Characteristics of publicly available skin cancer image datasets

Wen, David; Khan, Saad M; Xu, Antonio Ji; Ibrahim, Hussein; Smith, Luke; Caballero, Jose; Zepeda, Luis; de Blas Perez, Carlos; Denniston, Alastair K; Liu, Xiaoxuan; Matin, Rubeta N

DOI:
10.1016/S2589-7500(21)00252-1

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Wen, D, Khan, SM, Xu, AJ, Ibrahim, H, Smith, L, Caballero, J, Zepeda, L, de Blas Perez, C, Denniston, AK, Liu, X & Matin, RN 2021, 'Characteristics of publicly available skin cancer image datasets: a systematic review', *The Lancet Digital Health*, vol. 4, no. 1, pp. e64-e74. https://doi.org/10.1016/S2589-7500(21)00252-1

Characteristics of publicly available skin cancer image datasets: a systematic review

David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, Rubeta N Matin**

Publicly available skin image datasets are increasingly used to develop machine learning algorithms for skin cancer diagnosis. However, the total number of datasets and their respective content is currently unclear. This systematic review aimed to identify and evaluate all publicly available skin image datasets used for skin cancer diagnosis by exploring their characteristics, data access requirements, and associated image metadata. A combined MEDLINE, Google, and Google Dataset search identified 21 open access datasets containing 106 950 skin lesion images, 17 open access atlases, eight regulated access datasets, and three regulated access atlases. Images and accompanying data from open access datasets were evaluated by two independent reviewers. Among the 14 datasets that reported country of origin, most (11 [79%]) originated from Europe, North America, and Oceania exclusively. Most datasets (19 [91%]) contained dermoscopic images or macroscopic photographs only. Clinical information was available regarding age for 81 662 images (76·4%), sex for 82 848 (77·5%), and body site for 79 561 (74·4%). Subject ethnicity data were available for 1415 images (1·3%), and Fitzpatrick skin type data for 2236 (2·1%). There was limited and variable reporting of characteristics and metadata among datasets, with substantial under-representation of darker skin types. This is the first systematic review to characterise publicly available skin image datasets, highlighting limited applicability to real-life clinical settings and restricted population representation, precluding generalisability. Quality standards for characteristics and metadata reporting for skin image datasets are needed.

Introduction

Digital health innovation has the potential to improve health care by increasing access to specialist expertise.[1,2] Among the myriad of machine learning applications in health care, medical image classification, particularly for dermatology, has advanced substantially in recent years,[3] and includes diagnosis of skin cancers from dermoscopic or macroscopic photographs.[4–10] Advances in machine learning algorithm diagnostic accuracy have largely been driven by utilisation of deep learning architectures made possible through greater availability of computing power and large repositories of digital images for algorithm training.[11,12] For this purpose, large numbers of digital images easily accessible through publicly available datasets have been used in dermatology.[13] Publicly available datasets used for developing machine learning algorithms circumvent barriers to dataset procurement, such as having appropriate technological infrastructure, regulatory approvals, time, and financial investment for large-scale digital image acquisition from participants.[14] Furthermore, publicly available datasets can be used as a benchmark for direct comparison of algorithm performance.[15,16]

Skin cancer incidence continues to rise globally, placing increasing demands on health-care services.[17–21] Digital solutions to address this demand have been reflected by accelerated teledermatology adoption during the COVID-19 pandemic.[22,23] Machine learning algorithms have potential for automated diagnosis of skin malignancies through digital image analysis, and diagnostic accuracy of machine learning algorithms has been shown in the past 5 years to be comparable to, or even surpass, dermatologists in controlled experimental settings.[4–9]

Publicly available skin image datasets, such as those hosted through the International Skin Imaging Collaboration (ISIC) archive,[13] are increasingly used to develop machine learning algorithms for skin cancer diagnosis.[24–26] Additionally, although primarily aimed for use as educational resources, dermatology atlases (compilations of photographs of skin diseases) containing digital images are frequently used as a source of skin lesion images for algorithm development.[27] However, with training data from circumscribed populations, often curated retrospectively, machine learning algorithms are susceptible to overfitting, and their generalisability is heavily influenced by the participants and images used for training, which are prone to selection bias.[28] Algorithms used for skin lesion classification frequently underperform when tested on independent datasets.[26,29,30] Further examples of the susceptibility of machine learning algorithms to biases for clinical factors such as age, sex, ethnicity, and socioeconomic status have also been reported in diverse areas of health care and artificial intelligence.[31–35] This underlines the importance of detailing the exact composition of datasets through metadata reporting, to ensure the generalisability of algorithms to real-world populations. Furthermore, the recently developed concept of health data poverty—systematic data disparities leading to inequalities in health care—emphasises the need to ensure diversity, transparency, and usability of datasets.[14,36,37]

Demographic and clinical metadata can also greatly influence machine learning algorithm development and validation, with classification accuracy increasing for skin lesions when subject and lesion metadata are integrated into models.[38,39] This more closely reflects clinical practice, as dermatologists conduct in-person history taking and evaluation of skin lesions via naked eye, dermoscopic, and physical examination (eg, lesion

Lancet Digit Health 2021

Published Online November 9, 2021 https://doi.org/10.1016/S2589-7500(21)00252-1

*Joint last authors

Oxford University Clinical Academic Graduate School, University of Oxford, Oxford, UK (D Wen BMBCh); **Institute of Clinical Sciences, University of Birmingham, Birmingham, UK** (D Wen); **Royal Berkshire Hospital, Royal Berkshire NHS Foundation Trust, Reading, UK** (D Wen, S M Khan MBChB); **Department of Dermatology, Churchill Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK** (A Ji Xu BMBCh, R N Matin PhD); **University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (H Ibrahim MBChB, Prof A K Denniston PhD, X Liu PhD); **Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK** (H Ibrahim, Prof A K Denniston, X Liu); **Centre for Regulatory Science and Innovation, Birmingham Health Partners, Birmingham, UK** (H Ibrahim, Prof A K Denniston, X Liu); **Databiology, Oxford, UK** (L Smith BSc, J Caballero MSc, L Zepeda BSc, C de Blas Perez BSc); **Health Data Research UK, London, UK** (Prof A K Denniston, X Liu); **National Institute for Health Research Biomedical Research Centre for Ophthalmology, Moorfields Eye Hospital NHS Foundation Trust, London, UK** (Prof A K Denniston); **UCL Institute of Ophthalmology, London, UK** (Prof A K Denniston)

Correspondence to:
Dr Rubeta N Matin, Department of Dermatology, Churchill Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford OX3 7LE, UK
rubeta.matin@ouh.nhs.uk

palpation), and not through image analysis alone. Completeness of metadata provides information about the population, disease, and data types on which the algorithm was trained or validated, which is essential for extrapolating assumptions of generalisability of algorithm performance to other populations.

To the best of our knowledge, there are no guidelines or quality standards that characterise optimal skin image datasets that could be used to train machine learning algorithms, and these datasets have not previously been systematically characterised. This review aimed to identify publicly available skin image datasets used to develop machine learning algorithms for skin cancer diagnosis, categorise their data access requirements, and systematically evaluate their characteristics including associated metadata.

Methods

Search strategy and selection criteria

See Online for appendix

This systematic review adheres to PRISMA guidelines;[40] for the PRISMA checklist see the appendix (p 12). As the review did not evaluate a direct health-related outcome, it did not meet criteria for registration of the protocol with PROSPERO.[41] Searches were conducted on MEDLINE, Google, and Google Dataset Search, on Sept 4, 2020. The MEDLINE search was updated on Sept 1, 2021. The database was searched from inception with MeSH terms and keywords including "dataset" OR "database", "artificial intelligence" OR "machine learning", "skin cancer", and "imaging" OR "dermoscopy". The full search strategy is given in the appendix (p 2). Two independent reviewers (DW and AJX) screened titles and abstracts for articles describing skin cancer image datasets, studies using datasets to train or test machine learning algorithms for skin cancer diagnosis, or review articles detailing any skin cancer image datasets. Two independent reviewers (AJX, DW, HI, or SMK) subsequently reviewed full-text articles for publicly available datasets and attempted to access them at source. Full texts were reviewed for publications where abstracts were unavailable. Google Translate was used for seven non-English articles (four in German, one in French, one in Spanish, and one in Chinese).

Google and Google Dataset searches were completed by two independent reviewers (DW and AJX) to identify publicly available skin cancer image datasets. These searches were completed using the search term "skin cancer image dataset", and repeated with "melanoma image dataset" (appendix p 2). The number of new skin cancer image datasets found on each search result page were recorded. Reference lists of any online articles were also reviewed for named publicly available datasets. Once 20 consecutive search results no longer mentioned any new datasets, the number of search results for review was rounded up to the nearest multiple of 50, and no further webpages were reviewed beyond this point.

To be included for data extraction, datasets had to contain images of either cutaneous melanoma, basal cell carcinoma (BCC), or cutaneous squamous cell carcinoma (cSCC). Datasets could contain any form of non-radiological skin lesion images, such as macroscopic clinical photographs or dermoscopy. There was no restriction on geographical origin, patient population, or language.

Datasets were excluded if they did not contain images of skin cancers (eg, skin rashes or benign lesions only). Histopathological image and radiological image datasets were also excluded. Datasets containing text or numeric-only data and images of non-human subjects were excluded, as were inaccessible datasets that were described as open access but were either inactivated or unable to be found by two reviewers (DW and SMK). All datasets included in the review were agreed by consensus by three authors (DW, AJX, and SMK).

Data analysis

Identified datasets meeting inclusion criteria were grouped into categories based on a system used by Khan and colleagues.[14] These included open access datasets (freely accessible or easily accessible via registration or email) and regulated access datasets (requiring payment, formal institutional agreements, or ethical approval).

A data collection template was designed to evaluate dataset characteristics, including number of images, number of participants (ie, individuals from whom images were taken, regardless of whether their active consent was reported), country of origin, publication date, imaging modality, image capture device, image format, and number of skin lesion categories. Further items reviewed for reporting are listed in the appendix (p 3). Metadata labels associated with individual images were also reviewed for included datasets. Images and metadata of included datasets were manually reviewed by two independent reviewers (DW and SMK), together with corresponding articles or supplementary information describing the dataset. Discrepancies were resolved by discussion between reviewers, and a third party with methodological expertise (RNM) if required.

All images and accompanying data from open access datasets were imported into Databiology Lab (LS, JC, LZ, and CdBP), a data management and analysis platform. Wherever possible, image files were registered without downloading to reduce data duplication. Metadata was recorded in entities with a number of attributes and values. Search queries were created to evaluate metadata associated with images (collectively or by individual dataset).

We found that articles describing machine learning algorithms also used images from online dermatology atlases to train algorithms. Therefore, atlases were also included if they met the aforementioned inclusion criteria. Atlases were defined as mainly educational electronic resources that either included "atlas" in the title or displayed images on multiple webpages. This was in contrast to datasets, which contained multiple images

downloadable in one or a few compressed files (.zip or .rar).

As characteristics and metadata varied greatly within atlases and between atlases, a simplified data extraction template was used (appendix p 4). Images from atlases amenable to download or linkage were imported to Databiology Lab for search and examination of available metadata (LS, JC, LZ and CdBP). For atlases for which this was not possible, a sample of images from each atlas was reviewed by two reviewers (DW and SMK). If enough images were available, ten consecutive images each of BCC, cSCC, and melanoma were reviewed by each reviewer, either from the top (DW) or bottom (SMK) of webpages or atlas indexes. Atlases were deemed to have metadata reporting for an item (eg, lesion site) if more than 50% of reviewed images contained the metadata label.

Results

The MEDLINE search returned 1897 articles, which were screened by title and abstract (figure 1). Review of 584 full-text articles identified 26 datasets and 20 atlases meeting inclusion criteria. Review of 200 webpages from each of the Google and Google Dataset searches, along with supplementary screening, identified 25 datasets and 12 atlases. Combining these results, 22 duplicate datasets and 12 atlases were excluded, leaving 29 individual datasets and 20 individual atlases included for review.

Attempts to access data at source revealed eight regulated access datasets (table) and three regulated access atlases.[56–58] Of the eight datasets, six required ethical committee or institutional approval,[10,50,53–55] one required a £75 fee and a licencing agreement,[51] and one required a competition invitation.[52] Thumbnails of images from Asan, Hallym, SNU, and Severance datasets were available for download, but access to full-size images required formal approval from local data access or ethics committees,[50] or the originating hospitals.[10] All three regulated access atlases required payment for access to digital images (US$200; £145).[56–58] Characteristics of these regulated access datasets and atlases were gathered from information at source or accompanying publications; the images themselves or associated metadata files were not downloaded and reviewed.

Inaccessible datasets, which could not be analysed further, included one dataset from China (XiangyaDerm), which was described to be publicly released[59] but could not be found via the stated URL, and one atlas that had been inactivated (DermQuest). One further dataset might have been available upon reasonable request to the authors,[60] although we did not receive a response to our request for access. Datasets that contained images that were also included in another dataset, and therefore did not contain any unique images, were not classed as an additional independent dataset. All 1512 images in the ISIC 2018 task 3 challenge were included in the ISIC 2019 challenge, so the former was not included as an individual dataset. This also applied to SD-198 and SD-128 datasets,[61] images of which were included in SD-260.[49]

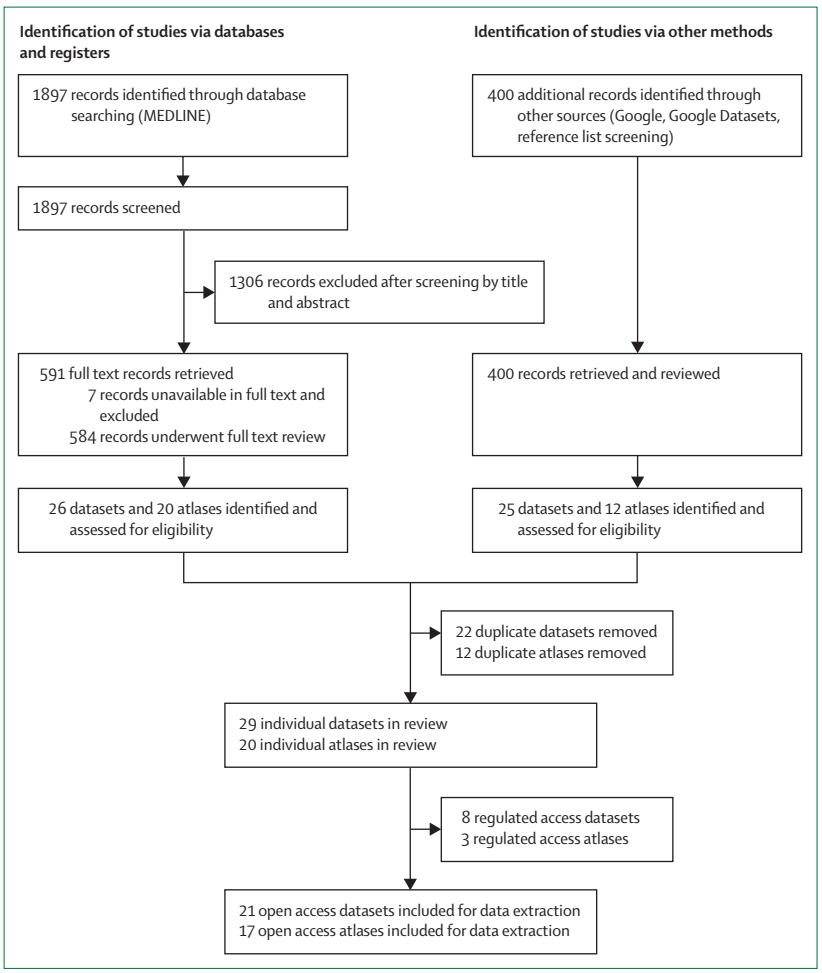

Figure 1: **PRISMA flow diagram outlining dataset and atlas identification**

As images from preceding ISIC challenges have been re-used in later challenges (appendix p 5), ISIC images were grouped by their original datasets, as presented in the ISIC archive image gallery (accessed March, 2021), to avoid duplicate images being counted more than once in analyses. Images not included in the ISIC archive image gallery and only available via ISIC challenges were listed separately (ISIC 2019 and 2020 challenge test sets, and ISIC 2018 challenge task 1–2 test set).

Overall, 21 open access datasets and 17 open access atlases included images that could be downloaded without barriers, and were included for data extraction. The access links to individual datasets are detailed in the appendix (p 6).

14 (67%) of 21 datasets detailed country of origin (table). Of these, 11 originated from one country only: six from Europe, two from Oceania, one from North America, one from South America, and one from Asia (appendix p 7). Three datasets contained images from multiple

	Country of origin	Year of dataset publication	Imaging modality	Image acquisition device	Image format	Number of skin lesion categories included	Number of participants	Number of images
Open access datasets								
ISIC archive								
ISIC 2020 Hospital Clinic Barcelona[42]	Spain	2020	Dermoscopic	MoleMax HD digital dermatoscopy system	DICOM or .jpg	2	356	7311
ISIC 2020 University of Queensland[42]	Australia	2020	Dermoscopic	Not reported	DICOM or .jpg	Not reported	304	8449
ISIC 2020 Medical University Vienna[42]	Austria	2020	Dermoscopic	MoleMax HD digital dermatoscopy system	DICOM or .jpg	2	432	4374
ISIC 2020 Memorial Sloan Kettering Cancer Centre[42]	USA	2020	Dermoscopic	Dermoscopic attachment to a digital single reflex lens camera or a smartphone	DICOM or .jpg	5	523	11108
ISIC 2020 Sydney Melanoma Diagnosis Centre and Melanoma Institute Australia[42]	Australia	2020	Dermoscopic	Dermoscopic attachment to a digital single reflex lens camera or a smartphone	DICOM or .jpg	8	441	1884
BCN20,000[43]	Spain	2019	Dermoscopic	Dermoscopic attachments to three high-resolution cameras	.jpg	9	Not reported	12413
HAM10,000[44]	Austria and Australia	2018	Dermoscopic	Various devices including: MoleMax HD, DermLite Foto (3Gen) camera, DermLite Fluid, DermLite DL3, and analogue cameras	.jpg	8	Not reported	10015
2018 JID editorial images[13]	Not reported	2018	Macroscopic	Not reported	.jpg	3	Not reported	100
MSK 1–5[13]	Not reported	MSK 1–2 2015; MSK 3-5 2017	Dermoscopic	Not reported	.jpg	15	Not reported	3918
UDA 1–2[13]	Not reported	UDA-1 2014, UDA-2 2015	Dermoscopic	Not reported	.jpg	7	Not reported	617
ISIC challenge only								
ISIC 2020 challenge test set[42]	Spain, Australia, Austria, USA, Greece	2020	Dermoscopic	Not reported*	DICOM or .jpg	Not reported	690	10982
ISIC 2019 challenge test set	Australia, Austria, Turkey, New Zealand, Sweden, Argentina[†]	2018 and 2019	Dermoscopic	Not reported	.jpg	Not reported	Not reported	8238
ISIC 2018 test set (tasks 1 and 2)[26]	Not reported	2018	Dermoscopic	Not reported	.jpg	Not reported	Not reported	1000
Non-ISIC datasets								
PAD-UFES -20[45]	Brazil	2020	Macroscopic	Smartphones	.png	6	1373	2298
PH[2] [15]	Portugal	2013	Dermoscopic	Tuebinger Mole Analyzer System	.bmp	3	Not reported	200
7-point criteria evaluation database[46]	Not reported	2018‡	Dermoscopic and macroscopic (paired)	Not reported	.jpg	15	1011	2013
MED-NODE[47]	Netherlands	2015	Macroscopic	Nikon D3 or Nikon D1x body and a Nikkor 2·8/105 mm micro lens	.jpg	2	Not reported	170
SKINL2[48]	Portugal	2019	Light field photographs, dermoscopic photographs (paired)	Raytrix R42 Galilean focused plenoptic camera, Ricoh 25 mm f/1·8 lens with custom-made housing	.png	8	Not reported	814
SNU dataset[10]	South Korea	2018	Macroscopic	Not reported	.png	81	Not reported§	240
University of Waterloo dataset¶	Not reported	Not reported	Macroscopic	Consumer level cameras	.jpg and .png for contours	2	Not reported	206
SD-260[49]‖	Not reported	2019	Macroscopic	Digital cameras and mobile phones	.jpg	260	Not reported	20600

(Table continues on next page)

	Country of origin	Year of dataset publication	Imaging modality	Image acquisition device	Image format	Number of skin lesion categories included	Number of participants	Number of images
(Continued from previous page)								
Regulated access datasets								
Asan dataset[50]**	South Korea	2017	Macroscopic	Not reported	Not reported	12	4867	17125
Hallym dataset[50]	South Korea	2017	Not reported	Not reported	Not reported	1	106	152
DERMOFIT Image Library: Edinburgh dataset[51]	UK	Not reported	Not reported	Canon EOS 350D DSLR	Not reported	10	Not reported	1300
IMA205[52]	Not reported	2018	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
MoleMapper app patient photos[53]	USA	2017	Macroscopic	iPhone (4S or newer)	Not reported	2	2069	2422
SNU dataset entire test set[10] (2201 images)	South Korea	2018	Macroscopic	Not reported	Not reported	134	1608	2201
Severance dataset (test subset)[54]	South Korea	2020	Macroscopic	Not reported	Not reported	43	10426	40331
Papadakis et al (2021) dataset[55]	Germany	2021	Macroscopic	Commercial digital camera	.jpg	1	156	156

ISIC=International Skin Imaging Collaboration. DICOM=Digital Imaging and Communications in Medicine. JID=*Journal of Investigative Dermatology.* *ISIC 2020 test set includes images from Andreas Syngros Hospital. Device used for image acquisition is not reported for this centre. †ISIC 2019 test set includes all images from ISIC 2018 test set (task 3). These images originated from Australia, Austria, Turkey, New Zealand, Sweden, and Argentina. Whether images from additional centres or countries are included in the ISIC 2019 test set is not reported. ‡Images in this dataset are from the Edra Atlas published in 2000.[56] These images were made publicly available in 2018. §The number of participants was not reported for the 240 full-size photos available for download. However, for the larger dataset of 2201 images, the number of participants was reported as 1608. ¶Images from this dataset include images from DermIS and DermQuest (deactivated) atlases. ||Images from this dataset include images from the deactivated DermQuest atlas. **A further iteration of this dataset containing additional images is described by Han and colleagues;[10] however, the images cannot be made available in totality due to privacy regulations.

Table: **Publicly available datasets and their characteristics**

countries, with two of these being recent ISIC challenge datasets. Among the 14 datasets, 11 (79%) were from Europe, North America, and Oceania exclusively.

Year of publication was available for 20 of 21 datasets and ranged from 2013 to 2020. The vast majority of images (101839 [95·2%] of 106950) were published in 2018 or later, and there was an increasing trend in the number of images published per year (figure 2A). 4905 images (4·6%) were published in 2017 or earlier, and 206 images (0·2%) had no publication date available. Imaging modalities included are shown in figure 2B; 19 datasets (91%) contained dermoscopic images or macroscopic photographs only.

Images were acquired as part of clinical care in 11 (52%) of 21 datasets, and for research in two datasets (10%). These images were all acquired in secondary care dermatology specialist settings. Three datasets (14%) used existing images from atlases and were created for the purpose of developing or testing of machine learning algorithms. Five datasets (24%) did not explicitly state the purpose for image acquisition. Devices used for data acquisition ranged from mobile phones to digital single-lens reflex cameras and dermoscopy systems (table). Images were available in .jpg, .png, .bmp, or Digital Imaging and Communications in Medicine (DICOM) formats.

Five datasets (24%) contained image segmentations and three (14%) included feature or lesion size annotations for images. The number of skin lesion classifications per dataset ranged from two to 260 (median seven, IQR 2·5–12). Methods of obtaining ground truth were described in 16 datasets (76%) and included review of medical records, clinical consensus between reviewers, diagnosis as described originally in an atlas, serial imaging, follow-up of patients over a specified time period, confocal microscopy, and histopathology.

Completeness of reporting for dataset characteristics is summarised in figure 3A. Reporting for individual datasets is displayed in the appendix (p 8). The number of individuals from whom images were taken was reported in eight datasets (38%), data collection period in eight (38%), inclusion or exclusion criteria in 12 (57%), device used for image acquisition in 12 (57%), and whether there was image processing or adjustment by a reviewer in four (19%). Ethical approval was reported in ten datasets (48%), and participant consent in seven (33%).

Overall, the open access datasets contained 106950 skin lesion images (median 2298, IQR 429–9232). 60189 (56·3%) were included in datasets hosted within the ISIC archive, 20220 (18·9%) are currently being used in ongoing ISIC challenge test sets, and 26541 (24·8%) were from non-ISIC datasets (figure 2C). No ground truth labels were available for the 20220 images currently being used in ongoing ISIC challenge test sets (table), although these might be released in future. Considering the eight datasets that reported number of participants, there were a total of 48419 images from 5130 participants.

In the SKINL2 dataset,[48] each light field comprised 81 different views. Each light field was counted as a single image for the purposes of counting images in each dataset. Light field images could not be downloaded in full and imported into the Databiology platform due to large file size. However individual images could be

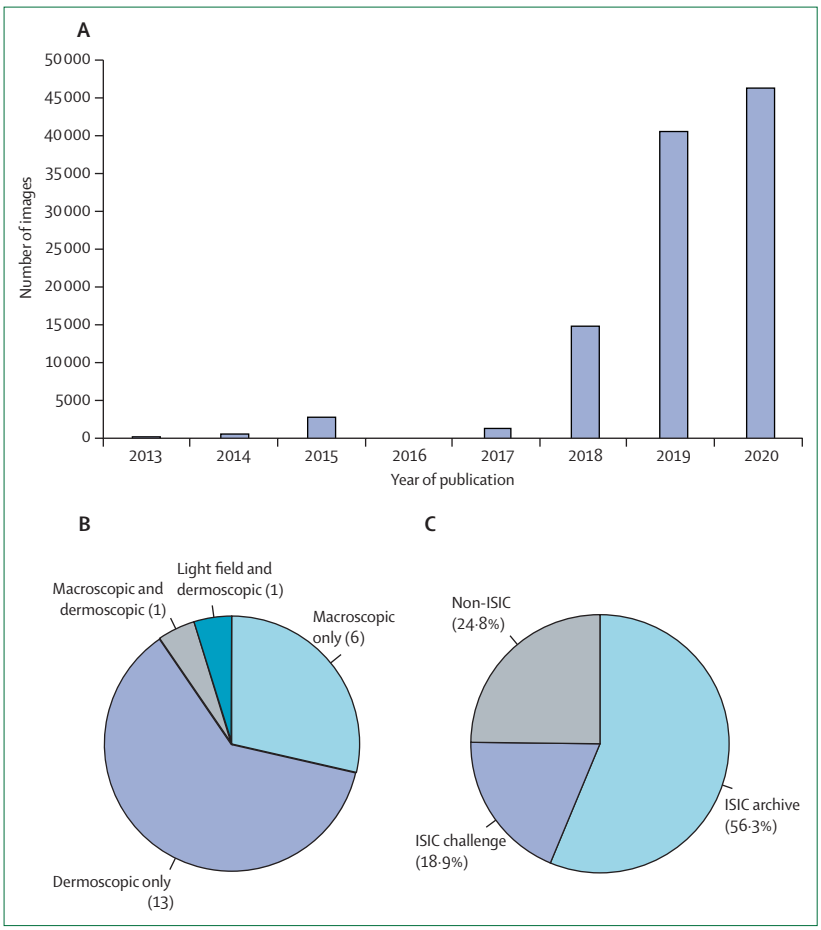

Figure 2: **Image publication year, imaging modality, and image source of open access datasets**
(A) Open access dataset skin lesion images according to year of publication. (B) Imaging modality of open access datasets. (C) Images according to dataset source (n=106 950 images). ISIC=International Skin Imaging Collaboration.

viewed and downloaded, and the dataset was examined manually from the accompanying metadata files. For the SNU dataset,[10] only the 240 full-size images were included; thumbnails of other images were excluded.

BCC was the commonest malignancy represented in datasets (6861 images), followed by melanoma (6802 images) and cSCC (873 images). Altogether, BCC, cSCC, and melanoma represented 13·6% of all images.

Reviewing individual image metadata (figure 3B and appendix p 9), clinical information was available for age (81 662 [76·4%] of 106 950 images), sex (82 848 [77·5%]), and body site (79 561 [74·4%]). Ground truth was confirmed through histopathology for 9995 (68·8%) of 14 536 malignant lesion images, and 21 857 (20·4%) of all images.

Patient ethnicity data were available for 1415 images (1·3% of all images), and Fitzpatrick skin type data for 2236 (2·1%). Skin type and ethnicity data associated with images is shown in the appendix (p 10) for datasets for which metadata labels were available. Aggregate data for ethnicity were available in one dataset containing 170 images (White participants only),[47] and skin type in another dataset that contained 200 images (all images were Fitzpatrick skin type II or III).[15] Aggregate metadata for the larger SNU dataset were given for age, sex, and ethnicity,[10] but further detail was not provided for the subset of downloadable 240 full-size images.

17 open access atlases were identified from the search (figure 1 and appendix p 11). These were mostly hosted by countries in Europe (seven), North America (six), and Oceania (two). One atlas was hosted from South America and one from Japan. 12 atlases contained macroscopic photographs only, one contained dermoscopic images only, and the remaining four contained both dermoscopic and macroscopic photographs, with three containing paired dermoscopic and macroscopic images. Regarding metadata, images were either presented alone (six of 17), with a caption containing variable clinical or educational material (three), with site metadata in the webpage name (three), with structured metadata labels (three), or both structured metadata and captions (two).

Four of 17 atlases were imported into the Databiology platform. Barriers to importing images and metadata from the remaining 13 atlases included images being spread across multiple webpages, aged websites, and presentation in non-English languages.

There were 1407 images from four atlases imported into the Databiology platform, comprising 771 BCC, 206 cSCC, and 430 melanoma images. Of these, 0% had associated metadata for age, 0% for sex, 54·4% (765 images) for site, 0% for ethnicity, and 0% for Fitzpatrick skin type. A few atlases included watermarks, some of which were in close proximity to the lesion;[62] this can affect algorithm performance by making the lesion more difficult to assess and by causing the algorithm to associate the watermark with the lesion.

Discussion

In this first review of publicly available dermatology datasets for skin cancer, 106 950 images from 21 open access datasets were identified, along with 17 open access atlases that are freely accessible to researchers and the public. Although this represents a rich data resource for innovation, lack of transparency in metadata reporting for clinically essential characteristics (such as ethnicity and Fitzpatrick skin type) limits the clinical utility of these images alone. These issues are not limited to dermatology datasets, but have also been reported in ophthalmology[14] and radiology.[63]

Additionally, we identified eight regulated access datasets containing at least 63 687 images, and three regulated access atlases. We also found that images are, in some cases, reused between datasets. Difficulties were encountered when importing images for the majority of atlases, with implications for usability.

Reporting of metadata was limited in many datasets and atlases, which, if used for training or validating machine learning algorithms, would have implications

for generalisability. Machine learning algorithms used for medical image classification are known to underperform on images collected from populations independent to those on which the algorithms were trained. An image classifier algorithm trained and validated predominantly on images of east Asian skin underperformed on skin lesion images of White patients from the USA.[29] Likewise, the majority of algorithms submitted to the ISIC 2018 task 3 challenge performed worse on test images from an external institution independent from the training dataset.[26] A limited number of studies of machine learning algorithms for medical image classification include validation with external datasets, highlighted by two recent systematic reviews which found that only 18–36% of studies validated their algorithms using externally sourced images, meaning the reported accuracy of algorithms might be generally overestimated.[3,64]

These findings highlight the dangers of implementing algorithms for widespread use on broad populations without dataset transparency, especially if algorithm training was undertaken using a restricted demographic cohort. Algorithm underperformance and misdiagnosis have serious implications for patients with skin cancer; they not only risk missing treatable malignancies, but can also result in avoidable surgical procedures and cause unnecessary anxiety.[65] Our review identified limited metadata reporting for datasets and atlases, therefore raising concerns about which populations are represented and to what extent any artificial intelligence algorithms developed using these are generalisable.

Limited metadata reporting could be due to the substantial time, effort, and approvals required to collect this information, but might also be because of the current lack of consensus and guidance regarding which metadata items are most important to report for skin image datasets. This is reflected in the large variability of metadata reporting between datasets, highlighting the need for consensus standards.

Although consensus recommendations for the development of imaging biomarkers in cancer and checklists aiming to improve algorithm development transparency have been proposed that highlight the importance of cohort characteristics and reproducibility to target populations,[66,67] to the best of our knowledge, no consensus standards or guidelines specifically exist for the publication of skin cancer image datasets that have the potential to be used in machine learning. Nevertheless, it is encouraging to note that organisations, such as ISIC, are exploring this area through formation of a metadata and DICOM working group, and have recently published a DICOM supplement for dermoscopy images.[68] In our opinion, the utility of atlases for clinical machine learning algorithm development might be limited, as we identified only five atlases that included systematic metadata (excluding the three atlases that contained site metadata only in the webpage name).

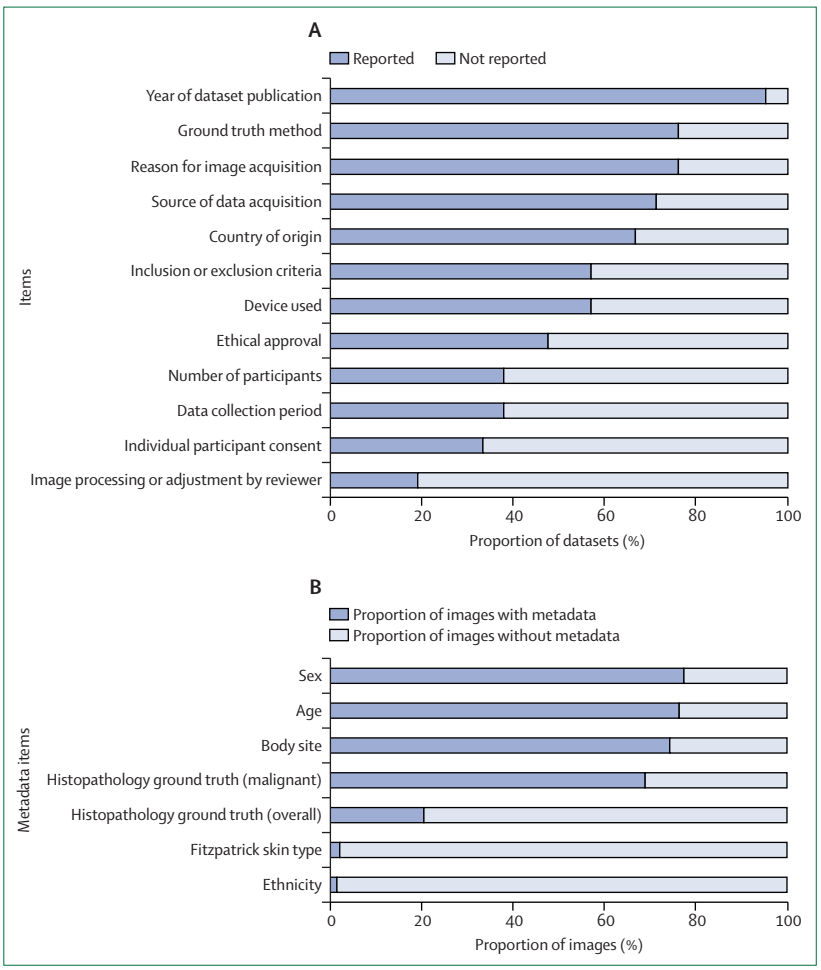

Figure 3: **Characteristic and metadata reporting of open access datasets and images**
(A) Proportion of open access skin image datasets (n=21) reporting general characteristics. (B) Proportion of open access skin images (n=106 950 images in 21 datasets) reporting image and subject metadata.

Relevant details from captions are also difficult to extract in an automated fashion.

To improve algorithm applicability, algorithms should be developed using metadata-rich datasets that accurately describe the skin image populations, with clearly stated inclusion and exclusion criteria. Ideally, datasets should be representative of and reflect the intended population where the algorithm will be deployed, to maximise generalisable performance.[69] Retrospectively collected images are frequently highly selected and might not represent the patient population to which they would be applied. In our review, inclusion and exclusion criteria were reported in only 57% of datasets, and the majority of datasets did not explicitly state obtaining ethical approval (52%) or individual patient consent (67%) for dataset publication. For these datasets, it was not known whether consent was waived, not sought, or not reported. Publication of images without consent can have implications for preservation of privacy. Nevertheless, for retrospectively collected images, consent might not

specifically be required if images were anonymised, for example through exclusion of identifying features.[42,44,47] However, selection bias can occur where specific groups are excluded (eg, facial images or identifiable tattoos). Systematic prospective data collection with associated metadata might be required to reduce selection bias and best represent the cohorts from which the data are acquired.[70] Encouragingly, more recently developed datasets are heterogenous, containing images from various sources, including those not used in the initial training set.[26,42] Nevertheless, merging of individual datasets into larger, composite datasets requires accurate reporting of characteristics and metadata; without this, there is a risk of bias being magnified.

Datasets containing paired images from different modalities better resemble clinical practice and could lead to more complex machine learning algorithms that can assess images of the same lesion from multiple modalities, in addition to demographic and clinical metadata.[38] In our review, datasets containing dermoscopic images only or macroscopic images only were the commonest, accounting for 19 (91%) of 21 datasets and 13 (81%) of 17 atlases. Only one dataset and three atlases contained paired macroscopic and dermoscopic images, and one dataset contained paired light field and dermoscopic images. We did not identify any datasets for other non-invasive imaging techniques to diagnose skin cancer, such as reflectance confocal microscopy, or optical coherence tomography. Light field image sets present interesting opportunities to characterise lesions in three dimensions (and at different depths of field),[48] although specialist equipment such as plenoptic cameras and training might be required for departments acquiring these images.

Health data poverty, defined as "the inability for individuals, groups, or populations to benefit from a discovery or innovation due to a scarcity of data that are adequately representative" is an emerging problem in digital health.[14,36] We found unequal geographical distribution of datasets and atlas origin, with 11 (79%) of 14 datasets and 15 (88%) of 17 atlases originating from Europe, Oceania, and North America exclusively. Only one dataset originated from Asia, two from South America, and none from Africa. In contrast to a recent review of publicly available ophthalmological image datasets, where a large proportion of datasets originated from China,[14] we identified none in this review, highlighting that different specialties might have disease-specific biases.

Fitzpatrick skin type was not only poorly reported but also very poorly representative. Of the 2436 images from three datasets where skin type information was available (either as an aggregate or for individual images),[15,45,48] only ten images were from individuals with Fitzpatrick skin type V, and only a single image was from an individual with Fitzpatrick skin type VI. The ethnicity of these individuals was either Brazilian or unknown. Of the two datasets containing data on ethnicity (1585 images in total),[45,47] no images were from individuals with an African, Afro-Caribbean, or South Asian background. Coupled with the geographical origins of datasets, there was massive under-representation of skin lesion images from darker skinned populations. This is also reflected in dermatology educational resources[71] and images of COVID-19-related rashes published in the literature.[72] This bias is especially concerning as prevalence, presentation, and types of skin cancer vary in skin of colour populations, with associated poorer outcomes,[73,74] and would be even more alarming when detecting skin cancers that preferentially affect patients of skin of colour (eg, Kaposi sarcoma).[75]

Geographical and ethnic disparities in digital health data have similarly been identified in other areas of digital health care where certain populations are under-represented by the amount of digital health data available.[14,36,63] Biased datasets can lead to digital solutions benefitting populations that have advanced data infrastructure that facilitates data collection, but can result in the exclusion or even harm of data-poor subpopulations. Skin cancer machine learning algorithms should be developed using inclusive and representative populations to reduce algorithmic biases. Strategies proposed to achieve this include increasing awareness of health data poverty within the machine learning and digital health communities, improving dataset transparency, investment into routine prospective collection of digital data from health-care systems, and transparent, effective communication to increase inclusion of all population groups.[36]

Our systematic review had several limitations. Although our search strategy incorporated searches from multiple sources, our search was limited to MEDLINE, Google, and Google Dataset Search, and our search terms were in English only. Additionally, seven non-English full-text articles were reviewed using Google Translate. Recent publications that were published but not yet PubMed indexed might have been missed. Similarly, our search was not specifically designed to find atlases, and therefore some atlases might not have been identified. Furthermore, due to variability within and between atlases that were included in this review, it was difficult to quantify their characteristics to the same degree of detail as the datasets. We undertook reasonable efforts to download and import atlas images into our platform, but for some atlases this was not feasible, and we did not pursue this further as we would argue that these atlases might have limited usability given the challenges in downloading and importing images. We attempted to reduce the number of duplicate images in our analyses by excluding datasets for which all images were used in other datasets, and by removing images with the same identifier within the platform. However, it is possible there were some images that overlapped between datasets. Furthermore, we could not assess if there were multiple images taken of the

same lesion (eg, from different views). The number of unique lesions, together with the number of subjects, could be helpful items to include in future open access datasets. Finally, reviews of this nature give a snapshot of the present situation, which can rapidly change without documentation of previous iterations. Datasets and atlases can be created, modified, or deactivated overnight, and in the future we expect more to emerge rapidly. Therefore, regular updates of systematic reviews and similar resources are needed to maintain usefulness to researchers and clinicians.

Future work could evaluate datasets containing images of all dermatoses, including those of rashes.[10] Factors such as positioning, rotation, colour balance, and even presence of surgical skin markings have all been shown to influence machine learning algorithm performance and any variability of these factors in datasets warrants further evaluation.[76–78] This has implications for the use of watermarked atlas images, which might require processing such as cropping prior to use. Furthermore, only 19% of open access datasets in our review stated whether image processing or adjustment by a reviewer occurred. Therefore, in addition to metadata standards, imaging standards are needed to ensure reproducibility of datasets and improve generalisability of the algorithms that they train. This is another area that is currently being explored by an ISIC working group,[11] along with use of DICOM.[79]

To the best of our knowledge, this is the first systematic review of pubicly available skin lesion images comprising predominantly dermoscopic and macroscopic images available through open access datasets and atlases. Key characteristics and metadata were variably reported, with inadequate applicability of datasets to real-life clinical settings, and restricted geographical distribution of datasets globally, limiting generalisability. This review highlights the need for quality standards for minimum characteristics and metadata reporting for skin image datasets. Ensuring inclusion of representative population cohorts in skin image datasets might require prospective image collection using defined criteria. Health data poverty is an under-recognised but fundamental cause of the growing digital health divide and has been demonstrated to be a problem for skin image databases. Ensuring equitable digital health includes building unbiased, representative datasets to ensure that the algorithms that are created benefit people of all backgrounds and skin types.

Contributors

DW, RNM, XL, and SMK were responsible for the conceptualisation of this systematic review. RM and XL devised the methods. DW, AJX, HI, and SMK did the dataset searches. DW, SMK, LS, JC, LZ, and CdBP performed the data extraction. LS, JC, LZ, and CdBP imported and reviewed data using the software. AKD, XL, and RM supervised the project. DW and LS wrote the original draft. SMK, AJX, HI, AKD, XL, and RM reviewed and edited the manuscript. DW, SMK, and LS accessed and verified the underlying data. All authors had access to the data presented in this paper and the appendix prior to submission.

Declaration of interests

LS, JC, LZ, and CdBP were employed by Databiology (Oxford, UK) when this work was completed. All other authors declare no competing interests.

Data sharing

The review protocol is available upon reasonable request to the corresponding author.

Acknowledgments

This report is independent research funded by NHSX and the Health Foundation and it is managed by the National Institute for Health Research (AI_HI200014). The views expressed in this publication are those of the author(s) and not necessarily those of NHSX, the Health Foundation, the National Institute for Health Research, or the Department of Health and Social Care. We would like to thank Adewole Adamson for his critical review of the manuscript.

References

1 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; **380:** 1347–58.
2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.
3 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1:** e271–97.
4 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.
5 Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018; **29:** 1836–42.
6 Brinker TJ, Hekler A, Enk AH, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur J Cancer* 2019; **111:** 148–54.
7 Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol* 2019; **180:** 373–81.
8 Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol* 2019; **20:** 938–47.
9 Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol* 2019; **155:** 58–65.
10 Han SS, Park I, Eun Chang S, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol* 2020; **140:** 1753–61.
11 Reiter O, Rotemberg V, Kose K, Halpern AC. Artificial intelligence in skin cancer. *Curr Dermatol Rep* 2019; **8:** 133–40.
12 Du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. *Br J Dermatol* 2020; **183:** 423–30.
13 International Skin Imaging Collaboration. ISIC Archive. http://www.isic-archive.com/ (accessed March 11, 2021)
14 Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021; **3:** e51–66.
15 Mendonca T, Ferreira PM, Marques JS, Marcal AR, Rozeira J. PH² - a dermoscopic image database for research and benchmarking. *Annu Int Conf IEEE Eng Med Biol Soc* 2013; **2013:** 5437–40.
16 Brinker TJ, Hekler A, Hauschild A, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. *Eur J Cancer* 2019; **111:** 30–37.
17 Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015; **136:** E359–86.

18 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; **71:** 209–49.

19 Donaldson MR, Coldiron BM. No end in sight: the skin cancer epidemic continues. *Semin Cutan Med Surg* 2011; **30:** 3–5.

20 Smittenaar CR, Petersen KA, Stewart K, Moitt N. Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer* 2016; **115:** 1147–55.

21 Godar DE. Worldwide increasing incidences of cutaneous malignant melanoma. *J Skin Cancer* 2011; **2011:** 858425.

22 Trinidad J, Kroshinsky D, Kaffenberger BH, Rojek NW. Telemedicine for inpatient dermatology consultations in response to the COVID-19 pandemic. *J Am Acad Dermatol* 2020; **83:** e69–71.

23 Kennedy J, Arey S, Hopkins Z, et al. Dermatologist perceptions of teledermatology implementation and future use after covid-19: demographics, barriers, and insights. *JAMA Dermatol* 2021; **157:** 595–97.

24 Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol* 2018; **78:** 270–77.e1.

25 Marchetti MA, Liopyris K, Dusza SW, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the International Skin Imaging Collaboration 2017. *J Am Acad Dermatol* 2020; **82:** 622–27.

26 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* 2019; published online Feb 9. https://arxiv.org/abs/1902.03368 (preprint).

27 Hosny KM, Kassem MA, Foaud MM. Classification of skin lesions using transfer learning and augmentation with Alex-net. *PLoS One* 2019; **14:** e0217293.

28 Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML. Artificial intelligence in dermatology: a primer. *J Invest Dermatol* 2020; **140:** 1504–12.

29 Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? *J Invest Dermatol* 2018; **138:** 2277–79.

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

31 Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019; **2:** e191095.

32 Chen IY, Szolovits P, Ghassemi M, Can AI. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics* 2019; **21:** E167–79.

33 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

34 Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. *PMLR* 2018; **81:** 77–91.

35 Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput* 2021; **26:** 232–43.

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

37 Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018; **154:** 1247–48.

38 Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol* 2018; **27:** 1261–67.

39 Pacheco AGC, Krohling RA. The impact of patient clinical information on automated skin cancer detection. *Comput Biol Med* 2020; **116:** 103545.

40 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; **372:** n71.

41 Booth A, Clarke M, Dooley G, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Syst Rev* 2012; **1:** 2.

42 Rotemberg V, Kurtansky N, Betz-Stablein B, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data* 2021; **8:** 34.

43 Combalia M, Codella NCF, Rotemberg V, et al. BCN20000: dermoscopic lesions in the wild. *arXiv* 2019; published online Aug 6. https://arxiv.org/abs/1908.02288 (preprint).

44 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.

45 Pacheco AGC, Lima GR, Salomão AS, et al. PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief* 2020; **32:** 106221.

46 Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J Biomed Health Inform* 2019; **23:** 538–46.

47 Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications* 2015; **42:** 6578–85.

48 de Faria SMM, Henrique M, Filipe JN, et al. Light field image dataset of skin lesions. *Annu Int Conf IEEE Eng Med Biol Soc* 2019; **2019:** 3905–08.

49 Yang J, Wu X, Liang J, et al. Self-paced balance learning for clinical skin disease recognition. *IEEE Trans Neural Netw Learn Syst* 2020; **31:** 2832–46.

50 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:** 1529–38.

51 Edinburgh Innovations. Dermofit image library. https://licensing.edinburgh-innovations.ed.ac.uk/i/software/dermofit-image-library.html (accessed April 13, 2021).

52 Kaggle. IMA205 Challenge. https://www.kaggle.com/c/ima205challenge/overview (accessed Feb 14, 2021).

53 Webster DE, Suver C, Doerr M, et al. The Mole Mapper Study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. *Sci Data* 2017; **4:** 170005.

54 Han SS, Moon IJ, Kim SH, et al. Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: a retrospective validation study. *PLoS Med* 2020; **17:** e1003381.

55 Papadakis M, Paschos A, Manios A, Lehmann P, Manios G, Zirngibl H. Computer-aided clinical image analysis for non-invasive assessment of tumor thickness in cutaneous melanoma. *BMC Res Notes* 2021; **14:** 232.

56 Argenziano G, Soyer HP, De Giorgio V, et al. Interactive atlas of dermoscopy. Milan: Edra Medical Publishing & New Media, 2000.

57 Marghoob AA, Braun RP, Kopf AW. Interactive CD-ROM of Dermoscopy. London: Informa Healthcare, 2007.

58 Menzies SW, Crotty KA, Ingwar C, McCarthy WH. An atlas of surface microscopy of pigmented skin lesions: dermoscopy, 2nd edn. Sydney: McGraw-Hill, 2005.

59 Xie B, He X, Zhao S, et al. XiangyaDerm: a clinical image dataset of Asian race for skin disease aided diagnosis. LABELS 2019, HAL-MICCAI 2019, CuRIOUS 2019; Oct 13 and 17, 2019. https://doi.org/10.1007/978-3-030-33642-4_3.

60 Soenksen LR, Kassis T, Conover ST, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci Transl Med* 2021; **13:** eabb3652.

61 Sun X, Yang J, Sun M, Wang K. A benchmark for automatic visual classification of clinical skin disease images. 14th European Conference on Computer Vision; Oct 11–14, 2016. https://doi.org/10.1007/978-3-319-46466-4_13.

62 Dermnet.com. Dermnet Skin Disease Atlas. http://www.dermnet.com/ (accessed April 11, 2021).

63 Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA* 2020; **324:** 1212–13.

64 Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 2021; **4:** 65.

65 Ferrante di Ruffano L, Takwoingi Y, Dinnes J, et al. Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults. *Cochrane Database Syst Rev* 2018; **12:** CD013186.

66 O'Connor JP, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* 2017; **14:** 169–86.

67 Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020; **26:** 1320–24.

68 DICOM Standards Committee, Working Group 19. Digital Imaging and Communications in Medicine (DICOM) supplement 221: dermoscopy. Nov 15, 2020. https://www.dicomstandard.org/News-dir/ftsup/docs/sups/sup221.pdf (accessed June 20, 2021).

69 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; **17:** 195.

70 van der Sommen F, de Groof J, Struyvenberg M, et al. Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. *Gut* 2020; **69:** 2035–45.

71 Ebede T, Papier A. Disparities in dermatology educational resources. *J Am Acad Dermatol* 2006; **55:** 687–90.

72 Lester JC, Jia JL, Zhang L, Okoye GA, Linos E. Absence of images of skin of colour in publications of COVID-19 skin manifestations. *Br J Dermatol* 2020; **183:** 593–95.

73 Cormier JN, Xing Y, Ding M, et al. Ethnic differences among patients with cutaneous melanoma. *Arch Intern Med* 2006; **166:** 1907–14.

74 Ward-Peterson M, Acuña JM, Alkhalifah MK, et al. Association between race/ethnicity and survival of melanoma patients in the United States over 3 decades: a secondary analysis of SEER data. *Medicine (Baltimore)* 2016; **95:** e3315.

75 Royse KE, El Chaer F, Amirian S, et al. Disparities in Kaposi sarcoma incidence and survival in the United States: 2000-2013. *PLoS One* 2017; **12:** e0182750.

76 Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol* 2019; **155:** 1135–41.

77 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:** 916–20.

78 Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science* 2019; **363:** 1287–89.

79 Caffery LJ, Clunie D, Curiel-Lewandrowski C, Malvehy J, Soyer HP, Halpern AC. Transforming dermatologic imaging for the digital era: metadata and standards. *J Digit Imaging* 2018; **31:** 568–77.