://www.ncin.org.uk/view?rid=3121

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Skin Analytics	83	9-10 44 11	Scientific Summary 3.2.4 Plain English summary	Diagnostic accuracy - we challenge the conclusion that more diagnostic accuracy data for DERM is required *'The diagnostic accuracy of DERM suggests that it has potential for use within a post-primary care referral setting. This could be either alongside assessment by dermatologists, or as an autonomous tool within the post-referral pathway within a subset of patients.'* *'These results suggest a high sensitivity when using DERM autonomously without assessment by a dermatologist is achievable, and may be higher than achievable using a standard diagnostic pathway without DERM.'* - We welcome the diagnostic accuracy results for DERM found within the report and the above statements highlighting the suitability for its use within autonomous post-referral pathways. - DERM has been used in NHS pathways that have seen >70,000 patients with an abundance of real world evidence generated supporting its use. Although not directly reflected within the report, the evidence we submitted for the review included our Q3 DERM Performance Report which summarised performance across >20,000 lesions from real world cases at 7 NHS sites demonstrated sensitivities of 95.7% (465/486) for melanoma; 98.1% (767/782) for SCC; and 97.3% (1534/1577) for BCC[6]. - Considering the abundance of data and above statements from the report, we struggle to understand why the summary and recommendations include further	Our concern is with the ability to reliably compare autonomous DERM or DERM with teledermatology to teledermatology alone; hence the need for a properly independent reference standard. We also note the need for diagnostic accuracy in people with darker skin tones. We note that much of this evidence could, in principle, be obtained within the existing studies, as we have stated.

[6] SA-003905-CS-2-_Overall_ DERM Clinical Performance Report Q3 2023

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				evidence of the diagnostic accuracy of DERM, '*The diagnostic accuracy of AI in a post-primary care referral pathway is uncertain and requires further evaluation' (Future research needs page 10)*. - The limitations cited seem to reflect a lack of evidence around comparator accuracy rather than DERM - *'a lack of key comparative data meant the relative clinical and cost-effectiveness of alternative pathways was necessarily based on often optimistic assumptions'; 'comparable diagnostic accuracy data describing current service provision is lacking, particularly for the teledermatology pathway'; 'Directly comparable evidence on the diagnostic accuracy of AI technologies and teledermatology in a post-referral setting compared with unassisted teledermatology is required to assess the potential value of AI technologies. This would require studies comparing AI with dermatologists assessments, recruiting a representative population and case-mix, use up-to-date versions of AI and dermoscopy, and with a robust independent reference standard for all patients'; 'limitations in the evidence on the diagnostic accuracy of AI technologies'* - Whilst we agree that comparator data is more limited, we note that no systematic review of comparator accuracy has been conducted, nor clear appraisal of the Cochrane reviews which we referenced within our	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				submission[7,8,9]. These included teledermatology demonstrating sensitivity of 94.9% (429/452) for skin cancer across ~4000 (4,057) lesions[7]. With the existing evidence base, Teledermatology has already been endorsed nationally[4]. - If the EAG assessment is that the Cochrane reviews are not sufficient, then while in an ideal world there would be an RCT comparing DERM, teledermatology and in-person dermatologist assessment, we have to question the feasibility and appropriateness of this recommendation: 1. Relevance - DERM is deployed in pathways with teledermatology and face-to-face downstream of it, not independently of them, and the EVA is focused on the ability of AI to detect benign lesions and reduce secondary care appointments, not the ability to detect every single lesion type. 2. Ethically - patients and lesions who are referred by DERM would all have to be biopsied in order to provide a robust and independent ground truth.	

[7] Chuchu N, Dinnes J, Takwoingi Y, Matin RN, Bayliss SE, Davenport C, Moreau JF, Bassett O, Godfrey K, O'Sullivan C, Walter FM, Motley R, Deeks JJ, Williams HC. Teledermatology for diagnosing skin cancer in adults. Cochrane Database of Systematic Reviews 2018, Issue 12. Art. No.: CD013193. DOI: 10.1002/14651858.CD013193

[8] Dinnes J, Deeks JJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, Wong KY, Aldridge RB, Abbott R, Fawzy M, Bayliss SE, Grainge MJ, Takwoingi Y, Davenport C, Godfrey K, Walter FM, Williams HC. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. Cochrane Database of Systematic Reviews 2018, Issue 12. Art. No.: CD011902. DOI: 10.1002/14651858.CD011902.pub2.

[9] Dinnes J, Deeks JJ, Chuchu N, Matin RN, Wong KY, Aldridge RB, Durack A, Gulati A, Chan SA, Johnston L, Bayliss SE, Leonardi-Bee J, Takwoingi Y, Davenport C, O'Sullivan C, Tehrani H, Williams HC. Visual inspection and dermoscopy, alone or in combination, for diagnosing keratinocyte skin cancers in adults. Cochrane Database of Systematic Reviews 2018, Issue 12. Art. No.: CD011901. DOI: 10.1002/14651858.CD011901.pub2.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				3. Cost - considering the supporting evidence that already exists for the diagnostic accuracy of DERM demonstrating safety through a high sensitivity, we are unclear on the additional return on public investment from funding this type of evaluation. 4. Capacity - There is a large, and growing mismatch between dermatology capacity and demand for skin cancer assessments. In that context, do trusts have the capacity to take on additional activities associated with an RCT and potential requirement to biopsy additional lesions for the purposes of such a study. Further, how would clinical variation across the country be accounted for. 5. Responsibility - whilst we are supportive of robust clinical evidence generation for DERM, we do not feel it is the medical device manufacturers' responsibility to correct for the lack of data on comparator clinical accuracy, particularly when Cochrane reviews already exist[7-9]. We feel that the omission of sample sizes in the relevant datasets referenced within the report may lead to a misrepresentation of how vast the evidence base actually is, particularly in the summaries.	
Skin Analytics	84	3 8, 10 54 62	Abstract Scientific summary 3.2.7.3 Figure 7 3.3.6.1	Patient opinion - we challenge the conclusions drawn around patient resistance to using AI *'Patient and clinical opinions showed substantial resistance to using AI without any assessment of lesions by a dermatologist.'*	All evidence on this is clearly set out in the relevant sections of the report. See response to point 4. We note there is no contradiction between accepting AI alongside human assessment, but

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
		64	Figure 11 3.4.1.3	*'...there was very substantial resistance, particularly among clinicians,* *to using DERM without any assessment of lesions by a dermatologist.' 'The substantial resistance from both patients and clinicians to using AI without any human dermatological assessment'* - We strongly challenge this conclusion and do not believe it is supported by evidence within the report. Indeed, it also seems to be contradicted later in the scientific summary, where the report reads *'Patient and clinical opinions of DERM were generally favourable towards accepting its use as part of the diagnostic pathway.'* - Whilst the references supporting the statements regarding patient resistance are not clear, we suspect the conclusion is likely to have been drawn from the data presented in Figure 11 page 62. This appears to be from reference 36, although we were not able to find any related data in the linked paper[10]. If the reference is correct, the data relates to a study conducted outside of the UK and we would challenge the relevance within the intended scope of this EVA, which is focused on use within the NHS. Patients outside the UK may have healthcare expectations that differ significantly from those familiar with the NHS. Moreover, the referenced study, if correct, appears to focus on patients who are high-risk for melanoma and underwent Sequential	being more concerned about AI WITHOUT human assessment.

[10] Winkler JK, Tschandl P, Toberer F, Sies K, Fink C, Enk A, et al. Monitoring patients at risk for melanoma: may convolutional neural networks replace the strategy of sequential digital dermoscopy? Eur J Cancer 2022;160:180-8. https://doi.org/https://dx.doi.org/10.1016/j.ejca.2021.10.030

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				Digital Dermoscopy (SDD)[10], a service not routinely available within the NHS. - *'Participants generally responded positively when considering AI as a tool to help doctors, but more cautiously when considering the use of AI to replace a dermatologist.'* It is unclear as to what data the latter point is based on in relation to DERM. As far as we. None of the patient feedback we have seen and submitted within our response suggests patients are not open to autonomous AI. - We are also uncomfortable with the unreferenced statement that *'up to 50% of patients indicated they preferred a face-to-face dermatology appointment' (3.4.1.3 Patient and clinician perspectives, page 64)*. This needs to be referenced as we cannot see supporting evidence in the 'Leicestershire study'[11] or DERM-005 study[12]. If the latter, the results are presented as median and IQRs so we are unclear how it would have been possible for the EAG to calculate this. Based on data we have seen from our existing pathways and included in the data submitted to the EAG in November, we have seen >95% out of 1,000 patients have given optional consent under GDPR for DERM to make autonomous discharge decisions on their care	
Skin Analytics	85	35 11 5	2.3 Plain English summary	Medical Device - we are concerned by the apparent lack of detail and knowledge relating to medical device regulations	This is taken from the NICE scope, and is outside the EAG's remit.

[11] An evaluation of AI Powered Tele Dermatology for Skin Cancer 2WW Pathway - Edge Health
[12] Kawsar A, Kalsi D, Marsden H. Patient perspectives of artificial intelligence as a medical device in a skin cancer pathway. Frontiers in Medicine.;10:1259595. https://doi.org/10.3389/fmed.2023.1259595

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
			Scientific Summary	*'AI technologies with Class IIa designation (DERM, Moleanalyzer Pro)'* - This represents one example of concerns we have regarding the authors' lack of familiarity with the regulatory landscape of medical devices and potentially misleading conclusions drawn as a result. - UKCA Class IIa designation and EU Class IIa designation have different underlying legal frameworks and are not equivalent. This point was made during scoping of the EVA. - DERM is a UKCA marked Class IIa device and has an intended use which has been reviewed by our notified body and supports its use within autonomous pathways. - The European Union Medical Device Regulations differ and autonomous clinical decision making would require a device to be approved with Class III CE marking. - The plain English summary does not describe the regulatory clearance and the requirement of an approval process for the technologies to be considered; which is of paramount importance for patient audiences to understand when reading this report. The background does not mention anything with respect to medical device regulation or the status of the devices included despite this being a key criterion in the scope. Mentions of AI ought to be 'AI as a medical device (AIaMD).	
Skin Analytics	86	3 7 42 40 33 50	Abstract Scientific summary 3.2.3 2.2 3.2.6.1	Bias - we challenge a number of the concerns around bias and applicability raised by the authors, which may in part relate to a lack of understanding of medical device requirements *'All recruited highly selected populations and raised concerns about bias and applicability.'*	See response to comment 2.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
		112	7.3.1	*'All studies excluded a substantial proportion of participants from assessment, which may produce biased results.'* *'the applicability of the diagnostic accuracy results for DERM to current practice is uncertain'* *'systematic exclusion of a significant proportion of participants who would normally be assessed in practice meant that the evidence base for both devices was considered to be at high risk of bias and raises concerns about its applicability to practice.'* - Another concern we have regarding the lack of understanding of medical device regulations relates to the misinterpretation of exclusion criteria as bias. All medical devices need to have a clear and stated intended use including the intended population within which they are appropriate. This seems to have been misinterpreted as a means of introducing bias when the exclusion criteria is a deliberate acknowledgement that DERM should not be used in circumstances that may introduce variation in performance. *'All studies were at high risk of selection bias due to the exclusion of a significant proportion of participants (15.6% to 27.4% where reported) that would have otherwise been eligible for assessment in clinical practice. The performance of AI is likely to be significantly improved by the exclusion of some of these lesions (e.g. images with body hair, tattoos, subungual lesions).'* - The exclusion of these lesion types is to ensure the appropriate and safe use of a regulated medical device in line with its intended use. Inclusion and exclusion criteria exist for all medical devices with common examples including X-ray imaging which is contraindicated for pregnant women in the first trimester	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				of pregnancy. In addition, teledermatology also has recommended inclusion/exclusion criteria many of which overlap with those from DERM and relate to limitations with the location of skin lesions and ability to capture useful images.For example, lesions that are larger than the dermatoscope or are covered by a tattoo/scar cannot be fully visualised and may be inappropriate for both DERM and teledermatology. - Whilst some exclusions are likely to remain indefinitely (e.g. where the lesions are too large to be fully captured within a dermoscopic image), there is the possibility that other exclusions could be removed when sufficient data is available. Making this change would need to follow our regulatory processes with adequate evidence captured within our regulatory documentation, namely our Clinical Evaluation Report (CER), and made available for review by our notified body - We challenge the position that exclusion of lesions ineligible for DERM assessment introduces bias or leads to an overestimation of DERM accuracy. *'All studies raised concerns with regards to the applicability of their populations; the high rate of exclusion of participants with suspected lesions that would normally be seen in practice is a significant limitation.' 'A notable issue was the substantial number of lesions that could not be assessed using DERM.'* - DERM is used within clinical pathways where we work with our NHS partners to ensure all patients receive timely care. Patients whose lesion(s) meet the stated exclusion criteria still receive a teledermatology or in-person assessment of their lesion. This is not reflected	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				within the report and it is incorrect to suggest they are not seen in practice. As noted above, we cannot use a medical device where it is contraindicated. - Regarding *Table 4 Quality assessment of DERM diagnostic accuracy studies*, we do not understand why the index test bias is felt to be low for version A but high version B as stated. Thomas et al 2023 was a real world, post deployment evaluation that reported on the performance of two versions of DERM deployed in the same settings across prospectively captured patients and lesions[13] and so we are unclear why the bias would change from low to high. - We want to emphasise that the exclusion rate is appropriately reflected in the later health economic modelling which incorporates the impact of the exclusion rate and the fact that these patients are routed to teledermatology or in-person assessment. We are confident with the accuracy of our exclusion assumptions used as these have been based on nearly 4 years of deployment data across tens of thousands of patients and multiple sites. *'The teledermoscopy devices used in two studies (Dermlite DL1 Basic (DermLite LLC) system) were considered out of date following clinical advice and therefore raised concern about their applicability to current practice.' 'the dermatoscopes used for*	

[13] Thomas L, Hyde C, Mullarkey D, Greenhalgh J, Kalsi D, Ko J. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. Frontiers in Medicine. 2023;10.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				clinical assessments were out of date; therefore, the applicability of the diagnostic accuracy results to current practice is uncertain.' - We would like to clarify how many opinions were sought to support this clinical advice and that adequate consideration has been given to the risk of bias that may introduce, as the report says that *'Input from a clinical expert on the applicability of the studies was sought where appropriate'*, suggesting there may only be a singular opinion. - The dermatoscope in question is in its own right a Class I medical device available for use on the market and remains referenced on the British Association of Dermatologists website[14] with no mention of concerns around its use. With this dermatoscope, the sensitivities documented in the report have been achieved by DERM, which are above that of clinicians as found in the Cochrane reviews[7-9].	
Skin Analytics	87	4 8-9 83-84 87	Abstract Scientific summary 5.1.3 5.1.5	Health economic conclusions *'Several issues with the modelling approach were identified, particularly the mechanisms by which value is driven and how diagnostic accuracy evidence was used.' '...did not include an executable model.' 'may be very sensitive to the use of alternative sources of diagnostic accuracy data.' '...the model imposed disincentives for the correct diagnosis and treatment of BCC; structurally imposed assumed sensitivity benefits for any strategy incorporating a triage step; used costs associated with biopsy and treatment which were inconsistent with sources generally used in NICE appraisals, and may*	No response required.

[14] Dermatoscope Comparison [Internet]. British Association of Dermatologists; 2023 Nov [cited 2024 Jan 11]. Available from: https://badmainstage.wpengine.com/wp-content/uploads/2023/11/Dermatoscope-Comparison-table.pdf

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				overvalue specificity in terms of generating cost-savings'; 'potentially optimistic assumptions around the diagnostic accuracy of comparators, and of the surrounding pathway.' 'This punishes diagnostic strategies with lower specificity and may inflate the potential cost savings associated with higher-specificity strategies.' 'This evidence was preliminary and did not include an executable model.' - In the body of the report, it is stated that *'The submitted [economic] model represents the most recent and complete attempt to represent the NHS urgent skin cancer referral pathway'* but this is not reflected in the abstract nor the scientific or plain English summaries. - We are committed to building a robust health economic model and case for the use of DERM. We will update the model as suggested, run the relevant sensitivity analyses signposted and publish an updated report. While the current timelines for the EVA do not allow us to make these changes in timeframe given for comments on this draft report, we hope to do so within the overall timeframe of the EVA (April 2024) and would gladly make this model available to NICE and the EAG. - We agree that the specificity of teledermatology used in the Exeter paper[15] is different to that seen in the Cochrane reviews[7]. We highlighted this fact in the Exeter paper and provided the rationale and supporting evidence from a number of real world teledermatology services including non-Skin Analytics data. We suggest you seek expert guidance on whether using the Cochrane review specificity is appropriate as it would	

[15] Cost-effectiveness of DERM v1.0

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				result in around 74% of patients being discharged from care after teledermatology review. This simply does not reflect real world performance. - The comments that the model is punitive towards lower specificity implies favour towards DERM; however the specificity assumptions for both DERM and teledermatology in the model are both lower than lesion-level diagnostic accuracy data suggests, in order to best reflect the actual impact of their specificity on the whole pathway. In the most recent DERM performance report[16], DERM correctly labelled >7/10 lesions confirmed as benign by telederm or histology as benign; however in the economic model, the specificity assumption for DERM was 42%. - We recognise the model challenge with BCC, and struggled with this during design. We made this assumption based on expert guidance in the absence of available data on progression and cost. We will however aim to improve this in line with the conceptual model proposals. However we disagree that based on the modelled assumptions this has a meaningful impact on the cost utility of DERM. To illustrate, there are 3.5 additional false negative (FN) BCC per 1,000 patients in the DERM vs face-to-face scenarios (Exeter report, page 15). Weighted cost for BCC treatment is £556. Even assuming the treatment cost for a FN BCC was 10x a true positive (TP) BCC, this would still equate to only circa £20 of additional cost per patient, against a saving of £51.89 (£556 x 10 x 3.5 / 1000).	

[16] SA-003905-CS-2-_Overall_ DERM Clinical Performance Report Q3 2023

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				We also note that DERM sensitivity for BCC in real world use is far above the 90% assumption used (the latter reflects our claim in our Instructions for Use). It is likely that there is a very limited difference in pathway sensitivity for BCC between DERM and other models of care.	
Skin Analytics	88	50	3.2.6.1	The following statement is not accurate and needs to be corrected *'However, between 4.3% and 6.4% of lesions were judged as eligible for discharge by DERM vA and, following a second read and referral to the trust, were subsequently diagnosed as malignant.'* This appears to have been derived from Figure 3 in Thomas et al 2023[13] but unfortunately the data has been incorrectly interpreted. Whilst 4.3% (UHB) and 6.4% (WSFT) of the total case volumes were discharged by DERM and then referred to the trust by the second read dermatologist and not immediately discharged by the Trust teledermatology review; **only 0.1-0.15%** of the total case volume were discharged by DERM and ultimately diagnosed as a skin cancer.	We will remove this statement on editing.
Skin Analytics	89	3 4 8 11	Abstract Abstract Scientific summary Plain English summary	*'DERM had a sensitivity of 96.1% to detect any malignant lesion (95% confidence interval (CI) 95.4 to 96.8); at a specificity of 65.4% (95% CI 64.7 to 66.1). Diagnostic accuracy was similar for melanoma, squamous cell carcinoma (SCC) and basal cell carcinoma (BCC). For detecting benign lesions the sensitivity was 71.5% (95% CI 70.7 to 72.3) for a specificity of 86.2% (95% CI 85.4 to 87.0). MoleAnalyzer Pro had lower sensitivity, but higher specificity for detecting melanoma than face-to-face dermatologists.'* *'promising accuracy'* *'Moleanalyzer Pro alone had a sensitivity of 84.4% (95% CI 73.9 to 91.0)'*	We are not suggesting that DERM and Moleanalyzer can or should be directly compared: that would seem inappropriate given current evidence. We consider all our statements to be reasonable.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				'Moleanalyzer Pro could identify about 85% of malignant melanomas' - It would be good to understand the rationale for presenting sensitivity data for DERM but not Moleanalyzer Pro within the abstract and scientific summary. We are concerned regarding the risk of performance being misinterpreted and would ask for this to be made more clear. - The report later refers to both technologies as demonstrating '*promising*' diagnostic accuracy. This is surprising considering there is ~10% sensitivity difference and the confidence intervals do not overlap suggesting a statistically significant difference (84.4%, 95% CI 73.9 to 91.0 compared to 96.1%, 95% CI 95.4 to 96.8). - 85% sensitivity is mentioned in the plain English summary and juxtaposed against 95% sensitivity, suggesting they are comparable. Further, rounding of the 84.4% sensitivity should be 84%. The positioning of this in the patient-facing section of the report is concerning. - This is followed by *'This appears to be a similar accuracy to that achieved by dermatologists alone'*. Stating 85% sensitivity as comparable to dermatologists only to later call for more diagnostic accuracy data for DERM which has higher sensitivity across much more data (>20,000 lesions in latest quarterly report data) seems inconsistent. We would challenge the authors to make it clearer if they feel 85% sensitivity for melanoma is the acceptable benchmark (as they seem to use for Moleanalyzer Pro) as it should be	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				acknowledged that specificity would be expected to increase with a reduction in sensitivity.	
Skin Analytics	90	3 5	Abstract Scientific summary	*'...as decision aids to triage and diagnose suspicious skin lesions following a primary care referral.'* *'...as decision aids to triage and diagnosis of suspicious skin lesions following a referral on the urgent suspected skin cancer pathway.'* DERM used in its full potential as a class IIa device is not a decision aid but rather is used for automated clinical decision making	Not an error
Skin Analytics	91	3	Abstract	*'Evidence on the clinical impact of the technologies was limited.'* - The Leicestershire study data included a section with effects on Dermatology Service Capacity[11]. We are unclear whether this has been included within one of the redacted sections. If it has not been considered, we would like to understand why. In addition, in November we also submitted evidence demonstrating DERM pathways can increase conversion rates 40-100% which does not seem to have been taken into consideration within this report.	This was included. We note that it compared a Teledermatology pathway with DERM to face-to-face assessment; so did not establish the clinical impact of DERM per se.
Skin Analytics	92	4	Abstract	*'No published assessments of the cost-effectiveness of the technologies were identified'* This is not strictly true - our report with the University of Exeter and the Leicestershire study[11] are available on request.	We stress the "published" in this sentence.
Skin Analytics	93	5	Scientific summary	*'In current practice, patients with suspicious skin lesions are referred to secondary care through the urgent suspected skin cancer referral pathway, where people attend a secondary care dermatology department for a face-to-face appointment with a consultant dermatologist.'*	It is still general standard of care.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				This is no longer standard of care at a number of sites where they have adopted a teledermatology model, which should be clarified.	
Skin Analytics	94	5	Scientific summary	*'AI systems are already used in the NHS in a research context'* This needs to be corrected. DERM has been deployed in live clinical pathways since 2020 in line with intended use and regulatory clearance and not limited to a research environment.	Not factually incorrect. For example we note the continued use of "Second read"
Skin Analytics	95	7	Scientific summary	*'use of DERM may slow progress to diagnosis.'* This statement is made without context in the summary, and we suspect it is based on the Leicestershire study. Importantly, the authors of that study emphasised that there were other factors contributing including known capacity issues with booking and admin teams and subsequent delays in booking teledermatology appointments and face-to-face appointments.[11]	This is based on the Leicestershire study results
Skin Analytics	96	8	Scientific summary	*'there was very substantial resistance, particularly among clinicians, to using DERM without any assessment of lesions by a dermatologist.'* - The source of this data is not immediately obvious; we feel it should be referenced. The statement insinuates that there is substantial resistance to the use of DERM autonomously from patients for which there is no evidence. In fact, >95% out of 1,000 patients have given optional consent under GDPR for DERM to make autonomous discharge decisions on their care. This datapoint was shared along with a number of other updates in November.	The evidence is fully described in our report. See response to comment 4.
Skin Analytics	97	9 48 64 108	Scientific summary 3.2.5.1 3.4.1.1 7.1.1.1	*'Current evidence for …is lacking with regards to the diagnostic accuracy of the whole diagnostic pathway (i.e. inclusive of subsequent steps).'* *'Diagnostic accuracy of the full pathway is largely unknown. Data on assessments by dermatologists after DERM assessment were not reported in publications.' 'The diagnostic accuracy of the whole teledermatology*	We accept this evidence, but note it is insufficient to properly estimate true diagnostic accuracy.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				*pathway including DERM could not be assessed because of a lack of any independent reference standard of diagnosis, but there was some evidence that a "second read" of lesions classified as benign by DERM **********. This is a key area of uncertainty in assessing the actual clinical value of using DERM.'* *'The diagnostic accuracy of the whole teledermatology pathway including DERM could not be reliably assessed because of a lack of any independent reference standard of diagnosis.'* The Thomas et al (2023) paper specifically had analyses on repeat presentations of patients and found that *'No lesions have been assessed by DERM-vA or -vB and discharged from these pathways with a subsequent re-presentation and diagnosis of cancer (service sensitivity 100% to date); however, there have been four lesions that presented twice to the UHB pathway before July 2021, with the second presentation resulting in a histologic diagnosis of skin cancer (melanoma, n = 2; BCC, n = 2)'*[13]. It is part of our post-market surveillance programme to look for such cases and at the time of submission these are the only instances identified. We therefore challenge the statements above.	
Skin Analytics	98	9 11	Scientific summary Plain English summary	*'the practical impact and clinical benefit (notably the burden on clinicians and clinical impact on patients) of using DERM in a post-referral setting is currently unclear.'* - It is unclear as to what is meant by burden on clinicians and clinical impact on patients here. We do not follow the logic behind an increased burden on clinicians as a result of implementing an autonomous AI tool with the potential to discharge benign lesions without relying on Dermatologist capacity. As per the data submitted in November, DERM pathways can increase conversion	We are not suggesting an increased burden on clinicians (indeed quite the opposite). We cannot comment on data that was not supplied during the report timeframe

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				rates by 40-100%, emphasising the reduction in non-cancerous cases requiring Trust input. *'DERM could potentially identify 95% of all skin cancers but would require about half of all patients to be referred to dermatologists'* - The reduction of case volume by 50% is framed negatively as if this would not be a substantial capacity gain for an already under-resourced and over-stretched dermatology workforce. - The potential benefit of earlier reassurance for patients who do not have skin cancer has not been mentioned. In our November updates, we submitted the following: - Although there may be a number of initiatives as well as Skin Analytics to improve Trust performance with respect to Cancer Waiting Times (CWT) targets, we have seen the following in nationally-available data[17]: - Birmingham, where we deployed in 2020 and over 80% of 2WW referrals go via the Skin Analytics pathway, have met the 2WW target every month since April 2021, have met the FDS target every month (up to August) in 2023 and have met the FDS target every quarter since 2022-23 Q2;	

[17] NHS England. Statistics» Cancer waiting times [Internet]. england.nhs.uk. 2019 [cited 2023]. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				- W Suffolk, where over 85% of 2WW referrals go via the Skin Analytics pathway, had 25.3% 2WW performance in the quarter before launch and 95.8% in 2023-24-Q1; - Chelsea, where the vast majority of the Chelsea site 2WW referrals go via the Skin Analytics pathway, had 88.1% 2WW performance in the quarter before launch and 98.2% in the 2023-24-Q1 and continue to meet the FDS target every month; and - <47% and <32% of Bristol and Ashford 2WW referrals respectively went through the Skin Analytics pathway, making it more difficult to assess without additional high quality data provided by the sites. - We have also received feedback from sites that: - W Suffolk average wait time for patients having first appointment from their referral date has decreased by 4.9 days; and Chelsea average wait time for patients having first appointment from their referral date has decreased by 11 days (from 14 to 3 days)	

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Skin Analytics	99	10	Scientific summary	*'truly comparative evidence may also be required.'* The DERM-005 study compared DERM and teledermatology[18] and as noted above there is a substantial evidence base for the accuracy of DERM and several impracticalities to consider if trying to generate this evidence. We welcome the suggestion that future research needs can be *'achieved through continuations and extensions of existing ongoing pilot studies of DERM'*.	No response needed.
Skin Analytics	100	26	1.4.2	*'Following DERM assessment, lesions classed as high-risk are triaged to urgent virtual review by a hospital dermatologist, whilst lesions classed as low-risk are sent for remote review by a second reader (consultant dermatologist), who will either discharge the patient if in agreement with AI, or overturn the AI risk assessment and proceed with an urgent referral to a hospital dermatologist.'* This ought to appear in the next section - *1.4.3 Potential positioning of AI technologies in the pathway*	This is a formatting error that will be corrected.
Skin Analytics	101	33	2.2	*'Input from a clinical expert on the applicability of the studies was sought where appropriate.'* - We also wanted to ensure that appropriate consideration had been given and reassurance received that there is no conflict of interest. - We note that NICE Policy on declaring and managing interests for NICE advisory committees[19] also includes expert commentators. This policy includes non-financial professional and personal interests including where they:	The clinical expert is acknowledged in the report, and is a member of the committee for this EVA.

[18] Marsden, H, Kemos P, Venzi M, Noy M, Maheswaran S, Francis N. Accuracy of an Artificial Intelligence as a medical device as part of a UK-based skin cancer teledermatology service. Pending publication

[19] Policy on declaring and managing interests for NICE advisory committees [Internet]. National Institute of Health and Care Excellence. 2022 [cited 2019 Dec 26]. Available from: https://www.nice.org.uk/Media/Default/About/Who-we-are/Policies-and-procedures/Declaring-managing-interests-for-advisory-committees.docx

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				- 'holds office or a position of authority in a professional organisation such as a royal college, a university, charity, or advocacy group' - 'has published a clear opinion about the matter under consideration' We seek reassurance that any statements attributed to either the individual or affiliated groups have been declared and considered.	
Skin Analytics	102	36	2.4, 3.2.5	- Section 2.4 suggests assessment of evidence submitted in November 2023 was more limited. Review could have helped to avoid some of the incorrect assumptions which have been made ○ E.g. Section 3.2.5 assumptions around overlap of data from DERM005 and Thomas 2023[13]. E.g. *'No evidence, published or unpublished, was identified for numbers of patients transferred to surgery, or test failure rates.' (p.50, section 3.2.6)* - this was sent in the November updates and there is data on this presented in the real world deployment paper[13].	The timeframe for the EVA was extremely limited. We have clearly stated the restrictions on assessing submitted evidence that resulted from this short timeframe.
Skin Analytics	103	42-43	3.2.3, Table 4, 3.2.4	*'As DERM-vB appears to have superseded DERM-vA we only report results for the more recent DERM-vB.'* Both sets of results are relevant given they are real world, post-deployment and prospective. Moreover, vA has been called low risk of bias but is subsequently removed from analysis. Both ought to be included. The update to vB can be considered a reflection of robust post-market surveillance processes that enable thorough, hypothesis-driven updates to DERM.	We note this that it is reasonable to base assessment on the current version of the tool in use and to discard findings from older versions. We note that this should favour the technology, as the aim is to exclude less accurate older versions.
Skin Analytics	104	42 112	3.2.3 7.3.1	*'In response to a clarification request, the company noted that the versions of DERM used in all three studies… were older than the current version used in the UK….Therefore, the applicability of the diagnostic accuracy results for DERM to current practice is*	See response to comment 103.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				uncertain.' 'The DERM versions (in particular, the set sensitivity/specificity thresholds)...therefore, the applicability of the diagnostic accuracy results to current practice is uncertain.' We are disappointed to see these statements included without any attempts made to understand the internal validation processes and documentation completed in line with our medical device responsibilities to ensure that the new model performance is well understood and the company is satisfied it reflects an improvement compared to older versions.	
Skin Analytics	105	42	3.2.3	*'One study (DERM-003) did not report sufficient details on the conduct of the reference standard and was at unclear risk of bias for this domain.'* It is unclear as to what the expected details were for this. The DERM-003 paper states that *'Where a biopsy was taken, the histopathology-confirmed diagnosis was collected and categorized as melanoma, SCC, BCC, IEC, Actinic Keratosis (AK), Atypical, Benign or other. When there was histopathological uncertainty in the diagnosis, investigators reported the most likely diagnosis. 'Other' diagnoses were reviewed by the Chief Investigator.'*[20] This is in line with other studies in this field.	Concerns regarding the reference standard did not relate to histopathological diagnosis, but to the final diagnosis of non-excised lesions. We understand that 188 final diagnoses were based on clinical judgment alone in DERM-003. For these lesions, was the final/reference standard diagnosis solely based on the same clinical assessment as per the non-AI/clinical assessment index test? Was the risk of incorporation bias addressed, and if so, how? The EAG would welcome further clarification from the company.
Skin Analytics	106	42	3.2.3	*'includes a different set of thresholds for sensitivity and specificity'* An updated algorithm will have different threshold values so these are not something you could or would keep exactly the same between versions; however the algorithms were all optimised to be above the sensitivity targets of 95% for melanoma and SCC and 90% for BCC, IEC and AK.	Not an error.

[20] Marsden H, Morgan C, Austin S, DeGiovanni C, Venzi M, Kemos P, Greenhalgh J, Mullarkey D, Palamaras I. Effectiveness of an image analyzing AI-based Digital Health Technology to identify Non-Melanoma Skin Cancer and other skin lesions: results of the DERM-003 study. Frontiers in Medicine. 2023;10.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
Skin Analytics	107	43	2.3.4	*West Suffolk not *'Sussex'*	We will correct this error.
Skin Analytics	108	45	3.2.4	*'In DERM-003, for detecting benign lesions, the sensitivity of DERM was significantly lower compared with face-to-face dermatologist assessment (DERM: 43.9% (95% CI 37.4 to 50.6); dermatologist: 73.9% (95% CI 67.6–79.4) although it had comparable specificity (DERM: 93.3% (95% CI 90.0–95.6); dermatologist: 93.7% (95% CI 90.5-95.9)). Hence around 56% of benign lesions were classified as not benign by DERM, compared with 26% for dermatologists, and approximately 7% and 6% of non-benign (but mostly pre-malignant) lesions were misclassified as benign by DERM and dermatologists respectively'* The DERM-003 paper says *'It should be noted that for non-biopsied lesions, the clinical diagnosis was used as the ground truth against which both the AIaMD and clinical diagnosis were compared. Clinical diagnosis therefore will appear more accurate in an all-lesion population, compared to a biopsy-only population, for those lesions where a high proportion do not have a histopathology diagnosis, specifically BCC, AK, and benign lesions.'*[20] However this limitation is not expressed in the EVA report.	This issue was considered in the quality assessment. See response to comment 105 for further details.
Skin Analytics	109	45	3.2.5	'*We assume this includes some patients from DERM-005, although the overlap is unclear.'* - There was no overlap of data between DERM-005[18] and Thomas et al 2023[13]. We would be happy to spend time helping to clarify any other assumptions relating to this data. We are also unclear why this section only makes reference to Birmingham and Chelsea & Westminster study centres as we	This sentence refers to overlap between the unpublished data and DERM-005, not Thomas 2023 and DERM-005 The Q3 2022 data we received includes data from 3 "care settings" not 7.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				also shared our overall Q3 2022 report which included performance across 7 NHS sites[6].	
Skin Analytics	110	44 46 47 45 49 50 64 108	3.2.4 Figure 5 3.2.5 3.2.5.2 Table 7 3.4.1.1 7.1.1.1	*'suggesting that DERM has some difficulty in distinguishing these types of cancer from benign lesions.'* - While we understand the desire to analyse the sensitivity and specificity of DERM for all lesion types is of interest; however the scope of this EVA and the deployment model of DERM are about the safe removal of benign lesions from the urgent suspected skin cancer pathway. This should mean a focus on high sensitivity for the cancers overall in terms of them staying on the pathway and a good specificity for identifying benign lesions that creates value and capacity for dermatology teams. - There is some additional analysis conducted by the EAG which we feel needs additional clarification to ensure the limitations are understood. *'When using the "Exact" classification there is a decrease in accuracy. For melanoma the sensitivity remains at near 95%, but for SCC and BCC the sensitivity declines substantially. This suggests that both SCC and BCC lesions may be being misclassified as more serious malignancies by DERM (i.e. SCC as melanoma and BCC as SCC or melanoma).' 'There was some evidence that DERM might tend to misdiagnose BCC, with many BCC cases being classified as SCC or melanoma.'* - We strongly challenge the relevance of the analysis conducted on "exact" diagnostic accuracy. This does not reflect how DERM is used, nor optimised and does not capture the patient pathway. In fact, the report even *'note[s] that this may not be exactly what might happen in practice'*. Without significant clarification there is a risk	See response to comment 1. We agree the "exact" approach may not be used in practice. It is included to fully assess the diagnostic accuracy of DERM.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				of DERM's performance being misinterpreted. Moreover the scope of the EVA is about identifying benign lesions in urgent skin cancer pathways and these analyses do not focus on nor reflect that. - DERM has consciously been built with a lesion hierarchy in place to prioritise patient safety and we optimise our threshold settings to achieve a high pathway sensitivity to make sure patients lesions are dealt with appropriately. - This is also given inordinate prominence in the 'statement of principal findings' including the statement that *'There was some evidence that DERM might misdiagnose BCC cases as SCC or melanoma.'*, which is not relevant to the focus and scope of the EVA as all of these cases are appropriately kept on skin cancer pathways for management by dermatology teams. *'Most referrals would be false positives, with around 64% of all referrals being benign lesions.'* All of these lesions have already been referred to the urgent suspected skin cancer pathway by GPs which should be noted. We would suggest that this also needs to be presented alongside some comparison to the current standard of care with false positives from GP + teledermatology also considered.	We agree.
Skin Analytics	111	50	3.2.6.1	*'There were some differences between the two locations in terms of rate of use of DERM and referral rates, suggesting that use of DERM may vary by location.'* - This is true of clinical referral rates too and ought to be juxtaposed against that. Local conversion rates are available in national datasets and the data in the real world deployment paper demonstrates this for teledermatology behaviour too.[13]	Not an error.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				'A notable issue was the substantial number of lesions that could not be assessed using DERM.' As covered above in comment #5, it seems excessive to consider this a notable issue when considering the rationale behind exclusion criteria (contraindications of a regulated medical device to ensure safety) and the fact these patients are still assessed by the pathway (via teledermatology or in-person). Table 8 could also make this more clear.	
Skin Analytics	112	64	3.4.1.2	*'The EAG identified very limited published evidence on any clinical impact'* - The Leicestershire study data included a section with effects on Dermatology Service Capacity. We are unclear whether this has been included within one of the redacted sections. If it has, then we would be disappointed that this evidence is still considered to be 'very limited'. If it has not been considered, we would like to understand why. - *Trend analysis suggests there is some evidence that increased usage of the post-intervention pathway has led to a reduction in 2WW referrals at UHL more generally. This **reduction in 2WW referrals appears to also be correlated with increased activity on other pathways suggesting that the intervention has helped increase capacity more generally**.* We also wonder whether the suggestion that further research is needed to ascertain the impact of releasing consultant time is excessive. It seems intuitive that reducing the number of benign lesions requiring assessment by Dermatologist will enable their time to be diverted to other patients, e.g. on routine pathways which have greater backlogs. The wide variation in how this capacity gain might be used makes it possibly impractical to quantify and seems beyond the scope of the EVA given it is supposed to be focused on whether the AIaMD can identify	We stress the word "published" in this sentence. This evidence should emerge naturally in any clinical study of DERM. It is important to collect this data so the economic impact of using DERM can be properly assessed.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				benign lesions safely and accurately in the suspected skin cancer pathway.	
Skin Analytics	113	64 103	3.4.1.3 6.3	*'there was a concern that AI used as a decision-aid was increasing patient' time on the diagnostic pathway. However, the evidence is limited to very small samples of responders.' 'the potential for lengthened waiting times as seen in the UHL pilot.'* We suspect this relates to survey data within the Leicestershire study, though we note has not been redacted unlike in comment #31. In that report, the authors were clear that they could not conclude causation with many factors at play including capacity issues with booking and admin teams.	This is based on Leicestershire data.
Skin Analytics	114	80	5.1.1	*'The model assumed that disease stage of melanoma at the point of diagnosis would also apply to SCC and other rare cancers. Evidence supporting this assumption was not in the presented the provided report.'* We appreciate this is a limitation of the model and could have been made clearer in the report; however this is rooted in the lack of data concerning non-melanoma skin cancers. Data for NMSC is notoriously poor and has so far not been included in UK or other national cancer data programmes (though this is starting to happen in the UK now - see cancerdata.nhs.uk).	No response required.
Skin Analytics	115	85	5.1.3.1	*'DERM with a second read may be the most costly approach, but may be associated with non-cash releasing benefits related to outsourcing of teledermatology review to Skin Analytics consultants.'* - No examples of such benefits are given and this seems to imply that there are benefits to outsourcing in general. Outsourcing is a solution that has been tested for many years that has not solved the problem of substantial skin cancer referrals compared to dermatologist capacity. In fact the GIRFT Dermatology report from 2022 stated	Reference to the potential (but unquantified) benefits of releasing consultant resource such as reduced waiting times and improvements in quality of care across wider dermatological indications are made throughout the EAG Report. These may be reiterated more explicitly in Section 8.1 on editing of the report. This response is noted - reference to potential benefits of outsourcing here simply speculates

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				that *'there were 659 consultant dermatologists working in the NHS in England (508 whole time equivalents), with 159 WTE consultant vacancies and more than 140 locums at the time of the review'*[2]. This is something that we feel ought to have been reflected in a description of the need in dermatology as per comment #1. If the suggestion is that the outsourced second read review might help downstream clinical decision making, this assumption is misplaced. Our experience suggests this can introduce friction in terms of difference of clinical opinion; for example, in our real world deployment paper, we saw that 39% of second read dermatologist referrals were discharged on hospital dermatologist review, which is a notable and potentially unnecessary increase in case volume.[13]	that NHS clinician time may be used more effectively elsewhere and thus generate some additional but potentially unquantifiable benefit. The provided reference may be acknowledged in a future iteration of the EAG Report.
Skin Analytics	116	87	5.1.5	*'DERM is less sensitive for BCC than teledermatology or face-to-face assessment.'* We want the committee to note that DERM's sensitivity for BCC in real world use is far above the 90% assumption used within the health economic model which reflects our claim in our IFU. Our Q3 DERM Performance Reports with >1500 BCCs demonstrate DERM was >97% sensitive for BCC while Cochrane reviews showed telederm to be 95% and F2F to be 93%. It is likely that there is a very limited difference in sensitivity for BCC between DERM and other models of care.	No response required.
Skin Analytics	117	88	5.1.5	*'The specificity of teledermatology reported in published sources is substantially higher than that observed in the pilot sites (which were largely not set up for teledermatology services).'* - The rationale for the assumptions regarding teledermatology specificity is thoroughly described in our health economic report and further in comment #6 above The comment here in brackets implies that only data from sites with DERM and teledermatology were used; however as shown	Noted – this point will be acknowledged when editing the report.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				in Table 2 of our report, there are 6 datapoints from case studies of teledermatology-only deployments, only 2 of which relate to Skin Analytics and neither of these 2 had DERM involvement.	
Skin Analytics	118	91	6.1.1	*'The modelled population should include all patients referred on the 2WW skin cancer pathway from primary care. The prevalence of cancer subtypes should be sourced from appropriate and recent UK national sources. Staging of disease at the point of entry into the model should be based on UK data if available. If there are differences by stage at presentation according to indication, this should also be reflected.'* - Our model does include all patients on the 2WW skin cancer pathway but takes account of ineligibility of patients for teledermatology or DERM assessment and subsequent routing of those patients to face-to-face assessment. There is some UK data on NMSC subtypes and staging however NMSC data is notoriously lacking and of poor quality. As UK data on this matures, so can the assumptions for these models.	This point is not a critique of the Skin Analytics model, and instead refers to the ideal data sources the proposed model should be using.
Skin Analytics	119	92	6.1.2	*'Given that these technologies may have limited experience of rare cancers, there remains uncertainty as to whether their high sensitivity to melanoma and SCC is maintained across these rarer indications'* - We appreciate this is an important point but it lacks the context we had shared in November: - DERM does not screen for any lesions not listed in the "Product Information - Device Inputs and Outputs" section of the Instructions for Use (e.g Merkel Cell Carcinoma)[21] therefore there is no target sensitivity for these; however we do monitor DERM performance for rare skin cancers.	This is a qualitative point about uncertainty and is equally applicable to all AI technologies. We appreciate the issues specific to DERM and the challenges associated with generating evidence in this area.

[21] SA-001165-LB-9-DERM Class IIa Instructions for Use

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				- These are often labelled as suspected cancer by DERM and the latest post-market surveillance reports based on 21,218 lesions assessed with ground truth outcomes across 7 sites shows that DERM has been 88.2% (30/34, 73.4-95.3%) sensitive for rare skin cancers since April 2022[22]. - This includes 6x Merkel cell carcinoma across 3 sites which were all labelled suspicious by DERM (100% sensitivity, 95% CI 61-100%). - The other rare skin cancers included were: 6x sebaceous carcinoma, 5x atypical fibroxanthoma, 4x dermal sarcoma, 3x chronic lymphocytic leukaemia / small lymphocytic lymphoma, 2x porocarcinoma, 1x spitzoid tumour of uncertain malignant potential, 1x angiosarcoma, 1x atypical cutaneous fibrous histiocytoma, 1x Sarcomatoid carcinoma / spindle cell carcinoma, 1x cellular spindle cell lesion, 1x myeloma, 1x follicular lymphoma and 1x sweat gland carcinoma. - Cochrane reviews of teledermatology and face-to-face diagnostic do not report on diagnostic accuracy of rare skin cancer due to lack of data[7-9] however a UK study looking at clinical recognition of Merkel Cell Carcinoma found that 4 Merkel cell carcinoma, grouped with adnexal neoplasms and lentigo maligna (due to small sample size) were accurately diagnosed in 30% of these combined cases and a Canadian study in which there were 65 Merkel cell carcinoma found dermatologist diagnostic accuracy to be 31%.	

[22] SA-003905-CS-2-_Overall_ DERM Clinical Performance Report Q3 2023

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				Given the low incidence of rare skin cancers (3.1 per 100,000, 95% CI 3.0-3.2, in the UK in 2018-2020) and even lower incidence of specific rare skin cancers (e.g. 0.62, 95% CI 0.58-0.66 for Merkel cell carcinoma in the UK in 2018-2020) as opposed to skin cancer as a whole (387 per 100,000, 95% CI 386-388), it is not feasible to run clinical trials over sufficient timelines and breadth to study DERM diagnostic accuracy of rare skin cancers; however we continue to monitor DERM performance in the real world in great detail such that we would eventually be able to reach a statistically significant sample size of cases seen to evaluate this.	
Skin Analytics	120	107	6.6	*'the clinical evidence identified in Section 3, was based on heterogenous pathways and settings and may not provide appropriate diagnostic accuracy inputs for the pathway described in this model.'* This is described only as a limitation but fails to acknowledge the context in the UK of heterogeneity of current clinical pathways (e.g. some have teledermatology and some do not while others have a combination) and fails to acknowledge this as an advantage having had diagnostic accuracy data from a variety of deployments with the same focus of safe removal of benign lesions from urgent suspected skin cancer pathways.	This is a limitation with respect to the modelling exercise in a NICE context as more granular and specific data may be required to generate truly representative comparisons of the alternative diagnostic models.
Skin Analytics	121	112	7.3.1	*'Most patients included in diagnostic accuracy had lighter skin colours (Fitzpatrick types II-III).'* This ought to reference that this is in line with skin cancer incidence as expected.	This is a simple statement of fact.
Skin Analytics	122	114	7.5	*'The evidence base for both technologies included few patients with non-white ethnicity or darker skin tones… Differences in diagnostic accuracy could lead to inequalities due to different diagnostic pathways, such as if some people have to wait for a*	This is a simple statement of fact, and was raised at scoping as an area of concern.

Stakeholder	Comment no.	Page no.	Section no.	Comment	EAG Response
				face-to-face appointment because an AI assessment was inconclusive.' - Currently, less than 0.5% of melanoma and non-melanoma skin cancer diagnosed in the UK are in Black and Asian patients[23]. Although the sample sizes are low, all cancers found in patients identified as having Fitzpatrick V or VI skin in our pathways have been identified by DERM[24]. In these pathways, patients with acral lesions, where melanoma are more common in darker skin tones, are not assessed by DERM and routed straight to Trust dermatology assessment in order to maximise patient safety. There is a lack of data with respect to the diagnostic accuracy of standard of care in patients with darker skin tones as well as other subgroups.	

[23] Delon, C., Brown, K.F., Payne, N.W.S. et al. Differences in cancer incidence by broad ethnic group in England, 2013–2017. Br J Cancer 126, 1765–1773 (2022). https://doi.org/10.1038/s41416-022-01718-5

[24] SA-003662-CS-2-DERM Equality and Health Inequalities Assessment