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

NHS
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 Jeffry 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

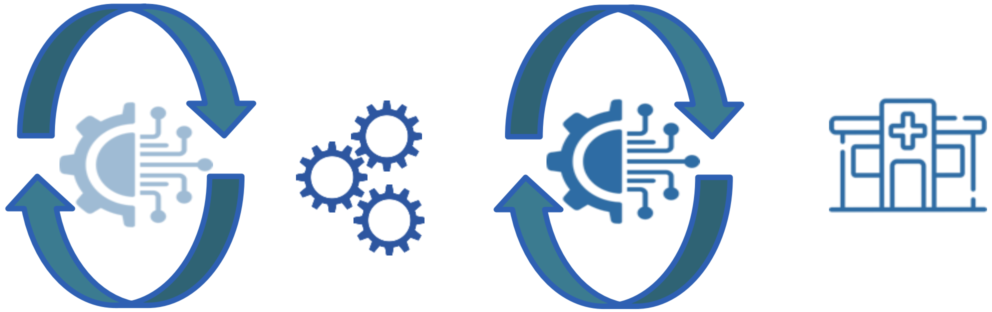
Jaspret Gill
Project Research Officer
STANDING Together

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

Tools for transparency

DATA **DEVELOPMENT** **DEVICE**

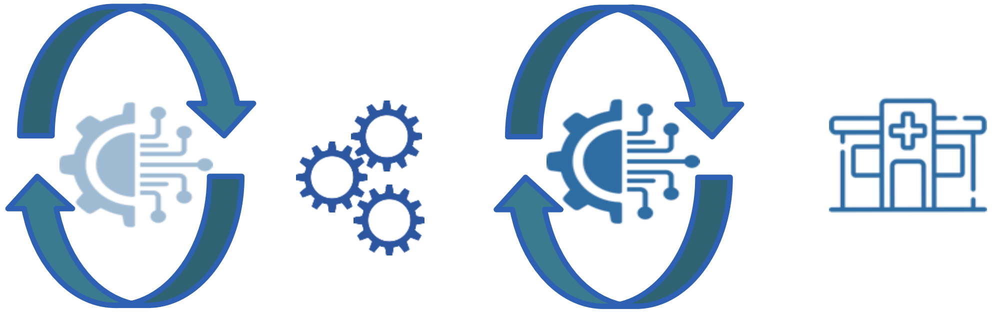
1010101000
1011110001
0101100010



Tools for transparency

DATA **DEVELOPMENT** **DEVICE**

1010101000
1011110001
0101100010



Transparency in data

- Datasheets for Datasets
- STANDING Together

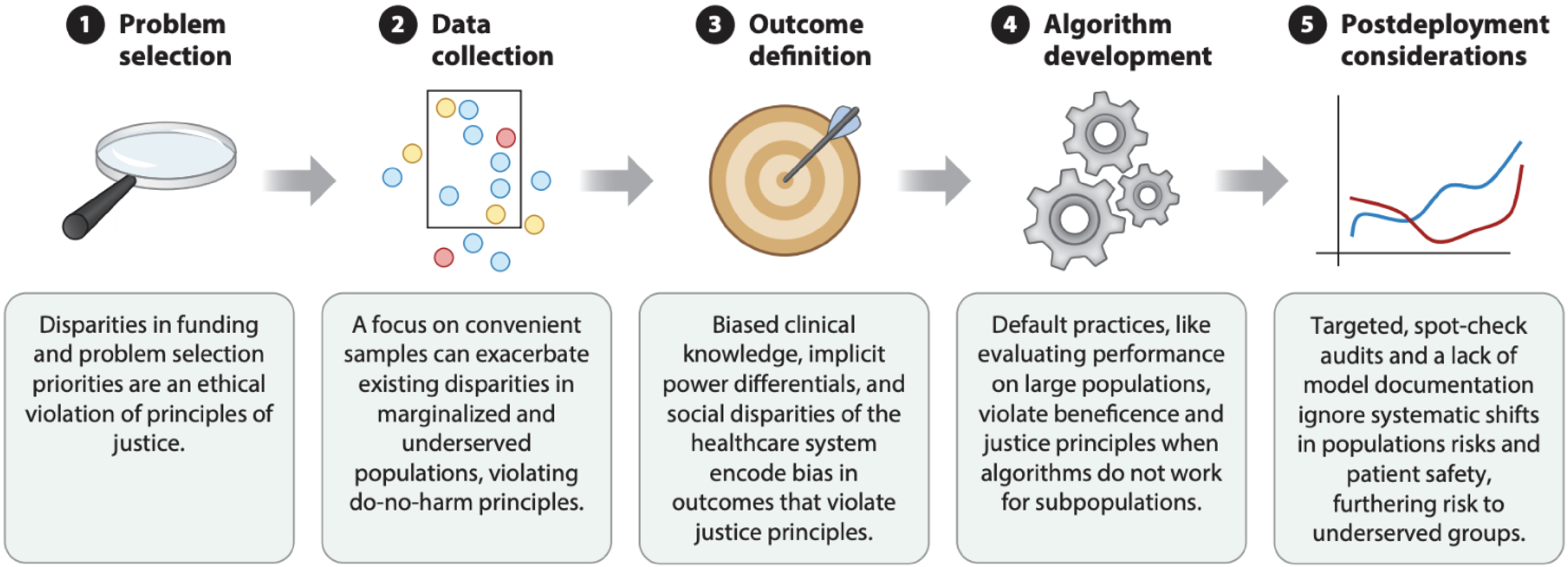
Transparency in development

- 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



Chen IY, Pierson E, Rose, S…Ghassemi M Ethical Machine Learning in Healthcare. Annual Review of Biomedical Data Science 2021 4:1, 123-144

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.

FDA U.S. FOOD & DRUG ADMINISTRATION

Health Canada Santé Canada

Medicines & Healthcare products Regulatory Agency

Good Machine Learning Practice for Medical Device Development: Guiding Principles

Tools for reporting data

Datasheets for Datasets arXiv:1803.09010

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Healthsheet: Development of a Transparency Artifact for Health Datasets

Negar Rostamzadeh, Google, Canada, nrostamzadeh@google.com
Diana Mincu, Google, United Kingdom, dmincu@google.com
Subhrajit Roy, Google, United Kingdom, subhrajitroy@google.com
Andrew Smart, Google, USA, andrewsmart@google.com
Lauren Wilcox, Google, USA, lwilcox@google.com
Mahima Pushkarna, Google, Canada, mahimap@google.com
Jessica Schrouff, Google, United Kingdom, schrouff@google.com
Razvan Amironesei, Google, USA, amironesei@google.com
Nyalleng Moorosi, Google, Ghana, nyalleng@google.com
Katherine Heller, Google, USA, 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 [1] **Jamie Morgenstern** [2] **Briana Vecchione** [3] **Jennifer Wortman Vaughan** [1] **Hanna Walla**
Hal Daumé III [1 4] **Kate Crawford** [1 5]

Healthsheet: Development of a Transparency Artifact for Health Datasets

Negar Rostamzadeh, Google, Canada, nrostamzadeh@google.com
Diana Mincu, Google, United Kingdom, dmincu@google.com
Subhrajit Roy, Google, United Kingdom, subhrajitroy@google.com
Andrew Smart, Google, USA, andrewsmart@google.com
Lauren Wilcox, Google, USA, lwilcox@google.com
Mahima Pushkarna, Google, Canada, mahimap@google.com
Jessica Schrouff, Google, United Kingdom, schrouff@google.com
Razvan Amironesei, Google, USA, amironesei@google.com
Nyalleng Moorosi, Google, Ghana, nyalleng@google.com
Katherine Heller, Google, USA, kheller@google.com

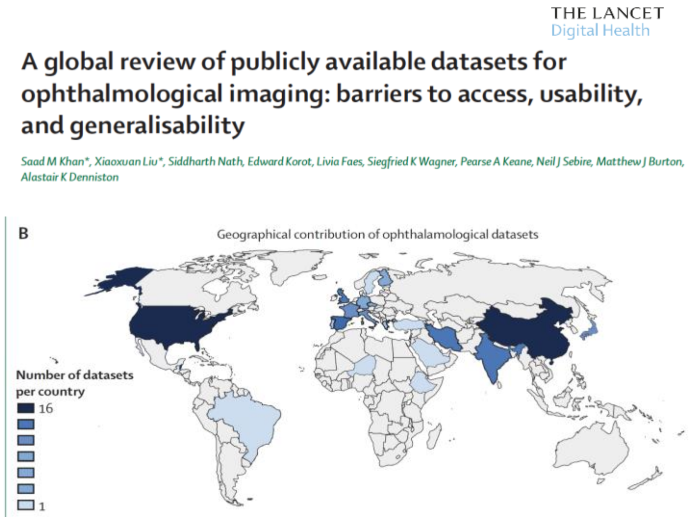
DOI: https://doi.org/10.1145/3531146.3533239
FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, **S**
of Korea, June 2022

- Motivation
- Composition
- Collection process
- Pre-precessing/cleaning/labelling
- Uses
- Distribution
- Maintenance

Growing evidence of patient harm caused or worsened by AI biases

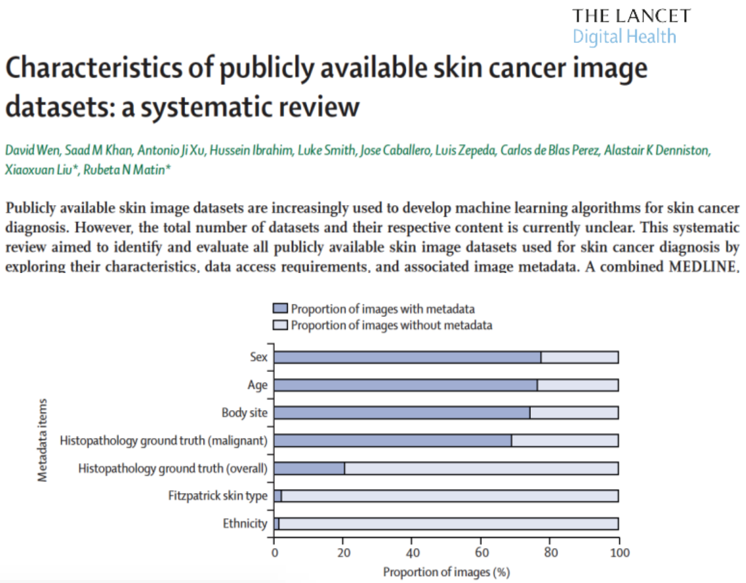
Obermeyer et al (2019)
Seyyed-Kalantari et al (2021)
Daneshjou et al (2022)
& others

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

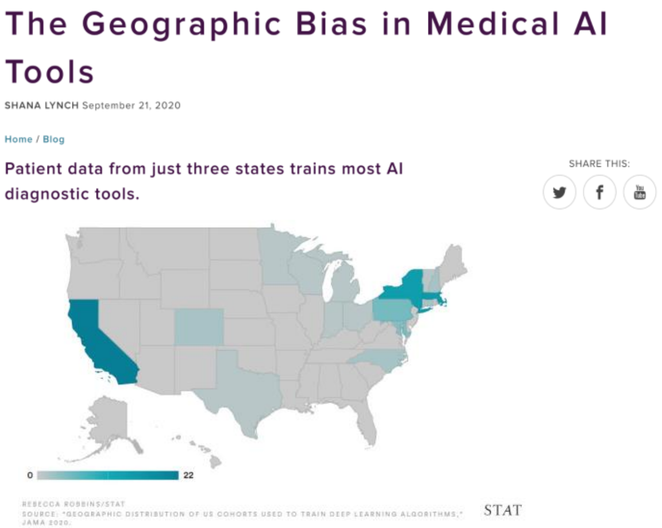
Saad M Khan*, Xiaoxuan Liu*, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, Alastair K Denniston


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,

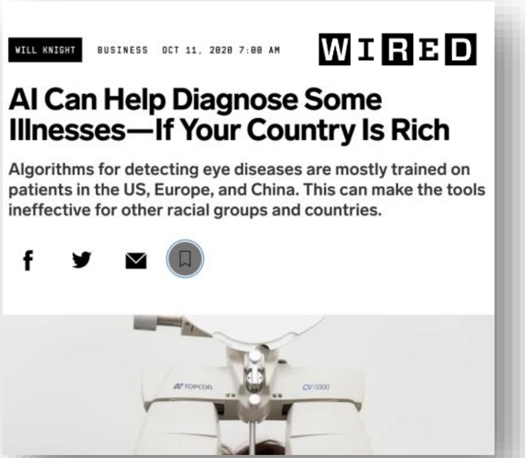

The Geographic Bias in Medical AI Tools

SHANA LYNCH September 21, 2020

Home / Blog

SHARE THIS:

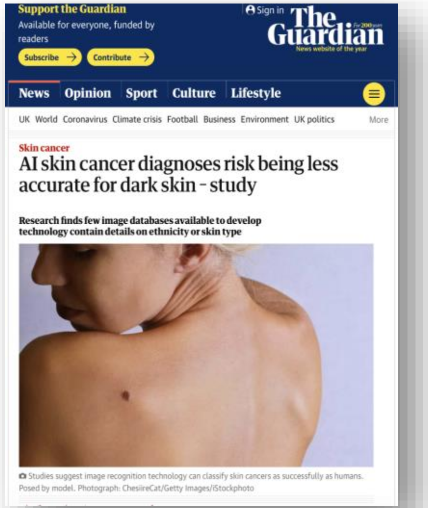
Patient data from just three states trains most AI diagnostic tools.


WILL 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.


Support the Guardian
Available for everyone, funded by readers
Subscribe → Contribute →
Sign in
The Guardian

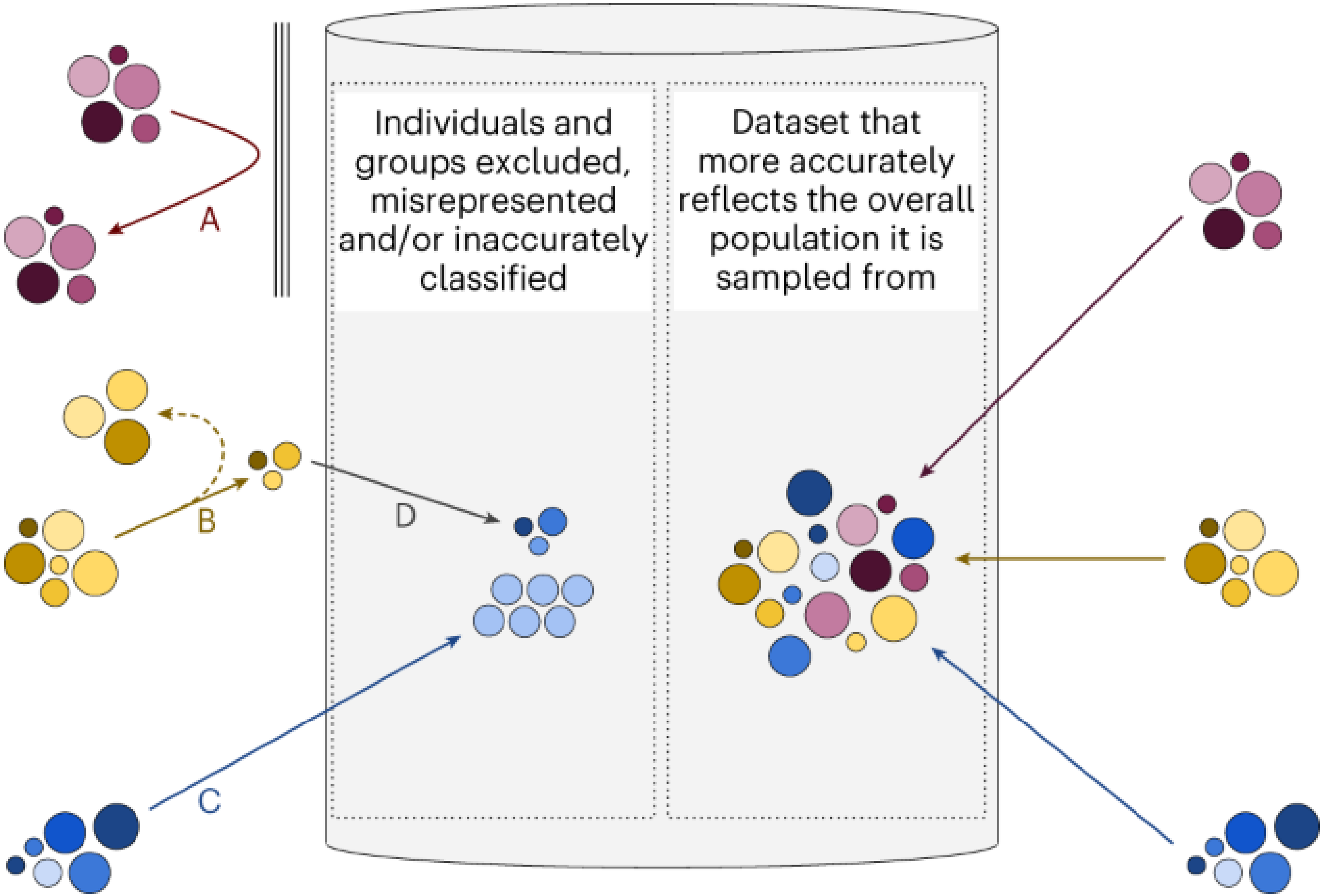
News Opinion Sport Culture Lifestyle

UK World Coronavirus Climate crisis Football Business Environment UK politics More

Skin cancer
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: ChesiireCat/Getty Images/iStockphoto



Arora, A., Alderman, J.E., Palmer, J…Liu X. The value of standards for health datasets in artificial intelligence-based applications. *Nat Med* **29**, 2929–2938 (2023)

STANDING Together

Developing standards for data Diversity, INclusivity, and Generalisability

www.datadiversity.org

The Health Foundation | NIHR National Institute for Health Research | NHS | HDRUK Health Data Research UK | University of Birmingham | NHS University Hospitals Birmingham NHS Foundation Trust | Medicines & Healthcare products Regulatory Agency | wellcome

U.S. FOOD & DRUG ADMINISTRATION | Health Canada Santé Canada | Medicines & Healthcare products Regulatory Agency

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

Medicines & Healthcare products Regulatory Agency

Guidance

Software and AI as a Medical Device Change Programme - Roadmap

Updated 14 June 2023

AIaMD 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 AIaMD into three broad challenges:

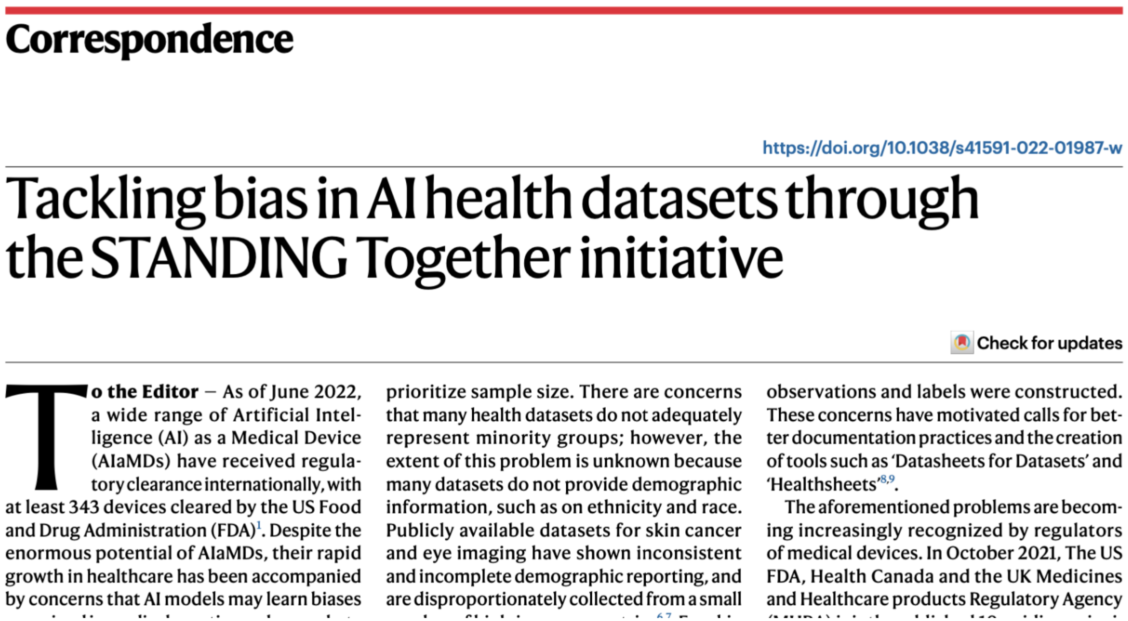
- Performance of AIaMD across populations and different real-world conditions
- Ensuring data are properly contextualised to avoid AIaMD perpetuating inequalities or leading to poorer performance in subpopulations
- Working to ensure that AIaMD meets the needs of the communities in which it is deployed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



Correspondence

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

Tackling bias in AI health datasets through the STANDING Together initiative

Check for updates

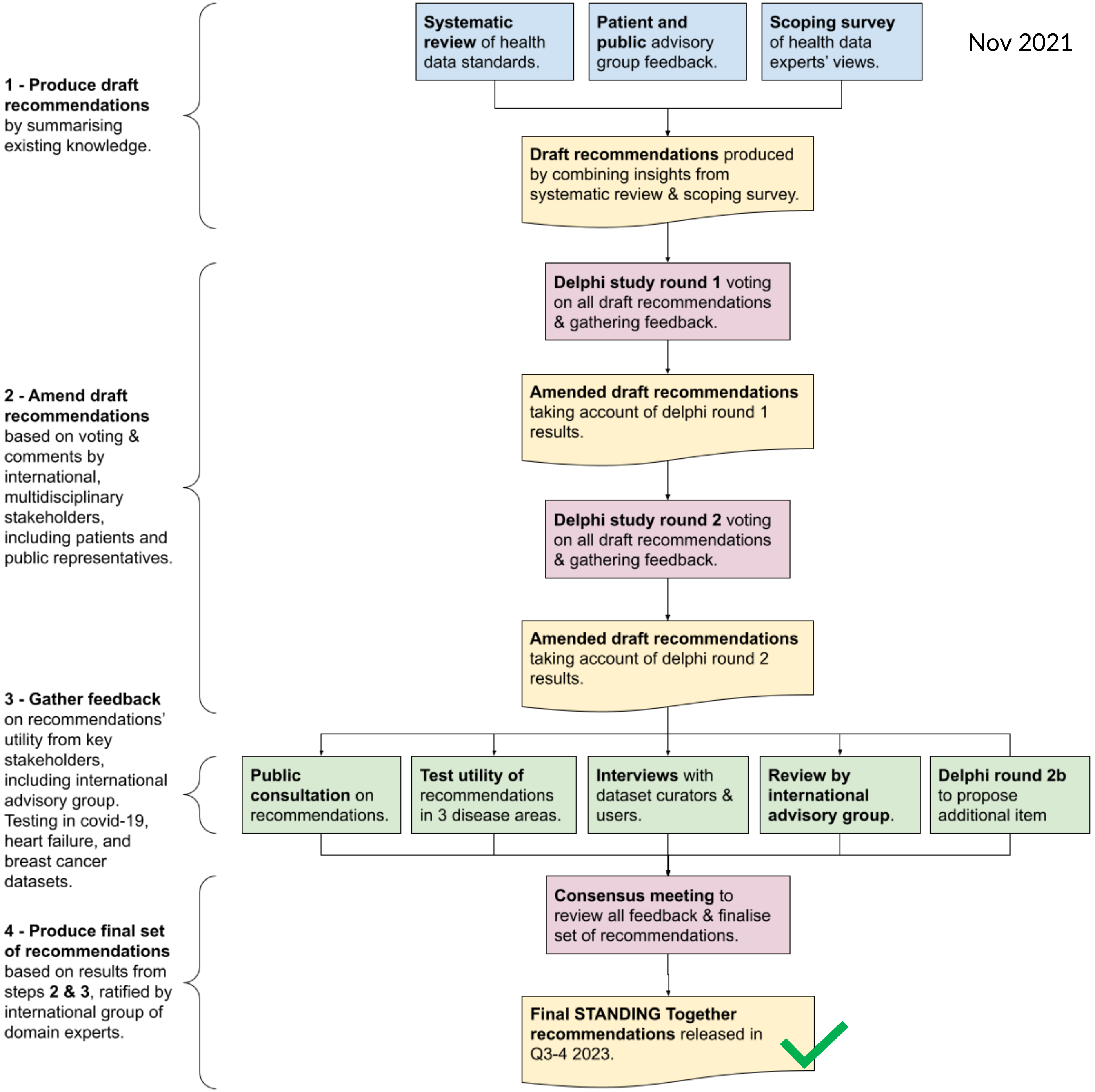
To the Editor — As of June 2022, a wide range of Artificial Intelligence (AI) as a Medical Device (AIaMDs) have received regulatory clearance internationally, with at least 343 devices cleared by the US Food and Drug Administration (FDA)[1]. Despite the enormous potential of AIaMDs, their rapid growth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

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

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'[8,9].

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

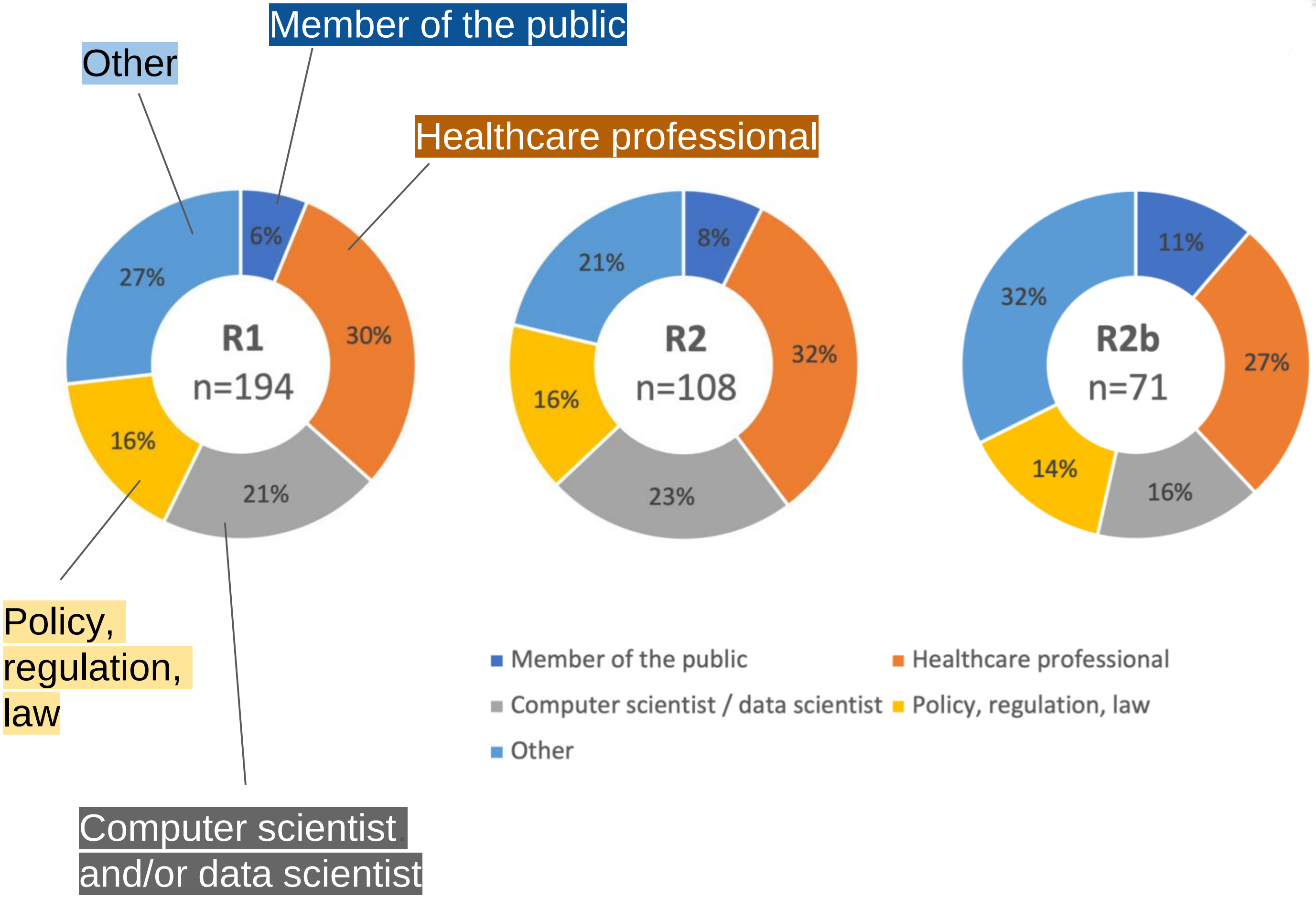
We will develop standards to promote **transparency** of bias in health datasets, and **mitigate the risk of health inequalities** caused by AI medical devices.

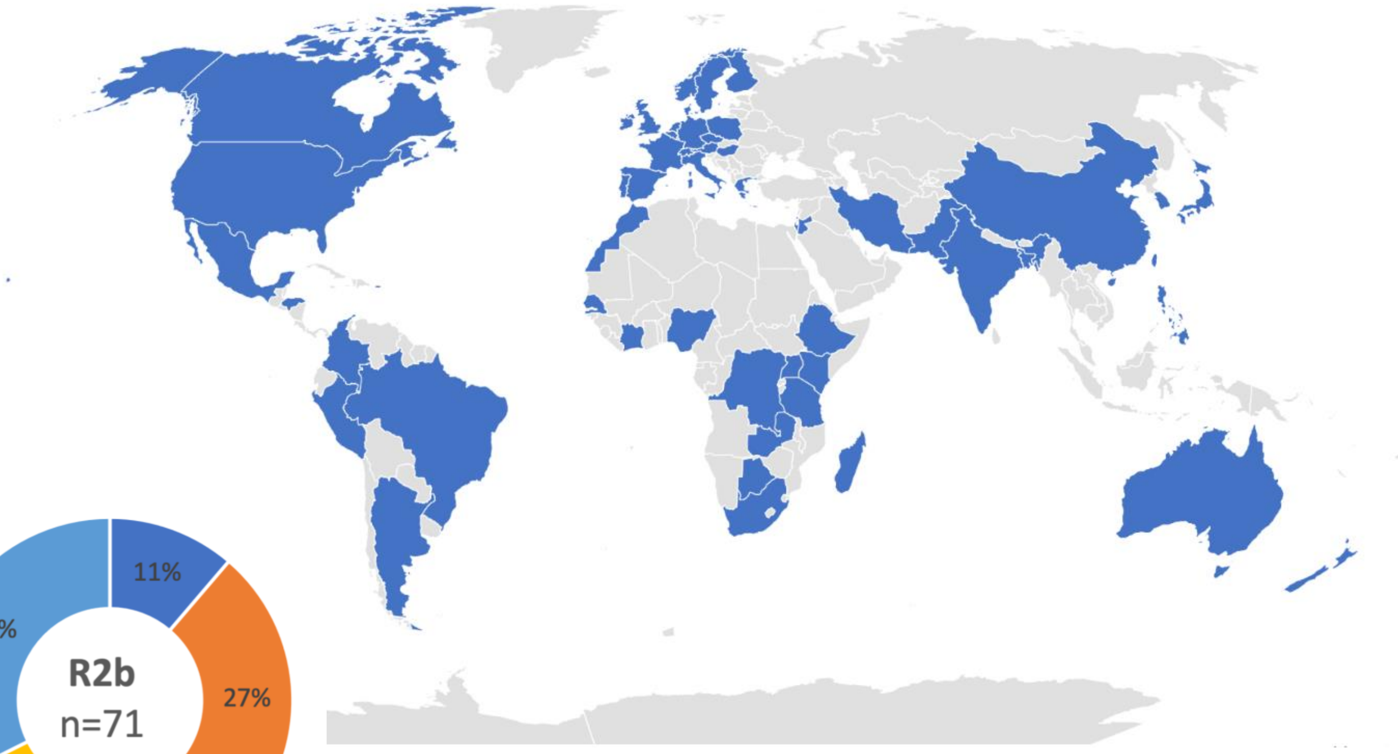


Funding and Support

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

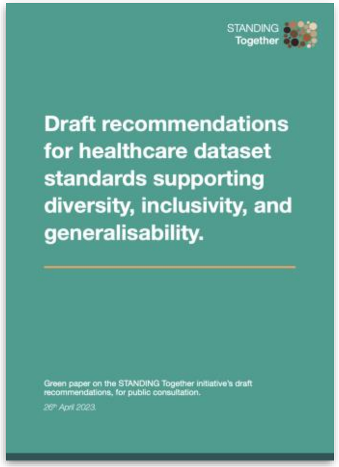
1.3f - Known or potential bias in assigned or derived Labels

1.4a - Ethics and governance

1.4b - Patient and public participation

1.4c - Bias and impact assessments

www.datadiversity.org

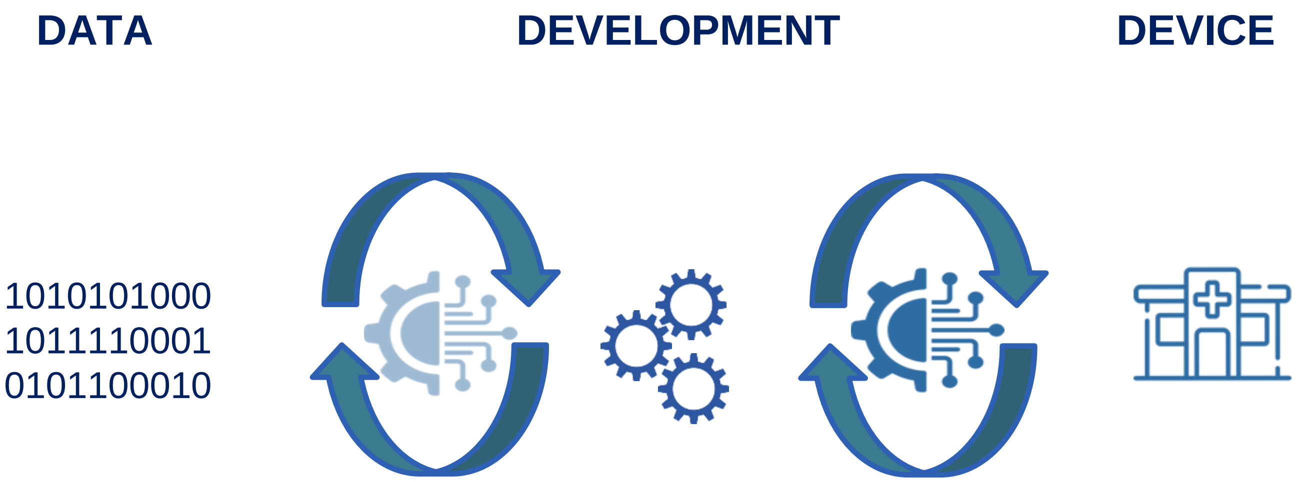

STANDING Together

Draft recommendations for healthcare dataset standards supporting diversity, inclusivity, and generalisability.

Green paper on the STANDING Together initiative's draft recommendations, for public consultation.

26th April 2023.

Transparency in development



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

- 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.

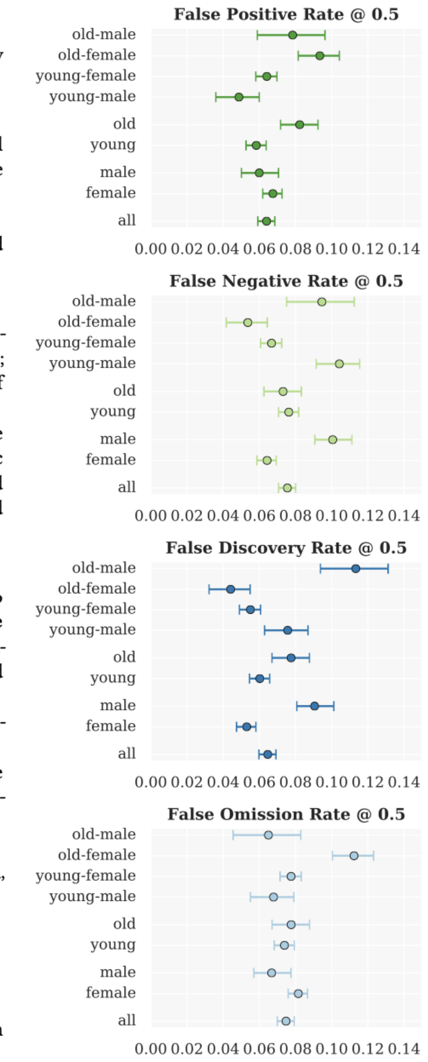
Evaluation Data

- 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.

Quantitative Analyses



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.

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.

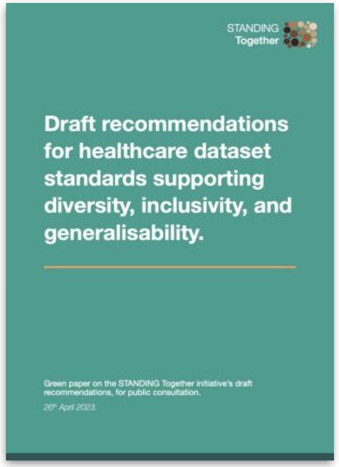
2.3a - Report limitations of datasets used, and any implications on the AI health technology

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

www.datadiversity.org

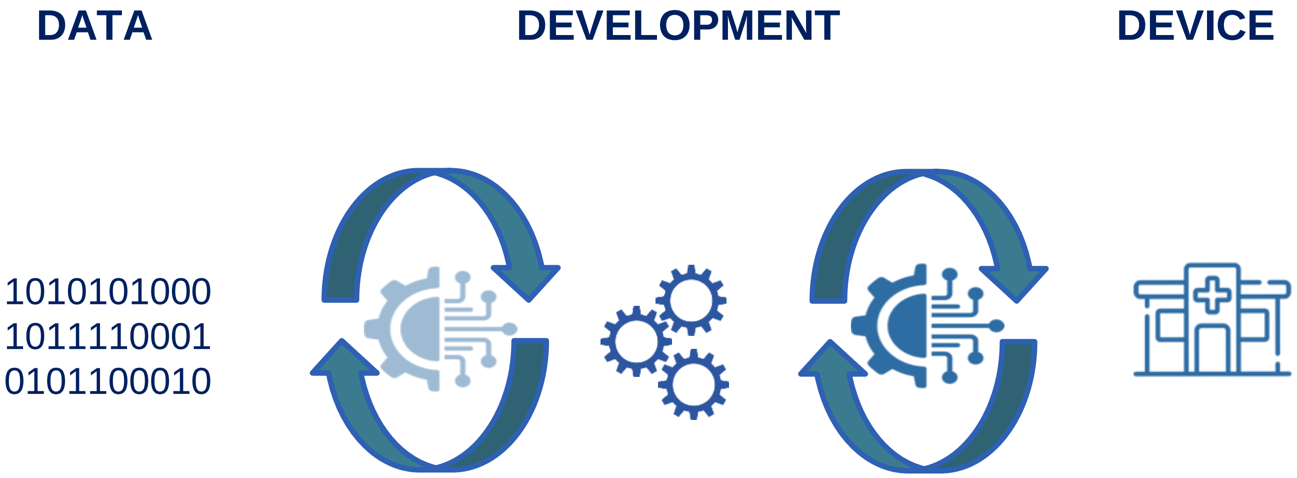

STANDING Together

Draft recommendations for healthcare dataset standards supporting diversity, inclusivity, and generalisability.

Green paper on the STANDING Together initiative's draft recommendations, for public consultation.

26th April 2023.

Transparency in device performance



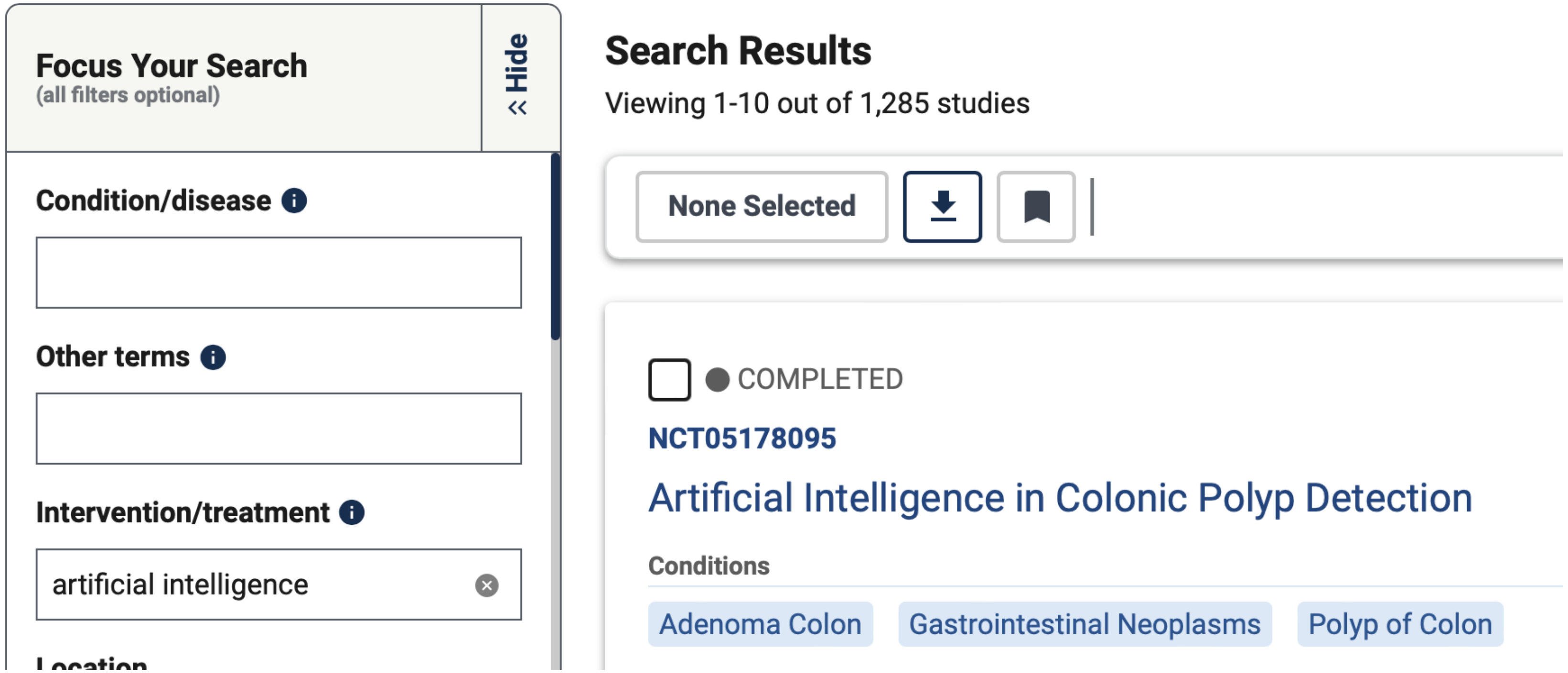
<u>Transparency in device performance</u>

- 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

NIH National Library of Medicine
National Center for Biotechnology Information

ClinicalTrials.gov



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.[1]

SPIRIT-AI

Reporting Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence

CONSORT-AI

Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence

www.clinical-trials.ai

Intended Use Statements

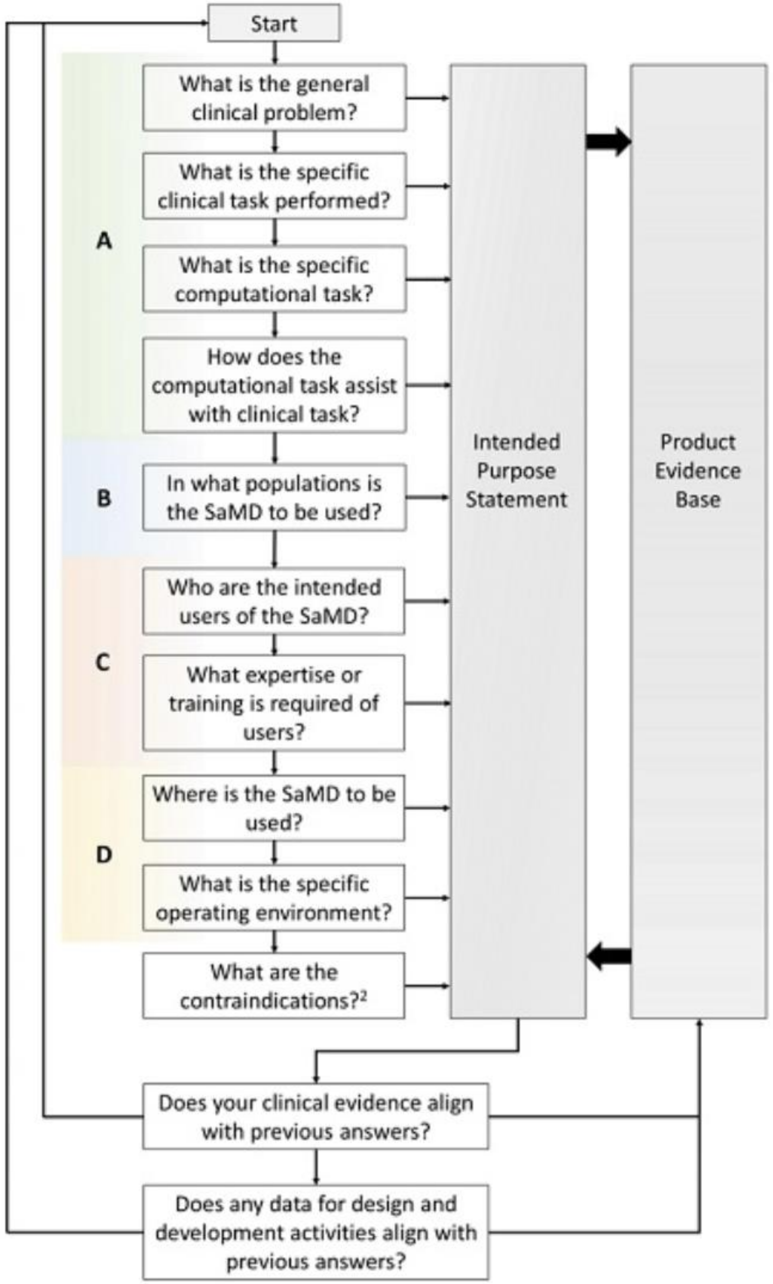
Medicines & Healthcare products Regulatory Agency

Guidance

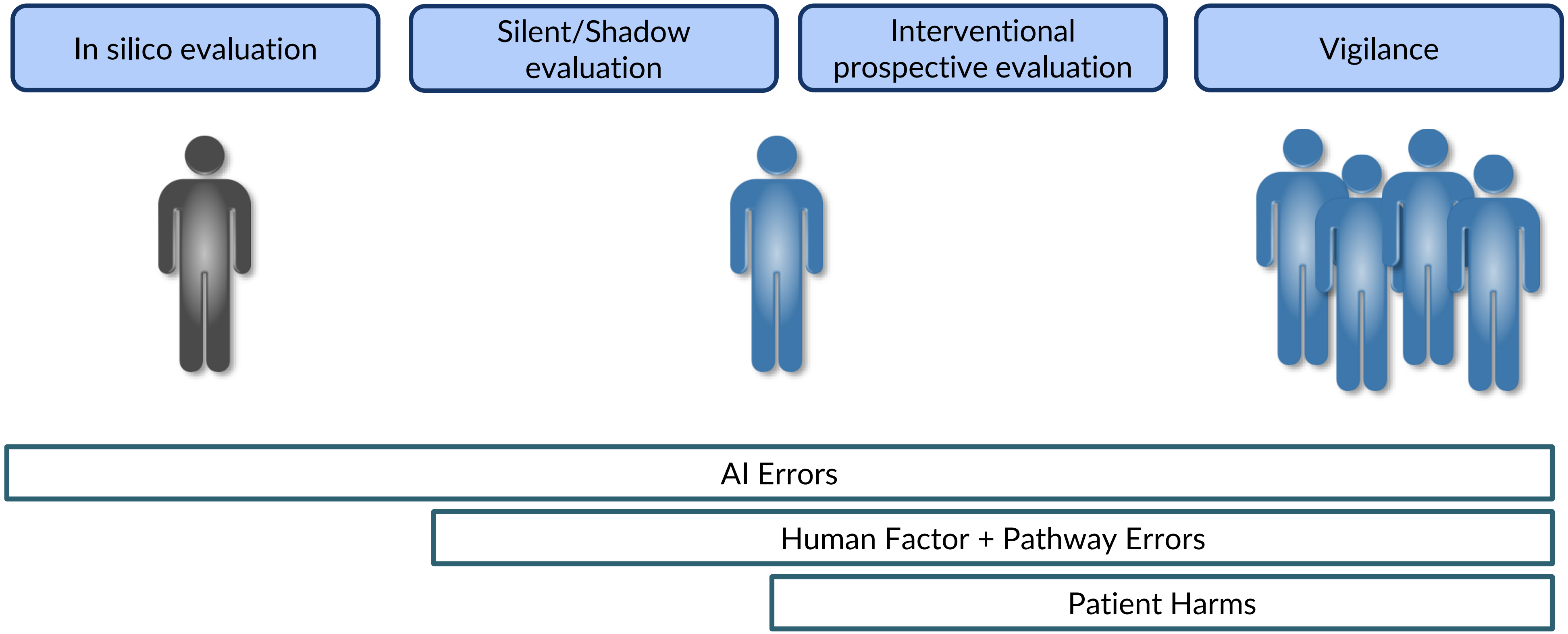
Crafting an intended purpose in the context of software as a medical device (SaMD)

Published 22 March 2023

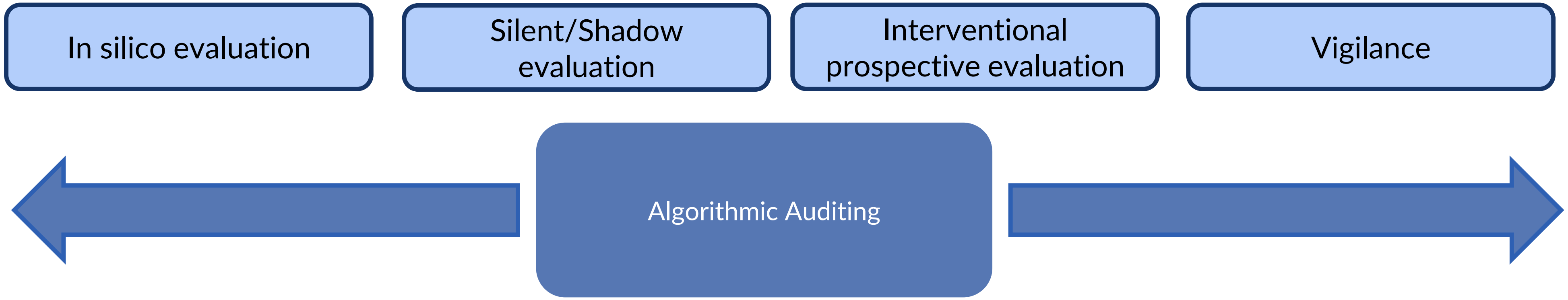
Creating a clear intended purpose is essential for successfully navigating the regulatory requirements for medical devices. In addition, the MHRA encourage manufacturers to maximise the benefits of a clear intended purpose by making this information publicly available. This clarity and transparency can 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



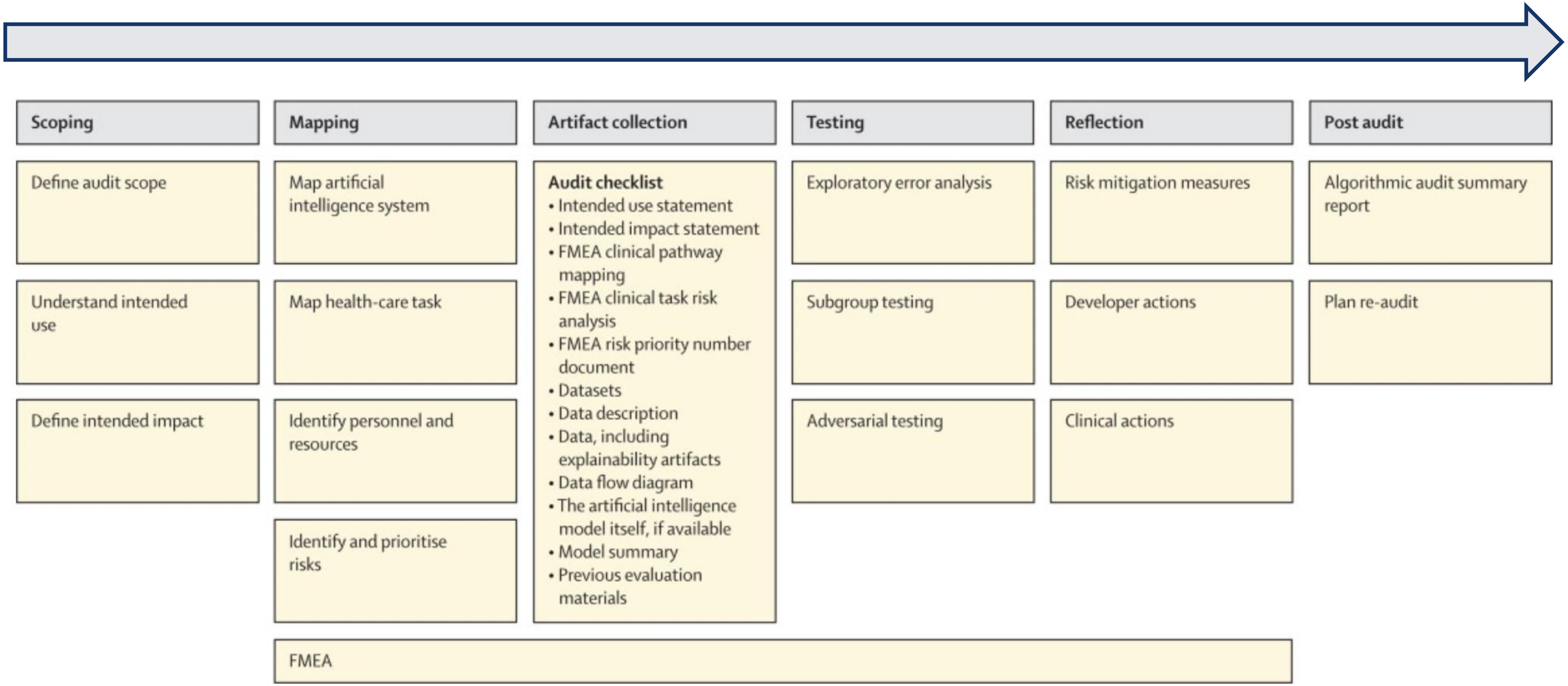
Medical algorithmic auditing



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

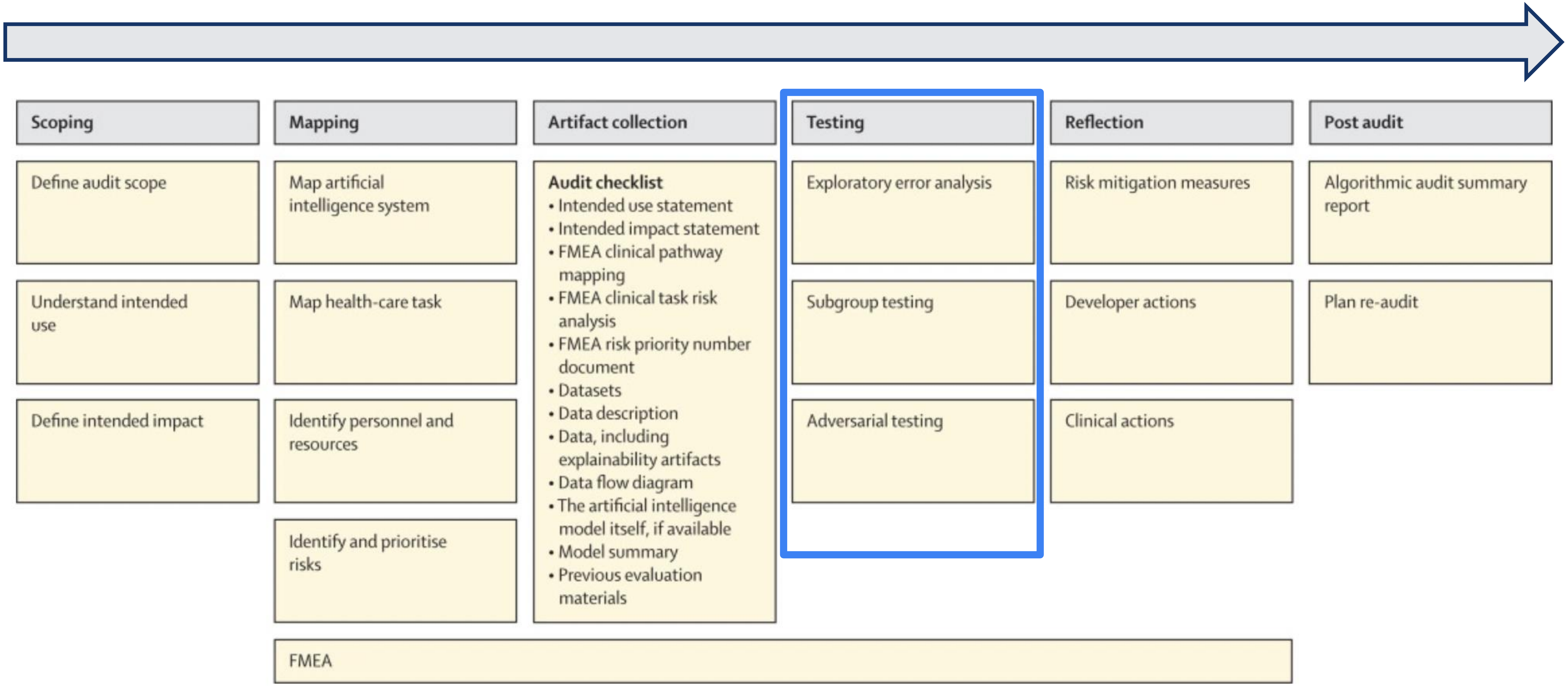
Lauren Oakden-Rayner et al. *Lancet Digital Health* 2022
Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department

Medical algorithmic auditing

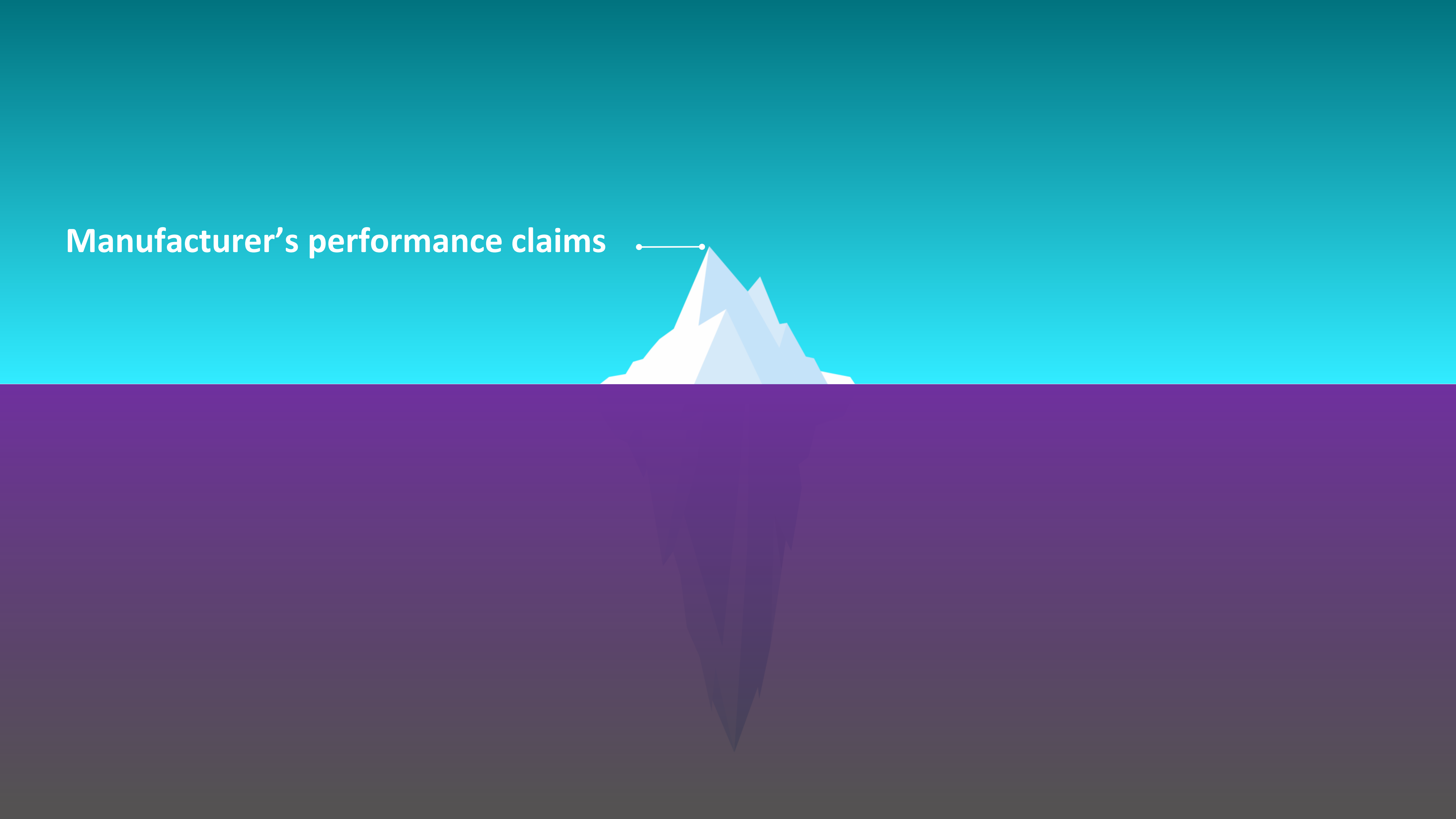


Liu et al. The Medical Algorithmic Audit *Lancet Digital Health 2022.*

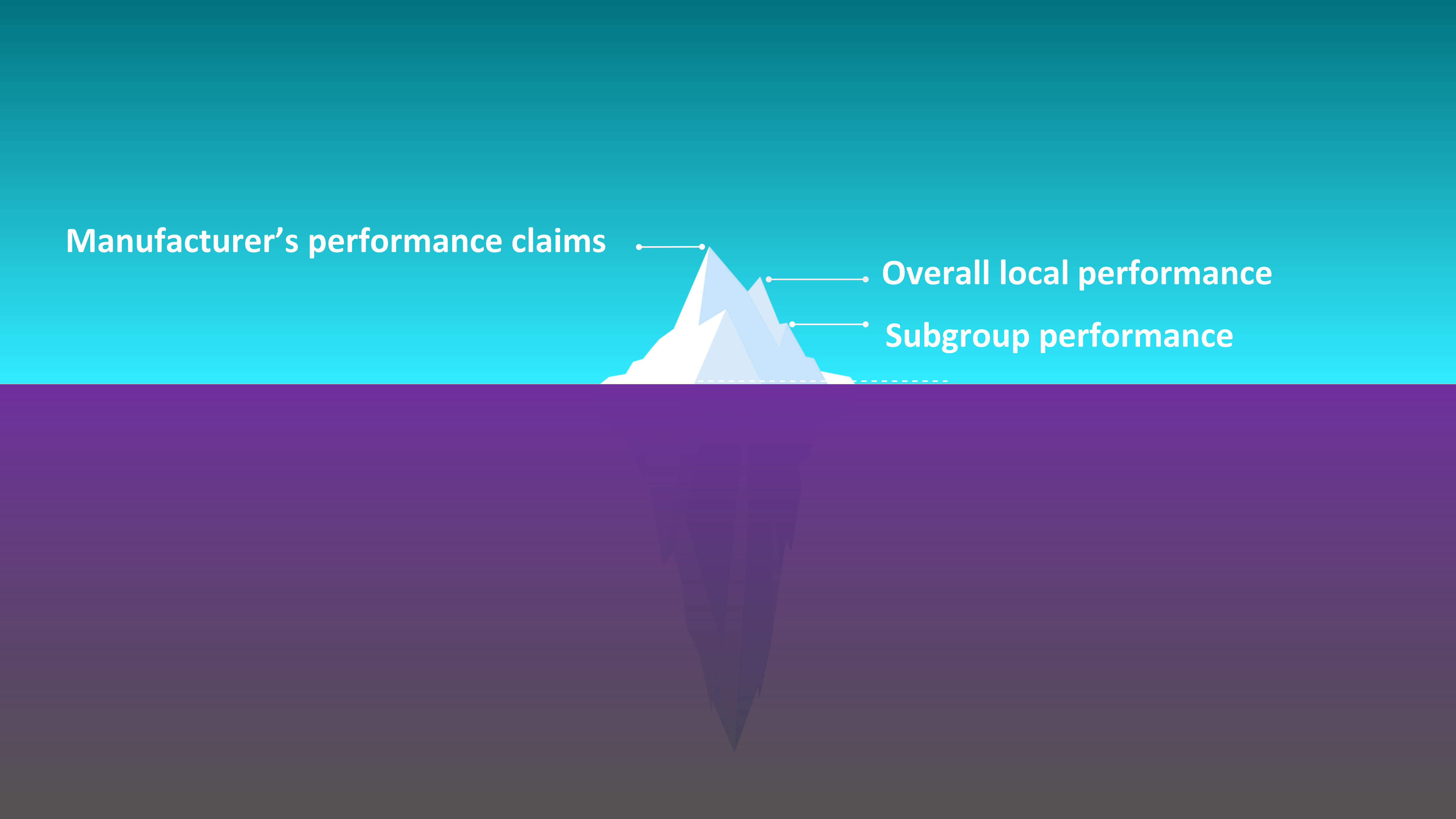
Medical algorithmic auditing

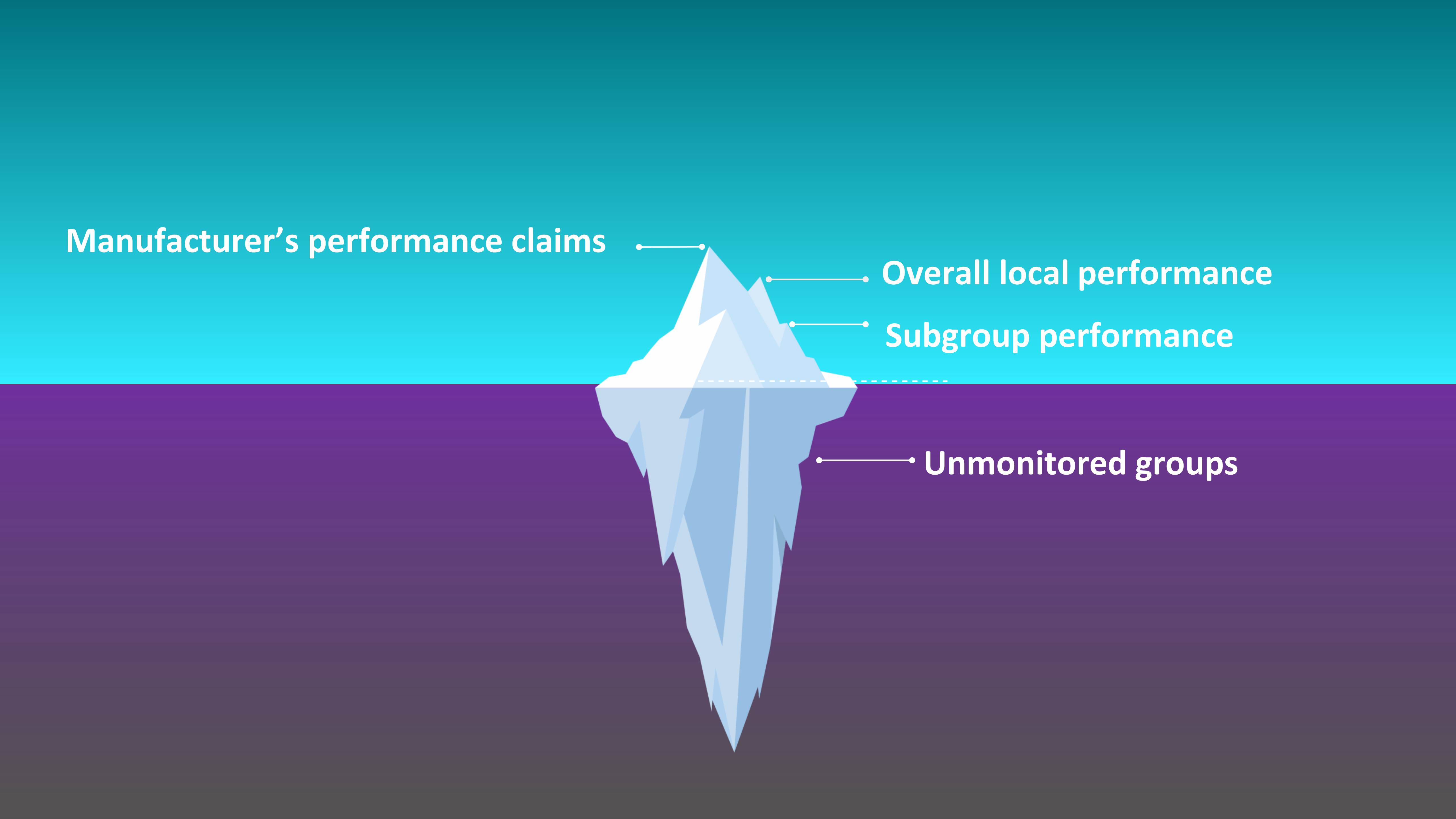


Liu et al. The Medical Algorithmic Audit *Lancet Digital Health 2022.*


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