

# Tools for transparency: From data to development to device

---

Prof Alastair Denniston & Dr Xiaoxuan Liu

AI & Digital Health Research & Policy Group, Birmingham UK



UNIVERSITY OF  
BIRMINGHAM



University Hospitals Birmingham  
NHS Foundation Trust



# Responsible Innovation in AI for Health...



**Dr Xiaoxuan Liu**  
Clinician Scientist  
Group Lead



**Prof Alastair Denniston**  
Clinician Scientist  
Group Lead



**Dr Aditya Kale**  
Medical Doctor, PhD student  
Regulation and Post Market Safety  
Monitoring of AI



**Dr Trystan MacDonald**  
Ophthalmologist, PhD student  
Target Product Profiles for AI in  
Diabetic Eye Screening



**Dr Sonam Vadera**  
Radiologist, AI Fellow  
Medical Algorithmic Audit in Chest  
Imaging and Stroke



**Dr Qasim Malik**  
Paediatrician, AI Fellow  
Medical Algorithmic Audit in  
Skin Cancer and Lung Cancer



**Charlotte Radovanovic**  
Programme Manager



**Dr Jeffrey Hogg**  
Ophthalmologist, PhD student  
Implementation and delivery of  
AI in the NHS



**Dr Jo Palmer**  
Postdoctoral Research Fellow  
STANDING Together



**Dr Elinor Laws**  
Medical Doctor, Research fellow  
STANDING Together  
Gender Bias in Large Language  
Models



**Dr Joe Alderman**  
Anaesthetics/Critical Care Dr, PhD student  
Bias in peri-operative risk  
prediction models  
STANDING Together



**Jaspret Gill**  
Project Research Officer  
STANDING Together

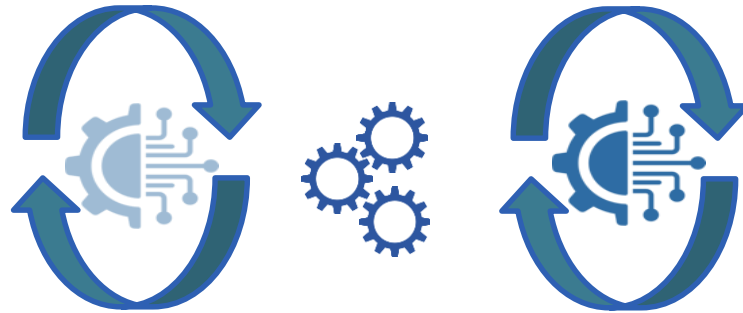
... through translation of scientific evidence into best practice in research,  
policy and regulation

# Tools for transparency

DATA

1010101000  
1011110001  
0101100010

DEVELOPMENT



DEVICE

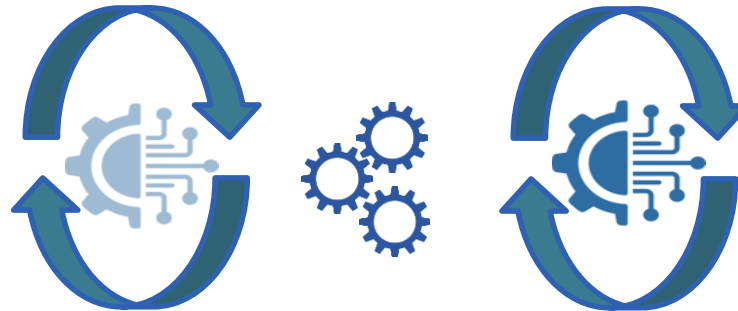


# Tools for transparency

## DATA

1010101000  
1011110001  
0101100010

## DEVELOPMENT



## DEVICE



### Transparency in data

- Datasheets for Datasets
- STANDING Together

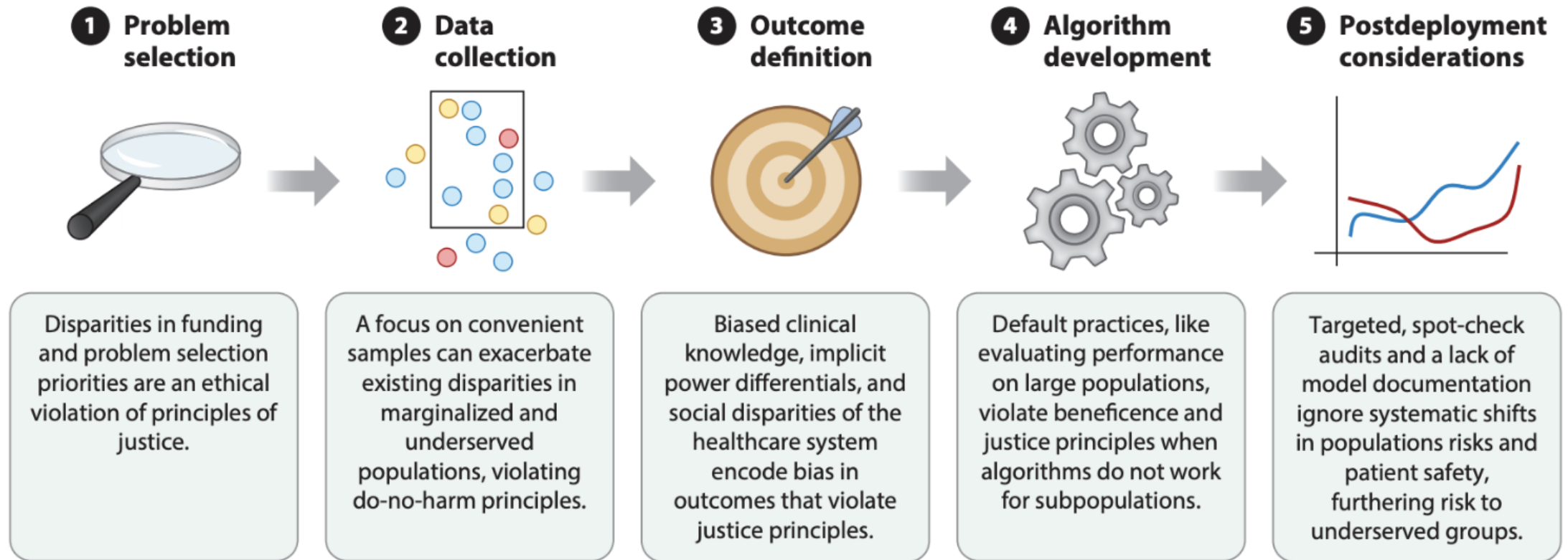
### Transparency in development

- Model cards
- STANDING Together

### Transparency in device performance

- Trial registration
- Reporting guidelines (CONSORT-AI etc)
- Intended use statements
- Medical algorithmic audit
- Local and national sharing of data in the post-market phase.

# Transparency as a tool for addressing risk of bias



# Transparency in data

9. **Users Are Provided Clear, Essential Information:** Users are provided ready access to clear, contextually relevant information that is appropriate for the intended audience (such as health care providers or patients) including: the product's intended use and indications for use, performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model. Users are also made aware of device modifications and updates from real-world performance monitoring, the basis for decision-making when available, and a means to communicate product concerns to the developer.



**Good Machine Learning Practice for Medical Device Development:  
Guiding Principles**

# Tools for reporting data

**Datasheets for Datasets**

[arXiv:1803.09010](https://arxiv.org/abs/1803.09010)

Timnit Gebru<sup>1</sup> Jamie Morgenstern<sup>2</sup> Briana Vecchione<sup>3</sup> Jennifer Wortman Vaughan<sup>1</sup> Hanna Wallach<sup>1</sup>  
Hal Daumé III<sup>14</sup> Kate Crawford<sup>15</sup>



## Healthsheet: Development of a Transparency Artifact for Health Datasets

[Negar Rostamzadeh](#), Google, Canada, [nroostamzadeh@google.com](mailto:nroostamzadeh@google.com)

[Diana Mincu](#), Google, United Kingdom, [dmincu@google.com](mailto:dmincu@google.com)

[Subhrajit Roy](#), Google, United Kingdom, [subhrajitroy@google.com](mailto:subhrajitroy@google.com)

[Andrew Smart](#), Google, USA, [andrewsmart@google.com](mailto:andrewsmart@google.com)

[Lauren Wilcox](#), Google, USA, [lwilcox@google.com](mailto:lwilcox@google.com)

[Mahima Pushkarna](#), Google, Canada, [mahimap@google.com](mailto:mahimap@google.com)

[Jessica Schrouff](#), Google, United Kingdom, [schrouff@google.com](mailto:schrouff@google.com)

[Razvan Amironesei](#), Google, USA, [amironesei@google.com](mailto:amironesei@google.com)

[Nyalleng Moorosi](#), Google, Ghana, [nyalleng@google.com](mailto:nyalleng@google.com)

[Katherine Heller](#), Google, USA, [kheller@google.com](mailto:kheller@google.com)

DOI: <https://doi.org/10.1145/3531146.3533239>

FAccT '22: [2022 ACM Conference on Fairness, Accountability, and Transparency](#), Seoul, Republic of Korea, June 2022



# Tools for reporting data

## Datasheets for Datasets

arXiv:

Timnit Gebru<sup>1</sup> Jamie Morgenstern<sup>2</sup> Briana Vecchione<sup>3</sup> Jennifer Wortman Vaughan<sup>1</sup> Hanna Walla  
Hal Daumé III<sup>14</sup> Kate Crawford<sup>15</sup>



## Healthsheet: Development of a Transparency Artifact for Health Datasets

[Negar Rostamzadeh](#), Google, Canada, [nroostamzadeh@google.com](mailto:nroostamzadeh@google.com)

[Diana Mincu](#), Google, United Kingdom, [dmincu@google.com](mailto:dmincu@google.com)

[Subhrajit Roy](#), Google, United Kingdom, [subhrajitroy@google.com](mailto:subhrajitroy@google.com)

[Andrew Smart](#), Google, USA, [andrewsmart@google.com](mailto:andrewsmart@google.com)

[Lauren Wilcox](#), Google, USA, [lwilcox@google.com](mailto:lwilcox@google.com)

[Mahima Pushkarna](#), Google, Canada, [mahimap@google.com](mailto:mahimap@google.com)

[Jessica Schrouff](#), Google, United Kingdom, [schrouff@google.com](mailto:schrouff@google.com)

[Razvan Amironesei](#), Google, USA, [amironesei@google.com](mailto:amironesei@google.com)

[Nyalleng Moorosi](#), Google, Ghana, [nyalleng@google.com](mailto:nyalleng@google.com)

[Katherine Heller](#), Google, USA, [kheller@google.com](mailto:kheller@google.com)

DOI: <https://doi.org/10.1145/3531146.3533239>

FAccT '22: [2022 ACM Conference on Fairness, Accountability, and Transparency](#), Seoul, South Korea, June 2022

Motivation

Composition

Collection process

Pre-processing/cleaning/labelling

Uses

Distribution

Maintenance



## RESEARCH ARTICLE

## ECONOMICS

## Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer<sup>1,2\*</sup>, Brian Powers<sup>2</sup>, Christine Vogel<sup>4</sup>, Senthil Mullaithan<sup>5\*</sup>

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm for predicting health care costs is considerably sicker than White patients, are considerably sicker than White patients, and are considerably sicker than White patients. Remedying this disparity would increase the help from 17.7 to 46.5%. The bias arises because illness, but unequal access to care means that for White patients. Thus, despite health care by some measures of predictive accuracy, convenient, seemingly effective proxies for bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them. Empirical work is increasingly supporting these concerns. For example, search ads for highly paid positions are more likely to trigger ads for arrest records for Black-sounding names (4), and image searches for professional CEO produce fewer images of women (5), and image searches for professional women more likely to trigger ads for arrest records (6). Facial recognition systems increase in law enforcement performance worse for Black and Hispanic people (7, 8), and natural language processing algorithms encode language in gendered ways. Empirical investigations of algorithms have been hindered by a key barrier: algorithms deployed on large scales are proprietary, making it difficult for researchers to dissect them. Our research seeks to “work from the outside with great ingenuity, and resort to the usual methods as audit studies. So document disparities, but understand why they arise—much less what to do about them—is difficult and greater access to the algorithm. Our understanding of a mechanism typically relies on theory or

<sup>1</sup>School of Public Health, University of Chicago, Chicago, IL, USA. <sup>2</sup>Department of Biostatistics, Harvard Medical School, Boston, MA, USA. <sup>3</sup>Department of Medicine, Brigham Young University, Provo, UT, USA. <sup>4</sup>Morgan Institute for Health Systems Research, Boston, MA, USA. <sup>5</sup>Morgan Institute for Health Systems Research, Boston, MA, USA. \*These authors contributed equally to this work. Corresponding author. Email: zobermeyer@chicago Booth

Obermeyer et al., *Science* 366, 4

that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique, but rather a generalizable ap-

## ARTICLES

https://doi.org/10.1038/s41591-021-01595-0

## nature medicine

Check for updates

## OPEN

## Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari<sup>1,2,3\*</sup>, Haoran Zhang<sup>1</sup>, Matthew B. A. McDermott<sup>2</sup>, Irene Y. Chen<sup>3</sup> and Marzyeh Ghassemi<sup>1,2,3</sup>

Artificial intelligence (AI) systems have increasingly achieved expert-level performance in medical imaging applications. However, there is growing concern that such AI systems may reflect and amplify human bias, and reduce the quality of their performance in historically under-served populations such as female patients, Black patients, or patients of low socioeconomic status. Such biases are especially troubling in the context of underdiagnosis, whereby the AI algorithm would incorrectly label an individual with a disease as healthy, potentially delaying access to care. Here, we examine algorithmic underdiagnosis in chest X-ray pathology classification across three large chest X-ray datasets, as well as one multi-source dataset. We find that classifiers produced using state-of-the-art computer vision techniques consistently and selectively underdiagnosed under-served patient populations and that the underdiagnosis rate was higher for intersectional under-served subpopulations, for example, Hispanic female patients. Deployment of AI systems using medical imaging for disease diagnosis with such biases risks exacerbation of existing care biases and can potentially lead to unequal access to medical treatment, thereby raising ethical concerns for the use of these models in the clinic.

As artificial intelligence (AI) algorithms increasingly affect decision-making in society, researchers have raised concerns about algorithms creating or amplifying biases<sup>1–4</sup>. In this work we define biases as differences in performance against, or in favor of, a subpopulation for a predictive task (for example, different performance on disease diagnosis in Black compared with white patients). Although AI algorithms in specific circumstances can potentially reduce bias<sup>5</sup>, direct application of AI has also been shown to systematize biases in a range of settings<sup>6–11</sup>. This tension is particularly pressing in healthcare, where AI systems could improve patient health<sup>12</sup> but can also exhibit biases<sup>13</sup>. Motivated by the global radiologist shortage<sup>14</sup> as well as by demonstrations that AI algorithms can match specialist performance particularly in medical imaging<sup>15</sup>, AI-based diagnostic tools present a clear incentive for real-world deployment.

Although much work has been done in algorithmic bias<sup>16</sup> and bias in health<sup>17</sup>, the topic of AI-driven underdiagnosis has been relatively unexplored. Crucially, underdiagnosis, defined as falsely claiming that the patient is healthy, leads to no clinical treatment when a patient needs it most, and could be harmful in radiology specifically<sup>18</sup>. Given that automatic screening tools are actively being developed in research<sup>19</sup> and have been shown to match specialist performance<sup>20</sup>, underdiagnosis in AI-based diagnostic algorithms can be a crucial concern if used in the clinical pipeline for patient triage. Triage is an important diagnostic first step in which patients who are falsely diagnosed as healthy are given lower priority for a clinician visit. As a result, the patient will not receive much-needed attention in a timely manner. Underdiagnosis is potentially worse than misdiagnosis, because in the latter case, the patient still receives clinical care, and the clinician can use other symptoms and data sources to clarify the mistake. Initial results have demonstrated

that AI can reduce underdiagnosis in general<sup>21,22</sup> but these studies do not deeply consider the existing clinical biases in underdiagnosis against under-served subpopulations. For example, Black patients tend to be more underdiagnosed in chronic obstructive pulmonary disease than non-Hispanic white patients<sup>23</sup>.

Here, we perform a systematic study of underdiagnosis bias in the AI-based chest X-ray (CXR) prediction models, designed to predict diagnostic labels from X-ray images, in three large public radiology datasets, MIMIC-CXR (CXR)<sup>24</sup>, CheXpert (CXP)<sup>25</sup> and ChestX-ray14 (US National Institutes of Health (NIH))<sup>26</sup>, as well as a multi-source dataset combining all three on shared diseases. We focus our underdiagnosis study on individual and intersectional subgroups spanning race, socioeconomic status (as assessed via the proxy of insurance type), sex and age. The choice of these subgroups is motivated by the clear history, in both traditional medicine and AI algorithms, of bias for subgroups on these axes<sup>27,28</sup>. An illustration of our model pipeline is presented in Fig. 1.

## Results

A standard practice among the AI-based medical image classifiers is to train a model and report the model performance on the overall population regardless of the patient membership to subpopulations<sup>18,19,22</sup>. Motivated by known differences in disease manifestation in patients by sex<sup>29</sup>, age<sup>30</sup>, race/ethnicity<sup>31</sup> and the effect of insurance type in quality of received care<sup>32</sup>, we report results for all of these factors. We use insurance type as an imperfect proxy of socioeconomic status because, for example, patients with Medicaid insurance are often in the low income bracket. Given that binarized predictions are often required for clinical decision-making at the individual level, we define and quantify the underdiagnosis rate based on the binarized model predictions. To assess model decision

## SCIENCE ADVANCES | RESEARCH ARTICLE

## HEALTH AND MEDICINE

## Disparities in dermatology AI performance on a diverse, curated clinical image set

Roxana Daneshjou<sup>1,2\*</sup>, Kailas Vodrahalli<sup>3\*</sup>, Roberto A. Novoa<sup>1,4</sup>, Melissa Jenkins<sup>1</sup>, Weixin Liang<sup>5</sup>, Veronica Rotemberg<sup>6</sup>, Justin Ko<sup>1</sup>, Susan M. Swetter<sup>1</sup>, Elizabeth E. Bailey<sup>1</sup>, Olivier Gevaert<sup>2</sup>, Pritam Mukherjee<sup>2\*</sup>, Michelle Phung<sup>1</sup>, Kiana Yekrang<sup>1</sup>, Bradley Fong<sup>1</sup>, Rachna Sahasrabudhe<sup>1,5</sup>, Johan A. C. Allerup<sup>1</sup>, Utako Okata-Karigane<sup>1</sup>, James Zou<sup>2,3,5,8\*</sup>, Albert S. Chiou<sup>1\*</sup>

An essential

People lack access to dermatological care globally. Artificial intelligence (AI) may aid in triaging skin diseases. However, most AI models have not been assessed on images of diverse skin tones. Thus, we created the Diverse Dermatology Images (DDI) dataset—the first curated, and pathologically confirmed image dataset with diverse skin tones. We show that dermatologists, who often label AI datasets, also perform worse on dark skin tones. These findings identify important weaknesses and biases in dermatology AI models. Fine-tuning AI models on the DDI images closes the performance gap for reliable application to diverse patients and diseases.

People have inadequate access to dermatology care globally. Even in developed countries, such as the United States, there is a shortage and unequal distribution of dermatologists, and long wait times for skin evaluation and decision support tools. Rapid development over the last few years has led to the use of AI to aid nonspecialist physicians in diagnosing potential malignancies (3–5), but cancer detection algorithms with

limited performance in dermatology AI, systematic biases in algorithm performance in this context, we curated the Diverse Dermatology Images (DDI) dataset—a pathologically confirmed benchmark dataset with diverse skin tones. The DDI was retrospectively selected from reviewing histopathology images from 2010 to the present in Stanford Clinics from 2010 to the present, a clinical classification scheme for skin tone, was determined using chart review of in-person visit and consensus review by two board-certified dermatologists. This dataset was designed to allow direct comparison between patients classified as FST V–VI (dark skin tones) and FST I–II (light skin tones) by matching patient characteristics. There were a total of 208 images of FST I–II (159 benign and 49 malignant), 241 images of FST III–IV (167 benign and 74 malignant), and 207 images of FST V–VI (159 benign and 48 malignant) (table S1).

## RESULTS

## Diverse Dermatology Images dataset

To ascertain potential biases in algorithm performance in this context, we curated the Diverse Dermatology Images (DDI) dataset—a pathologically confirmed benchmark dataset with diverse skin tones. The DDI was retrospectively selected from reviewing histopathology images from 2010 to the present in Stanford Clinics from 2010 to the present, a clinical classification scheme for skin tone, was determined using chart review of in-person visit and consensus review by two board-certified dermatologists. This dataset was designed to allow direct comparison between patients classified as FST V–VI (dark skin tones) and FST I–II (light skin tones) by matching patient characteristics. There were a total of 208 images of FST I–II (159 benign and 49 malignant), 241 images of FST III–IV (167 benign and 74 malignant), and 207 images of FST V–VI (159 benign and 48 malignant) (table S1).

## Previously developed dermatology AI algorithms perform worse on dark skin tones and uncommon diseases

We evaluated three algorithms on their ability to distinguish benign versus malignant lesions: ModelDerm [using the application programming interface (API) available at https://modelderm.com/] (12) and two algorithms developed from previously described datasets—DeepDerm (4) and HAM10000 (7). These algorithms were selected

Copyright © 2022 The Authors, some rights reserved; exclusive licensee: American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Downloaded from https://www.science.org at University of Birmingham on February 20, 2023

## Growing evidence of patient harm caused or worsened by AI biases

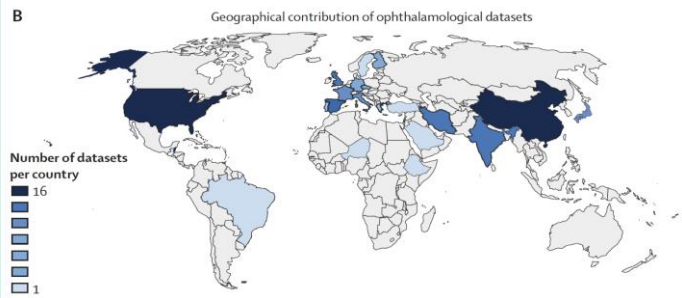
Obermeyer et al (2019)  
Seyyed-Kalantari et al (2021)  
Daneshjou et al (2022)  
& others

# Health data poverty – are you off the map?

THE LANCET  
Digital Health

**A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability**

Saad M Khan\*, Xiaoxuan Liu\*, Siddharth Nath, Edward Karot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, Alastair K Denniston

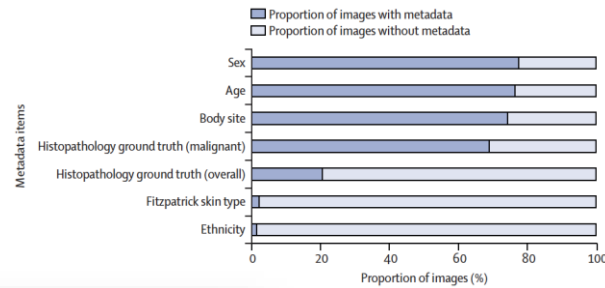


THE LANCET  
Digital Health

**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.



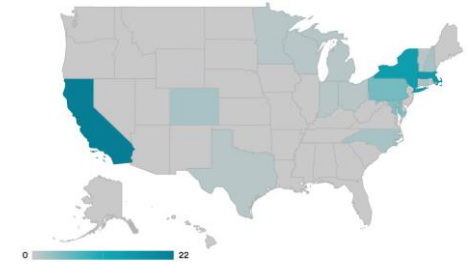
## The Geographic Bias in Medical AI Tools

SHANA LYNCH September 21, 2020

Home / Blog

Patient data from just three states trains most AI diagnostic tools.

SHARE THIS:



REBECCA ROBBINS/STAT  
SOURCE: "GEOGRAPHIC DISTRIBUTION OF US COHORTS USED TO TRAIN DEEP LEARNING ALGORITHMS," JAMA 2020.

STAT

WELL KNIGHT BUSINESS OCT 11, 2020 7:00 AM **WIRED**

**AI Can Help Diagnose Some Illnesses—If Your Country Is Rich**

Algorithms for detecting eye diseases are mostly trained on patients in the US, Europe, and China. This can make the tools ineffective for other racial groups and countries.

f t e

Support the Guardian Available for everyone, funded by readers. Sign in The Guardian

News Opinion Sport Culture Lifestyle

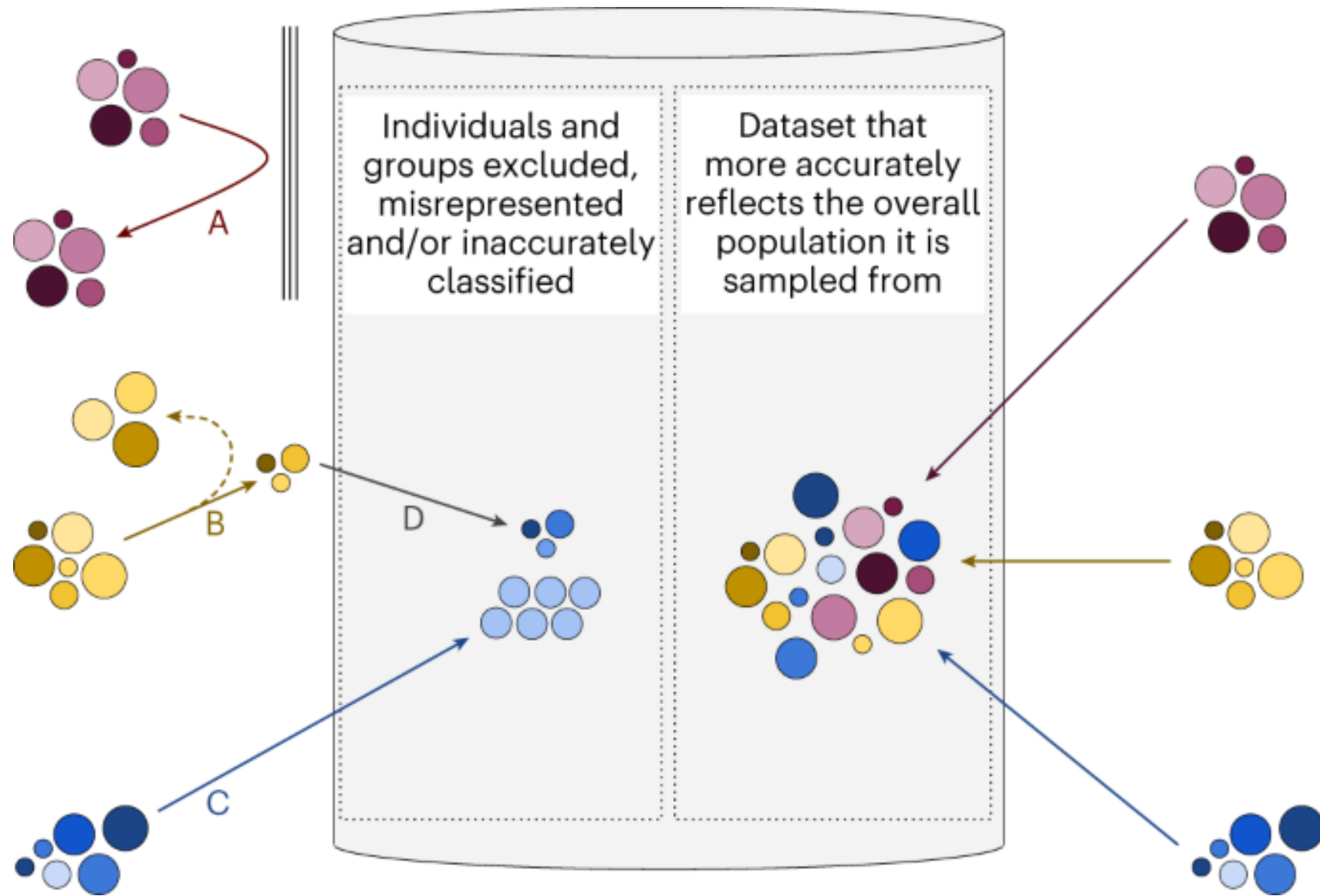
UK World Coronavirus Climate crisis Football Business Environment UK politics More

**Skin cancers**

**AI skin cancer diagnoses risk being less accurate for dark skin - study**

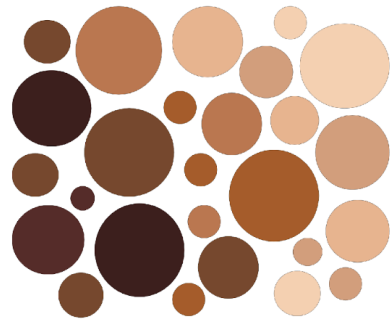
Research finds few image databases available to develop technology contain details on ethnicity or skin type

Studies suggest image recognition technology can classify skin cancers as successfully as humans. Posed by model. Photograph: ChesireCat/Getty Images/Stockphoto





# STANDING Together



Developing standards for data  
Diversity, INclusivity, and  
Generalisability



[www.datadiversity.org](http://www.datadiversity.org)

# Guidance Software and AI as a Medical Device Change Programme - Roadmap

Updated 14 June 2023

## AlaMD for all

This guidance will clarify and expand upon GMLP 3 “Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population”, going beyond the Good machine learning practice mapped guidance. Broadly, this guidance will break down bias in AlaMD into three broad challenges:

- Performance of AlaMD across populations and different real-world conditions
- Ensuring data are properly contextualised to avoid AlaMD perpetuating inequalities or leading to poorer performance in subpopulations
- Working to ensure that AlaMD meets the needs of the communities in which it is deployed in terms of verification and validation.

In addition, with respect to the first challenge, this guidance will provide a high-level framework to identify, measure, manage, and mitigate bias. We will endeavour to work with international partners to advance this work wherever possible.

## WP9-06 Standards Development

### Tools to identify bias

We will assist in the development of standards, frameworks, and tools to assist with the identification and measurement of bias. For example, we will work with the [STANDING Together project](#) which aims to establish standards for data inclusivity and generalisability via an international consensus process to ensure that datasets underpinning AI systems are representative and do not risk leaving underrepresented and minority groups behind through data gaps.

## Good Machine Learning Practice for Medical Device Development: Guiding Principles

October 2021

Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle	Good Software Engineering and Security Practices Are Implemented
Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population	Training Data Sets Are Independent of Test Sets
Selected Reference Datasets Are Based Upon Best Available Methods	Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
Focus Is Placed on the Performance of the Human-AI Team	Testing Demonstrates Device Performance During Clinically Relevant Conditions
Users Are Provided Clear, Essential Information	Deployed Models Are Monitored for Performance and Re-training Risks are Managed




---

## Correspondence

<https://doi.org/10.1038/s41591-022-01987-w>

# Tackling bias in AI health datasets through the STANDING Together initiative

 Check for updates

**To the Editor** – As of June 2022, a wide range of Artificial Intelligence (AI) as a Medical Device (AIaMDs) have received regulatory clearance internationally, with at least 343 devices cleared by the US Food and Drug Administration (FDA)<sup>1</sup>. Despite the enormous potential of AIaMDs, their rapid growth in healthcare has been accompanied by concerns that AI models may learn biases

and prioritize sample size. There are concerns that many health datasets do not adequately represent minority groups; however, the extent of this problem is unknown because many datasets do not provide demographic information, such as on ethnicity and race. Publicly available datasets for skin cancer and eye imaging have shown inconsistent and incomplete demographic reporting, and are disproportionately collected from a small

number of observations and labels were constructed. These concerns have motivated calls for better documentation practices and the creation of tools such as ‘Datasheets for Datasets’ and ‘Healthsheets’<sup>6,9</sup>.

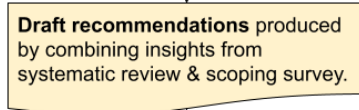
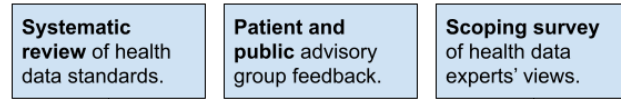
The aforementioned problems are becoming increasingly recognized by regulators of medical devices. In October 2021, The US FDA, Health Canada and the UK Medicines and Healthcare products Regulatory Agency

We will develop standards to promote transparency of bias in health datasets, and mitigate the risk of health inequalities caused by AI medical devices.

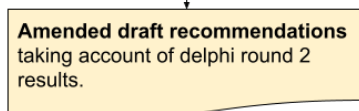
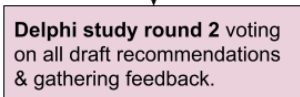
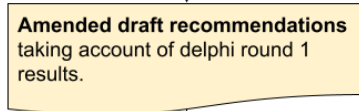
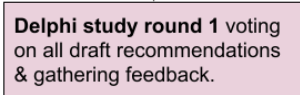


Nov 2021

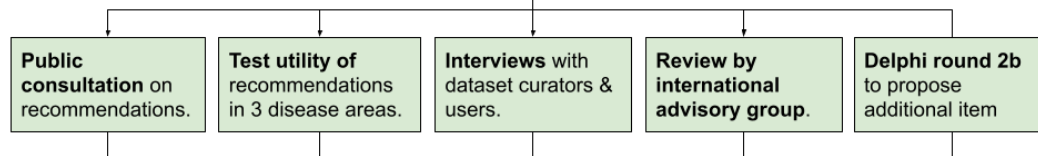
**1 - Produce draft recommendations** by summarising existing knowledge.



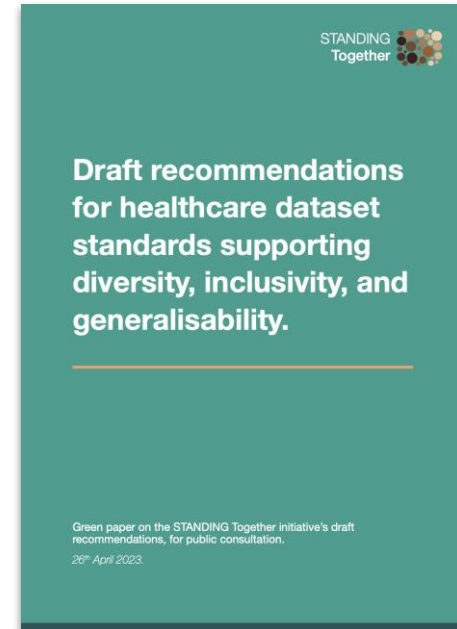
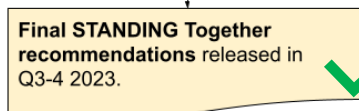
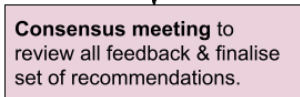
**2 - Amend draft recommendations** based on voting & comments by international, multidisciplinary stakeholders, including patients and public representatives.



**3 - Gather feedback** on recommendations' utility from key stakeholders, including international advisory group. Testing in covid-19, heart failure, and breast cancer datasets.



**4 - Produce final set of recommendations** based on results from steps 2 & 3, ratified by international group of domain experts.



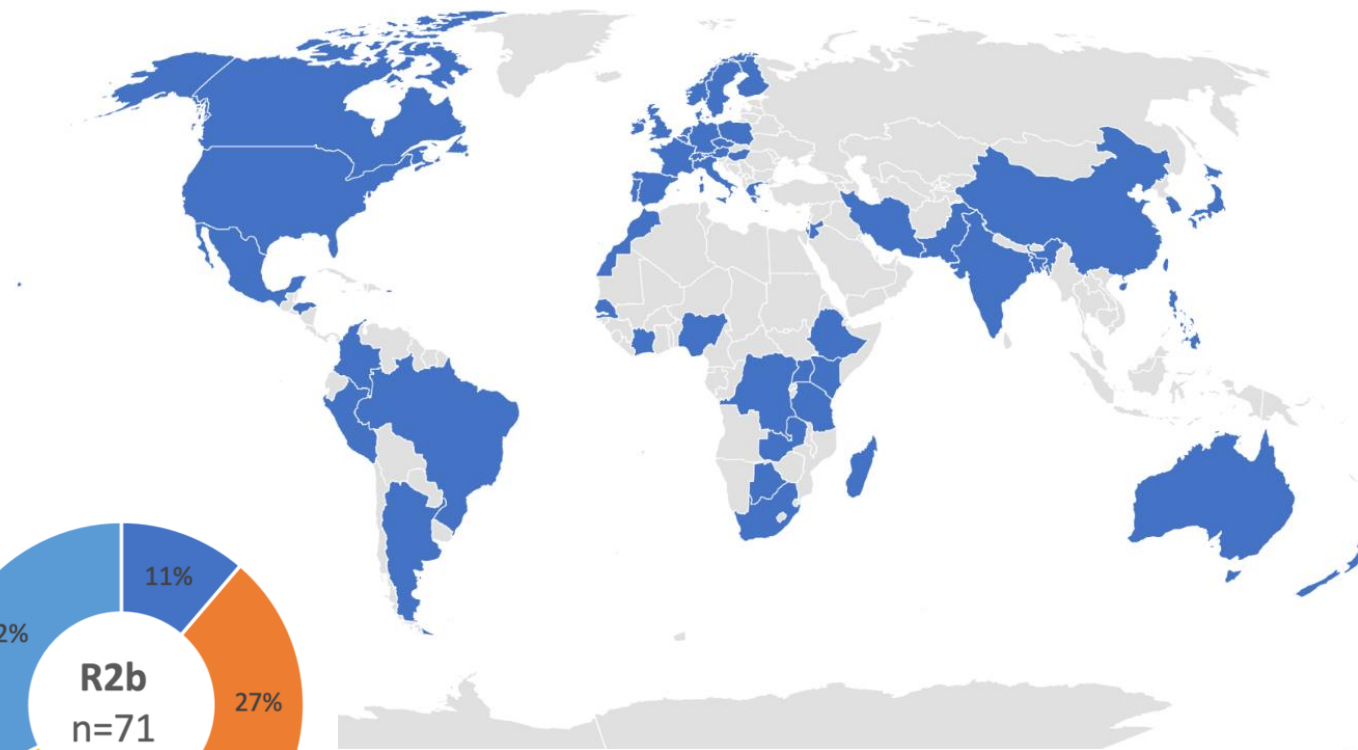
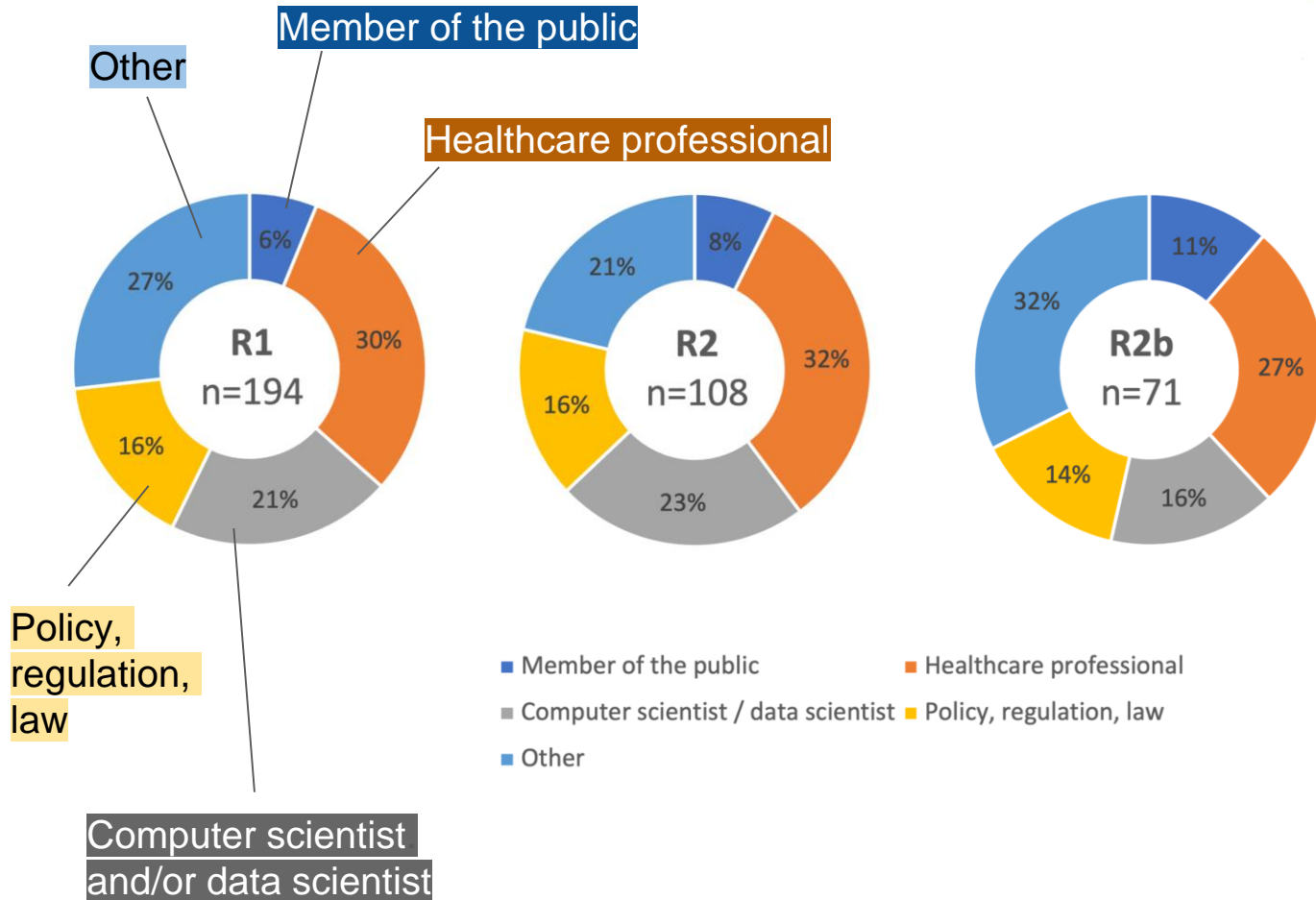
**Funding and Support**





# An International Consensus Study

## Delphi participants



58 countries  
>350 stakeholders

# 1 - Dataset documentation standards

1.1a - Dataset summary

1.1b - Dataset identity and access

1.1c - Reasons behind dataset creation and its purpose(s)

1.1d - Data Origin

1.1e - Data sampling and aggregation from multiple sources

1.1f - Data shifts over time

1.2a - Composition of groups within the dataset

1.2b - Recording of Individual Attributes

1.2c - Groups at risk of disparate health outcomes

1.3a - Limitations of the dataset

1.3b - Modifications made to the data

1.3c - Missing data

1.3d - Known or potential bias caused or exacerbated by data acquisition and processing

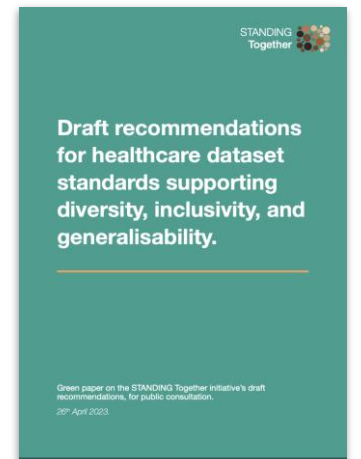
1.3e - Known or potential exclusion introduced by data collection

1.3f - Known or potential bias in assigned or derived Labels

1.4a - Ethics and governance

1.4b - Patient and public participation

1.4c - Bias and impact assessments

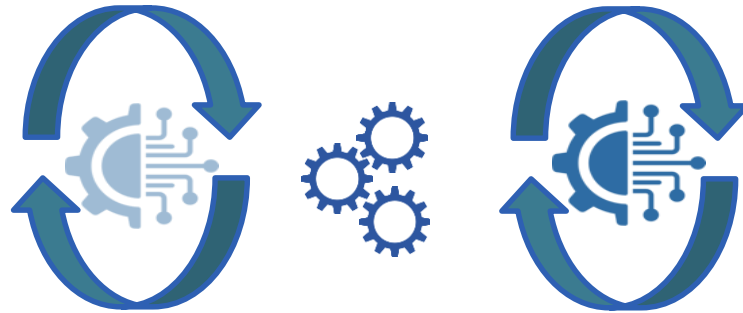


# Transparency in development

DATA

1010101000  
1011110001  
0101100010

DEVELOPMENT



DEVICE



## Transparency in development

- Model cards
- STANDING Together

# Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru  
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com  
deborah.raji@mail.utoronto.ca

[arXiv:1810.03993](https://arxiv.org/abs/1810.03993)

Details

Intended use

Factors

Metrics

Evaluation data

Training data

Quantitative analysis

Ethical considerations

## Model Card - Smiling Detection in Images

### Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

### Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

### Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

### Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of “fairness” in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

### Training Data

- CelebA [36], training data split.

### Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

### Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

### Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

### Quantitative Analyses

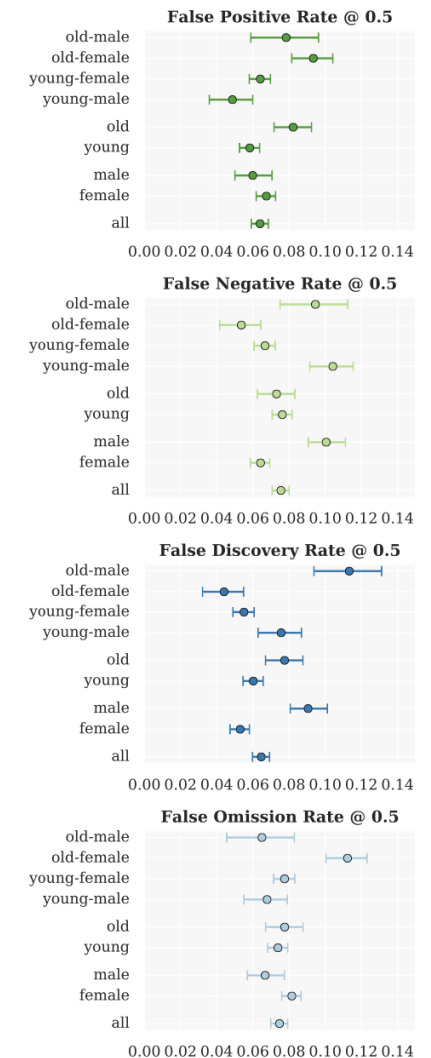


Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

## 1 - Dataset documentation standards

- 1.1a - Dataset summary
- 1.1b - Dataset identity and access
- 1.1c - Reasons behind dataset creation and its purpose(s)
- 1.1d - Data Origin
- 1.1e - Data sampling and aggregation from multiple sources
- 1.1f - Data shifts over time
- 1.2a - Composition of groups within the dataset
- 1.2b - Recording of Individual Attributes
- 1.2c - Groups at risk of disparate health outcomes
- 1.3a - Limitations of the dataset
- 1.3b - Modifications made to the data
- 1.3c - Missing data
- 1.3d - Known or potential bias caused or exacerbated by data acquisition and processing
- 1.3e - Known or potential exclusion introduced by data collection
- 1.3f - Known or potential bias in assigned or derived Labels
- 1.4a - Ethics and governance
- 1.4b - Patient and public participation
- 1.4c - Bias and impact assessments

## 2 - Dataset Use Standards

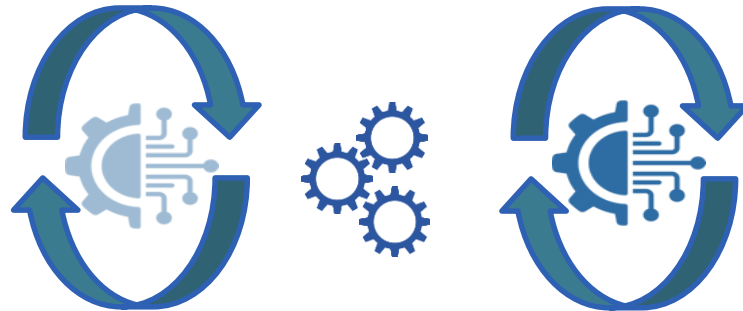
- 2.1a - Provide sufficient information about dataset(s) to allow traceability and auditability
- 2.2a - Identify Contextualised Groups of Interest in advance who may be at risk of disparate performance or harm from the AI health technology
- 2.2c Report the explicit and implicit use of Relevant Attributes during the lifecycle of the AI health technology
- 2.2d - Evaluate performance of the AI health technology for Contextualised Groups of Interest
- 2.2e - Identify disparate performance in any additional groups outside of the pre-specified contextualised groups of interest
- 2.2f Report any approaches or methods (including 'fairness' methods) used to intentionally modify performance across groups.
- 2.3a - Report limitations of datasets used, and any implications on the AI health technology
- 2.3b - Report differences between the intended purposes of the AI health technology and datasets used, including the implications of discordance
- 2.3c - Report findings from pre-existing assessments of the AI health technology and any datasets used
- 2.4a - Address uncertainties and risks with mitigation plans

# Transparency in device performance

DATA

1010101000  
1011110001  
0101100010

DEVELOPMENT



DEVICE



## Transparency in device performance

- Trial registration
- Reporting guidelines (CONSORT-AI etc)
- Intended use statements
- Medical algorithmic audit
- Local and national sharing of data in the post-market phase.

# Trial registration, design and reporting



National Library of Medicine  
National Center for Biotechnology Information

ClinicalTrials.gov

**Focus Your Search**  
(all filters optional)

**Condition/disease** ⓘ

**Other terms** ⓘ

**Intervention/treatment** ⓘ

**Location**

Hide <<

### Search Results

Viewing 1-10 out of 1,285 studies

None Selected [Download] [Bookmark]

● COMPLETED

**NCT05178095**

**Artificial Intelligence in Colonic Polyp Detection**

Conditions

Adenoma Colon    Gastrointestinal Neoplasms    Polyp of Colon

Only 11 of these specifically reference that they are using an FDA-cleared device



# Trial registration, design and reporting

Ensure studies are designed and reported according to best practice. Studies that fail to do this may hide significant bias, which could undermine the results.

Key guidelines are CONSORT-AI for RCTs, SPIRIT-AI for trial protocols and DECIDE-AI for earlier stage studies.<sup>1</sup>

The logo for SPIRIT-AI features the word "SPIRIT" in blue, a red hyphen, and "AI" in blue. A red checkmark is positioned over the "AI" part.

---

Reporting Guidelines for Clinical Trial Protocols for Interventions  
Involving Artificial Intelligence

The logo for CONSORT-AI features the word "CONSORT" in blue, a red hyphen, and "AI" in blue. A red checkmark is positioned over the "AI" part.

---

Reporting Guidelines for Clinical Trial Reports for Interventions  
Involving Artificial Intelligence

[www.clinical-trials.ai](http://www.clinical-trials.ai)

# Intended Use Statements



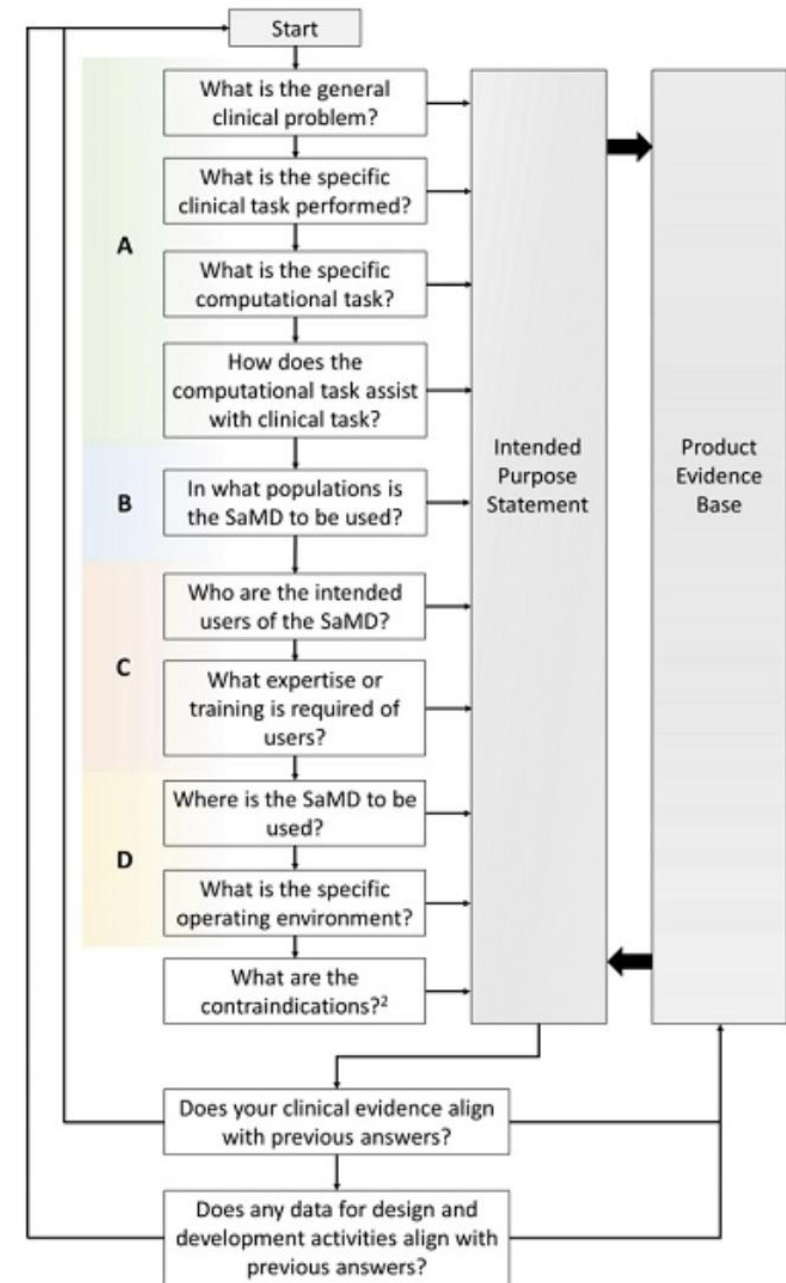
Medicines & Healthcare  
products  
Regulatory Agency

Guidance

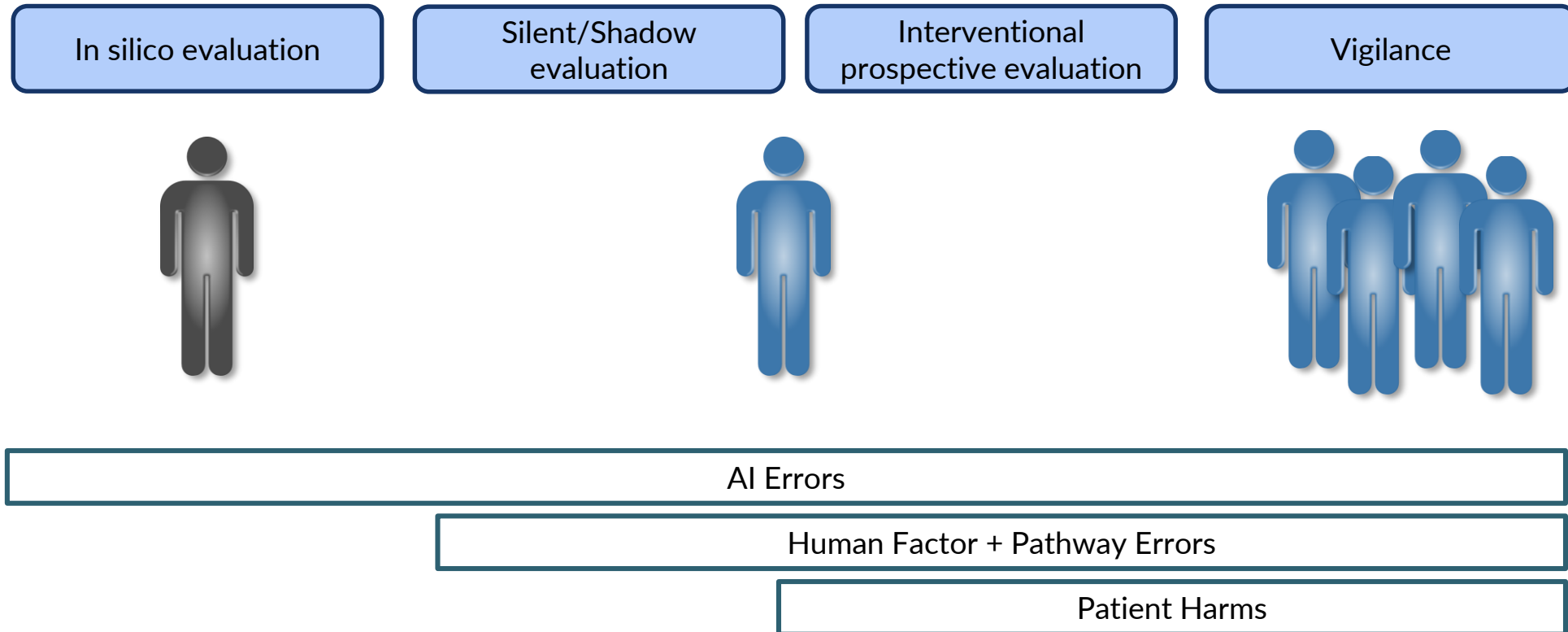
## Crafting an intended purpose in the context of software as a medical device (SaMD)

Published 22 March 2023

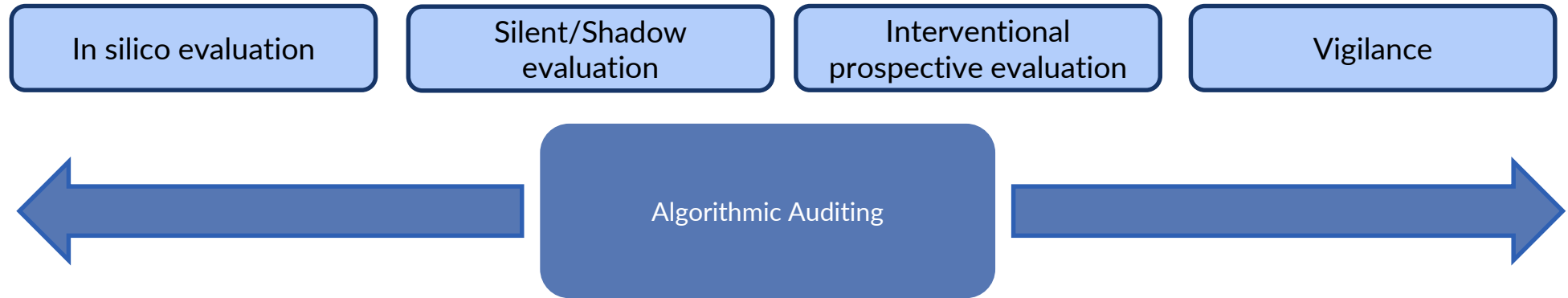
Creating a clear intended purpose is essential for successfully navigating the regulatory requirements for medical devices. In addition, the MHRA encourage manufacturers to maximise the benefits of a clear intended purpose by making this information publicly available. This clarity and transparency can have additional advantages for SaMD when looking to engage with other regulators, distributors, customers and more widely with the UK health and care system.



# Medical algorithmic auditing



# Medical algorithmic auditing

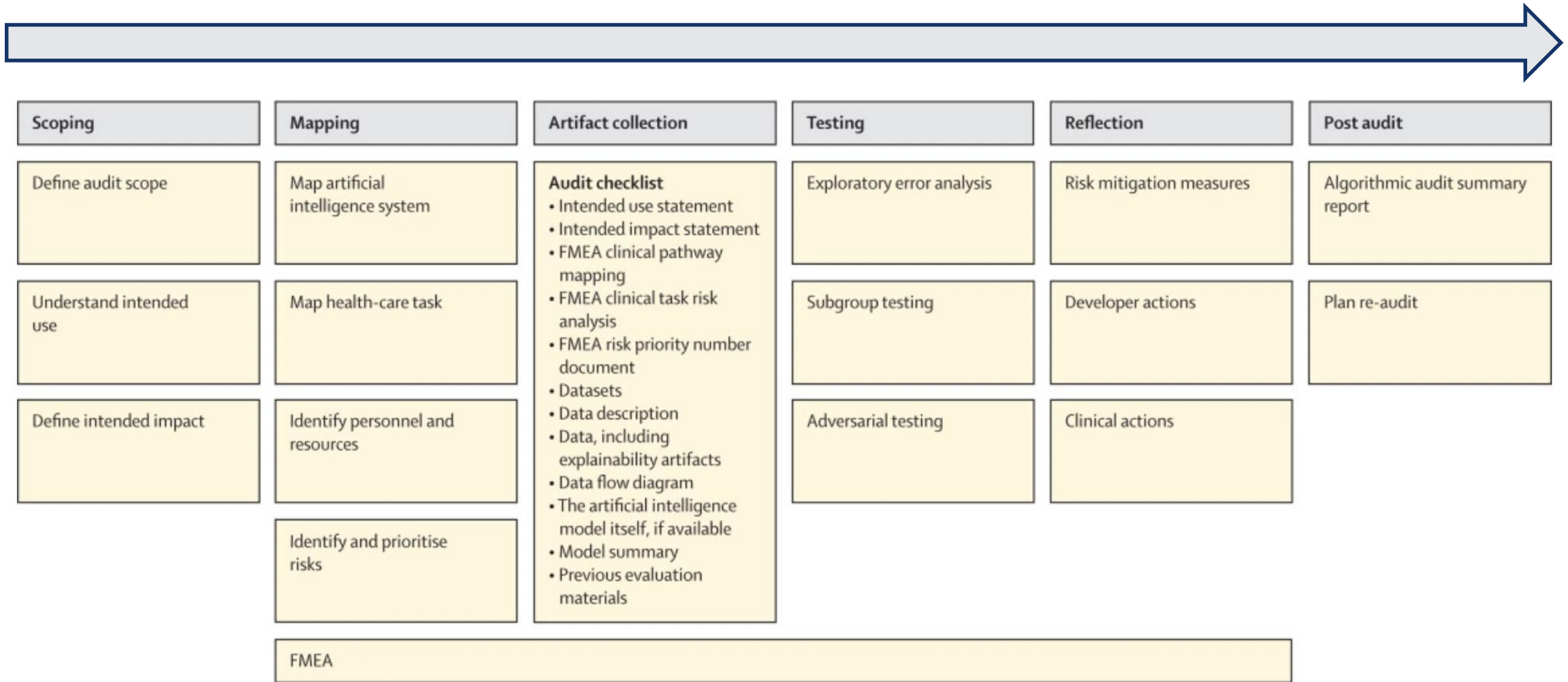


Liu et al. The Medical Algorithmic Audit *Lancet Digital Health* 2022.

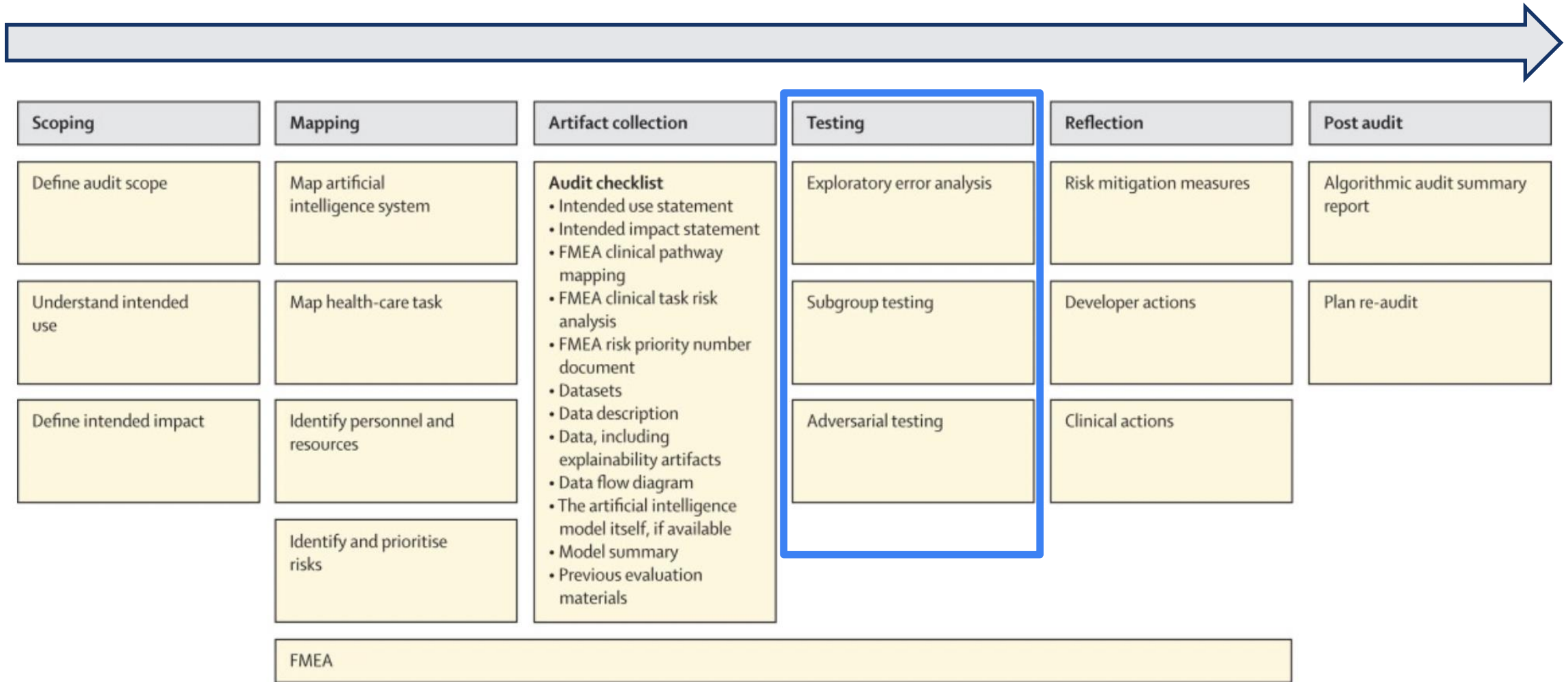
Lauren Oakden-Rayner et al. *Lancet Digital Health* 2022

Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department

# Medical algorithmic auditing



# Medical algorithmic auditing



Manufacturer's performance claims





Manufacturer's performance claims



Overall local performance



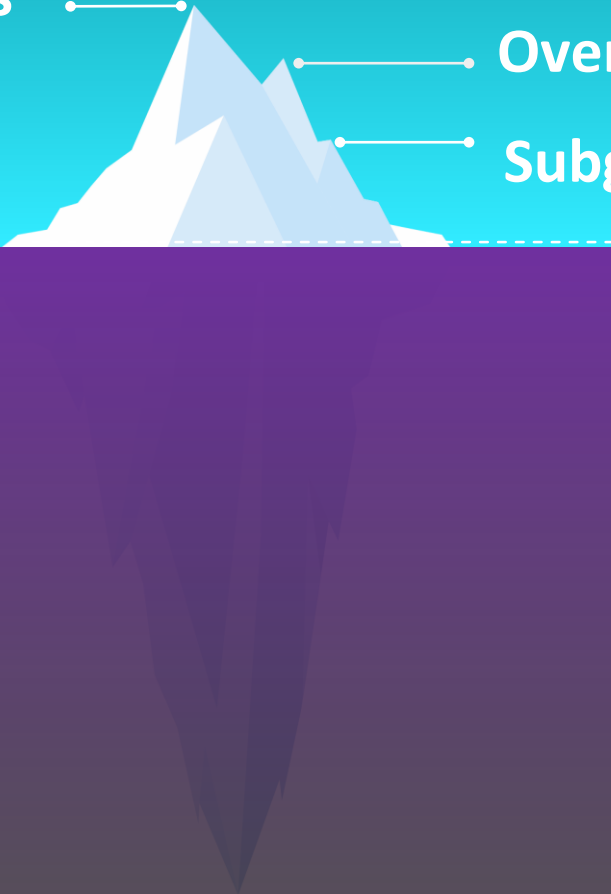
**Manufacturer's performance claims**



**Overall local performance**



**Subgroup performance**



**Manufacturer's performance claims**



**Overall local performance**



**Subgroup performance**



**Unmonitored groups**



# Responsible Innovation in AI for Health

Working together to ensure AI technologies are:  
**safe, effective, equitable and sustainable.**

Translating scientific evidence into best practice in  
research, policy and regulation

