

Do GPs trust Artificial Intelligence insights and what could this mean for patient care? A case study on GPs skin cancer diagnosis in the UK

Authors

Massimo Micocci^{1,2}, Simone Borsci^{1,2}, Viral Thakerar³, Simon Walne^{1,2}, Yasmine Manshadi⁴, Finlay Edridge⁴, Daniel Mullarkey⁴, Peter Buckle^{1,2}, George B. Hanna^{1,2}

¹ NIHR London In-Vitro Diagnostics Cooperative

² Imperial College London, Department of Surgery and Cancer

³ Imperial College London, Faculty of Medicine, School of Public Health

⁴ Skin Analytics Limited, London, United Kingdom

Abstract

Every year, General Practitioners (GPs) see over 13 million patients for dermatological concerns making dermatology the highest referring speciality. Artificial Intelligence (AI) systems could improve system efficiency by supporting clinicians in making appropriate referrals, but they are, like human clinicians, imperfect and there may be a trade-off between sensitivity and specificity that is likely to result in false negatives. In this paper, a study is presented to explore two areas. Firstly, the aptitude of GPs to trust appropriately (or not trust) the outputs of a fictitious AI-based decision support tool when assessing skin lesions. Secondly, to identify which individual characteristics could make GPs less prone to adhere to erroneous diagnostics results and to refrain from passive adherence to AI. Findings suggest that when the AI is correct, there is a positive effect on GPs' performance and confidence suggesting the potential to reduce referrals for benign lesions. However, when an inexperienced GP is presented with a false-negative result, they may passively deviate from their initial clinical judgement to accept the wrong diagnosis provided. AI systems will have a false-negative rate and, when adopting new technologies, this needs to be acknowledged and fed into risk-benefit discussions and considerations around additional safety measures.

1. Introduction

Artificial Intelligence-based (AI) technologies for medical purposes may have the ability to change the healthcare landscape, though to date only a few of these medical devices have made it through to real-world deployment.

The results of a national survey conducted by the AHSN (Academic Health Science Network) [1] in 2018 shows that major applications of AI are conceived to prioritise patients most at risk and to speed up the rate of diagnosis, being primarily used as diagnostics and screening tools. Specifically, solutions are being developed to achieve quicker diagnosis (79%), faster identification of care need (63%), and better experience of health services (63%). Further, 83% of diagnostics solutions are being developed for use by clinicians and 73% for use in secondary care, typically being images produced and evaluated in hospital settings by clinicians [2].

One of the more mature areas of application for machine learning in healthcare is image recognition, where evidence exists demonstrating the ability to identify the presence of malignant tumours [2]. A growing field of application of AI systems is dermatology in which early detection of melanoma may benefit patients [3-5]. Every year, GPs see over 13 million patients for dermatological concerns [6]; melanoma is one of the most dangerous forms of skin cancer, with the potential to metastasise to other parts of the body via the lymphatic system and bloodstream. The current standard of care for skin cancer is set by The National Institute for Health and Care Excellence (NICE) [7] who recommend adopting a ‘risk threshold’ in primary care to underpin recommendations for suspected skin cancer pathway referrals (2WW) and urgent direct access investigations in cancer. The threshold is determined by the positive predictive value – PPV – of a broad range of cancer and related risk factors (e.g., increasing age and a family history of cancer) and agreed on 3% to improve the early detection of cancer. General Practitioners (GPs) are expected to refer under the 2WW if the probability of cancer is 3% or higher. Referral rates are also influenced by factors beyond clinical suspicion of the lesion, such as a clinician’s personal risk tolerance and perceived patient expectations or concerns [8]. In 2019/20 more than 500,000 patients were referred under the 2-week wait (2WW) pathway (intended for the most urgent skin cancer conditions such as melanoma (MM) and squamous cell carcinoma (SCC)), making dermatology the highest referring speciality in the NHS [9]. However, of the half a million cases referred on this pathway, melanoma and squamous cell carcinoma (SCC) only made up 6.5% of referrals in 2019/20 [10]. This reflects accepted behaviour amongst clinicians to refer with a very low threshold to facilitate detection in the early stages of the disease, which are associated with better outcomes. The same data from the National Cancer Registration and Analysis Service (NCRAS) also indicates that only 64% of cancers are detected through 2WW referrals, suggesting several numbers of skin cancer cases are detected through alternative pathways, potentially representing missed diagnosis by a GP and risking a delay in diagnosis. These professionals, representing the first line of defence against skin cancer, may benefit from the support of an accurate AI solution to detect skin cancer early and to identify atypical presentations, with an overall beneficial impact for patients and the NHS, especially given their role as generalists rather than specialist dermatologists [11]. The introduction of AI systems into areas

that were the exclusive domain of human experts (i.e. clinical practice, translational medical research and, biomedical research) [12] has the potential to enhance the healthcare professional's ability to screen and diagnose.

There are many studies assessing the efficacy of intelligent systems [13-16], however, most lack prospective evidence and there are methodological limitations of these studies [17, 18] e.g. only including suspicious skin lesions that had undergone excision or biopsy and excluding lesions with histological and clinical features similar to melanoma [19] and lack of randomized trials in favour of in silico assessments of datasets [18]. Even the most sophisticated AI systems always come with uncertainties. They are dependent on the quantity and quality of training data [12] used to train them; the Academy of Medical Royal Colleges [20] states that insufficient, inaccurate or misrepresentative data could lead to poorly performing AI systems and misdiagnoses. However, performance may not be the limitation, and there may be a more active trade-off between sensitivity and specificity to optimise their performance within a given pathway i.e. the balance between false negatives and false positives.

The introduction of algorithm-based tools into this socio-technical system may create friction and conflict in the decision making; this is due to the intrinsic use of artificial intelligence to reach a certain 'conclusion' that may not be transparent to the human decision-makers, and to the consequent changing practices. Professionals may be used to discussing and interpreting (when needed) with colleagues the results of image and data acquisition to reach a decision; a clinical decision is, oftentimes, a team decision and there is limited evidence as to whether intelligent systems may be able to also fulfil this peer review. How to deal with the "agency" and the "reliability" of the AI systems is an open issue and healthcare applications require behavioural change and practice adjustments to establish a safe routine of AI-enhanced diagnostic decision making.

Ultimately, the key issue with AI is how much decision-makers will trust these medical devices once deployed in the market. It is particularly important for clinicians to critically appraise any clinical systems, including AI. In this sense, the inclusion of AI systems in the healthcare field should be supported by the awareness that these systems, like the existing workforce, are imperfect. For decision support tools, the resilience of the diagnostic process is in the hands of the clinicians even when an AI is involved, as they are the only ones who have a holistic view of each clinical scenario, and can decide to agree or disagree with an AI [21]. Beyond the issue associated with having a 'black box' AI or a fully transparent tool to support decisions [22], the main risk of having a second in-silico agent in a complex process such as the one of healthcare could be also that professionals might be willing to overestimate the insights provided by these tools especially when dealing with uncertainties due to lack of expertise or complexity around the cases [4, 23, 24].

In this paper, we present results from an online survey conducted on a pool of GPs who were presented with a combination of accurate and inaccurate results from a hypothetical AI-enabled diagnostic tool for the early detection of skin cancer.

1.1 Aims of the study

The primary aim of the study was to create a fictitious scenario of skin lesion assessment in which to test GPs decision-making performances (diagnosis and management plan) and their confidence in decisions made both before and after receiving a selection of correct and incorrect information presented as coming from an AI system.

The secondary aim was to understand which individual characteristics could enable GPs to be resilient (less prone to adhere) to erroneous diagnostics results and to refrain from passive adherence to AI. This explorative study hypothesised that GPs, being not specialists in skin cancer, may tend to adhere to the AI decision making (even when wrong) and that this may be modulated by their trust toward technology and level of expertise and confidence in dermatology.

1. Methods

2.1 Participants

A total of 73 GPs participated in the study. A number of 23 were excluded because they were not able to finalise or correctly complete the test. The final sample of 50 GPs (mean age 34.4, min = 26, max = 53; 76% female) completed the test online via Qualtrics^{XM} between April 2020 and the 10th of May 2020. Participants were directly informed of this study and recruited by email through a Clinical Lead in Primary Care Research at the NIHR LIVD; also, the link to the survey was posted on social media (Twitter and LinkedIn) and on a private Whatsapp group used by GPs and GPs with Special Interests working in the Greater London area.

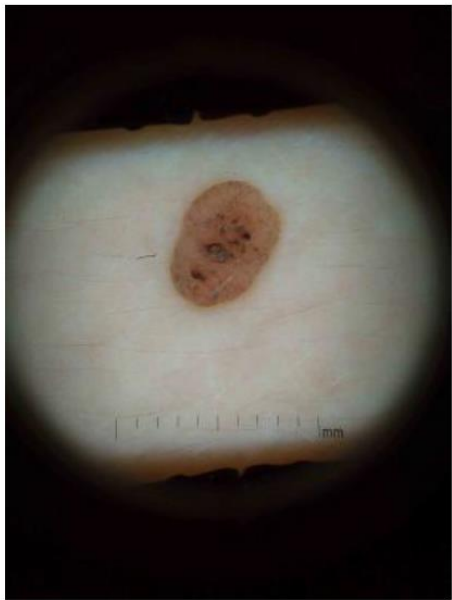
2.2 Ethics

Participants who accepted to complete the survey were asked to read the Participants Information Sheet and to sign the Consent Form with which they agree to take part in the study and to have their personal opinions reflected, anonymously, in reports and academic publications. Local approval for Service Evaluation was sought, and obtained, from Imperial College Healthcare NHS Trust (ICHNT) – registration no. 373.

2.3 Material

The online test was composed of the following sections:

- Demographics. This section was composed of 15 items. It included qualitative questions regarding individual characteristics (age, gender, years of practice etc.), questions regarding interest in dermatology and attendance to dermatology courses in the past three years, as well as their perceived confidence in dermatology, and familiarity with tools for early skin cancer diagnosis. Three questions considered the GPs overall trust attitude toward innovations in medical devices (McKnight et al 2002).
- Main test. This was composed of questions regarding 10 lesions (See **Appendix A**) purposively selected to be representative of commonly encountered lesions.



Details:
Age of the patient: 78
Gender: female
Duration of lesion: years
Evolution/changes: none
Sensory changes: none
Bleeds: none
Risk factors: none
Body location: back

Fig 1. Example of one lesion with only patient information (fictitious)

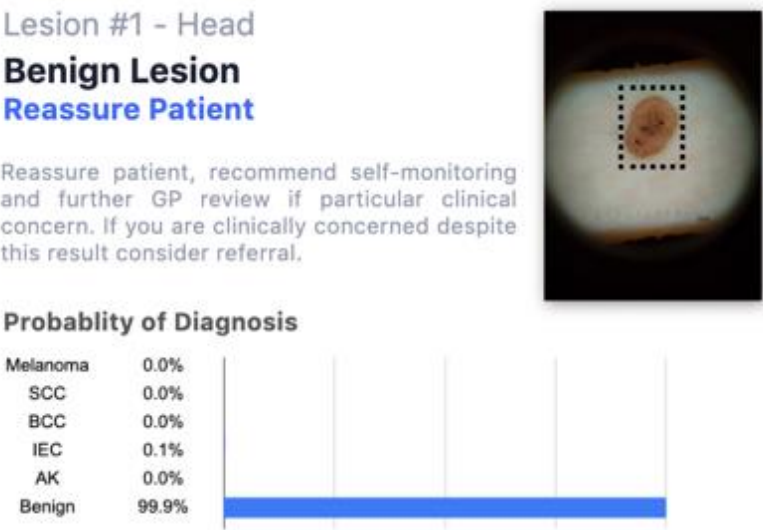


Fig 2. Example of one lesion with a fictitious AI assessment

Each lesion was accompanied by vignettes of hypothetical patients details likely to be asked in a routine GP consultation (*age, gender, duration of the skin lesion, evolution/changes of the lesion, sensory changes, bleeding, risk factors, body location*). Each lesion was associated with three questions regarding:

i) the diagnosis, with a range of seven options (Melanoma; Squamous Cell Carcinoma; Basal Cell Carcinoma; Intraepidermal Carcinoma; Actinic Keratosis; Bening, Other);

- ii) the management plan, with a range of four options (two-week referral 2WW; routine, but not 2WW; discharge with safety net advice; other);
- iii) the confidence in their decision making on a 5-points Likert scale.

Participants completed these questions regarding the diagnosis and the management plan, as well as their confidence, twice:

- 1) when they had access only to patient information and images of the skin lesion in Fig 1; and
- 2) when they had access to the AI insights regarding the images in Fig 2 in addition to this information

The skin lesions were divided in terms of the type of decision making and type of case (benign and malignant) as follows:

- i) Five of the ten skin lesions were considered everyday cases (EC), including lesions whose features are commonly observed in routine consultations and considered easy to interpret [25], two of these were benign and three were malignant skin lesions (cases 2 - 6).
- ii) Three of the ten skin lesions were considered cases with uncertainties (CU) i.e. cases in which the picture of the skin lesion is hard to interpret or it contains a bias (marked for biopsy) and for which GPs might be expected to ask for a second opinion. One of these CU cases was malignant and two were benign (cases 1, 7 and 8). For all the cases from 1 to 8 (EC and CU), the scenario was set up with the AI system presenting the correct diagnosis to the GPs.
- iii) Finally, two cases were misclassified, so the AI presented one benign case as being malignant, and one malignant case as being benign, thus presenting a dangerous scenario (DS, cases 9, 10).

2.4 Procedure

The study was presented to participants as a simulation -with fictitious patients' details- to assess their agreement with an AI system to better report diagnostic test results. Once the study was completed, a disclaimer email was sent to each participant clarifying that the provided combinations of lesions/diagnosis in the study were not always accurate; the study aim of assessing GPs performance and attitude with both accurate and inaccurate AI diagnosis was fully explained.

After the demographic survey, each participant received ten blocks of questions (each related to one lesion) in a fully randomised order. Each lesion was shown first without the fictitious AI decision, asking them to perform their decision making based on the image of the lesion and the patients' information (Fig 1). After GPs provided their opinion regarding diagnosis, management and confidence in the decision making of a lesion, the same lesion was

presented a second time with the addition of the AI diagnosis (Fig 2). GPs were then asked to decide whether to change or to maintain their answers regarding the diagnosis, management plan and confidence in their decision.

2.5 Data Analysis

Descriptive statistics were used to observe participants' characteristics, the frequency of correct diagnosis and management plan, and the GPs confidence in their decision making before and after receiving the AI-enabled information. Pre and post AI performance of GPs in terms of diagnosing and management plan were dichotomised (correct/incorrect) and a McNemar's Chi-square test was used to analyse the effect of AI information in each decision-making group (EC, UC, DS) by also accounting for the type of case (benign and malignant). The percentage of confidence was tested by a mixed general linear model.

Hit and false rates of GPs for the diagnostic and management decision making before and after the wrong AI insights were used to model GPs resilience when dealing with erroneous AI information (i.e., DS cases). In line with signal detection theory [26], the computation was used to compose a sensitivity index for when AI was wrong (d-AIW, see **Appendix B**). The more the index was higher than zero, the better the GPs' ability to ignore the wrong indication of the AI. The index was used to distinguish two groups: one included GPs who had a d-AIW over zero (hereafter called '*resilient group*') and the other included GPs with an index below or equal to zero (hereafter called '*non-resilient group*') for the management and diagnostics of patients with skin lesions. A Kruskal-Wallis test was performed to check if resilient and non-resilient GPs performed significantly differently when AI provided them with correct and incorrect answers, and to observe the differences between the two groups in terms of individual characteristics.

3. Results

3.1 Individual characteristics

The 76% of the participants had less than 5 years of experience, 16% from 5 to 10 years and 8% more than 10 years of experience. Overall, the GPs in our cohort declared an average level of confidence in dermatology of 51.5 out of 100 (SD: 16.2), despite 34% of them attended specialisation courses on the matter in the past three years. The 70% of the participants stated that they had not used a dermatoscope in the previous 12 months, with only 4% of the GPs declaring a weekly use of such an instrument. The 38% never used digital systems for skin lesions (e.g., taking pictures of patients' skin lesion to be uploaded into the system), while among the users of such digital systems for diagnostic purposes: 2% declared a daily usage, 10% a weekly and 50% at least once per month. The level of trust toward AI support systems for this application domain declared by GPs was sufficient (M:61.2%; SD:14.5%).

3.2 General practitioners’ correct decision making before and after AI insights

Table 1 shows the frequency of GPs performances before and after receiving the fictitious AI-enabled information, by suggesting that GPs tend to adhere to the indications of the AI. Specifically, when the AI is correct (EC and CU cases) there is a positive effect on GPs performance and confidence. Correct diagnosis, supported by a trustworthy AI, went up of 13.2 points for EC cases and 16.5 points for CU cases. Similarly, the selection of the correct management plan went up by 7.6 points (EC) and 18.5 points (CU). GPs confidence in their decision making went up of 12.7 for EC cases after the insights of the AI, while this aspect only increases by 1.5 points when dealing with CU. Conversely, when AI provided incorrect insights (DS cases), the correctness of diagnosis and management went down by 24 and 29 points respectively, with a positive boost of 5.7 points in the GPs confidence in their decision making after receiving AI insights.

Table 1 Frequency of GPs performances before and after receiving the information from AI

Decision making groups	Before AI			After AI		
	Correct Diagnosis (%)	Correct Management (%)	GPs Confidence (%)	Correct Diagnosis (%)	Correct Management (%)	GPs Confidence (%)
EC	73.6	82.4	66.8	86.8	90	79.5
Only Benign	68	62	63.5	89	84	82.7
Only Malignant	77.4	96	69.1	85.4	96	76.5
CU	37.5	44	61.8	54	62.5	63.3
Only Benign	9	8	61.7	42	41	62.5
Only Malignant	66	80	62.5	66	84	65
DS	46	54	60	22	25	65.7
Only Benign	32	32	58.5	10	4	67
Only Malignant	60	76	62.5	34	46	64

The McNemar's Chi-square test clarified how the AI insights affected the GPs decision making per each group:

- **Everyday cases:** GPs ability to correctly diagnose a skin lesion significantly improve after receiving the AI information from 73.6% to 86.8% ($X^2(1, N=50)= 21.787, p<001$) with significant effects for both the benign ($X^2(1, N=50)=21, p<.001$) and malignant cases ($X^2(1, N=50)= 4.654, p=.031$). The selection of the correct management plan is also positively affected by the AI information going from 82.4% to 90% ($X^2(1, N=50)= 3.78 p<.001$) which is particularly relevant for the plan regarding benign cases ($X^2(1, N=50)=22, p<001$), while no major improvement was observed for malignant cases. Confidence about decision making, independently from the type of skin lesion, significantly improved from 66.8% to 79.5% after receiving the AI information ($X^2 (1, N=48)= 107.2, p<.001$).

- **Cases with Uncertainties (CU):** GPs correct diagnosis improved significantly from 37.5% to 54% correct decision making when supported by AI ($X^2(1, N=50)=24.9, p<.001$). This difference was significant for benign cases ($X^2(1, N=50)=31.03, p<.001$) while no significant differences emerged in malignant cases before and after receiving AI information. Concurrently, the ability to correctly define a management plan significantly increased from 44% to 62.5% thanks to the AI ($X^2(1, N=50)=28.195, p<.001$) this effect was significant for benign cases ($X^2(1, N=50)=31, p<.001$). GPs confidence was not significantly affected by the information of the AI.
- **Dangerous Situations (DS):** When erroneous information was provided by the AI, it seems that GPs were significantly pushed to adhere to the erroneous suggestions of the AI. Correct diagnosis of the skin lesions significantly decreased from 46% to 22% ($X^2(1, N=50)=22.04, p<.001$). The adherence to the wrong AI insights was significant for both benign ($X^2(1, N=50)=9.08, p=.026$) and malignant cases ($X^2(1, N=50)=11.7, p=.009$). Similarly, the decision making about management is significantly affected by wrong AI insights decreasing from 54% to 25% the ability of GPs to correctly decide the plan for the patient ($X^2(1, N=50)=25.290, p<.001$). This significantly affected GPs decision making regarding both benign ($X^2(1, N=50)=12.07, p=.005$) and malignant cases ($X^2(1, N=50)=11.52, p=.007$). Confidence was not affected by the information provided by the AI.

3.3 Resilience to erroneous insights of the artificial agent

When the AI provided erroneous information (DS cases) only 10% of the GPs were able to correctly disagree with the indication of the AI in terms of diagnosis (d-AIW M: 0.12, SD: .37), and only 14% of participants were able to correctly decide the management plan despite the AI insights (d-AIW M:0.12, SD:.32). These GPs were categorized as the resilient ones (i.e., the ones able to correctly reject to adhere to the AI insights), as opposed to all the others which were categorized as less resilient to the wrong indications of the AI.

The Kruskal-Wallis test carried out on EC and CU cases (when the AI provided correct results) suggested that the performance of the GPs in the resilient group was not significantly different to the performance of the less resilient group. Conversely, when the AI provided erroneous diagnosis (DS cases) a significant difference was found between the two groups in terms of diagnostic decision making ($X^2= 12.4, p<.001$) and correct management plan ($X^2= 6.8, p=.009$).

The analysis of the difference between groups in terms of individual characteristics suggested that GPs who declared a regular usage of the dermatoscope are better at rejecting the wrong insights from AI and correctly diagnosis ($X^2= 7.8, p=.005$) and manage patients ($X^2= 5.1, p=.023$) compared to less resilient GPs. Some moderate, but still

significant effect also emerged regarding the overall confidence in dermatology, showing that resilient GPs were more confident than non-resilient doctors, and this may play a role in their ability to correctly diagnose ($X^2= 3.8$, $p=.049$) and define a management plan ($X^2= 5$, $p=.024$) even when AI is providing erroneous insights. The other individual factors (e.g., age, sex, training, predisposition to trust etc.) only showed some moderate tendencies.

4. Discussion

Results demonstrate high levels of trust by GPs towards results attributed to a fictitious AI system, a finding which has both positive and negative consequences on the healthcare system. Whilst an accurate clinical decision support tool may support GPs to correctly identify benign lesions, reducing the number of false positives referred to 2WW clinics, there is also a possibility that an erroneous result from the AI system could lead to a patient's case being under-triaged.

The adherence to an AI that is able to provide correct insights about cases, even when there are uncertainties, can significantly improve the correct decision making (diagnosis and plan) of GPs. The correctness and confidence of GPs in their decision making are significantly improved by using the AI when a case presents no uncertainties. Given the pressure on the 2WW pathway, this result may be convenient at ruling out negative cases at the triage stage with benefits on patient flow and for the individual patients who will avoid unnecessary anxiety associated with a suspected cancer referral. However, when dealing with some uncertainties (CU) or when AI is wrong (DS) the confidence of the GPs in the final decision is not affected by the AI insights. This might suggest that when GPs have doubts on how to treat a case (CU) or when they are not convinced by the insights of the AI (DS cases) they are not completely reassured by the use of the AI and, yet, a large majority of the GPs continue to adhere to the indication of the AI. These findings are aligned with previous studies [27] suggesting that over-reliance in automated systems may be triggered by confirmatory bias when participants direct their attention towards features consistent with the (inaccurate) advice. We also considered the variability of the personal expertise and attitude towards automate systems as having an influence to avoid passive adherence. The results suggest that the tendency to adhere, even when the AI is inaccurate, may be due to a lack of experience with the specific tasks or domain knowledge which may bring GPs to overestimate the insights of the intelligent systems. The small number of resilient GPs that were able to critically interpret the results of the AI declared, in fact, significantly higher usage of essential dermatological tools (i.e., dermatoscope) and confidence in the specific domain of dermatology compared to the ones who adhered to the suggestions of the mistaken AI.

The present study should be intended as an initial step in the understanding of the future relationship between AI and clinicians in the domain of dermatology. Three main limitations should be considered for future studies. First, the

small sample surveyed may not be representative of the variety of expertise, exposure to dermatology cases and experience with similar technologies that GPs may have. Secondly, the participants of the present study were aware that the test was a simulation, and they may have changed their behaviour because of the attention they received [28] and because of the absence of implications for patients. This effect may have implications on the generalisability of findings. Finally, how information from an AI system is presented may impact the end-user. In future studies, we advocate a larger group of GPs, with different expertise and familiarity with AI systems, should be involved to expand the current results. Concurrently, a larger number of cases should be tested with equal types of lesions in each group. This may bring further insights into the mechanism of adherence toward AI. Mixed-methods studies [29] could help mitigating bias and changes in behaviour of research participants as influenced by observation and measurement. The risk of a passive adherence to AI in the real world could emerge also due to the complexity of the healthcare system [21] and future longitudinal studies on real cases should be implemented to monitor such possibility. As well as the user interface, the role of training and documentation such as the ‘Instructions for Use’ (IFU) should be considered in future research both academically and from the perspective of regulatory applications.

5. Conclusions

Well designed and accurate, intelligent systems may be able to support GPs to manage patients confidently and appropriately in primary care with suspicious skin lesions, supporting them to not only refer suspicious lesions but also to manage other lesions in primary care relieving pressure on busy dermatology departments and saving patients from the anxiety of an unnecessary 2WW referral.

Whilst standards of clinical evidence for AI systems should continue to improve, with more emphasis on prospective clinical trials, it is fair to assume that much like the existing clinical workforce, no AI system will be 100% sensitive in a real-world deployment. Human expertise can be amplified by AI systems, but human decision-makers need to have the domain knowledge and confidence to disagree with such systems when it is necessary.

This, counterintuitively, suggests that AI tools are better suited in the hands of clinicians with certain domain knowledge (senior or specialist clinicians) rather than with less expert professionals and should perhaps be reflected in early deployments. For the specific case of skin cancer, results suggested that the more clinicians are practising the dermatological skills the more they were able to maximize the benefit of the AI systems.





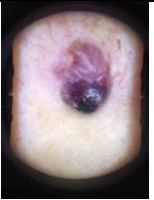
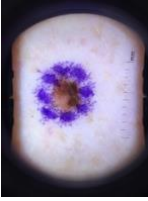

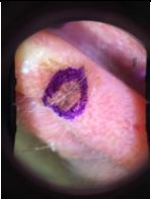
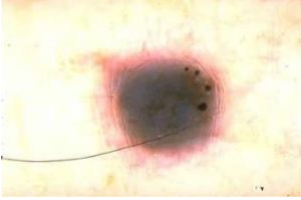

How the new relationship between healthcare professionals and AI systems will be regulated in future requires further exploration [30]. The risk of under-or over-estimating the usefulness of AI tools during clinical decision making might lead to severe consequences for patients.

To design safe, explainable, reliable, and trustworthy AI systems based on fair, inclusive and unbiased data is a key element to support the diffusion of such tools in the medical field. However, medical professionals will need to adapt, learn, and put in place behaviour and strategies to accommodate the unavoidable uncertainties around the interaction with intelligent systems. In this sense, the diffusion and adoption of AI in clinical practice will inevitably impact the training and education of clinicians who should learn how to interact with these systems, establish a practice to minimise and prevent system failure, and learn how to operate when the system fails, misbehaves or malfunctions.

Appendix A

Table A1 shows the ten lesions used in the simulation study and their classification.

Table A1 The ten lesions used in the simulation study and their classification.

Lesions			Classification
	 Case 2	 Case 3	Correct benign
 Case 4	 Case 5	 Case 6	Correct malignant
	 Case 1		Borderline – correct benign
	 Case 7		Borderline – correct malignant
	 Case 8		Borderline – correct benign
	 Case 9		Melanoma misclassified as benign
	 Case 10		Benign misclassified as Melanoma

Appendix B

Computation used to compose the sensitivity indexes

$d' = z^8 - z(\text{False})$

- decision wrong before and correct after AI insights= Hit rate
- decision correct before and Wrong after AI insights= False rate
- decision correct before and correct after AI insights= Correct rejection
- decision wrong before and wrong after AI insights= Miss

References

1. The AHSN Network. Accelerating Artificial Intelligence in health and care: results from a state of the nation survey. <https://wessexahsn.org.uk/img/news/AHSN%20Network%20AI%20Report-1536078823.pdf> (September 2020, date last accessed)
2. NHSx. Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care. https://www.nhs.uk/media/documents/NHSX_AI_report.pdf (September 2020, date last accessed)
3. Petrie T, Samatham R, Witkowski AM, et al.; Melanoma early detection: big data, bigger picture. *Journal of Investigative Dermatology* 2019;**139**(1):25-30.
4. Mar V, Soyer H; Artificial intelligence for melanoma diagnosis: how can we deliver on the promise? : Oxford University Press, 2018.
5. Kromenacker B, Maarouf M, Shi VY; Augmented Intelligence in Dermatology: Fantasy or Future? *Dermatology* 2019;**235**(3):250-252.
6. British Associations of Dermatologists. How can dermatology services meet current and future patient needs, while ensuring quality of care is not compromised and access is equitable across the UK? <https://www.bad.org.uk/shared/get-file.ashx?id=2348&itemtype=document>. (April 2021, date last accessed)
7. NICE. Suspected cancer: recognition and referral. <https://www.nice.org.uk/guidance/ng12/chapter/Introduction> (December 2020, date last accessed)
8. Foot C, Naylor C, Imison C; The quality of GP diagnosis and referral. London: The King’s Fund, 2010.
9. NHS England. Waiting Times for Suspected and Diagnosed Cancer Patients. <https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/07/Cancer-Waiting-Times-Annual-Report-201920-Final.pdf> (April 2021, date last accessed)
10. National Cancer registration and Analysis Service - NCRAS. Urgent Suspected Cancer Referrals: Conversion and Detection Rates. http://www.ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/tww_conversion_and_detection (April 2021, date last accessed)

11. British Association of Dermatologists. GP Trainees. <https://www.bad.org.uk/healthcare-professionals/education/gps/gp-trainees> (December 2020, date last accessed)
12. Yu K-H, Beam AL, Kohane IS; Artificial intelligence in healthcare. *Nature biomedical engineering* 2018;**2**(10):719-731.
13. Phillips M, Marsden H, Jaffe W, et al.; Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA network open* 2019;**2**(10):e1913436-e1913436.
14. Esteva A, Kuprel B, Novoa RA, et al.; Dermatologist-level classification of skin cancer with deep neural networks. *nature* 2017;**542**(7639):115-118.
15. Haenssle HA, Fink C, Schneiderbauer R, et al.; Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology* 2018;**29**(8):1836-1842.
16. Phillips M, Greenhalgh J, Marsden H, et al.; Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatology practical & conceptual* 2020;**10**(1).
17. Chuchu N, Takwoingi Y, Dinnes J, et al.; Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma. *Cochrane Database of Systematic Reviews* 2018(12).
18. Topol EJ; Welcoming new guidelines for AI clinical research. *Nature medicine* 2020;**26**(9):1318-1320.
19. Freeman K, Dinnes J, Chuchu N, et al.; Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *bmj* 2020;**368**.
20. Academy of Medical Royal Colleges. Artificial Intelligence in healthcare <https://www.aomrc.org.uk/reports-guidance/artificial-intelligence-in-healthcare/> (December 2020, date last accessed)
21. Lynn LA; Artificial intelligence systems for complex decision-making in acute care medicine: a review. *Patient Safety in Surgery* 2019(13).
22. Alufaisan Y, Marusich LR, Bakdash JZ, et al.; Does Explainable Artificial Intelligence Improve Human Decision-Making? *arXiv preprint arXiv:2006.11194* 2020.
23. Gilmore SJ; Automated decision support in melanocytic lesion management. *PloS one* 2018;**13**(9):e0203459.
24. Farmer ER, Gonin R, Hanna MP; Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. *Human pathology* 1996;**27**(6):528-531.
25. Erdmann F, Lortet-Tieulent J, Schüz J, et al.; International trends in the incidence of malignant melanoma 1953–2008—are recent generations at higher or lower risk? *International journal of cancer* 2013;**132**(2):385-400.
26. Macmillan NA, Creelman CD. *Detection theory: A user's guide*: Psychology press, 2004.
27. Gaube S, Suresh H, Raue M, et al.; Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 2021;**4**(1):1-8.
28. Sedgwick P, Greenwood N; Understanding the Hawthorne effect. *Bmj* 2015;**351**.
29. O’cathain A, Murphy E, Nicholl J; Three techniques for integrating data in mixed methods studies. *Bmj* 2010;**341**.
30. European Commission. White Paper: On Artificial Intelligence - A European approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (December 2020, date last accessed)