Edge Health

NHSE Outpatient Recovery and Transformation Programme

Evaluating Pathways for AI Dermatology in Skin Cancer Detection: A White Paper

31/07/2024

Contents

Executive summary

The integration of technology in the healthcare sector promises to enhance best practices, improve productivity, and bolster patient outcomes across numerous specialties. Dermatology is encountering advancements like its counterparts, with artificial intelligence as a medical device (AIaMD) set to reshape healthcare delivery and relieve the increasing mismatch between capacity and demand. This report aims to evaluate the adoption of autonomous AIaMD in suspected skin cancer pathways, focusing on performance, current implementation, and economic considerations. It is also the first to practically assess safety standards and recommendations for the post-market surveillance (PMS) of autonomously used AI.

AIaMD is already employed under human supervision in skin cancer pathways and can be used autonomously in the NHS if certified under classes UKCA IIa and CE III. While autonomous use offers the greatest productivity benefits, a lack of independently assessed real-world data and comparative clinical performance in excluding melanoma have hampered progress. To address this, we evaluated current clinical diagnostic performance by carrying out a meta-analysis of dermatologist accuracy in excluding melanoma, focusing on Negative Predictive Values (NPV). We then measured the current AIaMD approved for use within NHS skin cancer pathways (DERM, Deep Ensemble for the Recognition of Malignancy) against this standard, to assess its safety and set a precedent for evaluating future AI technologies.

Our meta-analysis revealed that dermatologists ruled out melanoma (malignant and in situ) with an NPV of 98.9%, in face-to-face clinical settings. As such, a benchmark NPV of 99% would be a pragmatic target for AIaMD intended for use in triage of skin cancer. DERM achieved an NPV of 99.8% at similar disease prevalence, demonstrating a performance at least as good as that of face-to-face dermatologist evaluations. This report provides a framework for the assessment of the safety and performance of current and future AIaMDs in view of any evolution within literature and tools available on the market.

The pathways analysis within this report provides a high-level review of current models of care and examines providers' experiences of deployment. We explored current implementation strategies and their differences and reported three providers' reflections on benefits, challenges, and practical solutions.

For AIaMD to be implemented safely, its deployment must be underpinned by a solid framework for PMS and use within approved regulatory guidance. We have outlined recommendations to support the adoption and ongoing surveillance of AIaMDs, including a balance of proactive and reactive monitoring activities, data-sharing practices, and an example of a practical auditing methodology. These recommendations are informed by extensive literature and example analysis of practical methods aimed at establishing AI as a reliable and safe adjunct in skin cancer diagnosis and care.

Finally, our illustrative budget impact analysis provides an overview of preliminary system-level savings, highlighting a return of up to £2.3 in savings for each £1 spent. Additionally, AIaMD can rapidly process initial assessments, which could reduce waiting times for secondary care reviews, thereby enhancing patient experience and service delivery.

1 Introduction

The application of artificial intelligence as a medical device (AIaMD) in healthcare has the potential to support clinical practice, improve service efficiency, and positively impact patient outcomes. Dermatology can particularly benefit from the support of AIaMD in managing highly sought-after diagnostic tasks and addressing the increasing demands placed on specialists.

The utilisation of AIaMD is of significant interest in skin cancer care, where melanoma incidence rates in the UK have already increased by 140% since the early 1990s and are expected to rise by 9% from 2023-2025 to 2038-2040[1], resulting in unprecedented demand for services.

Dermatology services are currently managing the dual challenge of a growing patient demand and consultant shortages, leading to extended waiting times for routine referral-to-treatment (RTT) pathways. As of April 2024, these averaged 17 weeks, with only 60% of pathways being completed in 18 weeks or less (against a national target of 92%)[2]. Additionally, the last years have seen a decline in the number of cancer patients diagnosed and treated within the target times of 28 and 62 days respectively, particularly following the summer when demand surges[3].

In 2018, 23% of all melanoma diagnoses arose from routine GP referrals, and more recent (unvalidated) National Disease Registration Service (NDRS) data suggests this figure may have risen[4]. These figures underscore the need to alleviate bottlenecks in both urgent-suspected cancer (USC) and routine dermatology referrals, which is central to improving patient care and service efficiency.

Skin cancer, while of significant concern, represents only a fraction of the dermatological landscape that specialists navigate. Dermatologists are tasked with caring for a multitude of skin conditions, highlighting the need for access to comprehensive dermatological services, which risk suffering when all attention is given to tackling the USC backlog. Within this landscape, several innovative solutions have been put forward to address rising demand and improve equity of care, including teledermatology, image-assisted advice and guidance, and AI tools.

AI has taken on an increasing role within skin cancer pathways since the COVID pandemic, as rising demand and the need to avoid patient attendance to hospitals provided grounds for innovation. Since then, the role of AI in Dermatology has expanded to include the first UKCA Class IIa AIaMD tool being approved for use in the NHS – which is currently live at 19 sites across England – as well as the development of several other devices for skin cancer triaging and detection, ranging from mobile phone apps to full-body scanning appliances.

As AIaMD moves from pilots to business-as-usual and shows promise for its autonomous use, the question of safety and regulation becomes increasingly relevant. There is a need to establish appropriate pathways and use cases, the standards it should perform at, and ensure

[1] Melanoma skin cancer statistics, Cancer Research UK. Accessed June 2024.
[2] Referral To Treatment Waiting Times, NHS England. Accessed June 2024.
[3] Cancer Waiting Times, NHS England. Accessed June 2024.
[4] NDRS, Get Data Out programme, and COVID-19 rapid cancer registration and treatment data. Accessed June 2024. [August 2024 Addendum: Correction to the paragraph to quote the validated Get Data Out data, rather than the unvalidated NDRS RCRD dataset.]

that surveillance methods are in place to monitor its outputs. In addition, integrating any AI technology into clinical workflows must be approached with attention to safeguarding patient safety.

This report is the first to set a precedent for the evaluation of the autonomous use of present and prospective AI solutions in skin cancer pathways and how to determine the standards of care they must meet. It will examine current implementation strategies, touch upon economic implications, and outline a recommended structure for post-market surveillance to monitor the long-term performance and reliability of autonomous AI technologies.

1.1 Rising Demand for Dermatology

In recent years, Dermatology services within the NHS have experienced a persistent rise in demand, affecting both routine consultations and the management of suspected cancer cases. This upsurge has been consistent over the past decade, indicative of an evolving healthcare challenge.

Figure 1 illustrates this trend by showing an 82% increase in the RTT waiting lists for Dermatology between April 2021 and March 2024. Furthermore, the rate of USC referrals in England, detailed in Figure 2, has escalated by 170% within the last ten years.

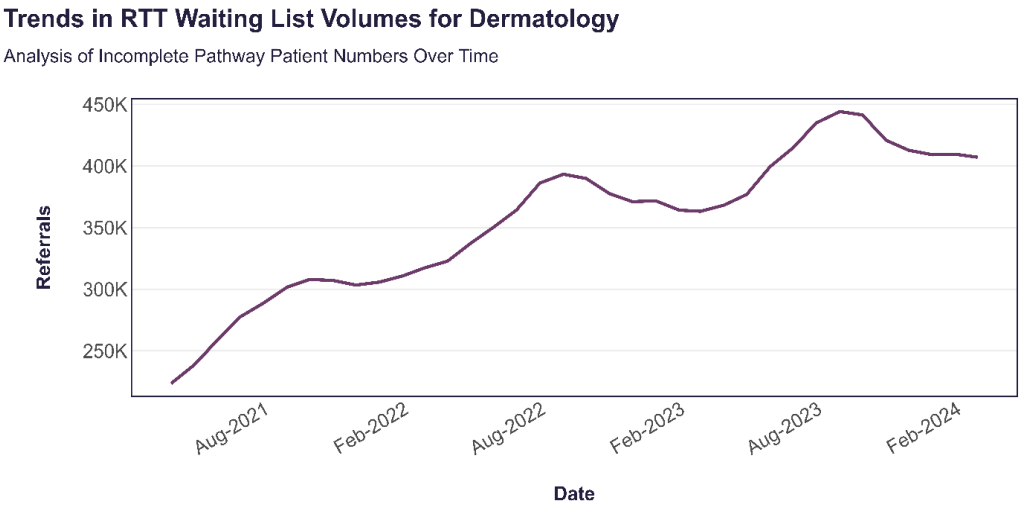


Figure 1. RTT Waiting List Volumes for Dermatology, April 2021 – March 2024

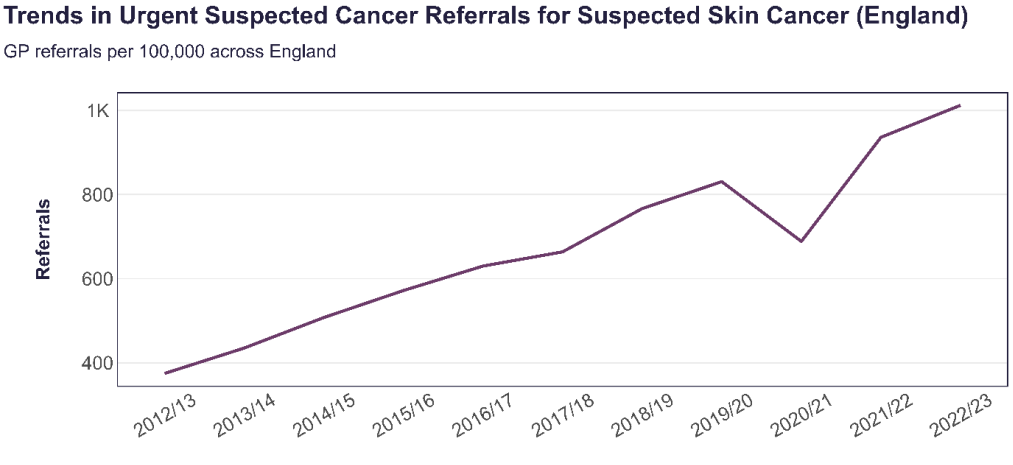


Figure 2. GP Referral Rate for USC Referrals for Skin Cancer, FY 2012/13 – FY 2022/23

Compounding this issue is a shortage of Dermatology professionals, as detailed in the 2021 Getting It Right First Time (GIRFT) Dermatology report[5]. At the time of the report, 24% of Dermatology consultant positions remained vacant, with only 508 whole-time-equivalent (WTE) consultants in posts and 159 WTE vacancies.

The report recommended an increase in Dermatology training posts, which temporarily increased to 30 in 2021 and 41 in 2022. Despite so, this number fell to 32 across England in 2023[6].

The shortfall in qualified dermatologists combined with rising demand has contributed to lengthier waiting times for patients. In 2016, the average wait for a Dermatology RTT was seven weeks, but by 2024, this had risen to 17 weeks[7]. While USC skin cancer referrals are among the few that regularly meet the 28-Day Faster Diagnosis Standard (FDS), the subsequent treatments struggle to keep pace, with only 80% of the 62-Day targets being met in October 2023 after a spike in referrals during summer, falling short of the national standard by 5%[8].

The protracted waits in routine referral lists are not without consequence; they increase the risk of adverse outcomes for cancer patients. Data from the NDRS reveals that, in 2018, 37% of cancers originated from routes other than USC (then two-week-wait), including routine GP referrals. More recent unvalidated data suggests this figure may now be as high as 48%[9]. Delays in USC pathways have a knock-on effect on routine Dermatology services, potentially disadvantaging patients with other skin conditions.

This context underscores the urgent need to address these challenges. A logical and sustainable approach would call for an increase in the number of training positions for dermatologists to meet the growing demand and to align with future projections concerning the prevalence of skin conditions. Nonetheless, it is worth noting the extensive timeline required to train a consultant dermatologist, which can extend beyond 15 years[10]. Therefore, while long-term solutions are essential, the implementation of immediate, short-term interventions is equally imperative to alleviate the current burden on Dermatology services.

1.2 The AIaMD Landscape for Skin Cancer Diagnosis

In the UK, the Medicines and Healthcare Products Regulatory Agency (MHRA), part of the Department of Health and Social Care, oversees the medical devices market. In September 2021, the MHRA initiated the Software and AI as a Medical Device (SaMD or AIaMD) Change Programme[11], outlining its strategy for future software and AIaMD regulations. By January 2024, the MHRA had released a regulatory roadmap projecting their approach to post-market surveillance (PMS) from 2021 to 2025. Under the Medicines and Medical Devices Act[12], they

[5] GIRFT Dermatology Report, September 2021
[6] Medical Specialty Recruitment Competition ratios, NHS England.
[7] Referral To Treatment Waiting Times, NHS England. Accessed June 2024.
[8] Cancer Waiting Times, NHS England. Accessed June 2024.
[9] NDRS. [August 2024 Addendum: Correction as per footnote 4]
[10] 6 years of medical school, followed by 2 years of Foundation training, 3 of Internal Medical Training, and a minimum of 4 years Dermatology training, without any interruptions.
[11] Software and AI as a Medical Device Change Programme – Roadmap, MHRA
[12] Medicines and Medical Devices Act 2021, UK Public General Acts

are updating existing legislation and will introduce new PMS requirements at the end of 2024[13]. The regulatory framework is swiftly progressing, leading to a potential lack of clarity regarding the appropriate regulatory status for AIaMD use. In February 2024, NHS England laid out a requirement for devices to possess either a UKCA Class IIa or CE[14] Class III certification and appropriate intended use for autonomous direct diagnostic use within the NHS[15].

The current range of AI tools that meet these regulatory demands while also providing evidence of real-world application is limited. Nonetheless, the anticipation is that the number of tools will grow in the future. For illustrative purposes, a snapshot of existing technologies reviewed by NICE[16] and developed since 2022 is provided below.

Technology	Intended Use Statement	MHRA status	Autonomous Approval
DERM (Skin Analytics)	DERM is an artificial intelligence (AI)--based skin lesion analysis device intended for use in the screening, triage and assessment of skin lesions suspicious of skin cancer. DERM will analyse a dermoscopic image of a skin lesion and return a suspected diagnosis and, if applicable, a referral recommendation for the lesion.[17]	UKCA Class IIa	Yes
DermaSensor	The DermaSensor device is indicated for use to evaluate skin lesions suggestive of melanoma, basal cell carcinoma, and/or squamous cell carcinoma in patients aged 40 and above to assist in the decision regarding referral of the patient to a dermatologist. [18]	FDA Class II but no UKCA/CE	No
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.[17]	CE Class IIa	No
Nomela (Moletest Scotland)	A non-invasive diagnostic aid to indicate the probability of melanoma in pigmented skin lesions (moles), is a software medical device installed on single-application iPads applying machine-learning AI to captured images; for use, after training, by medical professionals and intended as an adjunct screening technology in the clinical pathway of the management of suspect lesions.[19]	CE Class I	No

[13] Roadmap towards the future regulatory framework for medical devices 9th January 2024, MHRA
[14] Note that CE (European) certifications are approved for use in UK until 30 June 2030, MHRA.
[15] National Outpatient Recovery & Transformation programme – Expression of Interest in acquiring the use of AI in the urgent suspected skin cancer pathway, February 2024. FAQs
[16] Digital technologies for the detection of melanoma, NICE, November 2022
[17] NIHR Reference No. NIHR136014
[18] FDA. DermaSensor Regulation
[19] Nomela, Intended Use

Technology	Intended Use Statement	MHRA status	Autonomous Approval
SkinVision	The SkinVision Service is a software-only, over-the-counter (OTC), mobile medical application, which is intended for use on consumer mobile devices by laypeople. The SkinVision Service is not intended for use on persons under the age of 18 years old. The SkinVision Service does not diagnose skin cancer, nor does it provide any other diagnosis.	CE Class I	No

Table 1. Summary of AI technologies for skin cancer detection, their intended usage, and MHRA regulatory status

AI technology holds promises for skin cancer treatment pathways. Its autonomous use especially supports pressured departments and workforce shortages by decreasing the time clinicians need to spend on patient assessment and allowing for quicker reassurances to patients and more efficient use of specialist resources. Such a shift enables dermatologists to better balance their time between urgent and non-urgent cases, effectively narrowing the gap between healthcare demand and provision.

However, it is critical to appreciate that AI tools are meant to support, not replace, the expertise of dermatologists and the ongoing training of general practitioners. Collaborative approaches to enhance clinical assessments should go hand in hand with longer-term solutions to manage demand and capacity.

To date, DERM, developed by Skin Analytics (SA) is the only AIaMD that has satisfied regulatory standards for use in real-world settings as an autonomous screening, triage, or assessment tool (UKCA Class IIa or CE Class III mark), as shown in Table 1. Pilots of DERM began in 2020 with University Hospitals Birmingham pioneering its use and have since expanded to 19 organisations. Early deployments included the use of a dermatologist second read, reviewing the AI's findings in all cases, and providing a further level of scrutiny, with either Trust or contracted dermatologists carrying out reviews. Documented evaluations indicate benefits such as faster review times for lesions[20], and literature is already available to support DERM's autonomous use[21, 22, 23]. Nonetheless, there is a wish to further explore autonomous use within NHS settings and post-market surveillance implications.

Given that DERM offers substantial data from real-world use and is presently the sole tool with the necessary regulatory approval for autonomous use within the NHS, this report will focus on the performance and data of DERM. However, it is important to note that the safety protocols and PMS procedures discussed are generally relevant to any emerging AIaMD aiming to triage skin lesions autonomously within urgent suspected skin cancer pathways.

[20] An evaluation of AI Powered Tele Dermatology for Skin Cancer 2WW Pathway, Health Innovation East Midlands and Edge Health

[21] Phillips et al. (2019), Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions

[22] Marsden al. (2024), Accuracy of an Artificial Intelligence as a medical device as part of a UK-based skin cancer teledermatology service

[23] Thomas et al (2023), Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance

2 Assessment of Standards

In this section, we aim to establish appropriate safety standards for the use of AI as a tool to triage benign lesions. To this purpose, we first estimate the diagnostic performance of the dermatologists from available literature using meta-analysis. Afterwards, we provide an in-depth analysis of the diagnostic performance of DERM, an AIaMD currently approved for use in the NHS. Finally, we discuss the implications of the literature-derived diagnostic performance of dermatologists in setting a standard for AIaMD safety.

2.1 AIaMD Use Case and Diagnostic Accuracy

Although AIaMD has many potential use cases within Dermatology USC pathways, NHSE's priority for integration within services is to use it to accurately identify benign lesions that are suitable for discharge, so that dermatologists can focus their time on patients who need their expertise the most. In this context, where devices essentially function as triage tools, the AIaMD's ability to accurately discern benign lesions becomes the most important feature.

For AIaMD to autonomously triage skin lesions, its accuracy needs to be comparable to that of dermatologists. This assessment involves the use of epidemiological methods that hinge on specific measures of diagnostic test performance (Figure 3):

- **Sensitivity** is the measure of how well the system detects actual cases of a disease, such as melanoma, within those who have the disease. In simpler terms, it is the system's ability to correctly identify people who have the disease.
- **Specificity** gauges the system's capacity to correctly identify people who do not have the disease when they indeed do not. This is crucial for ensuring that benign, non-cancerous cases are not mistakenly identified as cancerous.
- **Positive Predictive Value** (PPV) reflects the proportion of positive test results that are truly positive. For melanoma, this would be the conversion rate of lesions diagnosed or managed as melanoma into confirmed melanoma diagnoses.
- **Negative Predictive Value** (NPV) indicates the proportion of negative test results that are truly negative. For melanoma, this would be the percentage of cases not diagnosed or managed as melanoma that were confirmed not to be melanoma.

		Predicted	
		Positive How many melanomas	**Negative** How many non-melanomas
Actual	**Positive** How many melanomas	**True Positive (TP)** Patients' lesions diagnosed as melanoma that are confirmed melanoma	**False Negative (FN)** Patients' lesions diagnosed as not melanoma but are confirmed melanoma
	Negative How many non-melanomas	**False Positive (FP)** Patients' lesions diagnosed as melanoma but are not melanoma	**True Negative (TN)** Patients' lesions diagnosed as not melanoma which are not melanoma

$$\text{FOR} = \frac{\text{FN}}{\text{TN} + \text{FN}} \quad \text{or} \quad 1 - \text{NPV}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Figure 3. Confusion matrix summarising the performance of a diagnostic model, serving as the basis for calculating diagnostic accuracy measures. Measures are further defined in Section 2.3.

From a statistical point of view, given the context for the use of AI in triage, the most relevant statistics for evaluating diagnostic performance are the NPV and the False Omission Rate (FOR). The NPV measures how reliably a negative result from the AI can confirm the absence of melanoma. The FOR, on the other hand, is the probability that the AI might incorrectly clear a case that was in fact melanoma. These metrics are critical in ensuring that benign lesions are accurately identified and that the AIaMD 's diagnostic capabilities align with those of skilled dermatologists.

Other metrics of performance such as sensitivity, specificity and PPV are still relevant within the context of refining the diagnostic performance of AI, which ensures the AI adds value by supporting clinical diagnoses, instead of triaging patients inaccurately and potentially overburdening pathways (such as by having too low a specificity). However, if the key question is whether AIaMD is safe to autonomously discharge benign lesions, NPV and FOR are the statistics of interest, and will be the focus of our analyses. We have included an analysis of sensitivity and specificity data in Appendices 7.1.4 (dermatologist) and 7.2.1 (AIaMD).

NPV and Prevalence

One important point to note is that the NPV is affected by the prevalence of a condition within a population[24]. A high prevalence typically leads to a reduction in NPV as the chance of obtaining a true negative result diminishes, thereby decreasing the overall proportion of true negatives in relation to all negative outcomes. Conversely, a lower prevalence rate increases the NPV due to a heightened likelihood of true negative results (Appendix 7.1.3). This relationship makes it essential to account for population prevalence when benchmarking the diagnostic performance of AI systems against that of dermatologists, ensuring a fair and contextually relevant comparison.

2.2 Literature Review Methodology

Currently, there is no single systematic review of the diagnostic performance of dermatologists in identifying benign lesions. For this reason, we undertook a rapid semi-systematic review and a meta-analysis to estimate dermatologists' diagnostic performance from multiple studies. We carried out this search in April 2024, guided by the standards set out in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol[25], ensuring a structured and transparent methodology in line with best practices.

We searched Cochrane and Google Scholar databases up until April 2024 which led to the identification of 153 studies detailing dermatologists' diagnostic performances. We included studies where dermatologist's assessment of the skin lesion utilised dermoscopy and/or teledermoscopy, in which melanoma diagnoses (including both malignant and in-situ[26] disease) were confirmed through histology. Consequently, we excluded 99 studies which did not meet our criteria, leaving us with 54 studies. We assessed the full texts of these studies, where a further 27 studies were omitted due to: a lack of raw data, non-dermatologist clinicians performing assessments, populations that were pre-selected for being high-risk, or images of

[24] Parikh et al (2008), Understanding and using sensitivity, specificity and predictive values

[25] Page et al (2021), The PRISMA 2020 statement: an updated guideline for reporting systematic reviews

[26] In-situ melanomas encompass stage 0 disease, when cancer cells are contained within the top layer of the skin (epidermis) and have not grown into deeper layers

exclusively malignant skin lesions, a high potential for conflict of interest, repeated study or a sample size smaller than 10 patients. The article selection process is depicted in Figure 4.

The remaining 27 studies formed the basis for our final meta-analysis on dermatologists' performance in recognising benign lesions. Of these, 19 were carried out in face-to-face (F2F) settings and 8 in teledermatology, providing us with a comprehensive overview of diagnostic accuracy across different practice environments.

We consulted a panel of four consultant dermatologists who agreed that the analysis should focus on melanoma. This is due to the scarcity of studies where other cancer types, such as Squamous Cell Carcinoma (SCC) are identified as the sole true positive lesion, as well as the fact that melanomas pose the greatest risk of harm if missed.

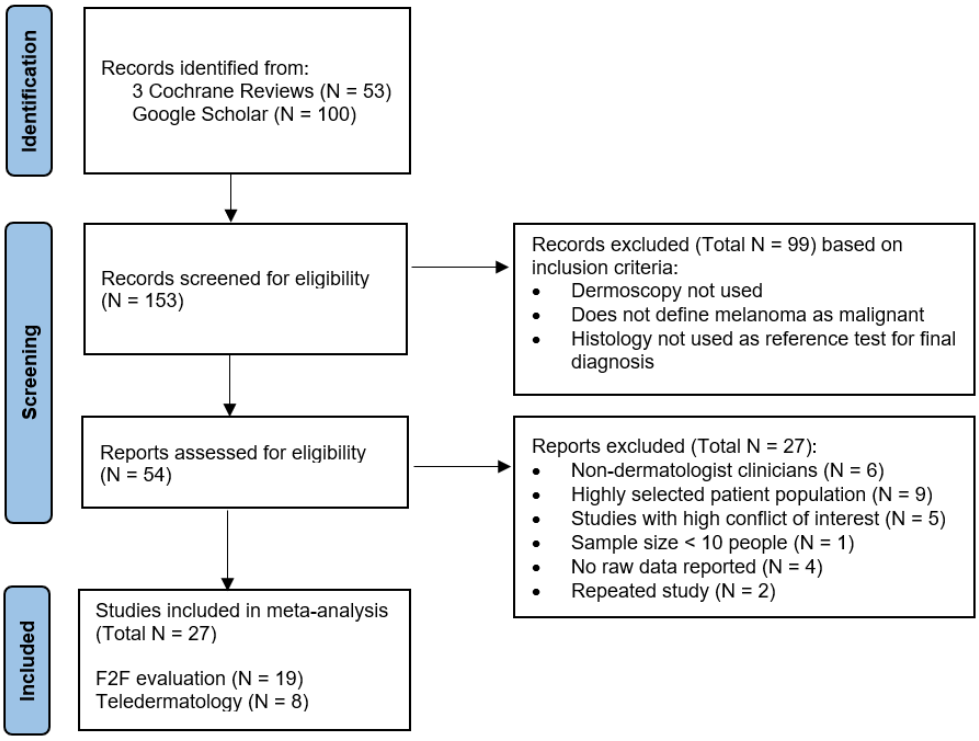


Figure 4. PRISMA flow diagram demonstrating articles selection process

2.2.1 Data Extraction and Analysis

We manually extracted study details and quantitative data from the 27 included studies. We recorded study characteristics including authorship, publication year, the study's country of origin, study design, and participant demographics. We selected studies where the positive definition was lesions confirmed to be melanoma (invasive, in-situ, and lentigo maligna). We also recorded the care settings, diagnostic methods for skin cancer identification, clinicians' experience, the criteria for inclusion and exclusion, and positive lesion detection. The characteristics of the included studies are outlined in the table in Appendix 7.1.1.

We recorded the number of True Negatives (TN) and False Negatives (FN) for NPV calculations and the prevalence of melanoma within each study population. While our focus remained on NPV and FOR, we also recorded sensitivity and specificity values (Appendix 7.1.2), recognising their significance in validating the safety of the diagnostic method. The meta-analyses of dermatologists' sensitivity, specificity and FOR are outlined in Appendices 7.1.4 and 7.1.5.

We used a random-effects model to estimate a pooled NPV, accounting for the heterogeneity between the study populations and the melanoma prevalence across the studies. The 95% Confidence Intervals (95% CI) for our pooled NPV were determined via the Clopper-Pearson exact method to maintain statistical integrity.

2.2.2 Results and Interpretation

Our meta-analyses provide insights into dermatologists' diagnostic accuracy in F2F and teledermatology settings available from the literature (Table 2). They demonstrated that the pooled NPV for dermatologists in F2F evaluations was 98.0% [95%CI 97.1%-98.9%], sampled from 8,909 lesions with an average prevalence rate of 8.1% (Figure 5). In teledermatology, the NPV was a marginally lower 97.6% [95%CI 95.5%-99.6%] based on 1,025 lesions, but with a higher average prevalence of 19.5% (Figure 6). The forest plots for the meta-analyses of dermatologists' FOR are outlined in Appendix 7.1.5.

	NPV [95% CI]	FOR [95% CI]
F2F Evaluation	**98.0%**	**2.0%**
(19 studies, 6,614 lesions)	[97.1%-98.9%]	[1.1%-2.9%]
Teledermatology	**97.6%**	**2.4%**
(8 studies, 1,025 lesions)	[95.5%-99.6%]	[0.4%-4.5%]

Table 2. Results from a meta-analysis of dermatologists' diagnostic accuracy for lesions confirmed not to be melanoma that were not diagnosed or managed as melanoma.

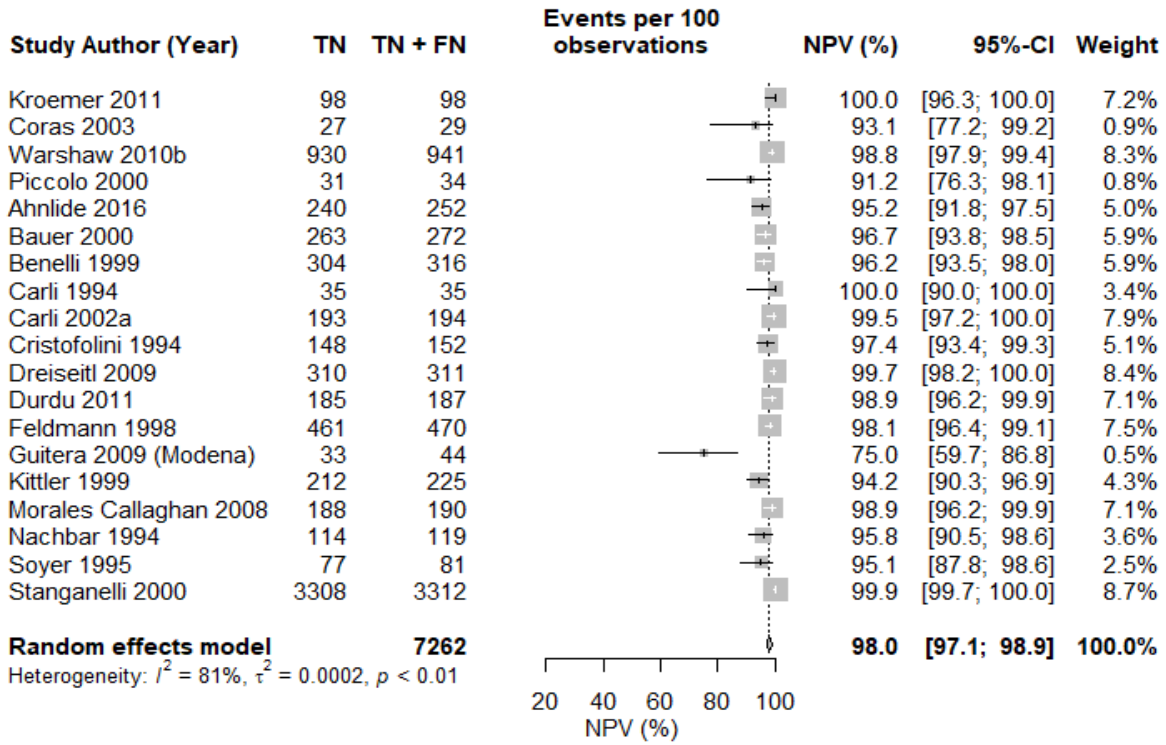


Figure 5. Forest plot for the meta-analysis of dermatologists' NPV performance in face-to-face settings (19 studies), with an 8.1% average prevalence of melanoma. TN: Confirmed non-melanomas; TN + FN: Total number of lesions marked as not melanomas; Weight: Weight assigned to each study determined by within-study variance and between-study variance.

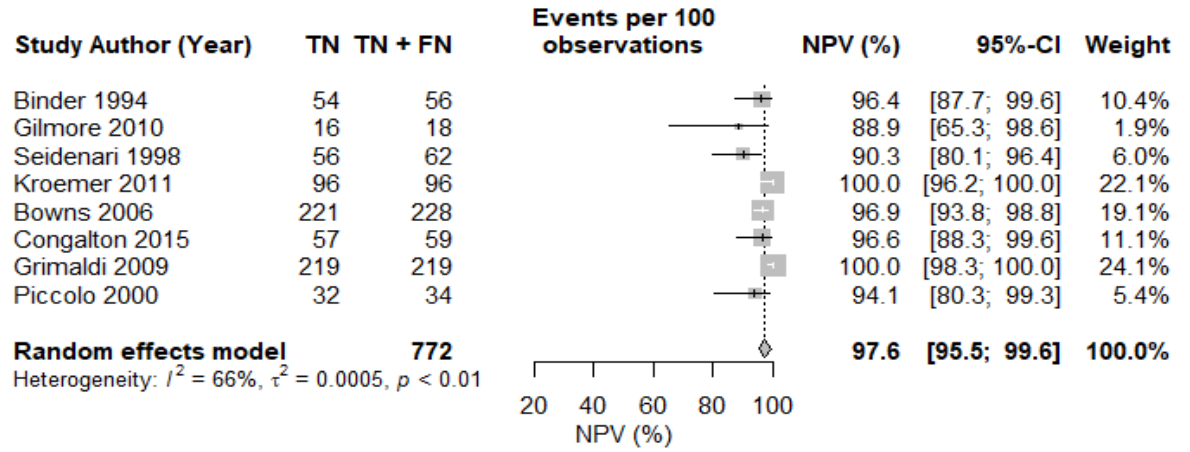


Figure 6. Forest plot for the meta-analysis of dermatologists' NPV performance in teledermatology settings (8 studies), with a 19.5% average prevalence of melanoma. TN: Confirmed non-melanomas; TN + FN: Total number of lesions marked as not melanomas; Weight: Weight assigned to each study determined by within-study variance and between-study variance.

As mentioned in Section 2.1, the disease prevalence has significant effects on NPV. For this reason, we carried out a separate meta-analysis to include only studies with populations of lower disease prevalence. These were chosen to match the confidence intervals for melanoma prevalence observed within the AI real-world population (2.5%), which is discussed in Section 2.3.1. Two F2F studies totalling 1,714 lesions reported similar prevalences of 2.7% and 3% and returned a combined prevalence of 2.7% [95% CI 2.02%-3.62%]. Since this rate falls within the confidence intervals for the AIaMD population prevalence, these studies were selected for the meta-analysis and resulted in an NPV of 98.9% [95%CI 98.2%-99.5%] (Figure 7).

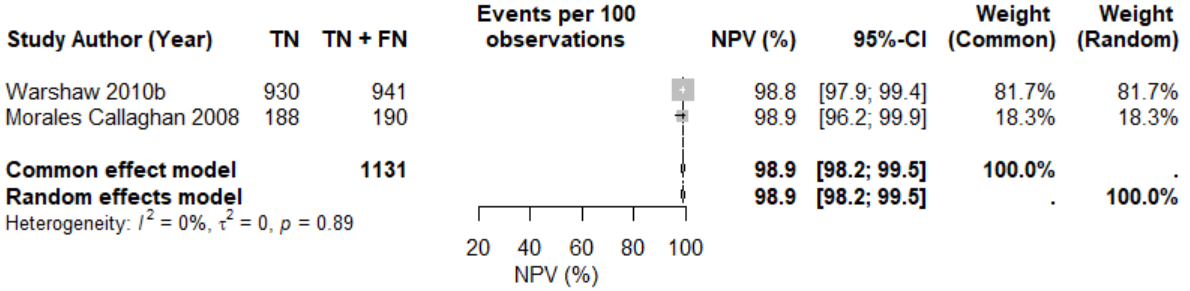


Figure 7. Forest plot for the meta-analysis of dermatologists' NPV performance in F2F setting (2 studies), with a 2.7% average prevalence of melanoma TN: Confirmed non-melanomas; TN + FN: Total number of lesions marked as not melanomas; Weight: Weight assigned to each study determined by the model. Note that both the Common-Effects Model[27] and the Random-Effects Model are included in this meta-analysis, accounting for the small sample size across studies. Both models returned the same pooled NPV.

2.2.3 Limitations of Literature-derived Standards

Our meta-analyses provide important insights for setting standards based on literature, yet it is crucial to be aware of its limitations. Firstly, dermatologists' reported performance may not always mirror real-world clinical practice, potentially limiting the representativeness of the results in everyday scenarios. Furthermore, there is often a lack of detailed information regarding the experience level of dermatologists in various studies, especially within

[27] The common-effects model pools the summary NPV estimate across studies assuming that all studies shared the same underlying effect size i.e. true NPV is consistent across studies.

international settings. It is difficult to fully gauge the impact these factors may have on diagnostic accuracy. To mitigate this, our study has only included literature assessing the performance of consultant dermatologists (or the international equivalent).

Another concern is the possibility of bias. Meta-analyses might draw from studies with inherent methodological biases, which have long been documented within literature, and range from patient selection, study design and interpretation of results[28]. For this work, we have applied stringent inclusion criteria, which excluded studies with highly pre-selected populations, or high conflict of interest risk.

Our review considers the highest quality evidence available up to April 2024. As the research landscape is ever-evolving, there might be emerging findings that necessitate additional analysis. It is important to approach these standards with a degree of caution and to understand the need for contextual validation. Recognising these limitations helps ensure that our conclusions are applied appropriately and remain relevant as further evidence comes to light.

2.3 Analysis of AIaMD Performance

In assessing the performance of DERM, we reviewed its NPV alongside sensitivity and specificity using patient-level data. This dataset encompassed AI and clinician diagnoses, supplemented by histological results, for 33,693 lesions from 29,778 cases evaluated between April 2022 and January 2024, covering the period when the latest versions of DERM (3.0 and above) were used.

Within the assessed cohort, Fitzpatrick skin types 2 and 3 formed the largest proportion, as shown in Figure 8.

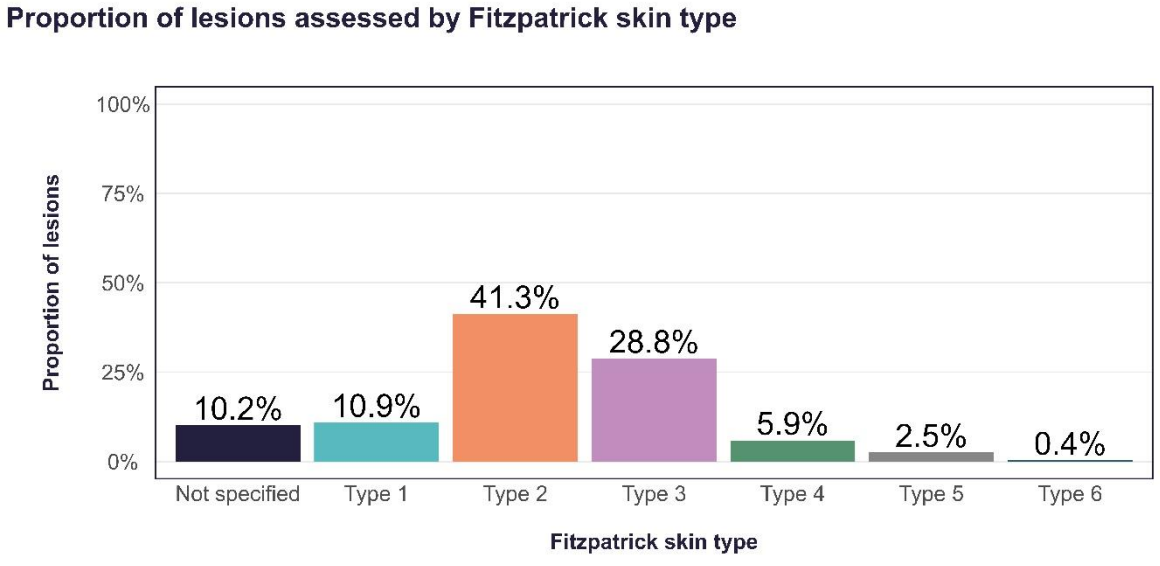


Figure 8. Distribution of Fitzpatrick skin type prevalence within skin lesions analysed in this report for the performance of AIaMD (n = 33,693).

[28] Kelly et al (1997), The identification of bias in studies of the diagnostic performance of imaging modalities

To match the standards established for dermatologists, we calculated DERM's NPV based on its accuracy in excluding melanoma. The following working definitions were employed to establish performance:

Key definitions for AIaMD Performance:

True Positives: Skin lesions marked as melanoma (including malignant, superficial spreading, in-situ and lentigo maligna) which were confirmed to be melanoma.
False Positives: Skin lesions marked as melanoma which were confirmed not to be melanoma.
True Negatives: Skin lesions marked as not melanoma which were confirmed not to be melanoma.
False Negatives: Skin lesions marked as not melanoma which were confirmed to be melanoma.
Negative Predictive Value: The likelihood that a lesion identified as not melanoma by the AIaMD is not melanoma
False Omission Rate: The likelihood that a lesion identified as not melanoma by the AIaMD is melanoma

Note that this definition of False Negatives leads to the inclusion of lesions flagged as suspicious, and therefore correctly sent for review[29]. As we do not expect the AIaMD to independently diagnose melanomas, this might be overcautious. However, we adopted it for consistency with our meta-analysis of dermatologist performance.

2.3.1 AIaMD's Negative Predictive Value

Our results highlighted that DERM performed at an NPV of 99.8% [95% CI 99.7%-99.9%], on a population with a melanoma prevalence of 2.5%, in ruling out melanomas (invasive, in-situ and lentigo maligna). Here we report raw data and summary metrics for this cohort, while further analysis on invasive melanoma-only cohorts is available in Appendix 7.2.2.

The confusion matrix below demonstrates that out of a cohort of 33,693 lesions obtained from ten secondary care pathways, DERM accurately identified 26,885 as not melanoma, while returning 59 false negatives. This results in an NPV of 99.8%, and a FOR of 0.2%. Care site-level data is available in Appendix 7.2.1.

Note that out of the 59 false negatives, 19 were identified as other high-risk lesions and were therefore sent for review by a dermatologist (one atypical naevus, one BCC, two Bowen's disease and 15 SCC). Therefore, 19 lesions would not be considered "missed" under the current pathway where a dermatologist performs all second reads of suspicious lesions. Out of the 40 melanomas not flagged for review, one was nodular melanoma, eight were Lentigo Maligna, 10 were superficial spreading, 17 were melanoma in situ, and 4 were marked as "other".

More analyses on DERM's performance by subgroups can be found in Appendix 7.2.1, this includes data by the following subgroups: Fitzpatrick skin types 1-4 and 5-6, care site and version of DERM.

Data on dermatologist specificity and sensitivity obtained from our meta-analysis is also available for comparison in Appendix 7.1.5.

[29] Suspicious lesions include all lesions marked for review by a dermatologist, i.e. Melanoma, SCC, BCC, Actinic Keratosis, Atypical Naevus and Bowen's Disease.

Confusion Matrix – Melanomas (invasive, in-situ and lentigo maligna)

		Predicted	
		Positive	Negative
Actuals	Positive	776	59 (40) *
	Negative	5,973	26,885

* Note that by the definition used, false negatives include lesions referred by the AI for review by a dermatologist marked as either SCC, BCC, IEC, AK or Atypical Naevus (n = 19). These would still be managed appropriately.

Summary Metrics:

Metric	Value [95% CI]
Negative Predictive Value	99.8% [99.7% - 99.8%]
False Omission Rate	0.2% [0.2% - 0.3%]
Sensitivity	92.9% [91.0% - 94.6%]
Specificity	81.8% [81.4% - 82.2%]
Prevalence	2.5%

2.3.2 Additional Melanoma Recognition

Within the assessed data, we identified 106 lesions that were marked by the AI as melanoma, where the Trust's suspected diagnosis on teledermatology review was benign (including atypical naevus, benign melanocytic naevus, seborrheic keratosis, solar lentigo, vascular lesions). In collating these cases, we ruled out instances where dermatologists could not put forward a diagnosis due to poor image quality or clinical ambiguity, and cases where dermatologists suspected other forms of cancer, such as SCC or BCC.

All 106 were eventually biopsied and correctly managed, though some experienced delays. Four lesions were marked for discharge on review but were seen F2F due to another suspicious lesion, and 11 were seen in routine clinics or had routine excisions. While it is possible that all lesions would have been seen in F2F clinics had they been reviewed in a standard teledermatology setting, we cannot exclude the possibility that the AI diagnosis might have contributed to escalating concerns for at least 91 that were sent for urgent onward review.

2.3.3 Repeat Presentations for Same Lesion

As part of our assessment of the pathway, we considered repeat presentations for the same lesion. This highlighted 24 lesions where patients re-presented following initial AI assessment and dermatologist discharge.

The average time for re-presentation was 16 months. Re-attending lesions included one melanoma, one unspecified benign, two atypical naevi, two Bowen's disease, four SCC, six BCC and eight actinic keratosis. Note that two Bowen's diseases, one SCC and the melanoma were biopsied after both attendances; these were initially found to be unspecified benign (for the first three lesions) and actinic keratosis (for the melanoma).

Due to the extended timeline for re-presentation, we cannot exclude whether lesions had transformed. For instance, seven lesions were biopsied on both occurrences and while the first

biopsy suggested a benign diagnosis, the second revealed a suspicious diagnosis. These findings emphasise the importance of adequate patient safety netting on discharge.

2.4 Standards for AI Performance

We have discussed the importance of setting performance standards for AIaMD by benchmarking against documented dermatologists' performance. In the context of identifying benign lesions, we have carried out a meta-analysis of dermatologists' NPV of correctly excluding melanoma. This revealed a pooled NPV in F2F settings of 98.0% and 98.9%, at prevalences of 8.1% and 2.7% respectively (Table 3).

DERM demonstrated an NPV of 99.8% at a 2.5% prevalence rate across a sample size of 33,693 lesions, in its intended real-world use case population. Considering the available evidence, these results suggest that the performance of DERM is at least as good as the documented accuracy of dermatologists. Note that the AIaMD performance is compared to F2F evaluations, as its intended use is to help reduce the need for specialist reviews of benign lesions. Additionally, there is a higher number of research studies available for F2F diagnostic accuracy, including those with a prevalence rate matching the 2.5% prevalence seen in the population assessed by the AIaMD, which is not the case for teledermatology studies.

Considering the role of NPV when employing AIaMD as a triage tool in skin cancer pathways, a safety standard for NPV at 99% would be a sensible target. However, it is crucial to acknowledge the limitations in recommending these standards, as discussed in Section 2.2.3, which include considering the complexities of real-world clinical practice, the experience of dermatologists as well as the changing landscape of available high-quality research.

Given these factors, recommended standards should be applied cautiously; they should be viewed as a starting point for evaluation and validation rather than as fixed benchmarks. Additionally, other performance metrics such as monitoring for repeat attendances and outcomes, accurate recognition of the nature of benign lesions and sensitivity and specificity should be considered to further evidence the safety of AIaMD.

	NPV [95% CI]	FOR [95% CI]
F2F Evaluation at 8.1% prevalence (19 studies, 8,909 lesions)	**98.0%** [97.1%-98.9%]	**2.0%** [1.1%-2.9%]
F2F Evaluation at 2.7% prevalence (2 studies, 1,714 lesions)	**98.9%** [98.2%-99.5%]	**1.1%** [0.5%-1.8%]
DERM at 2.5% prevalence (33,693 lesions)	**99.8%** [99.7%-99.8%]	**0.2%** [0.2%-0.3%]

Table 3. Summary table comparing NPV analysis of dermatologists, as obtained through meta-analysis, and DERM, as calculated from real-world data.

3 Implementation Pathways

This section offers an overview of the implementation of AI within secondary care, focusing on urgent skin cancer referral pathways, with the SA DERM tool being the only one utilised nationwide at present. Currently, there are two primary ways in which the AI is being integrated: pre-referral, which occurs before a patient consults a GP, and post-referral, which takes place after a GP has identified a potential cancer concern and referred the patient to USC pathways.

The motivations driving the adoption of AIaMD in both pre-and post-referral contexts aim to tackle different challenges within skin cancer care. The key benefit of the post-referral pathway is to reduce the strain on USC referral systems and secondary care, minimising unnecessary face-to-face consultations, and freeing up dermatologists' time and capacity. The pre-referral pathway additionally targets the burden on primary care, enhancing patient access to services.

While pre-referral pathways will be briefly discussed, it should be noted that fewer sites are undergoing funded pilots, and these are currently being evaluated by the Exeter Test Group. Consequently, this report will predominantly concentrate on post-referral pathways. It is also important to acknowledge that new pathways are being trialled, incorporating AIaMD in both routine and USC pathways. However, as the remit of this document is to evaluate pathways for suspected skin cancer, these will not be discussed further.

3.1 Overview of Current Implementation Pathways

Since 2020, several NHS Trusts have been conducting pilots of DERM, with University Hospitals Birmingham NHS Foundation Trust being the first to adopt the technology. Following this initial implementation, the pathways have been continually improved, resulting in 19 sites currently integrating DERM within their suspected skin cancer referral processes, with a further expansion in the number of partners planned for the future.

Sites that have currently adopted DERM within their skin cancer pathways

Pre-referral Pathways:

- Herefordshire and Worcestershire ICB (2023)
- Lancashire and South Cumbria ICB (2023)
- Suffolk and North East Essex ICB (2023)
- University Hospitals Birmingham (2020)

Post-referral Pathways:

- Ashford and St Peter's Hospital (2022)
- Buckinghamshire Healthcare Trust (2024)
- Chelsea and Westminster Hospital (2022)
- Dorset County Hospital (2024)
- Kingston Hospital (2024)
- Liverpool University Hospitals (2024)
- Manchester Foundation Trust (2024)
- Mid Cheshire Hospitals (2023)
- University Hospitals of Morecambe Bay (2023)
- Royal Devon and Exeter Hospital (2023)
- Tameside and Glossop (2024)
- University Hospitals Birmingham (2020)
- University Hospitals Dorset (2024)
- University Hospitals of Leicester (2022)
- West Suffolk Hospital (2023)

In addition to the Trusts currently using the tool, three others had initiated pilots but subsequently withdrew for reasons ranging from poor GP engagement and issues gathering outcome data, disruption due to changes in commissioning from Clinical Commissioning Groups (CCG) to Integrated Care Boards (ICB), staff shortages across clinical as well as programme management, issues reaching high-volumes of uptake within the pathway, unease as a consequence of initial statements from the British Association of Dermatologists (BAD).

The current implementation pathways are categorised into pre- and post-referral. In the pre-referral pathway, patients with concerns about their skin lesions are directed by GP practice staff to photography hubs, provided they meet the inclusion criteria and give their consent. In contrast, the post-referral pathway requires patients to have a GP appointment first, where concerns about the lesions lead to a USC referral; only after this referral can patients be booked into a photography hub, subject to similar criteria and consent.

The specific healthcare professional taking the photographs at the hub and whether extra images are captured can vary between sites.

Cases undergo further teledermatology review by either a Trust dermatologist or an SA-contracted consultant dermatologist (referred to here as "SA dermatologist" for brevity, who work on a contractor basis). In both pathways, lesions marked for continuation on the USC pathway by AIaMD, along with those that could not be assessed[30], are reviewed by a Trust dermatologist. Lesions that AI deems likely benign are currently reviewed by SA dermatologists, who may either concur with the AIaMD's discharge recommendation or overturn it. Disputed cases are sent for further teledermatology assessment by Trust dermatologists. Trust dermatologists then examine the overturned lesions to decide whether to discharge the patient or opt for an alternative management strategy.

The subsequent steps can vary across different Trusts and may include:

- Urgent referrals to face-to-face USC Dermatology clinics or other specialist clinics such as plastics.
- Direct to biopsy appointments.
- Re-routing to routine face-to-face clinics with Dermatology or other specialties.
- Telephone consultations.
- Scheduled follow-ups at a later date.

Pathways may be personalised to each Trust. Some Trusts have created specific pathways for conditions such as Actinic Keratosis (AK) or SCC. Other variations include who takes the clinical history, the professional responsible for capturing dermoscopy images, additional exclusion criteria, whether initial Trust reviews are conducted through teledermatology or face-to-face, management strategies for AK lesions, and whether GPs receive a PDF with images and discharge details.

[30] Reasons for non-assessment include lesions that are too large for dermoscopic examination, obscured by hair, tattoos, or scars, situated on challenging areas such as the nose, eyes, mucosal, or acral surfaces, or if there are multiple lesions (more than two). Additionally, lesions that are open or ulcerated, those that cannot be captured by dermoscopy, or those that have previously been biopsied or are currently under treatment are also not assessed.

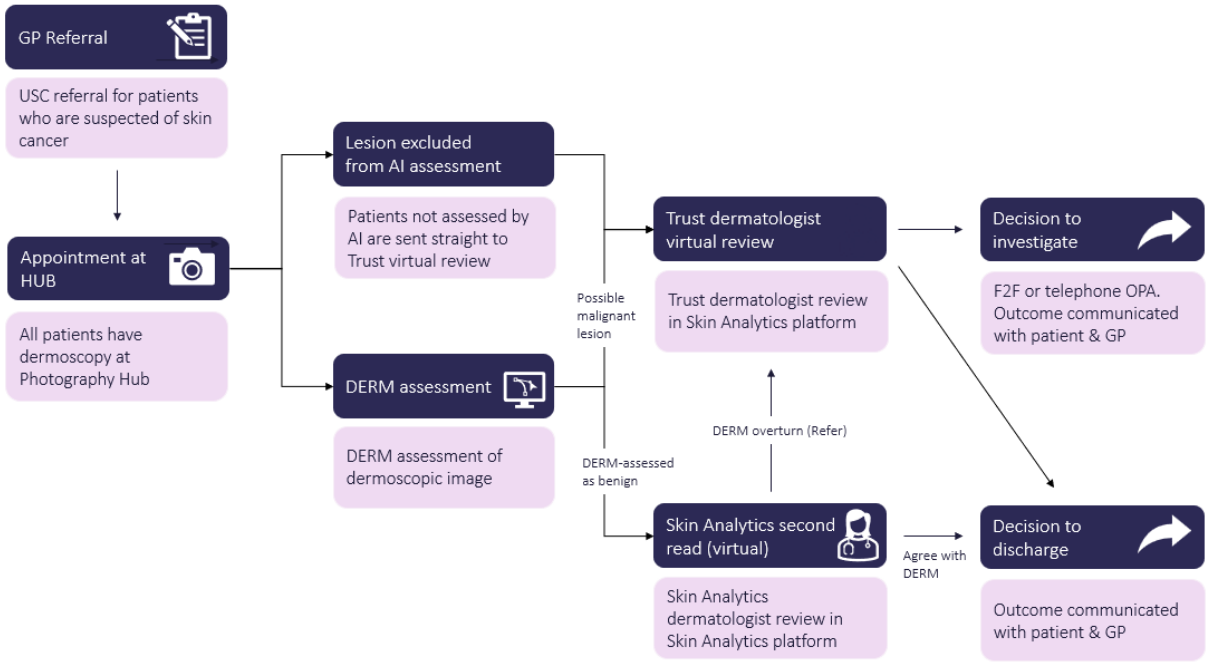


Figure 9. Diagram summarising the piloted Post-Referral Pathway

Patient Selection and Referral

The selection and referral of patients for AIaMD assessment are based on inclusion and exclusion criteria set according to the AIaMD regulatory requirements, with potential tailoring to local factors. The general criteria are as follows:

Inclusions:

- Adults aged 18 years and older.
- Individuals with 1 to 3 suspicious skin lesions

Exclusions:

- Individuals under the age of 18.
- Skin lesions that are not suspicious of malignancy, such as rashes, eczema, infectious diseases, or lupus.
- Lesions requiring disease staging.
- Non-dermoscopic images of skin lesions.
- Open or ulcerated skin lesions.
- Lesions too large to be fully captured by the dermoscopic device.
- Lesions obscured by hair, tattoos, or scars.
- Lesions located under the nail (subungual), on mucosal surfaces, genital areas, or on the soles or palms (palmoplantar).
- Lesions that have been previously subjected to a biopsy.
- Lesions under observation for treatment response.

All patients considered for the DERM pathway receive an information leaflet explaining the process and are asked to sign a consent form, as per GDPR article 22[31]. If a patient either does not consent or does not meet the referral criteria, photographs and dermoscopy images are still captured to facilitate review via teledermatology, ensuring the hub appointment is utilised effectively.

Looking ahead to a future where automated clinical pathways could be more prevalent, patients will also need to agree to this approach. Those who do not consent to AI decision-

[31] Art. 22 GFPR, Automated individual decision-making, including profiling

making would continue to follow the established pathway, with their cases reviewed by Trust dermatologists.

Photography Hub

Suitably trained staff at photography hubs operate using only approved hardware to capture images of skin lesions for AI assessment. Additional equipment may be used on request of the Trust, such as additional high-resolution photographs taken through a DSLR camera, or images taken on Trust iPads so that photographs can be included in the patient's record through the Trust's Electronic Patient Record (EPR) systems.

The hub staff are fully trained and have expressed a high level of confidence in their ability to perform their duties effectively, as evidenced by an evaluation conducted during the pilot at University Hospitals Leicester (UHL)[32]. Under the current protocol, staff at the photography hub do not communicate the results of DERM's assessments to patients during the appointment. All individuals discharged following the assessment are given a comprehensive safety netting statement to ensure they are aware of what steps to take should their condition change or if they have persistent concerns.

AI Assessments and Diagnoses

All imaged lesions are immediately assessed by the AI. DERM classified lesions as Melanoma, SCC, BCC, Bowen's disease / intraepidermal carcinoma (IEC), Actinic Keratosis (AK), Atypical Naevus, or Benign. The benign diagnoses are further subdivided into Benign Vascular Lesion, Seborrheic Keratosis, Dermatofibroma, Solar Lentigo and Melanocytic Benign Nevus.

Skin lesions are assessed by the AI using a risk hierarchy; if a lesion exhibits features of more than one possible type, the DERM diagnosis will reflect the higher risk type. More information on DERM's risk hierarchy is available from DERM's Instructions for Use[33].

At present, DERM diagnoses are recorded within the lesion's data and are visible to dermatologists performing a second read. In an autonomous pathway, DERM benign diagnoses could be included within patients' discharge information to provide further documentation of the assessment's outcome, and to be featured in patients' GP records.

Discharge Process

Upon discharge from the pathway without a face-to-face consultation, patients receive tailored safety netting guidance. This advice is provided to both GPs and patients, with the flexibility for each Trust to customise the information. In the context of an autonomous pathway, the discharge letter templates would be supplied by SA or co-developed in collaboration with them, incorporating patient-focused content informed by input from patient engagement groups.

[32] An evaluation of AI Powered Tele Dermatology for Skin Cancer 2WW Pathway, Health Innovation East Midlands and Edge Health

[33] Instructions For Use - Deep Ensemble for Recognition of Malignancy (DERM), Skin Analytics

Medicolegal Considerations

The medicolegal responsibilities for patients within the pathway are currently shared amongst the Trusts, the contracted dermatologists, and SA. While there have been no precedents for medicolegal disputes related to this pathway so far, the following protocols are advised for any future instances: contracted dermatologists, who are protected by their own indemnity insurance, are accountable for the clinical decisions they make. When it comes to patients assessed by the Trust – whether through teledermatology or in person – the Trust assumes medicolegal liability. SA's indemnity cover would be responsible for decisions made solely by the AIaMD without a human review.

3.2 Provider Interviews

To gain a deeper understanding of the implications associated with the integration of AIaMD in skin cancer pathways, we carried out a series of semi-structured interviews covering the topics summarised below. We engaged with three service providers, pseudonymised as Providers 1, 2, and 3, who are actively adopting AIaMD. We spoke with a total of 8 staff members, including consultant dermatologists and transformation managers, to gain a wide range of views on the motivations, benefits, challenges and practical implications of implementing AI within skin cancer pathways.

Motivations for the AIaMD pathway

A common theme across the three providers was the implementation of AIaMD to address the escalating patient backlogs and extended waiting periods for USC referrals, a situation exacerbated by the COVID-19 pandemic. The adoption of AI was envisaged to enhance both the efficiency and the quality of patient care within Dermatology services. Provider 2 experienced a coordinated managerial drive and corresponding investment to support this technological advancement. Of the three, Provider 1 had an existing teledermatology service in place, and integrating AIaMD appeared to be a logical progression to their pathways.

Pathways and Current Use

The current post-referral pathways are similar across all providers but there is some variation in the implementation (Figure 10).

Firstly, providers take different approaches for staff who perform image acquisition, largely dependent on existing staff availability. Specifically, Provider 2 has opted for skin lesions to be captured by medical photographers using DSLR cameras, in addition to the images taken with the manufacturer-provided equipment. In contrast, Provider 3's team has trained two Band 4 Healthcare Assistants to photograph the lesions using both the manufacturer-provided equipment and an iPad, so that images could be recorded within the patient records as well as the manufacturer's platform.

Secondly, there is a divergence in the approach to biopsies. While Provider 1 and Provider 3 have integrated a direct-to-biopsy route, Provider 2 seldom employs this method. Provider 3 offers direct referrals to oculoplastic services as part of their pathway.

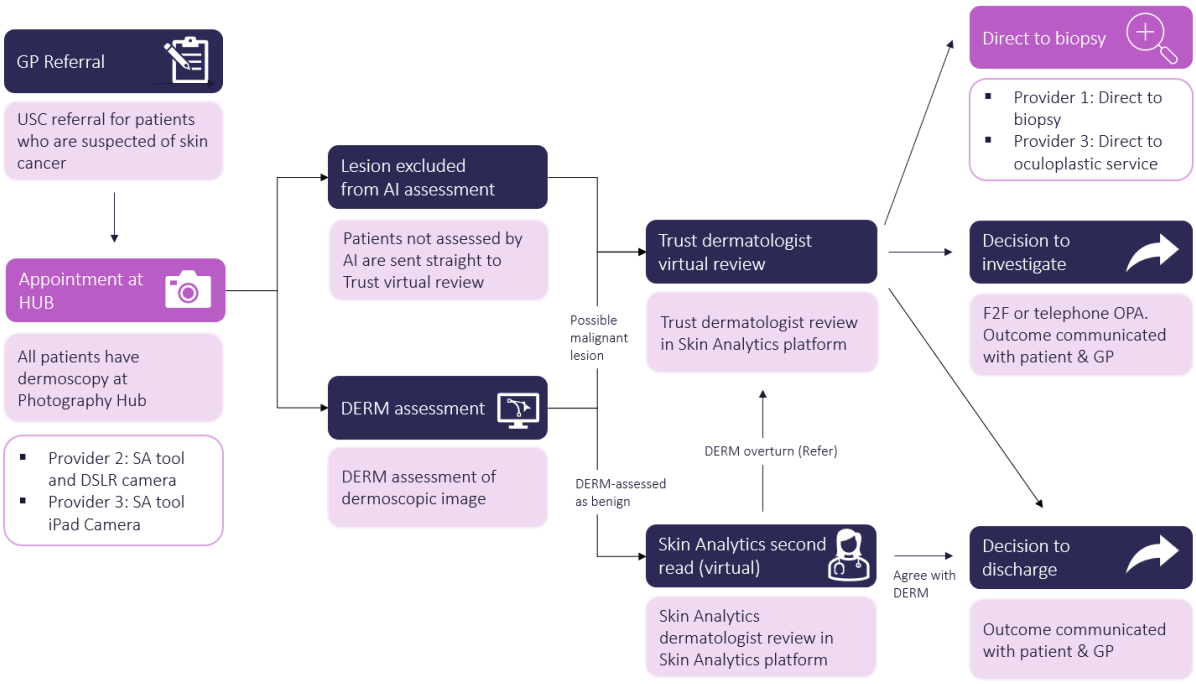


Figure 10. Diagram summarising the Post-Referral Pathway across providers. Highlighted boxes indicate where the implementation varies across the providers.

Benefits

All providers have acknowledged experiencing benefits, to varying degrees. The common benefits emerging from the implementation include enhanced operational efficiency, reduced need for in-person reviews, and the potential for immediate patient discharge.

A key benefit reported by all providers is the reduction in patient reviews and F2F appointments. Overall, all providers were able to discharge 20-25% of patients immediately without a F2F review.

Further benefits identified by providers included:

- Provider 1 has expressed that AIaMD has allowed them to meet and surpass cancer and performance targets, despite an increase in referral numbers.
- Provider 2 has expressed that the pathway has facilitated a 20-30% reduction in patient reviews. They noted that they had not observed greater capacity to see more patients, however, they expressed that in the past significant additional capacity was needed to address the backlog, in the form of Waiting List Initiative (WLI) clinics, which may have reduced.
- Provider 3 has observed positive patient feedback on the AI pathway, where patients expressed that they could be seen more promptly, reducing the need for multiple hospital visits.
- Two providers have expressed a reduction in their biopsy rates.
- One provider pointed out the added benefit of a user-friendly interface that comes with the AI tool.

Challenges

Two providers expressed common challenges due to IT infrastructure, such as Wi-Fi connectivity and integration with existing systems. Provider 1 faced these challenges during the initial phase of implementation of the pathways, where the set-up was delayed by about three months due to the lack of IT infrastructure within the trust.

Similarly, Provider 2 expressed that the AIaMD system has not been fully integrated with the existing clinical portal, which required navigating between separate systems. They highlighted that in an ideal scenario, the tool would be integrated within existing Electronic Patient Record (EPR) systems so that dermoscopy images and outcomes could be referenced later if the patient were to re-present. However, all Trusts have expressed that the teledermatology platform provided by the manufacturers was straightforward to use.

Additionally, since the implementation of the AI pathway, some consultant dermatologists expressed a change in their case mix, with more complex patients proceeding to F2F consultations. Provider 1 noted that this challenge might not be unique to AI pathways, with standard teledermatology pathways likely to yield similar effects, and that complex cases would have been part of the dermatologist's workload before the introduction of AI.

Contracting and commissioning

Provider 3 and Provider 2 did not face significant issues with commissioning. At Provider 3, it was largely due to support from relevant parties, such as the ICB and the skin cancer lead. Provider 2 found that the need to adapt care delivery during the COVID-19 pandemic allowed for expedited implementation. Additionally, proactive engagement with the cancer network facilitated the commissioning of the AIaMD and helped maintain an interest in its implementation.

Provider 1 has faced challenges in contract negotiations and procurement, as their service transitioned to business as usual from its initial funding through national grants. This transition faced delays in finalising the contract due to procurement issues. There were complications from the evolving AI policies within the local ICB and funding for additional post-market surveillance sitting at an ICB level. However, initial commissioning was smoother due to the pre-established teledermatology services. Provider 1 has now moved away from the block contract payment system, and the service agreement is based on a per-population level with a flat fee regardless of volume of activity.

Provider 2 highlighted the importance of clarifying the role of teledermatology appointments in funding models. They discussed the need for clear commissioning frameworks and the importance of understanding the cost-effectiveness of the AI-assisted pathway compared to traditional teledermatology or F2F consultations.

Methods for service evaluation and surveillance

The three providers are at different stages of readiness moving towards an autonomous AI pathway. Provider 1 is deploying an autonomous pathway and has received funding for a study on additional post-market surveillance covering both traditional dermatology as well as AI-enabled pathways. The provider will be trialling a text message follow-up system, where patients are sent text messages six months after discharge to check if they require further assessments and to address any ongoing concerns.

All providers are undertaking internal audits of the pathway. The dermatologists at Provider 3 conduct internal auditing to examine images of all benign lesions. During these reviews, the team cross-checks the patient's identity and the lesion's location in AIaMD's reports against

the hospital records, which dictates any patient correspondence when necessary. Furthermore, clinical images are maintained in a format compatible with the hospital's system for any subsequent case discussions in multidisciplinary team meetings. Similarly, Provider 2 carries out its internal review of a proportion of benign and malignant lesions to assess the alignment of diagnoses.

All Trusts reported that second reads provide an additional safety netting benefit, at least for a time before the implementation of autonomous use. Provider 1 elaborated on benefits such as mitigating human errors in the early phases of implementation, building confidence within clinical teams and assessing the AIaMD's performance within their local population if performance data does not already exist in sufficiently similar populations.

Summary of implementation advice to other Trusts

Lastly, we asked if providers had any advice for others planning to implement AIaMD in Dermatology pathways. These cover common themes such as building a good rapport with ICB and local GPs, having dedicated staff such as administrative staff and photographers to support implementation, and ensuring any additional work (such as teledermatology reviews of lesions) is incorporated into consultant job plans.

Provider 1 is working on an AIaMD implementation toolkit for other providers that is expected to be published in Summer 2024. Provider 2's operational manager highlights the importance of ensuring the availability of skilled clinical photographers, having the right equipment, engaging GPs from the outset, maintaining clear communication with patients and having proper documentation. Provider 3 emphasised the importance of building a good rapport with ICB, getting local GPs, administrative staff and photographers on board, and ensuring any additional work is adequately incorporated into their job plans.

Key Takeaways from Provider Interviews

Provider 1

- Encountered initial resistance with AI due to lack of national support and negative press from professional bodies.
- AI implementation resulted in efficiency gains, reducing face-to-face appointments and allowing immediate discharge for about 25% of patients, potentially rising to 40%.
- Improved pathway flow has allowed to consistently meet Cancer Waiting Times targets.
- Commissioning and IT infrastructure were initial challenges, but these were overcome by extensively liaising with ICB.
- Clinician buy-in was facilitated by sharing data early and allowing open discussion about the technology.
- There is a continued need for national endorsement and robust safety netting, including audits, internal reviews, and patient education regarding autonomous AI assessments.

Provider 2

- Teledermatology was adopted out of necessity during the COVID-19 pandemic, supported by leadership.
- There is some scepticism about the current capabilities of AI, with a preference for AI assistance in primary rather than secondary care.
- The AI is primarily used as a triage tool. The team commended the high sensitivity but expressed some concerns regarding low specificity as well as the potential of missing rare cancers.
- Challenges include the need for more accurate and user-friendly outcomes from AI, including more extensive information being given to patients regarding benign diagnoses when discharged.
- Benefits of AI include a reduction in the number of patients requiring review by dermatologists and an improvement in image quality for assessments.

Provider 3

- Rapid adoption of teledermatology due to a surge in skin cancer cases and long waiting times.
- Scepticism was initially high but quickly turned into support after seeing benefits.
- AI pathway allows patients to be seen more quickly and directly proceed to surgery if necessary.
- Challenges included initial resistance from some clinicians and concerns over potential missed diagnoses, which have now resolved after reviewing robust evidence.
- Trust clinicians see a benefit in being able to focus on potential cancers, with benign patients handled by the AI and SA dermatologists.
- Strong support and buy-in from managerial staff, the ICB, and GPs were crucial for successful implementation.
- Patient feedback has been positive and the provider is considering how to implement the AI autonomously within their pathways.

4 Post-Market Surveillance

After evaluating the AIaMD's effectiveness in identifying benign lesions by examining its NPV, we now turn our attention to post-market surveillance (PMS). This is an essential part of the post-deployment regulatory process that aims to ensure real-world safety and performance of the AIaMD and to validate ongoing performance, manage risks and ensure regulatory compliance through robust data collection and sharing in real-world clinical settings.

4.1 The Regulatory Landscape

Under the 2023 UK Medical Devices regulations[34] manufacturers must continually monitor the performance of medical devices as part of PMS. To legally deploy medical devices on the market, manufacturers should ensure compliance with two designated international standards, ISO 13485 covering quality management systems for medical device manufacturers and ISO 14971 covering risk management for medical devices, in line with the MHRA conformity

[34] The Medical Devices (Post-market Surveillance Requirements) Draft., draft legislation to amend the Medical Devices Regulations 2022 and incorporate new PMS requirements for medical devices

assessment process[35]. Furthermore, regulations require manufacturers to develop a PMS plan that is clear, organised and easily searchable that is maintained throughout the device's post-market lifespan. The PMS plan should outline the intended lifespan, the processes for systematically gathering and evaluating relevant data, mechanisms for communication with regulators and users, follow-up, risk assessment, and complaint investigation. Moreover, as part of the regulatory requirements, manufacturers are obliged to submit vigilance reports for any adverse events or incidents, done through the designated MHRA Yellow Card scheme[36].

PMS regulations are subject to ongoing changes. In January 2024, the MHRA published a strategic roadmap for an upcoming regulatory framework for medical devices[37]. This outlines the timeline for introducing essential new regulations, such as the laying of draft PMS regulations to parliament and the enactment of preliminary PMS regulations expected by the end of 2024. As a result of the changing landscape, the regulatory guidelines on PMS actions remain relatively open-ended.

4.2 Defining AI-Related Errors and Risks

Before delving into the specific practices required for effective PMS, it is necessary to define the errors and risks that could result from using AI as a clinical diagnostic tool in the post-deployment phase. Overall, the AI tool should be monitored for two categories of risks:[38]

1. **Changes in the performance of the AI algorithms**: The performance of the AI diagnostic tool may change over time and can occur for various reasons. For example, the changes in the performance of the AI tool (e.g. a drop in NPV), changes in the underlying patient population, and risk of data drift when the distribution of the data used in clinical practice is shifted from the original distribution of the dataset or algorithmic bias.
2. **Inappropriate or variable usage of the tool:** The risk of misdiagnosis could be heightened due to inappropriate use of the AIaMD. As such, monitoring how the instructions for use are applied in a clinical setting is essential. For example, failure to apply exclusion criteria correctly for image assessment by the AI, and clinical actions taken in response to the tool's outputs, would be essential in monitoring the overall clinical effectiveness of the tool.

The responsibilities for mitigating these errors could fall into either the manufacturers or the users (e.g. Trusts adopting the tool). Therefore, the post-market surveillance plan should involve not only monitoring the performance of the AI algorithm itself but also the appropriate uses of the tool.

In the context of diagnostic AIaMD in skin cancer, there is a risk of algorithmic biases which could amplify existing inequalities across ethnicity. It was found that algorithms for skin cancer detection are largely trained on biased datasets, such as the International Skin Imaging

[35] MHRA Guidance Medical devices: conformity assessment and the UKCA mark
[36] MHRA Guidance for manufacturers on reporting adverse incidents involving Software as a Medical Device under the vigilance system
[37] MHRA Roadmap towards the future regulatory framework for medical devices, 9th January 2024
[38] Evaluating AI-Enabled Clinical Decision and Diagnostic Support Tools Using Real-World Data, Margolis Centre for Health Policy

Collaboration, which mostly contains data from fair-skinned populations[39]. Additionally, the lower incidence of melanoma in darker-skinned populations[40] means that there is inherently less data covering these populations. It is crucial to acknowledge and address this risk to ensure that AIaMD can be effectively generalised to the diverse populations they serve across various deployment sites.

4.3 Real-world surveillance of AIaMD in the Post-market Phase

We have taken a dual approach in assessing recommendations for PMS of AIaMD:

- We provide a comprehensive review of existing literature on PMS best practices and condense these to a summary of high-level recommendations for deployment sites and manufacturers.
- Acknowledging the limited literature on practical PMS strategies, we propose a potential methodology for carrying out PMS in practice, that can form the basis of further discussion between deployment sites and manufacturers.

4.3.1 Review of Literature on Post-Market Surveillance

Our literature review focuses on identifying reactive and proactive monitoring frameworks and guidelines for post-market surveillance (see Appendix 7.3 for search strategy). These activities are grouped into three overarching themes: data collection and sharing practices, performance monitoring and evaluation of the AIaMD and transitioning to autonomous AIaMD use.

1 - Data Collection and Sharing Practices

The CLEAR Derm consensus, by the International Skin Imaging Collaboration AI Working Group, outlines best practices for image-based AI in Dermatology[41]. These include data collection, technical assessment, and monitoring use cases to evaluate AI performance in real-world settings (detailed in Appendix 7.4). Originally devised for pre-deployment stages, these guidelines are also relevant for post-market surveillance.

To ensure integrity and mitigate biases in the AIaMD's post-market phase, the guidelines recommend that manufacturers should follow clear documentation and examination of imaging modalities, artefacts, metadata, and dataset definitions:

- **Imaging and Artefacts**: Ongoing evaluation of imaging modalities and artefacts is essential, with any deviations documented to maintain AI robustness. Particular attention should be given to ensuring images used in the real-world setting are reflective of those in test datasets and ethical considerations should be described.

[39] Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts, European Parliamentary Research Service

[40] Delon et al (2022). Differences in cancer incidence by broad ethnic group in England, 2013–2017

[41] Daneshjou et al. (2022), 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

- **Metadata and Biases**: Continuous scrutiny of images and metadata is needed to identify and address biases, such as patient demographics and clinical settings, to identify any shifts in data distribution that could affect the algorithm's performance.
- **Dataset Definitions**: Clear criteria for image dataset inclusion and independence between training, validation, and test sets are vital to prevent data leakage, and strategies to mitigate this should be described.
- **Clinical Relevance of Test Dataset**: Test datasets should represent diverse patient characteristics and class distribution in test data should be stratified by patient characteristics, with procedures to address class imbalance clearly outlined.

The Medical Algorithmic Audit[42], following the SMACTR framework[43], provides a structured approach to identifying algorithmic discrepancies. It encompasses five auditing phases to ensure algorithm design aligns with organisational values. During the artefact collection, datasets and models, among other elements, are gathered for assessment.

The EADV AI Task Force[44] highlights transparency and user trust, through focusing on skin cancer diagnostic apps, their principles apply broadly to AIaMD for skin cancer diagnosis. Their principles include:

- **Transparent Validation**: Clear communication about algorithm validation, diagnostic accuracy, and health outcomes is required. Any unassessed populations should be explicitly disclosed.
- **Accountability and Traceability**: Error tracking and 'privacy by design' principles[45] are necessary for data protection and accountability. Explicit consent must be obtained if patient data will be used for any purpose, such as training the algorithm itself.
- **Inclusivity**: AI applications should perform consistently across skin tones and age groups.
- **Multidisciplinary Collaboration**: Stakeholder collaboration across disciplines is encouraged to advance AI diagnostics in Dermatology.

Building upon these recommendations, the Duke Margolis Centre for Health Policy[46] addresses the need for real-world data to evaluate and monitor AI-enabled clinical decision support systems. They highlight the importance of tracking performance and usage changes. They suggest:

- **Adaptive Monitoring**: AIaMD should adapt to maintain performance amidst changes in clinical practices, data entry, demographics, and care standards. Inappropriate or variable usage by healthcare professionals must be assessed.
- **Bias Assessment:** Investigating biases and their impact on clinical outcomes is essential, including reviewing algorithm outputs, data inputs, and demographic subgroup analyses.

[42] Liu et al. (2022), The medical algorithmic audit
[43] Raji ID, Smart A, White RN, et al. (2020), Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing
[44] Position statement of the EADV Artificial Intelligence (AI) Task Force on AI-assisted smartphone apps and web-based services for skin disease
[45] 'Privacy by design' is a set of proactive approach to ensure privacy and data protection throughout the lifecycle of a product (ref)
[46] Evaluating AI-Enabled Clinical Decision and Diagnostic Support Tools Using Real-World Data, Margolis Centre for Health Policy

For a thorough assessment of AIaMD devices, data elements should include the algorithm outputs (i.e. model recommendation), algorithm inputs (the data input), observed outcome, and any demographic subgroup analysis variable.

The Centre also notes the challenges of collecting real-world data, stressing the need for reliable data collection inclusive of socioeconomic factors and patient-reported outcomes.

2 - Performance Monitoring and Validation of AIaMD

Performance monitoring is crucial for the ongoing evaluation of AIaMDs, performance monitoring. The CLEAR guidelines stress that manufacturers must validate their selection of performance measures, which should align with those set during development and reflect the clinical application of AIaMDs. The performance criteria of AIaMDs are affected by the intended clinical use case; greater scrutiny will need to be in place for AIs used independently by patients due to the lack of clinician oversight, in contrast to those deployed in healthcare systems.

Performance analysis should consider demographic factors and image artefacts, and the impact of AIaMD on healthcare teams and patients should be subject to regular review, ensuring it aids the diagnostic process.

The medical algorithmic audit outlines a testing phase with three components:

- **Exploratory error analysis**: A systematic review of algorithmic errors, examining false positives and negatives in the classification systems to identify common elements among the errors
- **Subgroup testing:** To identify high-risk populations within the target groups, analysing AI performance across patient-specific (e.g. age, sex, ethnicity) and task-specific (e.g. lesion location, any other clinically relevant factors) subgroups
- **Adversarial testing**: Assessing AI model behaviours in high-risk or "worst-case" scenarios to understand the prevalence and sources of errors in these scenarios

Post-testing, the audit's reflection phase advises manufacturers and users to develop risk mitigation strategies. Developers might improve models with more diverse data, alter thresholds, or revise usage instructions, while clinical actions could include standardising image acquisition or increasing human oversight, especially for demographics more susceptible to AI errors. Where risks persist and performance is inadequate, withdrawal of the AIaMD system is an option.

This audit methodology has been applied in preclinical evaluations of deep learning systems in detecting proximal femoral fractures[47] and at University Hospitals Birmingham NHS Foundation Trust[48], offering a collaborative framework for planning PMS.

The existing literature on deploying AI applications is particularly rich in radiology. The Royal College of Radiologists AI Working Group[49] has put forward recommendations for AI evaluation, including contrastive analyses of AI and clinician diagnoses to identify both agreement and divergence, which helps pinpoint where AI errors are likely to occur. They also

[47] Oakden-Rayner, Lauren, et al. (2022), Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study
[48] Aditya Kale, University of Birmingham
[49] Ross et al. (May 2024) Royal College of Radiologists AI Working Group

advocate for monitoring data drift and automation bias to ensure AI performance consistency over time.

Their evaluation approach advocates for the examination of diagnostic complexity and clinician training, as well as AI's integration into clinical workflows. Understanding both positive and negative outcomes is key for gauging AIaMD's impact on clinical effectiveness and efficiency. They propose centralised audits across sites, with transparency in evaluation data and performance metrics, to facilitate the identification of common issues experienced in different deployment settings.

However, a multi-society commentary[50] highlights the lack of standardised guidelines for AI assessment in radiology, primarily due to difficulties in defining universal benchmarks. Their suggested solution is a tailored performance evaluation, involving periodic, random case reviews against set standards.

In Dermatology, one paper outlines suggested monitoring practices for the PMS of DERM, an AIaMD in use in the UK[51]. These practices include:

- **Quality Control and Root-Cause Analysis**: Quality control for AIaMDs includes detecting errors, monitoring updates and analysing performance drops. False negatives and near-misses should undergo root cause analysis, and findings should be recorded in a risk registry to identify common issues.
- **Bias Assessment:** To ensure data validity, losses to follow-up, ineligible assessments, and technical failures must be documented. Protocols should define time intervals to confirm no repeat presentations, which could indicate missed cancers. The use of AIaMD in final diagnosis should be clear to prevent bias.
- **Patient Outcomes:** Beyond AIaMD performance, clinically meaningful metrics such as time to diagnosis and treatment, lesion characteristics and longer-term outcomes like progression-free or overall survival should be monitored to assess the standards of care for patients.
- **Second-Read Review:** An appropriate safety netting measure at the initial phase of implementation is a second-read review of benign cases marked for discharge to mitigate human errors and build confidence in the usage of the tool

3 – Transitioning to Autonomous AIaMD Use

The Centre for Assuring Autonomy at the University of York is conducting ongoing research to support the safe introduction of autonomous technologies to health and social care. Key recommendations from their published framework of guidance on assurance and regulation for autonomous systems include[52]:

- **Upskilling Healthcare Professionals**: Training is essential to ensure that healthcare providers fully understand the AI technology's capabilities and limitations.
- **Defining Care Pathways**: Establish clear care pathways, including the roles and authority of AI systems, and delineate the specific capabilities of AI within these pathways.

[50] Brady et al, (2024) Developing, purchasing, implementing and monitoring AI tools in radiology
[51] Thomas L, et al (2023) Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance
[52] University of York, Centre for Assuring Autonomy, Guidance

- **Characterising AI Capabilities**: Contextualise the AI's capabilities within the care pathways, outlining its intended uses, limits of authority, and required monitoring mechanisms.
- **Monitoring for Potential Harm**: Implement a monitoring system to evaluate any potential harm to patients arising from the use of AIaMD.
- **Developing Standard Operating Procedures (SOPs)**: To create SOPs for proactive and reactive monitoring regimes to maintain efficiency and effectiveness and establish escalation protocols to address any issues that arise during AI operation.
- **Establishing Handover Procedures**: Define safe operational limits for AI handover to human operators and vice versa. Set criteria for re-engagement of the AIaMD, detailing the conditions and processes for its reactivation.
- **Maintaining Audit Logs**: Keep detailed records of AI operations, including incidents or near-misses, to continually refine safety management processes and enhance the understanding of AI technology.

4.3.2 Practical Post-Market Surveillance Methods

The current literature reveals a gap in practical clinical auditing methods for the long-term surveillance of AIaMD safety. To address this, we outline a potential methodology to conduct safety audits, exemplified through statistical analysis and simulation modelling. Central to this method is demonstrating that the AIaMD maintains a high NPV above a pre-defined standard. The meta-analysis presented in Section 2.2, which evidences an NPV for the detection of melanoma at 98.9% among dermatologists diagnosing a population with a similar disease prevalence to that of the DERM cohort, supports a recommended NPV standard exceeding 99%. It is important to note that this benchmark is derived from the best evidence currently at our disposal.

Conducting audits to assess NPV performance changes needs to account for two key elements:

1. **Sample selection:** A representative sample of lesions (and patients) diagnosed as benign, who would otherwise be discharged from care, are instead reviewed as if they were categorised under the high-risk pathway. The sample size needs to be sufficiently large to ensure enough data to determine AI's performance with the necessary statistical power and confidence levels.
2. **Frequency of auditing cycle**: Determining the optimal frequency to detect any significant decline in NPV promptly, considering the practicalities and resources available.

Additionally, consideration should be given to defining and obtaining the necessary data required for the calculation of NPV, including clinical diagnosis and histology reports.

The following text covers an example of quantitative methodology for monitoring NPV and provides a start for conversations on feasibility and responsibility.

1 – Sample Size Determination

To exemplify the steps required to assess a sufficient sample size to monitor and detect any significant decrease in NPV, we have conducted a one-sided power[53] analysis to determine the minimum sample size needed to detect a significant decrease in NPV. Specifically, the sample

[53] "Power" refers to the ability of a test to detect a meaningful effect or difference when it does exist. 80% power is the standard in most research applications. The power level relates to the chance of correctly identifying a significant decrease in performance if it occurs. "One-sided power analysis" determines the minimum sample size needed to detect a significant effect in only one direction (e.g., a decrease in performance) with a specified level of confidence.

size required to detect a drop of NPV from 99.8% (the current performance of the AI) to below 99% with an 80% power at a 2.5% significance level. The analysis was conducted in R Studio, using the package *pwr*. Cohen's h was chosen as the measure of effect size.

We set key parameters for the power analysis as follows:

- Prevalence of melanoma at 2.478%[54]
- Baseline NPV at 99.8%
- NPV detection threshold at 99%
- Power at 80%
- A significance level of 2.5%

Our statistical analysis suggests that an adequate sample for this context would be 660 lesions (approximately 570 patients). Note that this is not prescriptive, but rather an example of how this analysis may be carried out; revisions of this figure may be necessary, for instance, if there is a change in baseline NPV (due to new technologies or performance changes), or to consider factors such as specificity and sensitivity.

2 – Sampling Frequency

Determining the optimal frequency of NPV audits would be the next step in a potential surveillance methodology. While current MHRA regulations only require manufacturers to submit a performance report annually, we encourage a more proactive approach to assess the performance of NPV, particularly in the initial phases of autonomous AIaMD deployment. This also aligns with the monitoring approach already adopted by SA.

To assess the impact of sampling frequency on NPV checks, we have conducted simulation modelling for intervals of three months, four months, and biannually (six months):

1. We modelled synthetic cohorts of patients under three scenarios of performance issues, with NPV dropping from 99.8% to 98%, 97% and 96%
2. Sampling is modelled using a binomial test, which flags whether NPV performance has dropped below a 99% threshold, at a 2.5% significance level
3. Aligning with the sample size analysis, the simulation modelled the testing of 660 lesions every time the population was sampled
4. We ran these simulations 100,000 times to obtain an average time to detect an NPV drop

While this example focuses on NPV, the prevalence of disease across all lesions should also be monitored alongside NPV to ensure no significant changes, given the impact on NPV.

It is important to note that our modelling assumes NPV changes occur at random intervals, which may or may not match real-life scenarios. Our results, outlined in Table 4, demonstrate that more frequent testing leads to more rapid detection of any NPV decline. Specifically, the quickest detection occurs when sampling every three months, ranging from 45 to 109 days depending on the extent of the NPV drop. If the drop of NPV is larger, the time to detect the drop is shorter across all sampling frequencies.

[54] Note that we do not expect significant shifts in the prevalence of melanoma in this population, as it has maintained a similar level throughout the years of DERM data we have analysed. However, significant shifts in prevalence may make NPV comparisons less reliable and therefore further power analysis should be carried out.

Furthermore, at the chosen significance level, increased testing frequency does not inflate the rate of detecting false positives (detecting a drop in performance of the NPV when there is no change), which is a common risk with repeatedly running a statistical test (Appendix 7.5).

While greater frequency of checks would provide closer monitoring, there might be practical hurdles: feasibility of obtaining histology data, reduced benefits from avoided F2F appointments, added burden to analyse data for NPV and increased costs (see Section 5.4 for an estimated cost of one auditing cycle). This highlights the importance of tailoring this methodology to real-world operational contexts.

NPV Scenarios Dropping from 99.8%	Check every 3 months (91 days)	Check every 4 months (121 days)	Check every 6 months (182 days)
98%	109 days	174 days	280 days
97%	48 days	74 days	152 days
96%	45 days	64 days	122 days

Table 4. Results of simulating modelling, outlining the number of days to detect the drop of NPV performance across sampling frequencies and NPV scenarios

4.4 Recommendations for Surveillance

Drawing on previous sections, we provide a summary of recommendations, as informed by the existing literature discussed above, as well as potential practical strategies for PMS implementation.

4.4.1 High-Level Recommendations

Based on the literature discussed in Section 4.3.1, we have mapped out a series of recommendations relating to PMS that encompass both deployment sites and manufacturers and could be flexibly adapted to suit different sites and contexts.

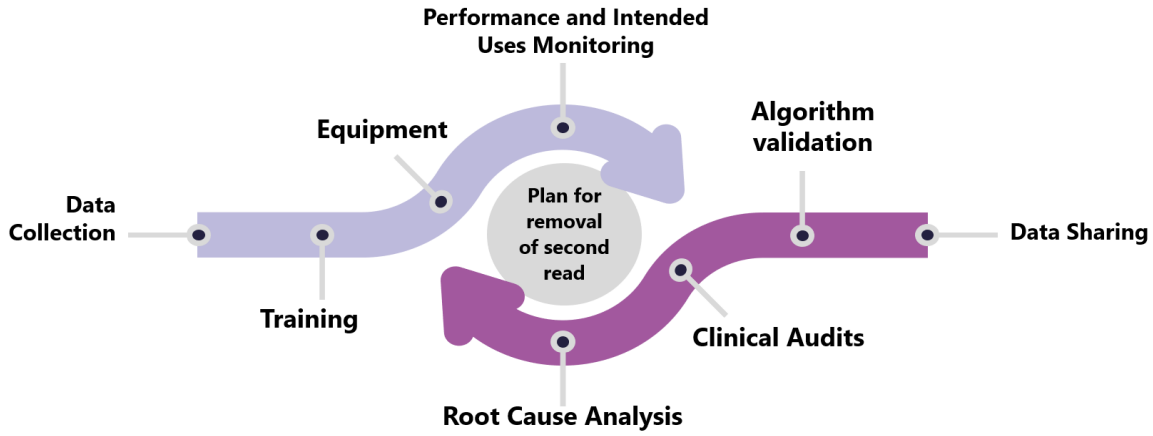


Figure 11. Diagram summarising the themes of the high-level recommendations for Post-Market Surveillance derived from existing literature

Data Collection and Data Sharing should be agreed upon before deployment, ensuring alignment with privacy and data sharing laws. The types of data shared for performance monitoring would typically include skin lesion images, patient-specific information for bias analysis, histology reports for AI diagnostic confirmation and standardised image assessment information. Protocols on data sharing need to be set, specifying the conditions under which data can be used for further AI training, and deployment sites should support the set-up of timely data-sharing mechanisms to support regular performance analysis. Manufacturers should commit to making AI performance data publicly available within an agreed timeframe.

Algorithm Validation is the manufacturer's responsibility; they must ensure ongoing training and validation of the AI using appropriate datasets. Regular reviews of the AI algorithm are necessary, following guidelines such as the CLEAR checklist to review development processes. Changes in algorithms should be completed in line with Standard Operating Procedures which may include informing a notified body. Each update should be communicated to sites and signed off. Manufacturers should ensure reliable AI performance through adversarial testing.

Equipment maintenance and communication on equipment updates are part of the manufacturer's duties, and they should provide comprehensive instructions for use and safety. Deployment sites are expected to adhere to these guidelines, use high-quality imaging equipment, and stay vigilant for any issues with the devices, ensuring AI performance is consistent across various clinical environments.

Training at the deployment site is imperative for staff to acquire images under standardised conditions and follow exclusion criteria. Training plans should be collaboratively established with manufacturers, emphasising the correct use of AIaMD and associated equipment.

Risk Management involves creating a risk registry database to track and analyse performance issues, maintained through collaboration between deployment sites and manufacturers. Regulatory reporting compliance is also recommended, including submitting reports for adverse events under the Yellow Card Scheme.

Performance and Intended Uses Monitoring requires ongoing oversight of the image acquisition process and regular feedback from users to manufacturers. Manufacturers must generate detailed performance reports, including benchmarks and subgroup analyses, to be shared with providers on an ongoing basis.

Clinical Audits / Service Evaluation recommends manufacturers to perform thorough regular audits of the AI's performance. Deployment sites, on their part, should assess the AI's impact on clinical practice and patient satisfaction, potentially conducting their performance evaluations if they have the capacity.

Root Cause Analysis should be conducted by manufacturers with input from dermatologists at deployment sites, focusing on understanding and mitigating false negatives. Deployment sites should convene consultant dermatologist panels to partake in these analyses to formulate strategies for risk mitigation.

Plan for Removal of Second Read outlines how deployment sites and manufacturers might agree on the introduction of autonomous AI use. It is ultimately a shared decision between manufacturers and deployment sites how this takes place; for instance, new sites may wish to rely on existing evidence of AI performance within similar populations to directly introduce an autonomous pathway within their locality. We note that providers interviewed in Section 3.2 reported benefits from the initial use of second reads, such as building confidence among staff and ensuring findings captured elsewhere to match the local population. However, this is not a strict requirement for future sites.

An example plan for the removal of second reads may be as follows:

- The AIaMD is deployed, either with initial second reads for all cases to support implementation and clinical buy-in or autonomously.
    - Once autonomous use is implemented (which may be immediately or after a settling-in period), an agreed number of benign cases continue to receive second reads to collect data for NPV analysis at an agreed interval frequency.
- Contingency plans are developed and signed off for scenarios where the NPV performance drops below the standard.
    - For example, deployment sites agree safety netting actions to be implemented if a drop in NPV is observed, which may be proportional to the degree of NPV drop. These may include additional retrospective reviews, increasing sampling frequency, or temporary introduction of a full second read of lesions if there is serious concern that the AI performance has dropped below safe levels.
    - At a minimum, full root-case analysis is recommended as per standard PMS approaches.

Though not explicitly mentioned as part of PMS strategies within the literature, patient education should be featured throughout the pathway. This includes educating patients on their condition, clearly communicating the outcomes of their assessment, and instructing them on when to seek further help. This ensures patients are well-informed about the implications of their assessment by an autonomous AI and are empowered to take responsibility for their health.

Summary High-level Recommendations for PMS

	Users / Deployment sites	Manufacturers
Data Collection	Maintain necessary IT infrastructure and capabilities to support data collection and integration with AI systems. Ensure accurate and complete data are captured during clinical use.	Agree with the deployment site on what data is required to ensure ongoing performance and any useful baseline values. Implement quality assurance processes for data collected from deployment sites and put in place plans for data audits.
Data Sharing	Agree with manufacturers on how data (e.g. histology reports) will be shared with consideration of data privacy and data sharing regulations. Agree on long-term data ownership.	Engage with healthcare providers to facilitate data sharing and address contractual obligations. Ensure NHS data is kept secure and long-term ownership is discussed with deployment sites.
Algorithm validation	Sign off clinical risk management documentation when the algorithm is	Perform rigorous validation of algorithms on appropriate training, validation and

	Users / Deployment sites	Manufacturers
	changed, before the implementation on local populations.	testing datasets and update them based on real-world performance.
Equipment	Ensure correct hardware (e.g. dermoscopy camera) and approved tools are used to collect skin lesion images.	Regular communication with trust when algorithms are updated and provide the most updated equipment.
Training	Facilitate appropriate training for relevant staff in image acquisition (e.g. images taken under standardised conditions), ensuring the application of standard protocols and adherence to appropriate exclusion criteria for image assessment by the AI.	Work with the deployment site to set out training plans for the use of the AIaMD, specifically the appropriate inclusion and exclusion criteria for AI image acquisition. Standard Operating Procedures could be created to ensure consistency across deployment sites.
Risk Management	• *Vigilance*: Risk registry database to identify performance issues and their causes. Maintain DCB0160 documentation. Proactive monitoring and adverse events reporting (e.g. MHRA Yellow Card Scheme). • *Criteria*: Deployment sites/hubs should monitor the application of inclusion and exclusion criteria in image acquisition. • *Repeat presentation*: Repeat presentation of the same lesion should be actively searched by the deployment site or manufacturer.	• *Performance*: Risk registry database to identify performance issues and their causes. • *Repeat presentation*: Repeat presentation of the same lesion should be actively searched by the deployment site or manufacturer. • *Vigilance and regulatory compliance*: Ensure compliance with regulatory requirements and guidelines, including performing proactive post-market monitoring and adverse events reporting (e.g. MHRA Yellow Card Scheme).
Performance and Intended Uses Monitoring	Deployment sites or hubs should monitor the application of inclusion and exclusion criteria in image acquisition.	Generate and disseminate performance reports, including benchmark data and subgroup analyses. • Benchmarks – Specificity, Sensitivity, PPV, NPV • Subgroup analyses (e.g. by lesion site, Fitzpatrick skin types, age ranges) This real-world evaluation of performance should be made publicly available.
Clinical Audits or Service Evaluation	Participate in auditing processes, including reviewing discordant cases, and evaluating clinical outcomes. Engage in service improvement, monitor user, and patient satisfaction and pathway implementation.	Work with deployment sites to determine the feasibility of sampling a set number of scans at a feasible frequency for AI performance evaluation.
Root cause Analysis	Put together a panel of consultant dermatologists to take part in the root cause analysis (RCA). Involve dermatologists in the RCA panels to	Set out a robust root cause analysis process for false negatives, involving relevant dermatologists from deployment site to ensure a comprehensive understanding of the clinical impact.

	Users / Deployment sites	Manufacturers
	investigate the causes of false negatives and develop mitigation strategies.	
Plan for removal of second read	Agree on process for implementing autonomous use (either de-novo or following initial second read review).	Agree on process for implementing autonomous use (either de-novo or following initial second read review).

Table 5. Summary of high-level recommendations for Post-Market Surveillance encompassing deployment sites and manufacturer responsibilities. These themes are collated from a literature review of existing recommendations for PMS of AIaMD and medical devices

4.4.2 Practical Post-Market Surveillance

In Section 4.3.2 we discussed an example methodology for practical PMS, which involves the selection of a suitable sample size of lesions to allow NPV analysis with sufficient power to detect small drops in NPV and establish adequate intervals for performance testing.

There are upsides and drawbacks to more frequent monitoring of NPV, namely a shorter detection interval for NPV drops on the one hand, and a greater draw on Trust resources on the other, including reduced savings from avoided F2F appointments and increased administrative and consultants' time for reviewing benign lesions.

A potential middle ground is an initial frequency of audits of 4 months, which could later be reduced to 6 months or yearly. However, deployment sites and manufacturers will ultimately need to agree on a suitable monitoring frequency, as well as who might be best placed to carry out analyses – whether manufacturers, central bodies, sites or external auditors.

It should be noted that monitoring NPV is not a substitute for other important steps to ensure patient safety on the pathway. For instance, patients should be adequately educated and empowered to take responsibility for their health, where possible, by providing sufficient information on discharge. Deployment sites may want to establish additional safety netting mechanisms, such as internal audits to search for repeat attendees and monitor long-term outcomes. Additionally, the high-level recommendations derived from the literature list other important practical PMS strategies, such as reporting faults through the MHRA Yellow Card Scheme and assessing any potential false negatives with AI adversarial testing and root cause analysis.

Moreover, the PMS methodology covered here relates specifically to the use of AI as a triage tool in skin cancer pathways, where the most important safety aspect is the accurate recognition of benign lesions. Should the AI be applied to a different use case, then other summary estimates of test accuracy may become more relevant, such as Sensitivity, Specificity or PPV.

5 Illustrative Budget Impacts

The below section provides an illustrative analysis of preliminary costs and savings associated with the implementation of AI within skin cancer pathways. This is intended as a high-level analysis to support the framing of the work contained within this report but does not constitute a comprehensive health economics analysis.

The Exeter Test Group is currently developing a comprehensive cost-effectiveness model for the implementation of DERM, which is undergoing validation by NICE. As such, the insights provided here should be regarded as a preliminary high-level perspective on the potential system-wide costs and savings associated with the use of DERM.

This analysis considers two scenarios: implementation of AIaMD with second reads (Scenario 1) and implementation of AIaMD with autonomous management of benign lesions (Scenario 2). We have relied upon several assumptions, outlined in Table 6.

Assumption	Value	Source
Population size	1,000,000	Example
Rate of USCR referrals for skin cancer by GPs	0.0101	USCR per skin cancer, rate per 100,000 (PHE)[55]
Average number of lesions per case	1.16	Calculated from SA data
Skin Analytics service cost per 10,000 population	£4,200	Provided by SA
Discount per 10,000 population	£250	Provided by SA
Cost per SA dermatologist review (per case)	£20	Provided by SA
Average Imaging Appointment Cost	£17.30	From data provided by SA and UHL Evaluation (2023)[56]
Post-Referral equipment cost per case	£0.86	Provided by SA
Consultant medical cost per working hour	£109	Unit Costs of Health and Social Care (2023)[57]
Average OP Appointment cost	£217	Unit Costs of Health and Social Care (2023)[55]
Average Trust Consultant time to review a lesion (mins)	7	Calculated from SA data[58]
Cost per Trust Consultant review (per lesion)	£13	Calculated from the above
Cost per Biopsy of Skin (per lesion)	£554	NIHR Unit Costs 2023/24 investigation code 11100[59]

Table 6. Assumptions used to develop the Health Economics model. Discount applies when Trusts agree to share histology outcome data with SA for training

We recognise several considerations for a comprehensive cost-effectiveness assessment of AIaMD implementation that are not covered by our analysis. These include the impact of cancer progression rates and the associated cost utilities with earlier cancer diagnosis.

[55] USCR per skin cancer, rate per 100,000. Public Health England

[56] An evaluation of AI Powered Tele Dermatology for Skin Cancer 2WW Pathway, Health Innovation East Midlands and Edge Health

[57] Unit Costs of Health and Social Care (2023)

[58] This was obtained from the difference in timestamps between a lesion review and the next on the same day. Lesions that took 45mins or more to review were excluded under the assumption that another task took place between reviews. Note that there are organisational variations in the time per review, ranging on average from 3.6 to 17 minutes.

[59] NIHR Unit Costs 2023/24

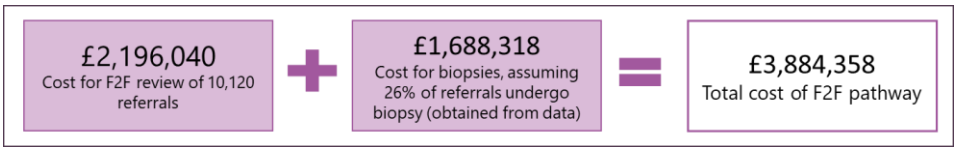
Additionally, the downstream costs and benefits related to cancer detection, such as those for follow-up and treatment are not included here. There are benefits such as cost utilities spared by reducing psychological distress associated with late cancer diagnosis, as well as the benefit of providing earlier reassurance to patients through the autonomous pathway. Furthermore, the wider societal costs including patient expenses, the economic impact of unemployment or time off work, travel time or cost to attend procedures such as biopsy will need to be evaluated.

5.1 Costs

Current Pathway

According to current pathways that have not adopted teledermatology, all patients are reviewed F2F within USC clinics. A proportion of patients undergo further biopsy and pathology investigation. Analysed data from Skin Analytics suggests that currently, 26% of referrals undergo biopsy. As all patients are currently reviewed by a dermatologist, we have assumed that this proportion is similar to that of current F2F USC pathways.

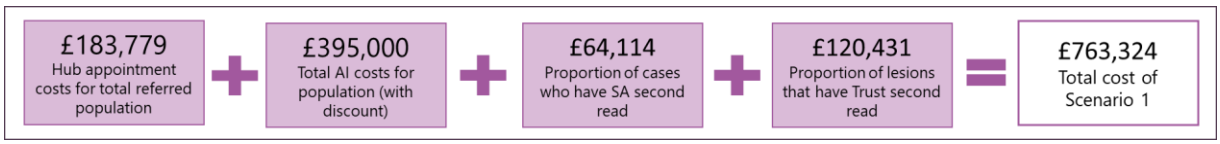
Using the above assumptions, the cost of current pathways for the example population amounts to **£3,884,358**.



Scenario 1: SA-dermatologists to perform benign lesions second reads and Trust dermatologists continue to triage all cases not discharged by SA

The above scenario describes the current implementation model, where an SA dermatologist reviews all benign lesions and either discharges or forwards to a Trust dermatologist, while all high-risk lesions, as well as the ones that could not be assessed by the AI, are first reviewed by a Trust dermatologist remotely before a decision to review F2F.

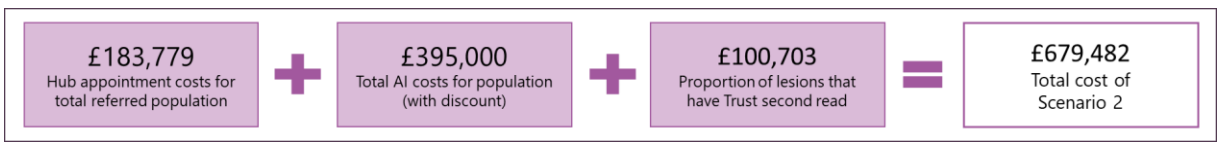
In this scenario, the total cost amounts to **£763,324.**



Scenario 2: All lesions assessed as benign by DERM are discharged

If no second reads of benign lesions are performed, Trust dermatologists will only need to review lesions marked as malignant by DERM, as well as lesions that were not assessed by DERM.

In this scenario, the total cost amounts to **£679,482.**



Note that for both scenarios, sites that employ Community Diagnostic Centres as photography hubs will incur a tariff of £102 per case, as reported by NHSE, though this figure is subject to revision in the near future.

5.2 Savings

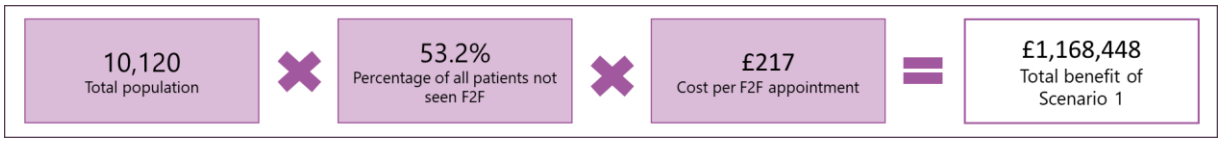
Financial Benefits

For this high-level, generalisable, analysis we have considered quantifiable benefits relating to cost-savings in F2F reviews as well as avoided biopsies for Scenario 2.

Scenario 1

At present, 53.2% of patients are not seen in F2F clinic following an initial AI assessment and teledermatology review by either a Trust dermatologist (suspicious lesions) or both a SA and Trust dermatologist (benign lesions that are overturned). Avoided F2F clinics attract direct savings as shown below.

The total costs saved for these scenarios amount to **£1,168,448**.



In this scenario, additional benefits are derived from Trust dermatologists' timesaving. As teledermatology reviews take on average less time than F2F reviews, and a proportion of F2F reviews is avoided altogether, this scenario results in a yearly Whole Time Equivalent (WTE)[60] of 0.14 WTE, i.e. 241 hours of clinical time or 60 programmed activities.

Scenario 2

In the absence of a second read of benign lesions, further benefits are seen through a higher volume of avoided 2WW F2F appointments, as no patients flagged as benign by the AI would be overturned. Additional savings are enabled by further reductions in biopsies: currently, 3.2% of benign lesions are biopsied after being overturned by an SA dermatologist and reviewed by Trust dermatologists and are still found to be benign. This represents an additional saving.

The total costs saved for this scenario amount to **£1,553,495.**

Autonomous AI use enables further dermatologist time savings, equivalent to 0.36 WTE across a year, i.e. 628 hours of clinical time or 157 programmed activities.

Wider Benefits

Apart from financial savings, the deployment of AI in the pathway has yielded several non-quantifiable benefits.

[60] NHS-BSA definition of a medical WTE as 40 hours (or 10 programmed activities) across 44 working weeks a year.

Our analysis includes benefits from reduced biopsies in Scenario 2, though two providers reported already experiencing a reduction in biopsies even without autonomous AI use. It could be assumed, therefore, that savings from avoided biopsies might be greater. Note that reduction in biopsies also frees downstream surgical and pathology capacity, as well as sparing patients from a procedure and the need to attend hospital.

Providers have noted that the use of AI has streamlined operational workflows, leading to a reduction in waiting lists and better prioritisation of urgent cases. Additionally, patient experience has been enhanced through faster diagnosis, as evidenced by the results of Provider 3's patient satisfaction surveys. At University Hospitals of Leicester, reductions in F2F reviews for USC cases resulted in a direct increase in F2F reviews for routine patients, suggesting that the pathway can release capacity[61].

5.3 Summary of Economic Analysis

Cost Benefit Analysis

By comparing the costs and benefits discussed above, we calculated the cost-benefit ratio in each Scenario. This describes the monetary return to the health system for each £1 invested.

	Total Costs	Total Benefits	Cost Benefit Ratio	Net Savings	Net Savings Per Case
Scenario 1	£763,324	£1,168,448	1.5	£405,123	£40
Scenario 2	£679,482	£1,553,495	2.3	£874,014	£86

Table 7. Summary of Cost Benefit Analysis for the implementation of DERM as a triage tool in skin cancer pathways

Cost Comparison Analysis

We also performed a cost comparison analysis, to enable a more direct comparison of costs for Scenarios 1 and 2 to the current standard of care, defined above in the Current Pathway costs.

	Total Costs	Cost Savings to Current	Cost Savings Per Case
Current (all seen F2F)	£3,884,358	-	-
Scenario 1	£3,479,234	£405,123	£40
Scenario 2	£3,010,344	£874,014	£86

Table 8. Summary of Cost Comparison Analysis for the implementation of DERM as a triage tool in skin cancer pathways

It is worth noting that, while this report explores system-level costs and savings, further considerations should be made on adequate commissioning arrangements and sustainability. The AI-enabled pathway essentially allows for lower system costs for each patient referred on

[61] An evaluation of AI Powered Tele Dermatology for Skin Cancer 2WW Pathway, Health Innovation East Midlands and Edge Health

an USC pathway by reducing the need for F2F reviews and biopsies. It also improves patient experience by allowing faster diagnosis or reassurance of worrying skin lesions.

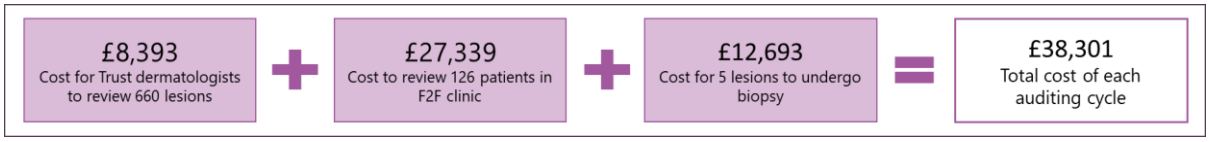
In practice, however, given the demand for Dermatology, freed-up capacity is rapidly taken up by other patients on waiting lists, masking reduced costs. Appropriate commissioning arrangements should reflect the enhanced service provided to patients, which is likely to contribute to preventing harm from long waits for both USC and routine referrals.

5.4 Illustrative Savings Reduction for PMS

In Section 4.3.2 we explored an example methodology for practical PMS of AIaMDs. There, we raised the consideration of costs to carry out PMS that would reduce total savings from autonomously implementing AIaMD. Below, we illustrate potential costs that the example methodology might attract by requiring additional teledermatology reviews as well as potentially leading to additional F2F reviews.

To calculate these costs, we employed relevant assumptions from Table 6 on consultant hourly costs and time for teledermatology review, appointment and biopsy costs.

According to the sampling methodology, each PMS audit cycle would require approximately 570 patients (660 lesions) to be reviewed, who would otherwise be discharged. Using current proportions for benign cases who are seen face-to-face (25.7%), and the proportion of benign lesions that undergo biopsy (3.2%), we estimate that the nationwide savings reduction of each auditing cycle would amount to **£38,301**.



It is important to note that these reductions in savings are not individual to each provider, who would be contributing cases for PMS in proportion to their activity, as long as audits are carried out centrally or by the manufacturer. Additionally, savings reductions would only affect sites implementing autonomous AIaMD, as other sites that continue to second read all benign cases would already incur these costs.

If we assume costs are shared between the ten pathways whose data we have analysed, savings reductions per site, per auditing cycle, might range from £76 to £12,754, if we assume a proportional contribution of cases for PMS based on the total activity carried out at each site.

6 Conclusions

AI holds considerable promise for improving the efficiency and effectiveness of skin cancer pathways, addressing present challenges within a shorter timeframe than much-needed long-term solutions such as workforce planning. These technologies have the potential ability to reduce clinician workload by accurately triaging benign lesions, thereby allowing specialists to focus their expertise on urgent and complex cases. The successful incorporation of AIaMD into routine dermatological practice, however, hinges on ensuring patient safety and maintaining the high standards of care that the public expects from the NHS.

Our analysis has delineated the current dermatological landscape, characterised by a rise in melanoma incidence rates and exacerbated by consultant shortages that contribute to increasing waiting times. We have detailed the deployment of AIaMD in skin cancer pathways—with a focus on DERM as the only current AIaMD with appropriate regulatory clearance for use in the NHS—and provided an assessment of standards of care for dermatologists documented within the literature. We have measured DERM's performance against these standards, concluding that DERM's NPV is at least as good as that of dermatologists as reported within available literature, setting a precedent for the evaluation of similar technologies in future.

This report has also considered the economic implications of different AIaMD implementation strategies, outlining the potential cost-effectiveness of integrating AIaMD into skin cancer pathways. Our budget impact analysis indicates that AIaMD deployment could yield net benefits and cost savings compared to the current pathway, with the added advantage of reduced wait times for patients and avoiding unnecessary biopsies.

It is imperative, however, to approach the integration of AIaMD with a clear strategy for PMS. The regulatory landscape is evolving, so close monitoring of AIaMD performance and adopting robust PMS practices are essential. We have explored a dual approach, encompassing high-level strategic recommendations informed by comprehensive literature reviews and exemplified practical methods for real-world surveillance.

The strategic implementation of AIaMD in Dermatology offers a pathway to address current and forecasted demand. This report has laid the groundwork for the safe and effective use of AI technologies in skin cancer pathways, providing a blueprint for healthcare providers, policymakers, and industry stakeholders. As the integration of AIaMD in healthcare continues to evolve, it will be paramount that all stakeholders remain adaptable, responsive, and committed to the principles of patient safety, clinical excellence, and equitable access to care.

7 Appendices

7.1 Rapid Meta-Analysis

7.1.1 Study Characteristics

Study	Dermatology Setting	Study Design	Care Setting	Country	Age or gender	Ethnicity	Inclusion Criteria	Exclusion Criteria	Clinicians experience	Positive Lesion Definition	Diagnostic Method
Binder 1994	Teledermatology	Case-Control	Secondary	Austria	Not Reported	Not Reported	Images of pigmented skin lesions randomly selected from a pigmented skin lesions image database	Not Reported	High experience	Melanoma (in-situ and invasive, or not reported): 4	Dermoscopy (Modified) pattern analysis. Single observer (n = 3). Dermatologist. High experience
Gilmore 2010	Teledermatology	Case Series	Secondary	Austria	Not Reported	Not Reported	Polarised dermoscopic images of atypical melanocytic lesions	Not Reported	High experience	Melanoma (in-situ and invasive, or not reported): 36	Dermoscopy: No algorithm; dermoscopic method of diagnosis Not Reported. Single observer (n = 1). Dermatologist.
Seidenari 1998	Teledermatology	Case-Control	Secondary	Italy	Not Reported	Not Reported	Melanomas and benign pigmented skin lesions from a larger series of pigmented skin lesions were used to develop a new automated classifier; all melanomas with x20 magnification images were included plus a random sample of benign lesions with the same magnification. For the larger series, lesions were referred by dermatologists or general physicians because of 1 or more pigmented skin lesions that were difficult to interpret on clinical grounds alone, numerous pigmented skin lesions, or because the patients were at increased risk for melanoma or had had a	Not Reported	Mixed. 1 high 1 low	Melanoma (in-situ and invasive, or not reported): 31	Dermoscopy No algorithm. Single observer. 2 dermatologists

Study	Dermatology Setting	Study Design	Care Setting	Country	Age or gender	Ethnicity	Inclusion Criteria	Exclusion Criteria	Clinicians experience	Positive Lesion Definition	Diagnostic Method
							malignant pigmented skin lesions in the past				
Kroemer 2011	Teledermatology	Not Reported	Secondary	Austria	Mean: missing; median: 69; range: 3–93 years	Missing: not stated, but they were Austrian	People self-referred or referred by a local doctor for evaluation of a skin tumour. Men or women with benign or malignant (or both) skin tumours of melanocytic or non-melanocytic origin	3 declined participation. In 33% of cases, no history could be obtained. Clinical and 18 dermoscopic pictures were inadequate, so 104 tumours from 80 participants were included	High experience (board certified with clinical expertise in teledermoscopy and dermoscopy)	Melanoma (invasive): 2; melanoma (in-situ): 1; lentigo malignant: 3	Dermoscopy: No algorithm. Clinical photographs and dermoscopic images. Single observer. Dermatologist
Bowns 2006	Teledermatology	Case Series	Specialist clinic	UK	Classified by age band; 61% were aged > 55 years. male: 46.9%	Not Reported	People (with skin lesions) who were either referred to the 2-week wait or 'target' clinics or those initially referred to the normal outpatient service but who were diverted by the consultant based on the referral form.	Not Reported	Not Reported, dermatologists	Melanoma: 19; MiS: 5; BCC: 29; cSCC: 16; other malignant: 1	Clinical and dermoscopic images (or both)
Congalton 2015	Teledermatology	Case Series	Secondary	New Zealand	Median: 58 (range: 15–92) years. male: 142; female 168	White: 242 (78%); black or African American: 12 (4%); Hispanic or Latino: 3 (< 1%); Asian: 16 (5%); other: Maori 16 (5%), Pacific islanders 12 (3%); missing: 12 (4%)	People referred from primary care with skin lesions suspicious of melanoma were triaged via a VLC instead of being seen FTF at a hospital clinic. Referrals that indicated 1–6 lesions of concern were included	Difficult to diagnose lesions – location/site of lesion skin lesions on scalp and genitals were generally excluded, as were those where body site was not identified in the referral	High experience dermatologist	Melanoma: 47; melanoma metastases: 1	Dermoscopy. Clinical and dermoscopic images. Single observer. 2 dermatologists
Grimaldi 2009	Teledermatology	Case Series	Secondary	Italy	Not Reported	Not Reported	Cutaneous pigmented lesions with digital images forwarded by primary care physicians to a referral centre for confirmation of diagnosis	Not Reported	High experience or 'Expert'. dermatologists; plastic surgeons	Melanoma: 5	Dermoscopy and clinical photographs
Piccolo 2000	Teledermatology	Case Series	Unspecified	Austria	Median: 39.5 years; range: 3–91 years	Not Reported, pigmented skin lesions but unclear skin types	People with pigmented skin lesions were selected because of their diagnostic difficulty and subsequently excised for a histopathological evaluation.	poor-quality index test image (all images scoring 4 were excluded from the study)	High experience	Melanoma: 11	Dermoscopy and clinical photographs
Kroemer 2011	Face to Face	Not Reported	Secondary	Austria	Mean: missing; median: 69; range: 3–93 years	Missing: not stated, but they were Austrian	People self-referred or referred by a local doctor for evaluation of a skin tumour. Men or women with benign or malignant (or both) skin tumours of melanocytic or non-melanocytic origin	3 declined participation. In 33% of cases, no history could be obtained. Clinical and 18 dermoscopic pictures were inadequate, so 104 tumours from 80 participants were included	Not Reported	Melanoma (invasive): 2; melanoma (in-situ): 1; lentigo maligna: 5	Not clear from the paper how an in-person assessment was conducted but most likely VI of the skin (± use of dermoscopy) no

Study	Dermatology Setting	Study Design	Care Setting	Country	Age or gender	Ethnicity	Inclusion Criteria	Exclusion Criteria	Clinicians experience	Positive Lesion Definition	Diagnostic Method
											algorithm was described
Coras 2003	Face to Face	Case Series	Secondary	Germany	Not Reported	Not Reported	Pigmented skin lesions undergoing excision due to diagnosis of melanoma or atypical nevus, to rule out melanoma or at the participant's request	Not Reported	High experience	Melanoma (in-situ and invasive, or not reported): 16	Dermoscopy: pattern analysis
Warshaw 2010b	Face to Face	Case Series	Secondary	USA	Mean: pigmented: 66; non-pigmented: 71; range: pigmented: 23–94; non-pigmented: 21–94 years	White: pigmented: 97.1%; non-pigmented: 98.9%; black or African-American: pigmented: 1.3%; non-pigmented: 0.7%; other: pigmented: 1.5%; non-pigmented: 0.4%	People enrolled at the Department of VA dermatology clinic who required (or requested) removal of ≥ 1 skin neoplasm ('high-risk group') and participants who were referred to general dermatology clinic by non-dermatology healthcare providers for evaluation of a skin neoplasm (lower-risk group). Biopsied lesions only were included.	Individuals requesting or referred for skin tag removal only or with papulosquamous or eczematous conditions (non-neoplastic), previous biopsy of the lesion and inability to comprehend and give informed consent	Not Reported	Melanoma: 41	Dermoscopy and visual inspection
Piccolo 2000	Face to Face	Case Series	Multicentre	Austria (Graz)	Median age 39.5 years, (range 3–91 years). Male: 21 (53%); female 19 (47%)	Not Reported	Pigmented skin lesions were selected because of their diagnostic difficulty and were excised for a histopathological evaluation	Poor-quality index test images	Dermatologists (n = 1; exp High)	Melanoma (invasive): 11	Dermoscopy (no algorithm)
Ahnlide 2016	Face to Face	Case Series	Secondary	Sweden	Not Reported	Not Reported	Excised melanocytic skin lesions with recorded dermoscopy ABCD score and clinician's preliminary diagnosis	Previously biopsied lesions and wide excisions not included; other exclusion before enrolment included: invalid report or missing data (n = 34); visiting residents' data (n = 66); non-melanocytic on histology or benign melanocytic lesions with special patterns (e.g. papillomatous, congenital naevi and mucosal lesions) (n = 658)	Dermatology residents (n = 6; "residents were encouraged to consult the specialists in difficult cases"); dermatologists (n = 7)	Melanoma (invasive): 23; melanoma (in-situ): 23	Dermoscopy: no algorithm (clinician's preliminary diagnosis); ABCD
Bauer 2000	Face to Face	Case Series	Secondary	Italy	Not Reported	Not Reported	pigmented skin lesions examined and excised during a campaign for the early	Not Reported	High	Melanoma (invasive): 30;	Dermoscopy: no algorithm; possibly

Study	Dermatology Setting	Study Design	Care Setting	Country	Age or gender	Ethnicity	Inclusion Criteria	Exclusion Criteria	Clinicians experience	Positive Lesion Definition	Diagnostic Method
							diagnosis of cutaneous melanoma			melanoma (in-situ): 12	based on pattern analysis
Benelli 1999	Face to Face	Case Series	Dermatologic Surgery Department	Italy	Not Reported	Not Reported	All pigmented skin lesions were observed and excised at the Dermatologic Surgery Department	Not Reported	Not Reported. Dermatologist	Melanoma (invasive): 54 (13.5%); melanoma (in-situ): 6 (1.5%)	Dermoscopy 7FFM
Carli 1994	Face to Face	Case Series	Secondary	Italy	Mean age 36 years; median age 33; all > 20 years; male: 31%	Not Reported	Clinically suspicious melanocytic lesions undergoing excision for diagnostic purposes	Not Reported	Not Reported; likely dermatologist	Melanoma (invasive): 3; melanoma (in-situ): 2	Dermoscopy: pattern analysis; criteria derived from several other studies
Carli 2002a	Face to Face	Case Series	Secondary	Italy	Not Reported	Not Reported	Clinically suspicious melanocytic lesions undergoing excision for diagnostic purposes	Not Reported	Not Reported; likely dermatologist	Melanoma (invasive): 3; melanoma (in-situ): 2	Dermoscopy: pattern analysis; criteria derived from several other studies
Cristofolini 1994	Face to Face	Case Series	Secondary	Italy	Not Reported	Not Reported	Patients with pigmented lesions presenting during a campaign for the early diagnosis of cutaneous melanoma at the Dermatology Department in Trento	Not Reported	High experience. Dermatologist n = 4	Melanoma (in-situ and invasive, or not reported): 33	Dermoscopy: pattern analysis
Dreiseitl 2009	Face to Face	Case Series	Specialist clinic	Austria	Not Reported	Not Reported	Patients presenting at the pigmented skin lesions clinic	Not Reported	Dermoscopy: no algorithm	Melanoma (in-situ and invasive, or not reported): 27 participants; 31 lesions	Dermoscopy: no algorithm
Durdu 2011	Face to Face	Case Series	Secondary	Turkey	Mean age: 48 years (4-85 years). male: 64; 36.4%	Not Reported	Pigmented skin lesions that could not be diagnosed with only dermatologic physical examination	Not Reported	Dermatologist Not Reported	Melanoma (in-situ and invasive, or not reported): 10; BCC: 34; 1 pigmented mammary Paget disease; 1 pigmented metastatic melanoma carcinoma	Dermoscopy: ABCD single observer; n = 2; 1 for dermoscopy diagnosis and 1 for Tzanck smear
Feldmann 1998	Face to Face	Case Series	Secondary	Austria	Not Reported	Not Reported	Melanocytic lesions examined by dermatoscopy before excision	Not Reported	Not Reported	Melanoma (invasive): 25; melanoma (in-situ): 5	Dermoscopy (ABCD)
Guitera 2009 (Modena)	Face to Face	Case Series	Secondary	Italy	Median age: 42 (7-88 years); IQR	Not Reported	Lesions suspicious of melanoma based on dermatoscopic diagnostic criteria or lesion change;	Not Reported	Dermatologist. High experience	Melanoma (invasive): 61; melanoma (in-situ): 18	Dermoscopy: pattern analysis. In-person diagnosis

Study	Dermatology Setting	Study Design	Care Setting	Country	Age or gender	Ethnicity	Inclusion Criteria	Exclusion Criteria	Clinicians experience	Positive Lesion Definition	Diagnostic Method
					32y, 59y; male: 51.3%		included only a random sample of 50% of benign naevi observed during the period				
Kittler 1999	Face to Face	Case Series	Secondary	Austria	Mean age 52 (SD 17 years); male: 49%	Not Reported	Pigmented skin lesions < 1 cm in diameter, consecutively excised	Lesion size ≥ 1 cm	Not Reported likely dermatologists	Melanoma (invasive): 55 (51 superficial spreading, 4 nodular, 15 lentigo maligna, 3 otherwise non-classified melanomas); melanoma (in-situ): 18	Dermoscopy: ABCD; ABCDE (developed in this study)
Morales Callaghan 2008	Face to Face	Case Series	Secondary	Spain	Mean age 33.7 years (SD 14.5), range 8-84 years; male: 64 (38.6%);	Fitzpatrick phototype II (44%); type III (41.5%)	Randomly selected melanocytic lesions; melanocytic on both clinical and dermoscopic criteria	Exclusion criteria: palms, soles, mucous membranes of face, under nails; non-melanocytic appearance	Dermatologist. High experience	Melanoma (in-situ and invasive, or not reported): 6 (3%)	Dermoscopy: pattern analysis
Nachbar 1994	Face to Face	Case Series	Secondary	Not Reported (Germany / USA)	Not Reported	Not Reported	Pigmented melanocytic skin lesions consecutively excised	Unequivocal appearance/diagnosis criteria used to exclude non-melanocytic	Assumed dermatologists. High experience or 'Expert	Melanoma (in-situ and invasive, or not reported): 69	Dermoscopy: ABCD. > 5.45 (determined based on retrospective analysis of the data)
Soyer 1995	Face to Face	Case Series	Specialist clinic	Austria	Not Reported	Not Reported	Pigmented skin lesions are difficult to diagnose on clinical grounds alone	Not Reported	Dermatologist. Not Reported, high	Melanoma (invasive): 50; melanoma (in-situ): 15	Dermoscopy: pattern analysis
Stanganelli 2000	Face to Face	Case Series	Specialist clinic	Italy	Not Reported	Not Reported	Patients with pigmented skin lesions referred by dermatologists and GPs either for pre-surgical assessment or consultation	Non-melanocytic appearance	Not Reported likely dermatologist	Melanoma (in-situ and invasive, or not reported): 55; BCC: 43	Dermoscopy: pattern analysis

BCC, Basal cell carcinoma, sCC: Squamous cell carcinoma, MiS: Melanoma in-situ

7.1.2 Study Raw Data

Study	Setting	TP	FP	FN	TN	Prevalence	Sensitivity	Specificity	NPV	FOR
Binder 1994	Teledermatology	38	6	2	54	40.0%	95.0%	90.0%	96.4%	3.6%
Gilmore 2010	Teledermatology	34	17	2	16	52.2%	94.4%	48.5%	88.9%	11.1%
Seidenari 1998	Teledermatology	25	3	6	56	34.4%	80.6%	94.9%	90.3%	9.7%
Kroemer 2011	Teledermatology	5	3	0	96	4.8%	100.0%	97.0%	100.0%	0.0%
Bowns 2006	Teledermatology	17	11	7	221	9.4%	70.8%	95.3%	96.9%	3.1%
Congalton 2015	Teledermatology	46	23	2	57	37.5%	95.8%	71.3%	96.6%	3.4%
Grimaldi 2009	Teledermatology	5	11	0	219	2.1%	100.0%	95.2%	100.0%	0.0%
Piccolo 2000	Teledermatology	9	0	2	32	25.6%	81.8%	100.0%	94.1%	5.9%
Kroemer 2011	Face to Face	5	1	0	98	4.8%	100.0%	99.0%	100.0%	0.0%
Coras 2003	Face to Face	14	2	2	27	35.6%	87.5%	93.1%	93.1%	6.9%
Warshaw 2010b	Face to Face	30	543	11	930	2.7%	73.2%	63.1%	98.8%	1.2%
Piccolo 2000	Face to Face	8	1	3	31	25.6%	72.7%	96.9%	91.2%	8.8%
Ahnlide 2016	Face to Face	34	23	12	240	14.9%	73.9%	91.3%	95.2%	4.8%
Bauer 2000	Face to Face	33	10	9	263	13.3%	78.6%	96.3%	96.7%	3.3%
Benelli 1999	Face to Face	48	37	12	304	15.0%	80.0%	89.1%	96.2%	3.8%
Carli 1994	Face to Face	5	28	0	35	7.4%	100.0%	55.6%	100.0%	0.0%
Carli 2002a	Face to Face	53	9	1	193	21.1%	98.1%	95.5%	99.5%	0.5%
Cristofolini 1994	Face to Face	29	39	4	148	15.0%	87.9%	79.1%	97.4%	2.6%
Dreiseitl 2009	Face to Face	26	121	1	310	5.9%	96.3%	71.9%	99.7%	0.3%
Durdu 2011	Face to Face	8	5	2	185	5.0%	80.0%	97.4%	98.9%	1.1%
Feldmann 1998	Face to Face	16	14	9	461	5.0%	64.0%	97.1%	98.1%	1.9%
Guitera 2009 (Modena)	Face to Face	68	83	11	33	40.5%	86.1%	28.4%	75.0%	25.0%
Kittler 1999	Face to Face	60	71	13	212	20.5%	82.2%	74.9%	94.2%	5.8%
Morales Callaghan 2008	Face to Face	4	6	2	188	3.0%	66.7%	96.9%	98.9%	1.1%
Nachbar 1994	Face to Face	64	11	5	114	35.6%	92.8%	91.2%	95.8%	4.2%
Soyer 1995	Face to Face	61	17	4	77	40.9%	93.8%	81.9%	95.1%	4.9%
Stanganelli 2000	Face to Face	51	9	4	3308	1.6%	92.7%	99.7%	99.9%	0.1%

TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative, NPV: Negative Predictive Value, FOR: False Omission Rate, Prevalence: Prevalence of Melanoma

7.1.3 Effect of Prevalence on NPV

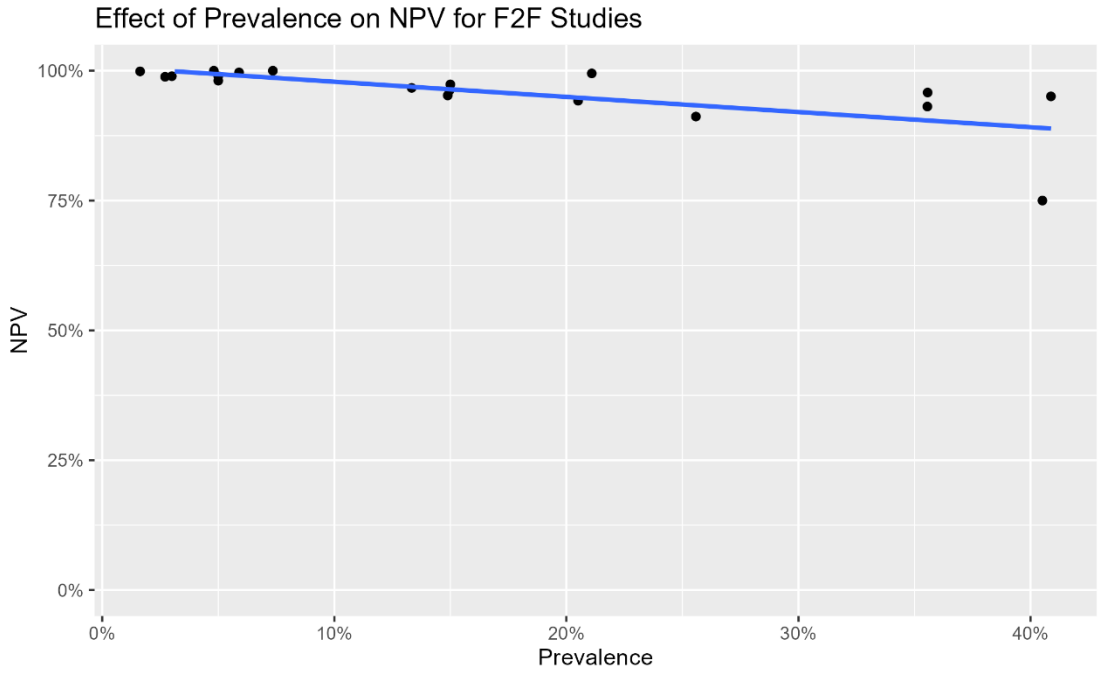


Figure 1. The relationship between NPV and prevalence for 19 studies conducted in a face-to-face setting

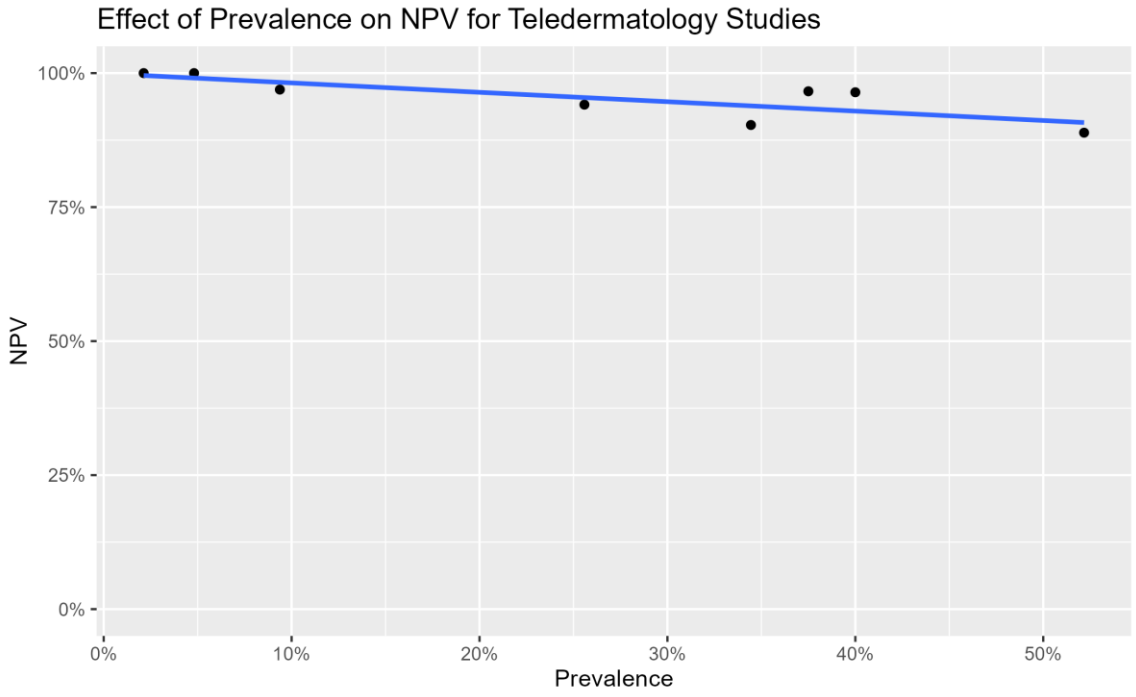


Figure 2. The relationship between NPV and prevalence for 8 studies conducted in a face-to-face setting

7.1.4 Dermatologists' Sensitivity and Specificity for Melanoma (Meta-analysis results)

Setting	Sensitivity [95% CI]	Specificity [95% CI]
Face-to-Face (19 studies)	86.8% [82.4%, 91.1%]	82.4% [76.2%, 92.6%]
Teledermatology (8 studies)	92.2% [87.3%, 97.0%]	87.9% [77.1%, 98.6%]

Note: Values pooled using random-effects model

Note that sensitivity and specificity are not affected by the prevalence of disease.

7.1.5 Dermatologists' FOR for Melanoma (Meta-analysis results)

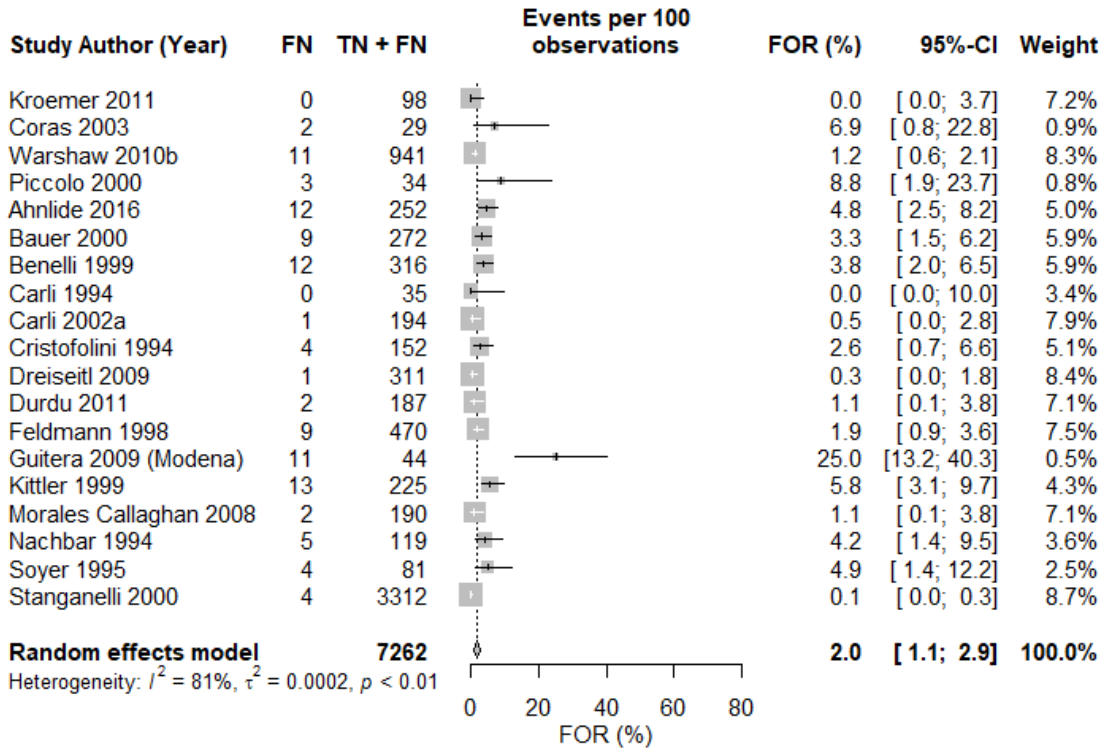


Figure 1. Forest plot for the meta-analysis of dermatologists' FOR performance in a face-to-face setting (19 studies). FN: The number of melanomas marked as non-melanomas; TN + FN: Total number of lesions marked as not melanomas; Weight: Weight assigned to each study determined by within-study variance and between-study variance.

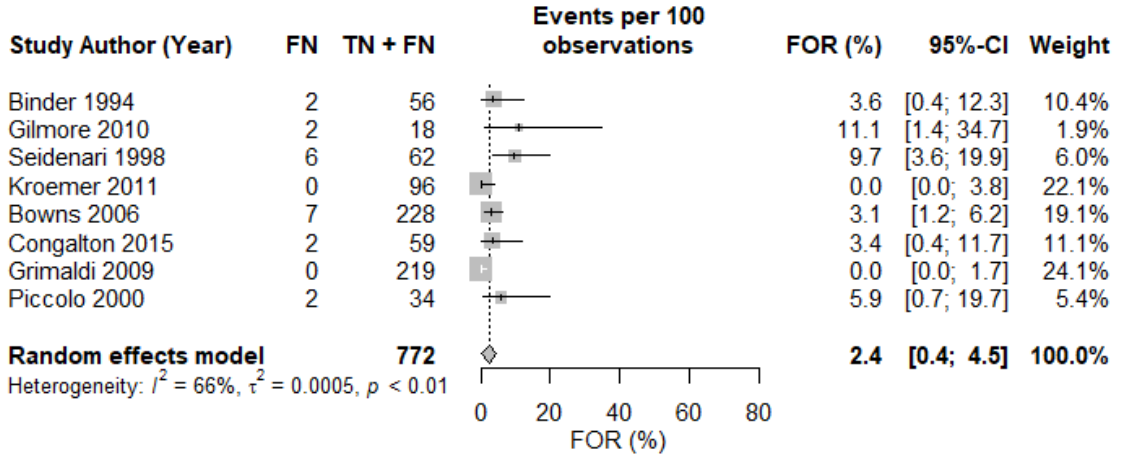


Figure 2. Forest plot for the meta-analysis of dermatologists' FOR performance in teledermatology setting (8 studies). FN: The number of melanomas marked as non-melanomas; TN + FN: Total number of lesions marked as not melanomas; Weight: Weight assigned to each study determined by within-study variance and between-study variance

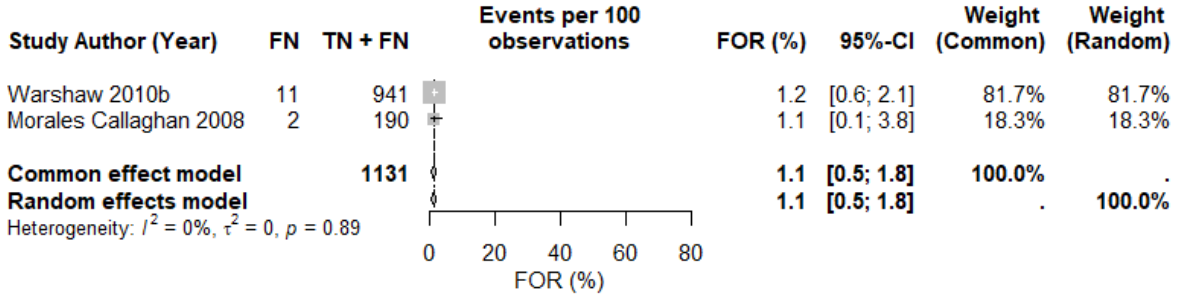


Figure 3. Forest plot for the meta-analysis of dermatologists' FOR performance in face-to-face settings with comparable prevalence (2 studies). FN: The number of melanomas marked as non-melanomas; TN + FN: Total number of lesions marked as not melanomas; Note that both the Common-Effects Model and the Random-Effects Model are included in this meta-analysis, accounting for the small sample size across studies. Both models returned the same pooled FOR.

7.2 DERM Performance – Extended Analysis

7.2.1 DERM NPV Subgroup Analyses

The below tables apply to the positive definition of melanoma (invasive, in-situ and lentigo maligna)

1. Fitzpatrick Skin type

Skin Type	NPV [95% CI]	FOR [95% CI]
Types 1-4, n = 29,292	99.8% [99.7%,99.8%]	0.2% [0.2%,0.3%]
Types 5 & 6, n = 976	100.0% [99.4%,100.0%]	0.0% [0.0%,0.6%]

2. Care Site

Care Site	NPV [95% CI]	FOR [95% CI]
West Suffolk NHS Foundation Trust, n = 2,668	99.9% [99.7%,100.0%]	0.1% [0.0%,0.3%]
University Hospitals Birmingham, n = 10,934	99.7% [99.5%,99.8%]	0.3% [0.2%,0.5%]
University Hospitals of Leicester NHS Trust, n = 8,934	99.8% [99.7%,99.9%]	0.2% [0.1%,0.3%]
Chelsea & Westminster NHS Foundation Trust, n = 4,723	99.9% [99.7%,100.0%]	0.1% [0.0%,0.3%]
University Hospitals Bristol and Weston NHS Foundation Trust, n = 2,258	99.8% [99.4%,99.9%]	0.2% [0.1%,0.6%]
Ashford and St Peter's Hospitals NHS Foundation Trust, n = 2,334	99.7% [99.4%,99.9%]	0.3% [0.1%,0.6%]
Mid-Cheshire Hospitals NHS Foundation Trust, n = 67	100.0% [93.8%,100.0%]	0.0% [0.0%,6.2%]
University Hospitals of Morecambe Bay NHS Foundation Trust, n = 1,507	99.7% [99.3%, 99.9%]	0.3% [0.1%,0.7%]
RDUH Eastern Services, n = 38	100.0% [88.4%,100.0%]	0.0% [0.0%,11.6%]
Ashford and St Peter’s Hospitals (Routine), n = 230	100.0% [98.1%,100.0%]	0.0% [0.0%, 1.9%]

Note that 23 out of all 40 melanomas not sent for review by DERM were identified at University Hospitals Birmingham (UHB). These included 14 in situ melanomas, 6 lentigo maligna and 3 superficial spreading melanomas. The cause for this higher proportion is not identifiable from the available data, and their distribution of Fitzpatrick skin types is comparable with other sites. UHB is a tertiary referral centre, which might be a contributing factor.

3. DERM Version

DERM version	NPV [95% CI]	FOR [95% CI]
3.0.1, n = 1,817	99.9% [99.5%, 100.0%]	0.1% [0.0%, 0.5%]
3.0.2, n = 17,070	99.8% [99.7%, 99.8%]	0.2% [0.2%, 0.3%]
3.0.4, n = 14,204	99.8% [99.7%, 99.9%]	0.2% [0.1%, 0.3%]
4.0.1, n = 586	100.0% [99.2%,100.0%]	0.0% [0.0%, 0.8%]

7.2.2 DERM NPV for Other True Positive Definitions

Performance where positives: only invasive melanomas (not in-situ or lentigo maligna).

Confusion Matrix – Invasive Melanomas (total n = 33,693)

		Predicted	
		Positive	Negative
Actuals	Positive	776	34 (15) *
	Negative	5,973	26,910

TP: Invasive Melanoma; TN: Not invasive melanoma.
* Note that by this definition, false negatives include lesions referred by the AI for review by a dermatologist marked as either SCC, BCC, IEC, AK or Atypical Naevus (n = 19). These would still have been managed appropriately.

Summary Metrics:

Metric	Value [95% CI]
Negative Predictive Value	99.9% [99.8% - 99.9%]
False Omission Rate	0.1% [0.08% - 0.2%]
Sensitivity	95.8% [94.8% - 97.1%]
Specificity	81.8% [81.4% - 82.3%]
Prevalence	2.4%

Performance where positives: Melanomas (all degrees), Squamous Cell Carcinoma (SCC).

Confusion Matrix – Melanomas, SCCs (total n = 33,693)

		Predicted	
		Positive	Negative
Actuals	Positive	1,999	89 (59) *
	Negative	11,843	19,762

TP: Melanoma or SCC skin cancer; TN: Non-Melanoma or Non-SCC skin cancer.
* Note that by this definition, false negatives include lesions referred by the AI for review by a dermatologist marked as either BCC, IEC, AK or Atypical Naevus (n = 30). These would still have been managed appropriately.

Summary Metrics:

Metric	Value [95% CI]
Negative Predictive Value	99.6% [99.5% - 99.7%]
False Omission Rate	0.4% [0.3% - 0.5%]
Sensitivity	95.7% [94.8% - 96.6%]
Specificity	62.5% [62.0% - 63.1%]
Prevalence	6.2%

Note that a higher prevalence of disease will cause a reduction in NPV (Appendix 7.1.3).

7.3 Post-Market Surveillance Literature Search

The search strategy for the literature review for PMS literature includes the following search terms: "post-market" AND "surveillance" OR "evaluation" OR "monitoring" AND "AI" OR "artificial intelligence" OR "AIaMD" OR "AI as a medical device" OR "Software as a Medical Device" OR "SaMD" OR "Medical Device"

7.4 CLEAR Derm Checklist

Domain	Checklist items
Data	Describe imaging modalities, confounding artefacts, and pre/post data processing (Items 1–6)
	Describe the metadata on images used for AI development. Comment on potential biases that may arise as a result (Items 7–9).
	Define image datasets (training, validation, test) used during AI algorithm development (Items 10–12).
	Describe how the test dataset relates to the proposed clinical setting, with special attention to out-of-distribution classes (Items 13–15)
Technique	Develop new algorithms using standard labels of reference (Items 16–18).
	Describe algorithm development (Item 19)
	Provide a method for the AI algorithm or algorithm output to be publicly evaluable (Item 20)
Technical Assessment	Describe how performance measures and benchmarks are consistent with the proposed clinical translation (Items 21–23).
Application	Describe intended use cases and target conditions (inside distribution, Item 24)
	Discuss potential impacts on the healthcare team and patients (Item 25).

Checklist for Evaluation of Image-Based Artificial Intelligence (AI) Algorithm Reports in Dermatology (CLEAR Derm) – directly taken from the consensus guidelines by Daneshjou et al. (2022)

7.5 Analysis of False Positive Rates in Practical PMS

False Positive Rate

To complement our simulation model for NPV auditing, we calculated the expected false positive rates per year based on 100,000 simulation runs. This is to account for statistical errors deriving from repeating the same statistical test over time (Type 1 errors). In this context, type 1 errors are exemplified by changes in the false positive rate.

Here, the false positive rate refers to the likelihood of incorrectly reporting a negative change in the AI's NPV performance, when in fact there is no change. We have calculated this through the same modelling approach taken for the estimation of the timeframe for detecting NPV drops. We divided the number of times that the binomial test returned a false positive result by the total number of days elapsed within the synthetic populations.

The table below shows that sampling at a higher frequency (3 months) has a higher chance of false positives, but this remains low. This means that the chance of incorrectly reporting a drop in performance is low.

False positive rates (per year)

NPV Scenario	Check every 3 months (91 days)	Check every 4 months (121 days)	Check every 6 months (182 days)
No change from 99.8%	0.00015	0.00000	0.00000

Edge Health

If you want to read more about our work, or contact us. Please visit our website:
www.edgehealth.co.uk

or email us at:
info@edgehealth.co.uk