JAMA Dermatology | Consensus Statement

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

Roxana Daneshjou, MD, PhD; Catarina Barata, PhD; Brigid Betz-Stablein, PhD; M. Emre Celebi, PhD; Noel Codella, PhD; Marc Combalia, MSc; Pascale Guitera, MD, PhD; David Gutman, MD, PhD; Allan Halpern, MD; Brian Helba, BS; Harald Kittler, MD; Kivanc Kose, PhD; Konstantinos Liopyris, MD, PhD; Josep Malvehy, MD; Han Seung Seog, MD, PhD; H. Peter Soyer, MD; Eric R. Tkaczyk, MD, PhD; Philipp Tschandl, MD, PhD; Veronica Rotemberg, MD, PhD

IMPORTANCE The use of artificial intelligence (AI) is accelerating in all aspects of medicine and has the potential to transform clinical care and dermatology workflows. However, to develop image-based algorithms for dermatology applications, comprehensive criteria establishing development and performance evaluation standards are required to ensure product fairness, reliability, and safety.

OBJECTIVE To consolidate limited existing literature with expert opinion to guide developers and reviewers of dermatology AI.

EVIDENCE REVIEW In this consensus statement, the 19 members of the International Skin Imaging Collaboration AI working group volunteered to provide a consensus statement. A systematic PubMed search was performed of English-language articles published between December 1, 2008, and August 24, 2021, for "artificial intelligence" and "reporting guidelines," as well as other pertinent studies identified by the expert panel. Factors that were viewed as critical to AI development and performance evaluation were included and underwent 2 rounds of electronic discussion to achieve consensus.

FINDINGS A checklist of items was developed that outlines best practices of image-based AI development and assessment in dermatology.

CONCLUSIONS AND RELEVANCE Clinically effective AI needs to be fair, reliable, and safe; this checklist of best practices will help both developers and reviewers achieve this goal.

JAMA Dermatol. 2022;158(1):90-96. doi:10.1001/jamadermatol.2021.4915
Published online December 1, 2021.

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Veronica Rotemberg, MD, PhD, Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, 530 E 74th St, 9th Floor, Dermatology, New York, NY 10021 (rotembev@mskcc.org).

Artificial intelligence (AI) has the potential to transform clinical care and workflows in dermatology; however, achieving fair, reliable, and safe algorithms is necessary for clinical implementation.[1,2] While the pace of AI development is accelerating in all areas of medicine, dermatology is particularly accessible for image-based AI owing to the widespread use of photography as an assessment tool, including on consumer devices such as smartphones and tablets. Guidelines have been proposed for prospective clinical trials of AI in medicine and dermatology through Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT)-AI and Consolidated Standards of Reporting Trials (CONSORT)-AI.[3] However, many key decisions are made during algorithmic development and initial evaluation. There is a clear need for comprehensive assessment guidelines of AI algorithms as they are being developed and reviewed prior to clinical trials.[3-6]

Most AI publications in dermatology describe the development and initial testing of new AI algorithms. While other specialties such as radiology and cardiology have proposed guidelines for reviewing articles that use AI, dermatologists and researchers have thus far not proposed an evaluative framework.[7-9] We propose a framework that builds on the Standards for Reporting of Diagnostic Accuracy (STARD-15) guidelines for diagnostic accuracy studies. The STARD-AI, Developmental and Exploratory Clinical Investigation of Decision-Support Systems Driven by Artificial Intelligence (DECIDE-AI), Prediction Model Risk of Bias Assessment Tool (PROBAST)-AI, and Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis or Diagnosis (TRIPOD)-AI guidelines are still pending and are unlikely to address dermatology-specific aspects, such as image source, lack of standardization, skin tone, and considerations of bias.[10] We propose dermatology-specific considerations for AI algorithms in dermatologic practice, clinical trials, or reviewing dermatology AI development literature.[5,9,11-13]

Dermatology image-based AI algorithms must consider the unique features of dermatology data, which currently include a lack of standardization among imaging modalities and the risk of bias from noisy labels or demographically unrepresentative data.[2,14-16] These guidelines are intended as requirements for the consideration of study design and the publication of articles and products that de-

scribe AI-based computer vision tasks for dermatology applications, including diagnosis, triage, monitoring, segmentation, and decision support to provide needed context for more general guidelines around AI studies.

Methods

All 19 members of the International Skin Imaging Collaboration (ISIC) AI working group volunteered to be part of a 2-round virtual consensus process. A PubMed search was performed of English-language articles published between December 1, 2008, and August 24, 2021, for "artificial intelligence" and "reporting guidelines," as well as other pertinent studies identified by the expert panel. In total, 650 articles met the search criteria, of which 17 were reported specific guideline recommendations.[6,8,12,17-30] An additional 34 articles were suggested by the expert panel as specific to factors that influence AI diagnosis, which informed development of the criteria.

Prior to initiation, all 19 experts independently proposed considerations for the guidelines, and these were compared with relevant factors noted in the literature. Factors that were viewed as critical to AI development and performance evaluation were included. All suggestions that pertained to AI reports outside of clinical trials were included and summarized into draft guidelines by R.D. and V.R. for round 1. In round 1, all 19 members of the ISIC AI working group reviewed the draft guidelines and provided written feedback and suggestions, including for the checklist items. In round 2, 14 members of the group (73.6%) provided written feedback on the guideline document; the checklist headings remained unchanged, and only 1 clarifying item in the checklist was added. The other 5 members provided assent via email, achieving unanimous agreement. The final document was presented and approved at the ISIC Annual Meeting (June 7, 2021).

Key Points

Question How should artificial intelligence (AI) algorithm reporting in dermatology be assessed?

Findings In this consensus statement, key recommendations for developers and reviewers of imaging-based AI reports in dermatology were formulated and grouped into the topics of (1) data, (2) technique, (3) technical assessment, and (4) application. Guidelines are proposed to address current challenges in dermatology image-based AI that hinder clinical translation, including lack of image standardization, concerns about potential sources of bias, and factors that cause performance degradation.

Meaning The recommendations provided will support algorithm development and assessment, with specific emphasis on dermatologic considerations and intended use scenarios.

Recommendations

These recommendations are intended to support existing mechanisms of review that include analyzing the strengths and limitations of any AI algorithm. The recommendations are summarized in checklist form in the **Table**.

Data

Describe Imaging Modalities, Confounding Artifacts, and Data Processing (Items 1-6)

Given the wide variety of acquisition devices and techniques in dermatology, the descriptions of images used for AI reports require significantly more detail than other medical imaging applications.[15] Image artifacts and their distributions in the data used should be described, particularly for artifacts that have been previously shown to affect performance. For photography, these include the type of camera used; whether images were taken under standardized or varying conditions; whether they were taken by professional photographers, laymen, or health care professionals; and image quality.[31] Other artifacts to consider if relevant to the particular application include pen markings, rulers, hair, other physical perturbations (eg, injury, surgical effects, tattoos), illumination source and lighting conditions (eg, natural light, clinic light for clinical photos), distance from the patient (overviews or close-ups), type of clinical site (eg, academic practice, community private practice), and color calibration performed, as those may influence model performance.[14-16] If using dermoscopic images, the mode of acquisition (polarized vs nonpolarized) should be reported. If there is doubt on whether an artifact should be regarded as potentially confounding, it should be reported if possible. For specialized imaging modalities (eg, confocal microscopy, low-coherence imaging, elastography), any relevant technical details must be reported (eg, frequency/wavelength spectrum of energy source). Acquisition metadata, such as that available in EXIF (exchangeable image file format) headers, should be retained in provided data. All information should be in alignment with legal/privacy data protection and be addressed with appropriate consents to permit openness and scientific rigor.

Any other aspects of the images, such as preprocessing (eg, color normalization) and postprocessing (crop, manual selection, filtering), should also be detailed.[7,32] If images are synthetic (algorithm generated), the authors should state the motivation for their use, how the images were generated, and how they were used in model development.[33] Synthetic images should be made public if they are not subject to patient privacy concerns.[34] If images from publicly available data sources are used (eg, the ISIC archive or public websites), the images used should be specified.[35,36] Privately sourced images, where possible, should be shared through a public repository, such as the ISIC archive, and ethical considerations of data capture and use should be clearly described.[37,38]

Describe the Metadata on Images Used for AI Development and Comment on Potential Biases That May Arise as a Result (Items 7-9)

Patient-level image metadata should be described. Such metadata may include the clinic, hospital, or geographic location of patients from which the data were generated; anatomic sites (of solitary lesions); sex and gender; age; ethnicity and/or race; and skin tone.[14,39-41] The procedure for assessing skin tone should be described, such as the scale used for labeling (eg, Fitzpatrick, individual topology angle), and whether labeling was done in person or through a photograph. Any limitations related to the procedure used for skin tone assessment should also be conveyed. This includes discussing limitations of the skin tone scale used; for example, the

Table. Checklist for Evaluation of Image-Based Artificial Intelligence (AI) Algorithm Reports in Dermatology (CLEAR Derm)

	Checklist for image-based AI algorithm development in dermatology	Description is present/absent
Data		
1	Image types	
2	Image artifacts (eg, image quality, pen markings, anatomic site for photography)	
3	Technical acquisition details	
4	Preprocessing procedures	
5	Synthetic images made public if used	
6	Public images adequately referenced	
7	Patient-level metadata: geographic location of patients, sex and gender distribution, ethnicity and/or race, and how it was extracted	
8	Skin tone information and procedure by which skin tone was assessed	
9	Potential biases that may arise from use of patient information and metadata	
10	Data set partitions	
11	Sample sizes of training, validation, and test sets	
12	External test set	
13	Multivendor images	
14	Class distribution and balance	
15	Out-of-distribution images	
Technique		
16	Labeling method	
17	References to common/accepted diagnostic labels	
18	Histopathologic review for malignant neoplasms	
19	Detailed description of algorithm development	
Technical assessment		
20	How to publicly evaluate algorithm	
21	Performance measures	
22	Benchmarking, technical comparison, and novelty	
23	Bias assessment	
Application		
24	Use cases and target conditions (inside distribution)	
25	Potential impacts on the health care team and patients	

commonly used Fitzpatrick scale does not adequately capture human skin diversity.[42] If metadata are unavailable, describe the potential drawbacks of not having this information and the potential for bias in the data set.[2] If reported metadata are weighted toward a certain population, discuss how this may affect generalizability of the algorithm and the potential for bias.

Additionally, some studies may include clinical metadata, such as medical history or history of present illness, in algorithm development.[43] If such clinical metadata are incorporated into the algorithm, the source of this information and how it was used in algorithm development should be described.

Define Image Data Sets (Training, Validation, Test) Used During AI Algorithm Development (Items 10-12)

Clearly indicate any inclusion or exclusion criteria for images.[7] Discuss any reasoning behind the size of the training, validation, and test sets and how they were partitioned.[7] Indicate information regarding statistical distributions of metadata or imaging artifacts described earlier (eg, same clinical site, image capture device, patient population, presence of artifacts) and whether the independent test set comes from similar distributions as the training and validation data or whether it includes samples drawn from different distributions. As AI algorithms are prone to overfitting, test sets that include samples drawn from distributions that vary from training are preferred to measure how well the algorithm generalizes beyond the training distribution.[44] The training, validation, and test sets must be independent to avoid data leakage. Potential sources of data leakage between partitions (such as lack of consistent patient labels) and applied mitigation strategies should be described.[7,8]

Describe How the Test Data Set Relates to the Proposed Clinical Setting, With Special Attention to Out-of-Distribution Classes (Items 13-15)

Authors should consider any differences between the image characteristics used for algorithm development and those that might be encountered in the real world. Out-of-distribution (OOD) "classes" are defined as those class categories or diagnoses that were not included in algorithm training data. For example, if an algorithm is trained to differentiate nevi vs melanomas, any image showing a diagnosis outside of nevi and melanomas would be OOD. Describe if images with classes that are OOD were included in the study test set, and report findings.[45] If images with OOD classes were not assessed, explain the drawbacks to clinical application (ie, undefined behavior when presented with classes outside of those studied). In some cases, OOD data may be subtle—for example, beyond classes not represented in training data, OOD may include unique combinations of other characteristics, such as clinical site, camera used, lighting, and patient demographics, of which some combinations may be underrepresented in algorithm training data.[42,44] To improve generalizability, multivendor and multisource images should be clearly labeled and included in algorithm development and evaluation.[7,15] The distribution of "classes" (eg, diagnoses or other label) in test data, stratified by patient characteristics such as ethnicity, age, and sex, should be clearly described. If there is any class imbalance (overrepresentation or underrepresentation) across classes, explain any procedures used to rectify class imbalance (such as oversampling or reweighting).[7]

Technique

Develop New Algorithms Using Standard Labels of Reference (Items 16-19)

The method used for image labeling should be clearly described with the reasoning behind the method selected. For malignant neoplasms, histopathological diagnosis should be considered the gold standard in diagnostic tasks.[1,37] However, note that even histopathology-based labels can be quite noisy given poor interobserver agreement for some diagnoses, which adds an additional challenge to establishing gold standard diagnoses (eg, melanoma).[46,47] If an alternative method is used for diagnosing malignant neoplasms, the potential for biases should be discussed (eg, level of label noise expected). For cases where histopathology is not available (eg, benign lesions, inflammatory disorders), there should be a clear description (eg, monitoring for change, consensus diagnosis) and justification of the labeling method. Additional research is needed to establish gold standards for labeling these classes of images. For choosing terms for diagnoses, labels and diagnostic groups

used in data repositories as well as public ontologies (*International Classification of Diseases, 11th Revision* [*ICD-11*], AnatomyMapper, SNOMED-CT) should be used whenever possible.[42] For histopathologic diagnoses of tumors, histopathologic extension codes of *ICD-11* can be used as an aid. Describe how terms were selected. For nondiagnostic tasks (eg, lesion monitoring, triage, predicting patient outcomes), how data were labeled and the rationale for the labeling scheme should be described.

Describe Algorithm Development (Item 19)

Methods, workflows, and mathematical formulas previously described elsewhere can be referenced but should be described in such manner to allow replicability. Reiterating known terms for metrics or loss functions by formulas only for the sake of suggesting technical height should be avoided; however, any new developments in methodologies should be described. Include information on how hyperparameters (eg, learning rate) were tuned and any limitations (eg, concerns about generalization—the ability of the algorithm to apply broadly across multiple data sets).

Recently, substantial research interest has been focused on interpretable and explainable AI algorithms. Interpretable algorithms are ones where causes for an output can be understood—for example, algorithms that can identify what parts of an input image helped with generating the output (eg, saliency maps) or are based on content-based image retrieval approaches.[48-50] Explainable AI algorithms generate information on the importance of each feature for each particular output; explainable algorithms allow us to describe in human terms how any algorithmic decision is made.[48,49] Interpretability and explainability may help with AI transparency but are still an active area of research.[48] Moreover, the end user (eg, patient, dermatologist, nonspecialist) is an important consideration for how interpretable or explainable features are presented. For reviewing purposes, we prefer that the authors include interpretability features such as saliency maps for appropriate evaluation. While these may help interpret algorithm results, clinical relevancy has yet to be determined.[51,52]

Technical Assessment

Provide a Method for the AI Algorithm or Algorithm Output to be Publicly Evaluable (Item 20)

Ideally, the AI algorithm would be made publicly available with a reference implementation available via open source code (eg, in a DOI-granting resource such as figshare, or domain-specific archives such as GitLab, GitHub, or BitBucket) or containerized for external testing. Alternatively, a public-facing test interface can be made available for external testing on individual images.[53,54] When possible, algorithms should be evaluated on standardized public test data sets and leaderboards for comparability and reproducibility against previously top-performing algorithms.

Describe How Performance Measures and Benchmarks Are Consistent With Proposed Clinical Translation (Items 21-23)

Authors should state why the performance measure chosen is appropriate to the algorithm task (eg, average precision, free-response receiver operating characteristic for detection tasks). In this context, the use case for the algorithm should be clearly described—who are the intended users and under what clinical scenario are they using the algorithm.[52] For example, an algorithm may be intended to be used by patients at home without a physician in the loop. Such a patient-facing algorithm may have more stringent expectations than an algorithm designed to support a dermatologist in clinic, where a human expert makes the final decision. If using frequently published metrics such as area under the curve, balanced accuracy, or sensitivity and specificity for classification tasks, the authors should consider implications of population-based screening for rare diseases. Reported performance and accuracy should be stratified according to demographic information and image artifacts if possible.

In addition to performance measures, diagnostic algorithms should be benchmarked against experts in their intended use setting, and the benchmarking process should be outlined.[15] Ideally, comparisons should also be made against the current reasonable standard of care as well. For example, patients are not usually treated by a panel of expert dermatologists, but by 1 dermatologist or general practitioner in the real-world setting. If there is a public benchmark or a previously published algorithm applicable to the task, it should be used for comparison. For example, tasks involving ISIC challenge data should include comparisons against previously developed algorithms. Some algorithms perform tasks such as predicting patient outcomes or risk stratification; such tasks may not have clearly defined expert comparators or previously defined benchmarks. In these cases, clear descriptions of intended applications are important (discussed in the next section).

Application

Describe Intended Use Cases and Target Conditions (Inside Distribution, Item 24)

For models to be used in the setting they were intended for, clearly describe the use case for the model (eg, diagnosis, triage) and the primary intended users (eg, patients, nurses, physician extenders, clinicians) and health care setting (eg, home, primary or secondary care, specialized centers).[55] Indicate how the information is intended to be used (eg, decision support or without supervision) and describe where in the health care workflow the model may fit.[56] Describe how the intended user or setting was incorporated into model development. For example, if a model is intended to be used by physicians in a telemedicine setting, model development should include physicians in reviewing the data, and the data should be representative of what is generated by telemedicine.

Discuss Potential Impacts on the Health Care Team and Patients (Item 25)

The goal of developing AI models for dermatology is eventual clinical application with benefits to health care teams, the health care system, and community. However, shortcomings and potential for harm must also be anticipated and evaluated prior to implementation.

Preliminary assessments of the algorithm's performance in conjunction with its intended user should be reported. For example, if an algorithm is meant to be used by a primary care physician to decide whether to refer to a dermatologist, researchers should assess performance of the target group with and without the algorithm. The desired outcomes should be clearly defined, and any biases assessed. The preliminary assessment does not need to be in the form of a prospective clinical trial but rather can demonstrate the value-

add using retrospective data and identify any early concerns prior to a larger prospective clinical trial.

The impact on patients should also be assessed in line with the algorithm's intended use. For example, an algorithm with a false-negative rate of 5% for diagnosing melanoma has a different impact if it is used as a decision support system by a clinician who can overrule the algorithm based on clinical judgment vs the same algorithm in the hands of patients directly, where the false reassurance may cause harm.

Ethical considerations and impact on vulnerable populations should also be considered and discussed. For example, an algorithm suggesting aesthetic medical treatments may have negative effects given the biased nature of beauty standards. An algorithm that diagnoses basal cell carcinomas but lacks any pigmented basal cell carcinomas, which are more often seen in skin of color, will not perform equitably across populations.

Prospective studies are recommended and should be performed prior to clinical implementation but may not be present in preliminary model descriptions. Please refer to SPIRIT-AI and CONSORT-AI for recommendations regarding AI clinical trials.[3,4,6]

Conclusions

In this consensus statement, we outline recommendations for the appropriate evaluation of AI algorithms for dermatology image applications and provide a checklist for addressing them. These recommendations inform all aspects of AI development, including data set curation, model building, and evaluation. We highlight areas where special attention to ethical considerations and potential sources of bias unique to clinical photography must be considered. While we propose guidelines for clinical and peer-review evaluation of AI, these recommendations are also relevant for a regulatory framework and should be considered for any automated dermatology algorithm that may affect the wider community.

ARTICLE INFORMATION

Accepted for Publication: October 1, 2021.

Published Online: December 1, 2021. doi:10.1001/jamadermatol.2021.4915

Author Affiliations: Stanford Department of Dermatology, Stanford School of Medicine, Redwood City, California (Daneshjou); Stanford Department of Biomedical Data Science, Stanford School of Medicine, Stanford, California (Daneshjou); Institute for Systems and Robotics, Instituto Superior Tecnico, Lisboa, Portugal (Barata); The University of Queensland Diamantina Institute, The University of Queensland, Dermatology Research Centre, Brisbane, Australia (Betz-Stablein, Soyer); Department of Computer Science and Engineering, University of Central Arkansas, Conway (Celebi); Microsoft, Seattle, Washington (Codella); Melanoma Unit, Dermatology Department, Hospital Clínic Barcelona, Universitat de Barcelona, IDIBAPS, Barcelona, Spain (Combalia, Malvehy); Melanoma Institute Australia, the University of Sydney, Camperdown, Australia (Guitera); Sydney Melanoma Diagnostic Centre, Royal Prince Alfred Hospital, Camperdown, Australia (Guitera); Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia (Gutman); Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York (Halpern, Kose, Rotemberg); Kitware, Inc, Clifton Park, New York (Helba); Department of Dermatology, Medical University of Vienna, Vienna, Austria (Kittler, Tschandl); University of Athens Medical School, Athens, Greece (Liopyris); Department of Dermatology, I Dermatology Clinic, Seoul, Korea (Seog); IDerma, Inc, Seoul, Korea (Seog); Dermatology Service and Research Service, Tennessee Valley Healthcare System, Department of Veterans Affairs, Nashville (Tkaczyk); Vanderbilt Dermatology Translational Research Clinic, Department of Dermatology, Vanderbilt University Medical Center, Nashville, Tennessee (Tkaczyk); Department of Biomedical Engineering, Vanderbilt University, Nashville, Tennessee (Tkaczyk).

Author Contributions: Drs Rotemberg and Daneshjou had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Daneshjou, Barata, Betz-Stablein, Codella, Combalia, Gutman, Halpern, Helba, Kittler, Kose, Liopyris, Soyer, Tkaczyk, Rotemberg.
Acquisition, analysis, or interpretation of data: Daneshjou, Celebi, Codella, Guitera, Gutman, Helba, Malvehy, Han, Tkaczyk, Tschandl, Rotemberg.
Drafting of the manuscript: Daneshjou, Betz-Stablein, Celebi, Codella, Liopyris, Malvehy, Tschandl, Rotemberg.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Rotemberg.
Obtained funding: Rotemberg.
Administrative, technical, or material support: Gutman, Helba, Tkaczyk, Rotemberg.
Supervision: Codella, Combalia, Gutman, Liopyris, Tkaczyk, Rotemberg.

Conflict of Interest Disclosures: Dr Daneshjou reported grants from Stanford Medicine Catalyst and UCB and personal fees from DWA, Pfizer, and VisualDx outside the submitted work. Dr Barata reported grants from Fundação para a Ciência e Tecnologia (FCT) during the conduct of the study; and a 2021 Google Research Award from Google Research outside the submitted work. Dr Codella reported investments in technology and health care outside the submitted work; in addition, Dr Codella had a patent for surgical skin lesion removal (US10568695) issued, a patent for surface reflectance reduction in images using nonspecular portion replacement (US10255674) issued, and a patent for category oversampling for imbalanced machine learning (US Patent App. 14/500,023) pending. Mr Combalia reported personal fees from IDIBAPS during the conduct of the study. Dr Guitera reported personal fees (honoraria) from MetaOptima outside the submitted work. Dr Halpern reported personal fees from Canfield Scientific, Inc, Scibase, and Lloyd Charitable Trust and an equity position from HCW LLC and SKIP Derm LLC outside the submitted work. Dr Kittler reported equipment and personal fees from Fotofinder; equipment from Heine and Derma Medical; and nonpersonal fees from MetaOptima outside the submitted work. Dr Kose reported grants from National Institutes of Health/National Cancer Institute (P30 CA008748) during the conduct of the study. Dr Han reported being the founder, CEO, and CTO of IDerma, Inc, during the conduct of the study. Dr Soyer reported grants from National Health and Medical Research Council (APP1137127) during the conduct of the study; personal fees (medical reporting fee, medical consultant, minor shareholder) from MoleMap NZ Limited, personal fees (medical consultant) from Canfield Scientific, and shareholder and medical reporting from E-Derm Consult GmbH outside the submitted work; in addition, Dr Soyer had a microbiopsy device patent for PCT/AU/2013.000394 US 9, 662,095 B2 issued and served as a member of the Digital Health Committee of the Australasian College of Dermatologists. Dr Tkaczyk reported grants from Department of Veterans Affairs (Career Development Award Number IK2 CX001785) during the conduct of the study. Dr Tschandl reported grants from Lilly and MetaOptima and personal fees from Lilly, Silverchair, FotoFinder, and Novartis outside the submitted work. Dr Rotemberg reported nonfinancial support (expert adviser) from Inhabit Brands, Inc, outside the submitted work. No other disclosures were reported.

Funding/Support: Dr Soyer holds a National Health and Medical Research Council Medical Research Future Fund Next Generation Clinical Researchers Program Practitioner Fellowship (APP1137127). This work is partially supported by Career Development Award Number IK2 CX001785 from the US Department of Veterans Affairs Clinical Science R&D (CSRD) Service to Dr Tkaczyk. Dr Barata is funded by FCT under the scope of project (CEECIND/00326/2017) and by a 2021 Google Research Award. Dr Celebi's work was supported by the National Science Foundation under grant No. 1946391. Dr Daneshjou is supported by National Institutes of Health (T32 5T32AR007422-38). Dr Rotemberg is supported by the Melanoma Research Alliance, the National Institutes of Health/National Cancer Institute (Cancer Center Support Grant P30 CA008748), and the Charina Fund.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study;

collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Daneshjou R, He B, Ouyang D, Zou JY. How to evaluate deep learning for cancer diagnostics—factors and recommendations. *Biochim Biophys Acta Rev Cancer*. 2021;1875(2):188515. doi:10.1016/j.bbcan.2021.188515

2. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform*. 2021;28(1):e100289. doi:10.1136/bmjhci-2020-100289

3. Taylor M, Liu X, Denniston A, et al; SPIRIT-AI and CONSORT-AI Working Group. Raising the bar for randomized trials involving artificial intelligence: the SPIRIT-Artificial Intelligence and CONSORT-Artificial Intelligence guidelines. *J Invest Dermatol*. 2021;141(9):2109-2111. doi:10.1016/j.jid.2021.02.744

4. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x

5. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. 2021;27(2):186-187. doi:10.1038/s41591-021-01229-5

6. Charalambides M, Flohr C, Bahadoran P, Matin RN. New international reporting guidelines for clinical trials evaluating effectiveness of artificial intelligence interventions in dermatology: strengthening the SPIRIT of robust trial reporting. *Br J Dermatol*. 2021;184(3):381-383. doi:10.1111/bjd.19616

7. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029. doi:10.1148/ryai.2020200029

8. Sengupta PP, Shrestha S, Berthon B, et al. Proposed Requirements for Cardiovascular Imaging-Related Machine Learning Evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. *JACC Cardiovasc Imaging*. 2020;13(9):2017-2035. doi:10.1016/j.jcmg.2020.07.015

9. Kovarik C, Lee I, Ko J; Ad Hoc Task Force on Augmented Intelligence. Commentary: position statement on augmented intelligence (AuI). *J Am Acad Dermatol*. 2019;81(4):998-1000. doi:10.1016/j.jaad.2019.06.032

10. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. doi:10.1136/bmjopen-2020-048008

11. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. doi:10.1136/bmjopen-2016-012799

12. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat Med*. 2020;26(6):807-808. doi:10.1038/s41591-020-0941-1

13. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6

14. Bissoto A, Valle E, Avila S. *Debiasing Skin Lesion Datasets and Models? Not So Fast. Computer Vision and Pattern Recognition*: IEEE; 2020.

15. Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20(7):938-947. doi:10.1016/S1470-2045(19)30333-X

16. Du-Harpur X, Arthurs C, Ganier C, et al. Clinically relevant vulnerabilities of deep machine learning systems for skin cancer diagnosis. *J Invest Dermatol*. 2021;141(4):916-920. doi:10.1016/j.jid.2020.07.034

17. Campbell JP, Lee AY, Abràmoff M, et al. Reporting guidelines for artificial intelligence in medical research. *Ophthalmology*. 2020;127(12):1596-1599. doi:10.1016/j.ophtha.2020.09.009

18. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. 2020;2(10):e549-e560. doi:10.1016/S2589-7500(20)30219-3

19. Kelly B, Judge C, Bollard SM, et al. Radiology artificial intelligence, a systematic evaluation of methods (RAISE): a systematic review protocol. *Insights Imaging*. 2020;11(1):133. doi:10.1186/s13244-020-00929-9

20. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health*. 2020;2(10):e537-e548. doi:10.1016/S2589-7500(20)30218-1

21. Pfau M, Walther G, von der Emde L, et al. Artificial intelligence in ophthalmology: guidelines for physicians for the critical evaluation of studies. Article in German. *Ophthalmologe*. 2020;117(10):973-988. doi:10.1007/s00347-020-01209-z

22. Chiang S, Picard RW, Chiong W, et al. Guidelines for conducting ethical artificial intelligence research in neurology: a systematic approach for clinicians and researchers. *Neurology*. 2021;97(13):632-640. doi:10.1212/WNL.0000000000012570

23. Ibrahim H, Liu X, Rivera SC, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials*. 2021;22(1):11. doi:10.1186/s13063-020-04951-6

24. Kundeti SR, Vaidyanathan MK, Shivashankar B, Gorthi SP. Systematic review protocol to assess artificial intelligence diagnostic accuracy performance in detecting acute ischaemic stroke and large-vessel occlusions on CT and MR medical imaging. *BMJ Open*. 2021;11(3):e043665. doi:10.1136/bmjopen-2020-043665

25. Meshaka R, Pinto Dos Santos D, Arthurs OJ, Sebire NJ, Shelmerdine SC. Artificial intelligence reporting guidelines: what the pediatric radiologist needs to know. *Pediatr Radiol*. Published online July 1, 2021. doi:10.1007/s00247-021-05129-1

26. Omoumi P, Ducarouge A, Tournier A, et al. To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur Radiol*. 2021;31(6):3786-3796. doi:10.1007/s00330-020-07684-x

27. Parums DV. Editorial: artificial intelligence (AI) in clinical medicine and the 2020 CONSORT-AI study guidelines. *Med Sci Monit*. 2021;27:e933675.

28. Petzold A, Albrecht P, Balcer L, et al; IMSVISUAL, ERN-EYE Consortium. Artificial intelligence extension of the OSCAR-IB criteria. *Ann Clin Transl Neurol*. 2021;8(7):1528-1542. doi:10.1002/acn3.51320

29. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Health Care Inform*. 2021;28(1):e100385. doi:10.1136/bmjhci-2021-100385

30. Sounderajah V, Ashrafian H, Golub RM, et al; STARD-AI Steering Committee. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11(6):e047709. doi:10.1136/bmjopen-2020-047709

31. Katragadda C, Finnane A, Soyer HP, et al; International Society of Digital Imaging of the Skin (ISDIS)-International Skin Imaging Collaboration (ISIC) Group. Technique standards for skin lesion imaging: a Delphi consensus statement. *JAMA Dermatol*. 2017;153(2):207-213. doi:10.1001/jamadermatol.2016.3949

32. Barata C, Celebi ME, Marques JS. Improving dermoscopy image classification using color constancy. *IEEE J Biomed Health Inform*. 2015;19(3):1146-1152. doi:10.1109/JBHI.2014.2336473

33. Ghorbani A, Natarajan V, Coz D, Liu Y. DermGAN: synthetic generation of clinical skin images with pathology. NeurIPS ML4H Workshop, 2019. *arXiv*. Posted online November 20, 2019. https://arxiv.org/abs/1911.08716

34. Bissoto A, Valle E, Avila S. GAN-based data augmentation and anonymization for skin-lesion analysis: a critical review. ISIC Skin Image Analysis Workshop at CVPR 2021. *arXiv*. Posted online April 20, 2021. https://arxiv.org/abs/2104.10603

35. International Skin Imaging Collaboration. ISIC archive. Accessed July 7, 2020. https://www.isic-archive.com/

36. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol*. Published online September 22, 2021. doi:10.1001/jamadermatol.2021.3129

37. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161. doi:10.1038/sdata.2018.161

38. Codella N, Rotemberg V, Tschandl P, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv*. Posted online February 9, 2019; revised March 29, 2019. https://arxiv.org/abs/1902.03368

39. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep

learning algorithms. *JAMA*. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067

40. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol*. 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348

41. Kinyanjui NM, Odonga T, Cintas C, et al. Estimating skin tone and effects on classification performance in dermatology datasets. NeurIPS 2019 Workshop on Fair ML for Health; 2019.

42. Okoji UK, Taylor SC, Lipoff JB. Equity in skin typing: why it is time to replace the Fitzpatrick scale. *Br J Dermatol*. 2021;185(1):198-199. doi:10.1111/bjd.19932

43. Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900-908. doi:10.1038/s41591-020-0842-3

44. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. 2021;27(4):582-584. doi:10.1038/s41591-021-01312-x

45. Roy AG, Ren J, Azizi S, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *arXiv*. Posted online April 8, 2021. https://arxiv.org/abs/2104.03829

46. Elmore JG, Barnhill RL, Elder DE, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ*. 2017;357:j2813. doi:10.1136/bmj.j2813

47. Elder DE, Piepkorn MW, Barnhill RL, et al. Pathologist characteristics associated with accuracy and reproducibility of melanocytic skin lesion interpretation. *J Am Acad Dermatol*. 2018;79(1):52-59.e5. doi:10.1016/j.jaad.2018.02.070

48. Das A, Rad PS. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. *arXiv*. Posted online June 16, 2020. https://arxiv.org/abs/2006.11371

49. Codella NCF, Lin C-C, Halpern A, Hind M, Feris R, Smith JR. Collaborative Human-AI (CHAI): evidence-based interpretable melanoma classification in dermoscopic images. MICCAI 2018, Workshop on Interpretability of Machine Intelligence in Medical Image Computing (IMIMIC); 2018.

50. Barata C, Santiago C. Improving the explainability of skin cancer diagnosis using CBIR. Presented at: 24th International Conference on Medical Image Computing Computer Assisted Intervention (MICCAI); 2021; virtual.

51. Saporta A, Gui X, Agrawal A, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*. Posted online March 2, 2021. doi:10.1101/2021.02.28.21252634

52. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229-1234. doi:10.1038/s41591-020-0942-0

53. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol*. 2018;138(7):1529-1538. doi:10.1016/j.jid.2018.01.028

54. Abid A, Abdalla A, Abid A, Khan D, Alfozan A, Zou J. Gradio: hassle-free sharing and testing of ml models in the wild. *arXiv*. Posted online June 6, 2019. https://arxiv.org/abs/1906.02569

55. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Presented at: FAT* '19: Conference on Fairness, Accountability, and Transparency; 2019; Atlanta, Georgia.

56. Janda M, Soyer HP. Can clinical decision making be enhanced by artificial intelligence? *Br J Dermatol*. 2019;180(2):247-248. doi:10.1111/bjd.17110