Controlling AI: regulators and other gate-keepers

Prof Alastair Denniston

Professor of Regulatory Science & Innovation
University of Birmingham, UK

Director, Centre of Excellence for Regulatory Science & Innovation
in AI & Digital Health (CERSI-AI), UK

UNIVERSITY OF BIRMINGHAM

NHS
University Hospitals Birmingham
NHS Foundation Trust

CERSI-AI
Smarter Regulation for Better Innovation
in AI & Digital Health

The Centre of Excellence for Regulatory Science & Innovation
in **AI & Digital Health**

CERSI-AI

The UK's Centre of Excellence working with
innovators,
regulators, researchers,
patients and the NHS
to **optimise**
the **regulation** of
AI & Digital Health Technologies
so as to **accelerate innovation**
that will **improve people's lives**.

www.cersi-ai.org

CERSI-AI
Smarter Regulation for Better Innovation
in AI & Digital Health

UK Research and Innovation

Medicines & Healthcare products Regulatory Agency

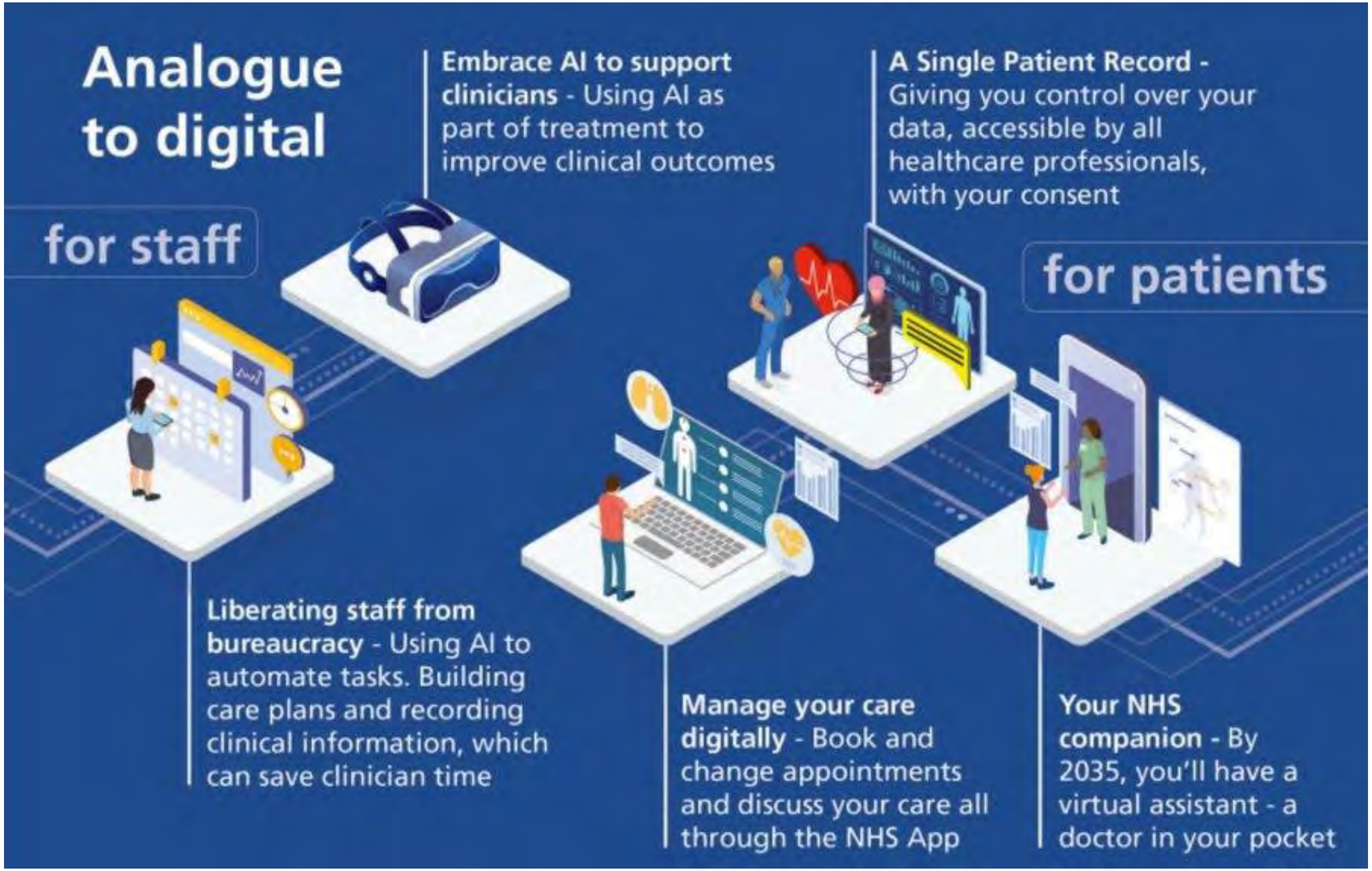
Office for Life Sciences

Department of Health & Social Care

Department for Science, Innovation & Technology

Destination


UK Government
NHS
FIT FOR THE FUTURE
10 Year Health Plan for England
Executive Summary
July 2025


Analogue to digital
for staff
for patients
Embrace AI to support clinicians - Using AI as part of treatment to improve clinical outcomes
A Single Patient Record - Giving you control over your data, accessible by all healthcare professionals, with your consent
Liberating staff from bureaucracy - Using AI to automate tasks. Building care plans and recording clinical information, which can save clinician time
Manage your care digitally - Book and change appointments and discuss your care all through the NHS App
Your NHS companion - By 2035, you'll have a virtual assistant - a doctor in your pocket

It’s already happening

Self-help chatbots
Home triage
Digital therapeutics

Community diagnostics
Decision support systems
Automated advice & guidance
Diagnosis
Prediction

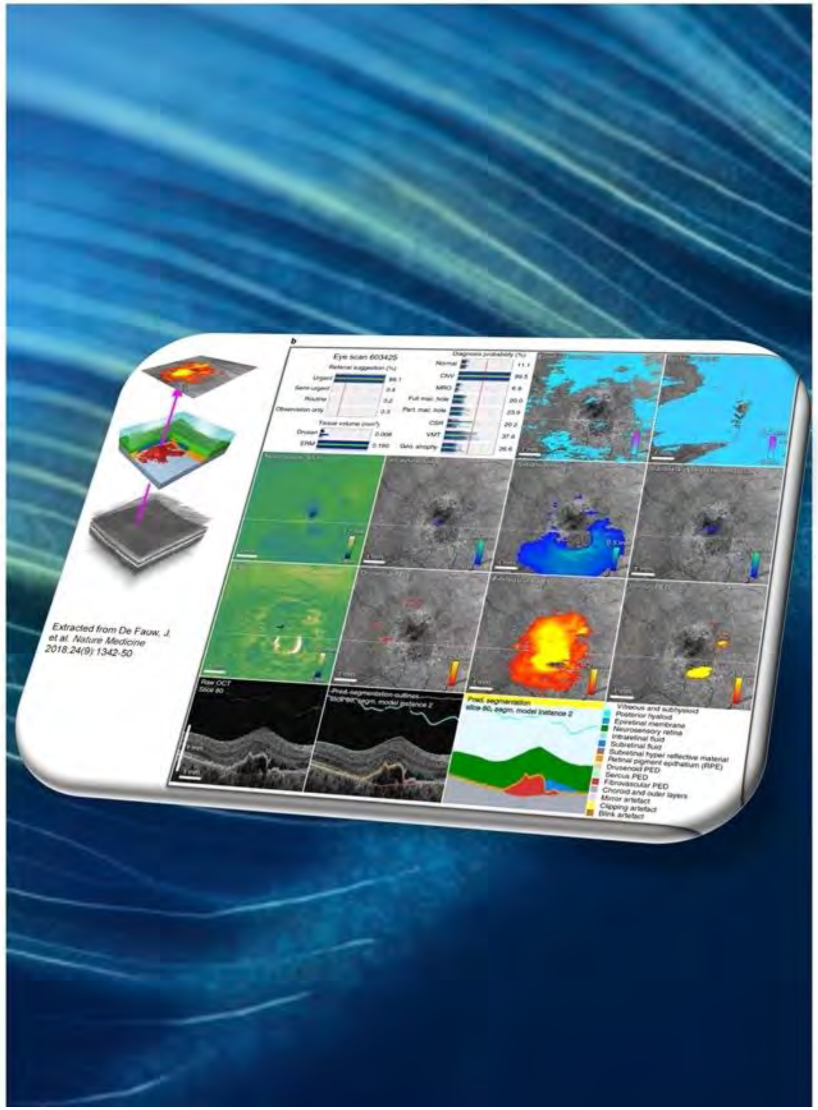
Intelligent history taking
Decision support systems
Diagnosis
Prediction

Segmentation
Quantification
Diagnosis
Prediction

The need and opportunity

If we get this right…

- Widespread, rapid, 24-7 services that are highly accurate and safe across the diverse population of the UK



The need and opportunity

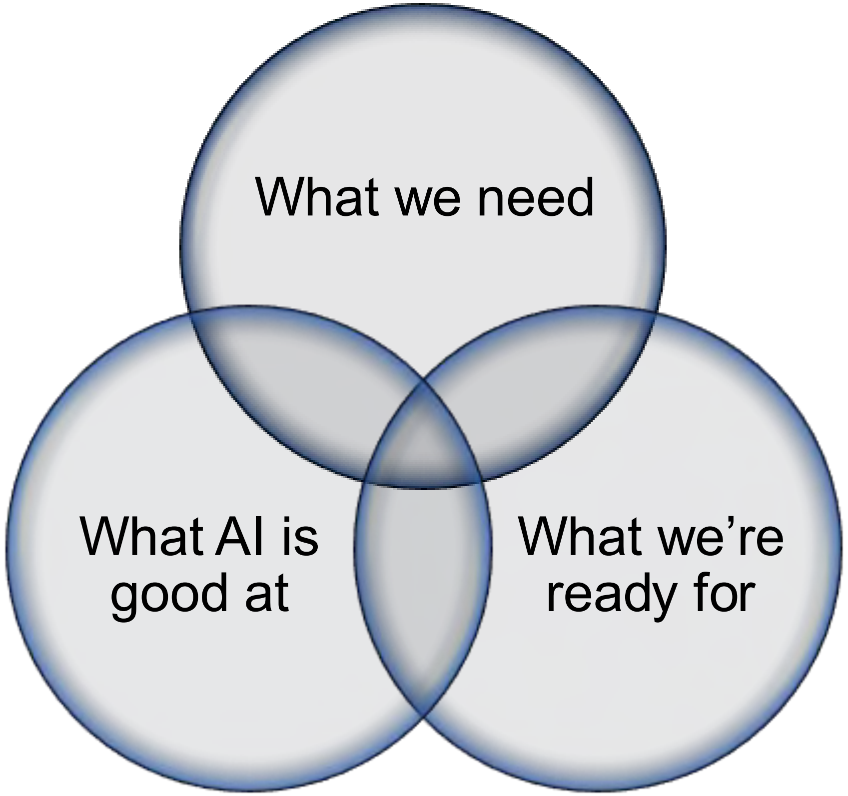
If we get this right…

- Widespread, rapid, 24-7 services that are highly accurate and safe across the diverse population of the UK

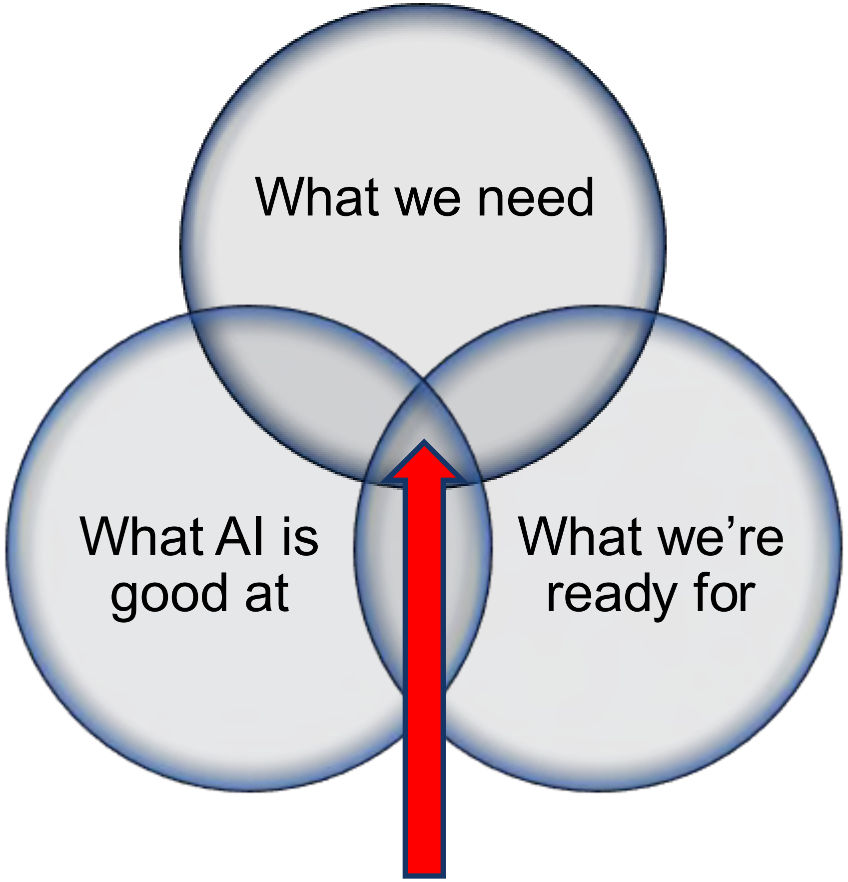
If we get this wrong…

- Highly variable services that are brittle, widely distrusted and only accurate and safe in majority groups

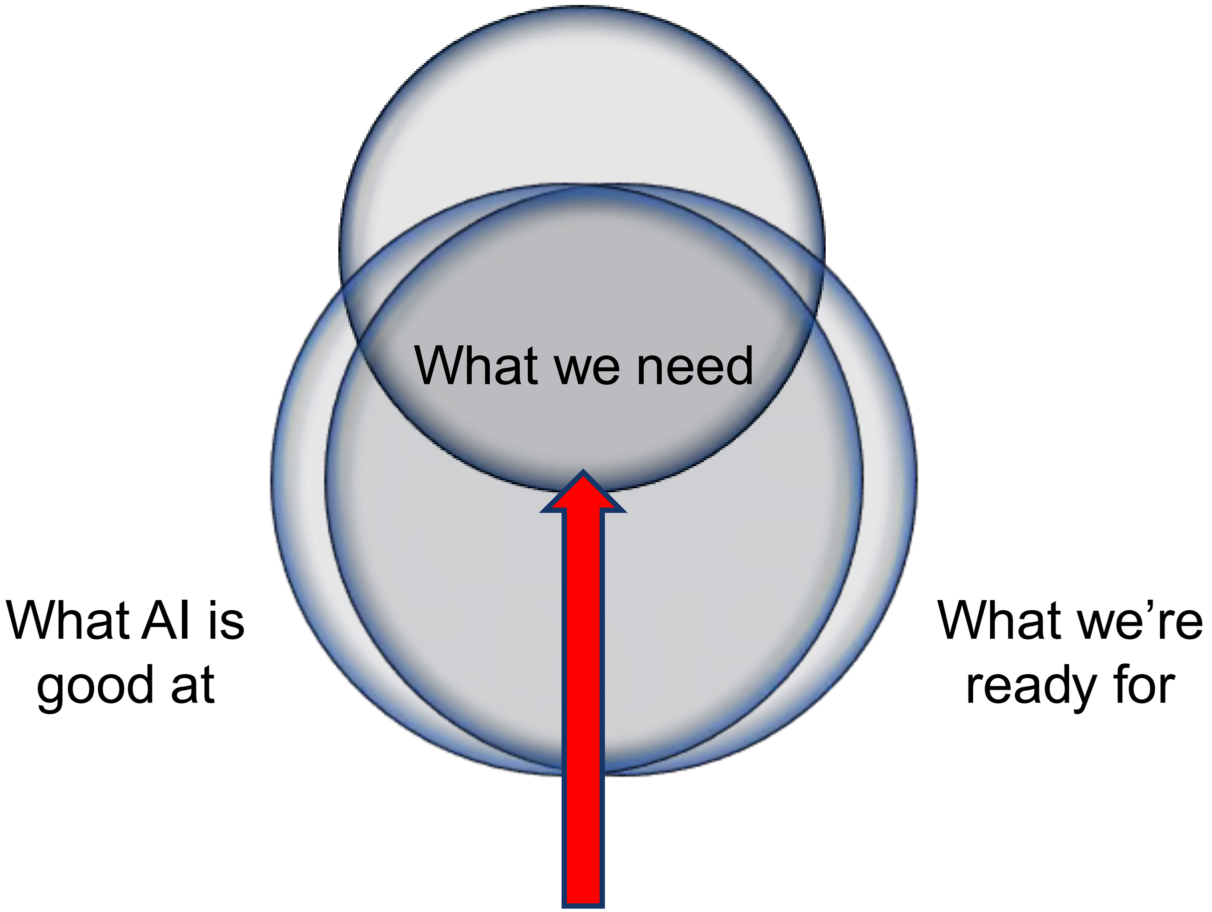
Find the value intersect



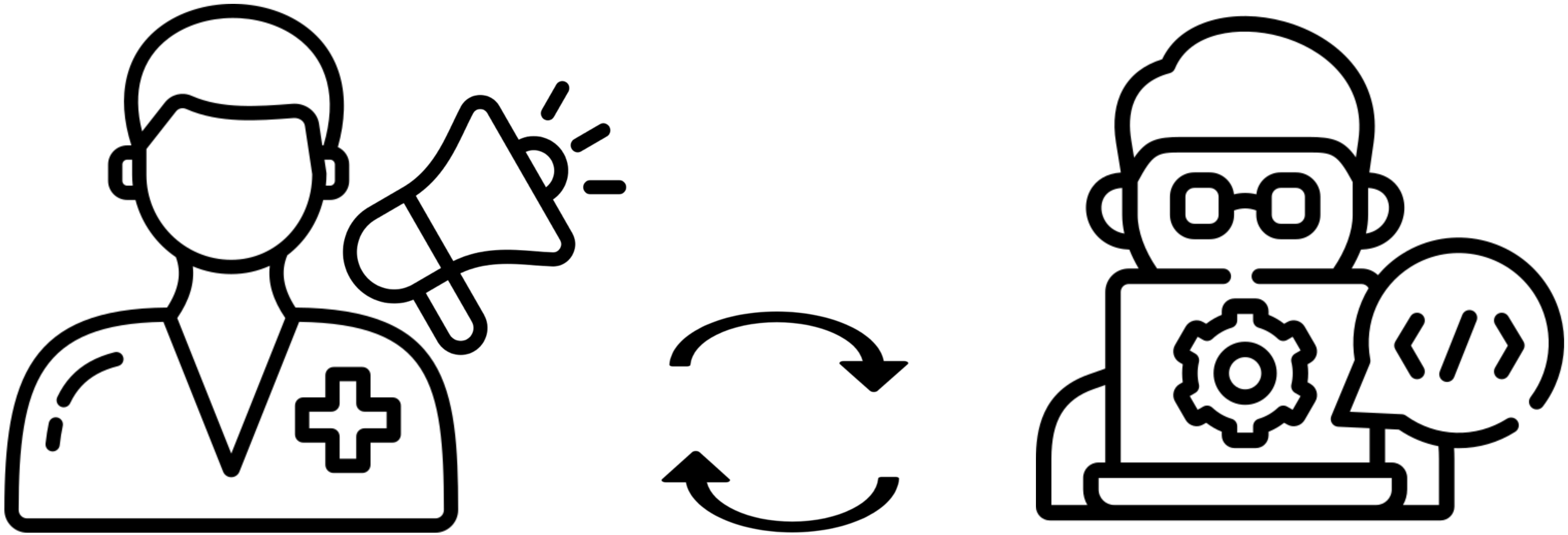
Find the value intersect



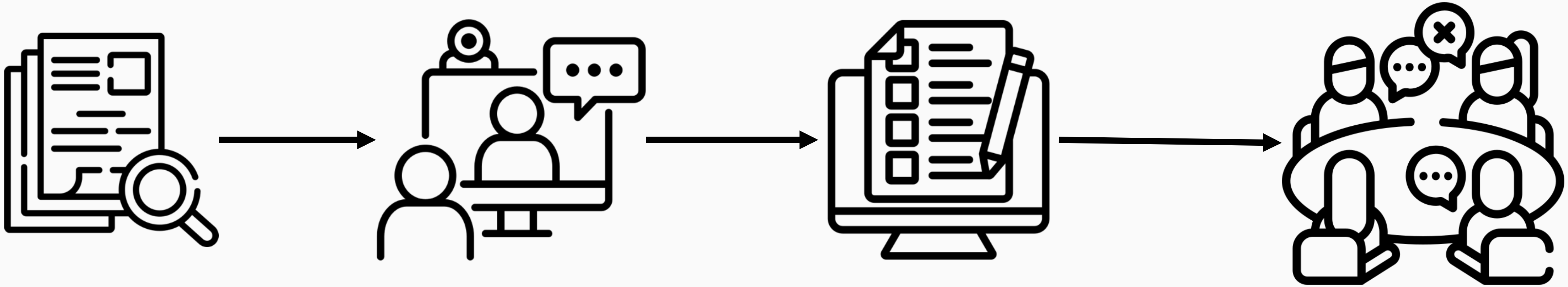
Find the value intersect



Send/receive ‘demand signals’



Send/receive ‘demand signals’



Literature review | Interviews | Delphi (Survey) | Delphi (Consensus meeting)

Development of a Target Product Profile for an Artificial Intelligence for Use in English Diabetic Eye Screening, a Modified Delphi Consensus Study

122 Pages • Posted: 18 Jun 2025

Trystan Macdonald

NHS England

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5293401

1. Intended Use	1.1 Intended population
	1.2 Intended user
	1.3 Level of professional oversight
	1.4 Intended use environment
	1.5 Scalability
	1.6 Pathway position
2. Clinical validity & utility	2.1 Diagnostic accuracy
	2.2 Clinical efficacy and effectiveness
	2.3 Subgroup performance
3. Data management	3.1 Input data
	3.2 Output data
	3.3 Data dictionary
	3.4 Data governance
4. Value and costs	4.1 Cost
	4.2 Effect on service outcomes
	4.3 Data sharing

5. Environmental & societal impact	5.1 Environmental sustainability
	5.2 Social value
	5.3 Safeguarding vulnerable groups
6. Safety & security	6.1 Safety
	6.2 System malfunction protection
	6.3 Security
7. Regulatory	7.1 Regulatory requirements
	7.2 Post-deployment monitoring
8. Acceptability	8.1 Acceptability with stakeholders
	8.2 Training
	8.3 Consent
	8.4 Interface
	8.5 Language
	8.6 Product support
9. Infrastructural requirements	9.1 Compatibility to existing hardware
	9.2 … to existing software
	9.3 Technology type

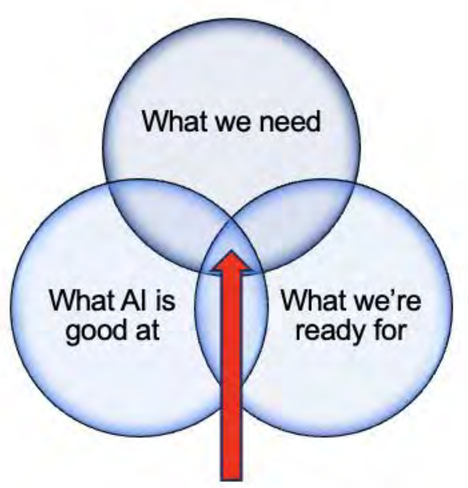
TPP: Essential characteristics & specifications

Consensus meeting

Section	Characteristic	Specification	Explanatory text
1. Intended use	1.1 Intended population	**Essential** - must be able to process the diabetic eye screening photos of people with diabetes aged 12 or over.	***Explanatory text:*** *all People Living With Diabetes (PLWD) 12 or over are eligible for eye screening.*[20] *Diabetic Eye Screening (DES) AI (Artificial Intelligence) use in those under 18 will place an additional compliance burden on vendors and Diabetic Eye Screening Programmes (DESPs) in terms of medical device regulation, information governance, and data ethics considerations. However, implementing DES AI contraindicated in PLWD under 18 would likely require significant changes to existing service designs, reducing or negating any potential benefits from DES AI use. Whilst there has historically been a paucity of data for model training and evaluation in paediatric populations, increasing amounts is now being made available for AI training and testing, including from UK health populations.*[31,32,58]
	1.2 Intended user	**Essential** - diabetic eye screening professionals involved in grading or quality control. This includes screeners, graders, ophthalmologists, relevant IT staff, and screening management staff.	***Explanatory text:*** *defining an intended user is essential to comply with medical device regulation*[46,59] *and restricts deploying healthcare institutions as to who can use the device 'on label' post-deployment due to considerations around liability and patient consent.*[60] *DES AI use as outlined in these recommendations would require a range of individuals to interact with it as primary, secondary, and tertiary users. Primary users can be considered those inputting data into the AI, screeners in the case of DES. Secondary users, those receiving data downstream of the AI, would be graders and relevant management staff. Tertiary users, those with oversight of the AI or interacting with it indirectly (such as those tasked with maintaining technical functionality or compliance with regulations and policies), are likely to be IT and management staff. Guidance on the roles, training, and qualifications of staff involved in DES has been published by NHS England.*[61]

Know the destination

Build to the intersect



Knowing the route

The road of AI innovation in health

Design → Proof-of-concept → Clinical evaluation → Regulation → Implementation

The road of AI innovation in health

£ Venture $

UKRI Medical Research Council

NIHR

UKRI Innovate UK

NHS Health Research Authority

Medicines & Healthcare products Regulatory Agency

NICE

NHS

Design → Proof-of-concept → Clinical evaluation → Regulation → Implementation

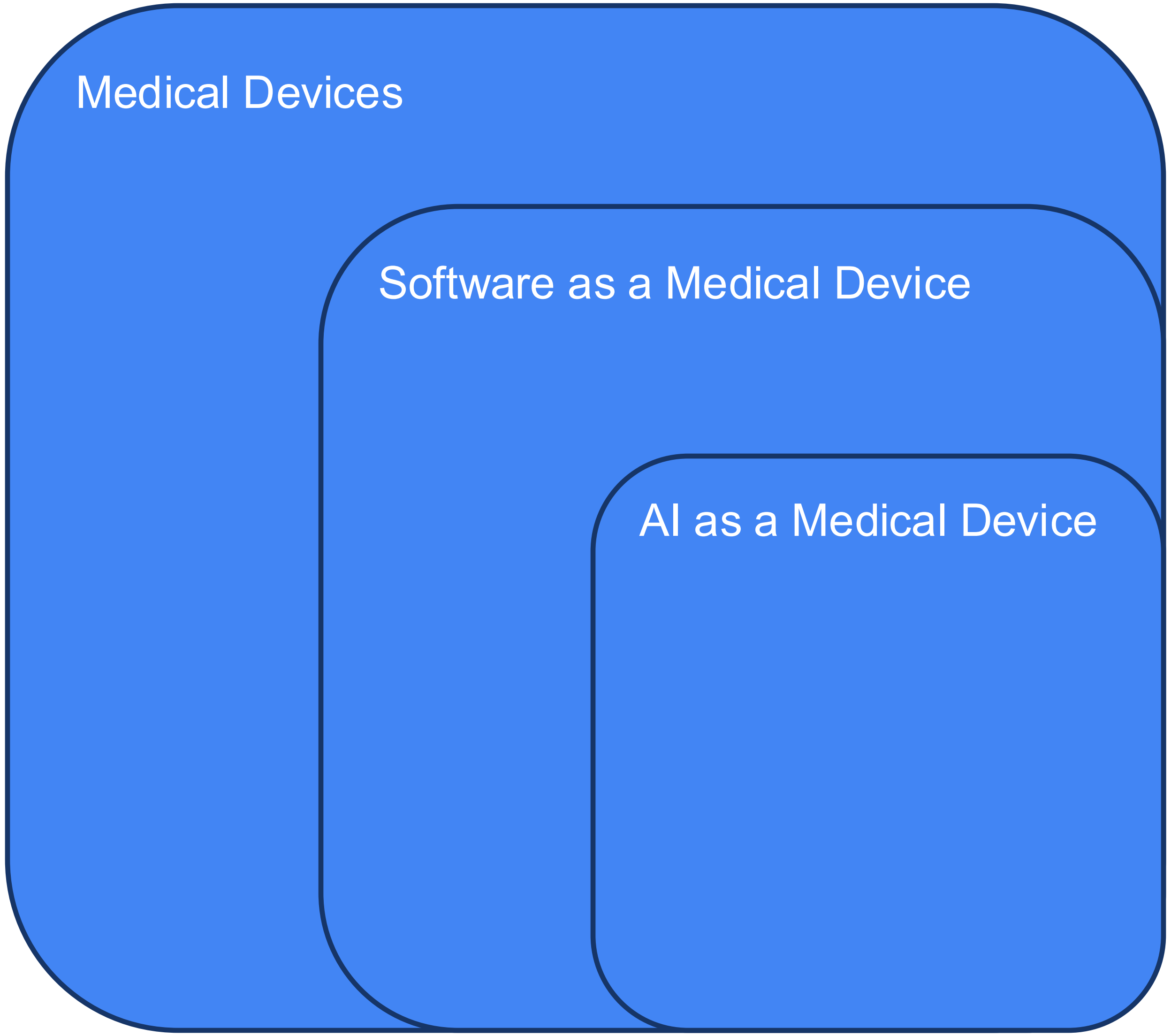
Funding

Permission to research

Permission to sell

Permission to buy

Decision to buy



Medicines &
Healthcare products
Regulatory Agency

Informing the gate-keepers

We work with:

Medicines & Healthcare products Regulatory Agency · Care Quality Commission · NHS Health Research Authority · NICE · bsi. · Department for Science, Innovation, & Technology · Department of Health & Social Care · Regulatory Horizons Council · World Health Organization · FDA · Health Canada · TGA

Setting UK and international policy:


CNN health

AI may be as effective as medical specialists at diagnosing disease

By Jack Guy, CNN

Updated 1123 GMT (1923 HKT) September 25, 2019

First global analysis of AI diagnostics


Search jobs Sign in Search UK edition

The Guardian

ion Sport Culture Lifestyle More

avirus Football Environment UK politics Education Society Science Tech Global de

AI standards launched to help tackle problem of overhyped studies

New guidelines aimed at ensuring AI research is of same quality as that in other fields

International standards for trials of AI


WIRED

WILL KNIGHT BUSINESS OCT 11, 2020 7:00 AM

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.

International standards for inclusive data for equitable AI

The Evidence Gap

THE LANCET
Digital Health

A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis

Xiaoxuan Liu*, Livia Faes*, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, Alastair K Denniston

Summary

Background Deep learning offers considerable promise for medical diagnostics. We aimed to evaluate the diagnostic accuracy of deep learning algorithms versus health-care professionals in classifying diseases using medical imaging.

Methods In this systematic review and meta-analysis, we searched Ovid-MEDLINE, Embase, Science Citation Index, and Conference Proceedings Citation Index for studies published from Jan 1, 2012, to June 6, 2019. Studies comparing the diagnostic performance of deep learning models and health-care professionals based on medical imaging, for any disease, were included. We excluded studies that used medical waveform data graphics material or investigated the accuracy of image segmentation rather than disease classification. We extracted binary diagnostic accuracy data and constructed contingency tables to derive the outcomes of interest: sensitivity and specificity. Studies undertaking an out-of-sample external validation were included in a meta-analysis, using a unified hierarchical model. This study is registered with PROSPERO, CRD42018091176.

Of 20 503 studies…

fewer than 1%
were sufficiently well-designed
to provide reliable evidence
of performance

Liu X et al. Lancet Digit Health. 2019

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.

Table 1 | Reporting guidelines for healthcare studies using machine learning

Reporting guideline	Scope
STARD-AI	Studies evaluating the diagnostic accuracy of an artificial intelligence based test (in preparation)[67]
TRIPOD+AI	Studies developing or evaluating the performance of a prediction model, using artificial intelligence, including machine learning methods
CLAIM	Medical imaging studies using artificial intelligence[68]
DECIDE-AI	Early stage clinical evaluation (including safety, human factors evaluation) of decision support systems driven by artificial intelligence[69]
CHEERS-AI	Studies describing health economic evaluations to estimate the value for money (cost effectiveness) of artificial intelligence interventions[70]
SPIRIT-AI	Protocols for clinical trials evaluating an intervention with an artificial intelligence component[71]
CONSORT-AI	Clinical trial reports evaluating an intervention with an artificial intelligence component[72]
PRISMA-AI	Systematic reviews and meta-analyses of artificial intelligence interventions (in preparation)[73]

STARD=Standards for Reporting of Diagnostic Accuracy; TRIPOD=Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; AI=artificial intelligence; CLAIM=Checklist for Artificial Intelligence in Medical Imaging; DECIDE=Decisions in health Care to Introduce or Diffuse innovations using Evidence; CHEERS=Consolidated Health Economic Evaluation Reporting Standards; SPIRIT=Standard Protocol Items: Recommendations for Interventional Trials; CONSORT=Consolidated Standards of Reporting Trials; PRISMA=Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

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

Intended Use Statements

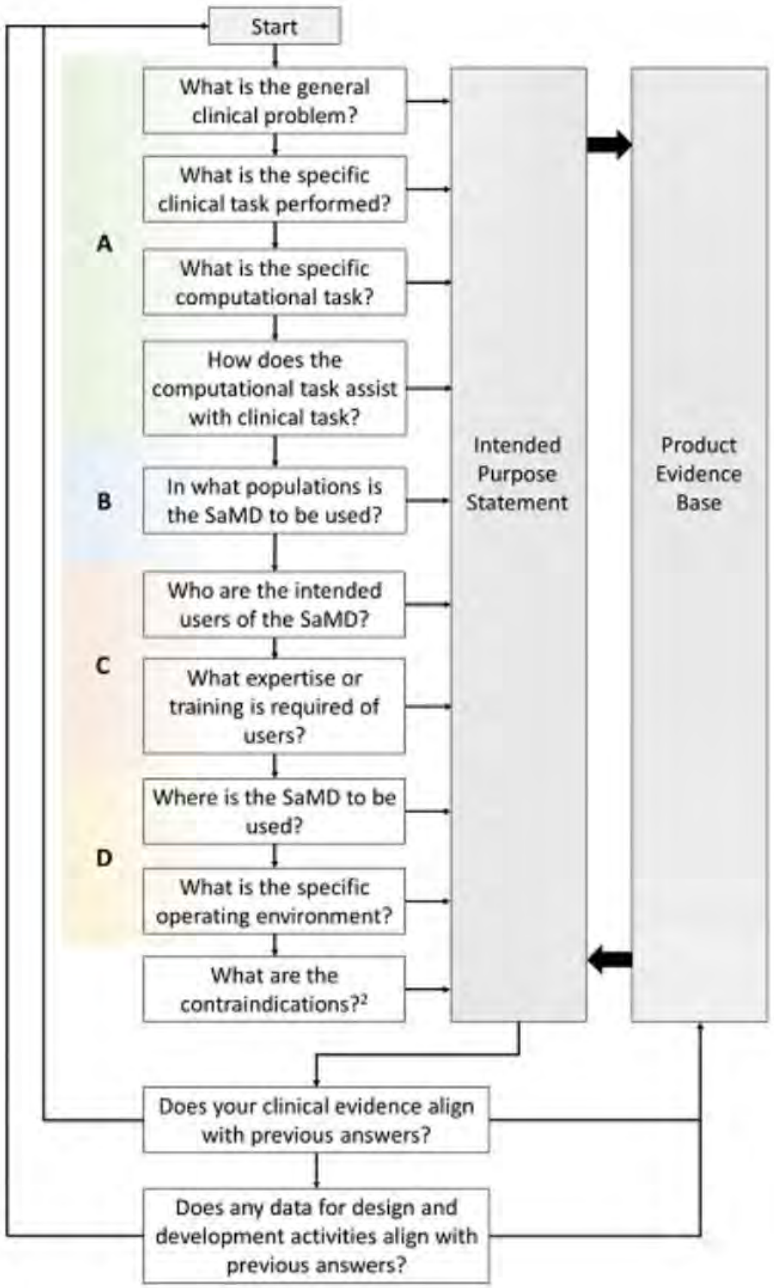
Medicines & Healthcare products Regulatory Agency

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.

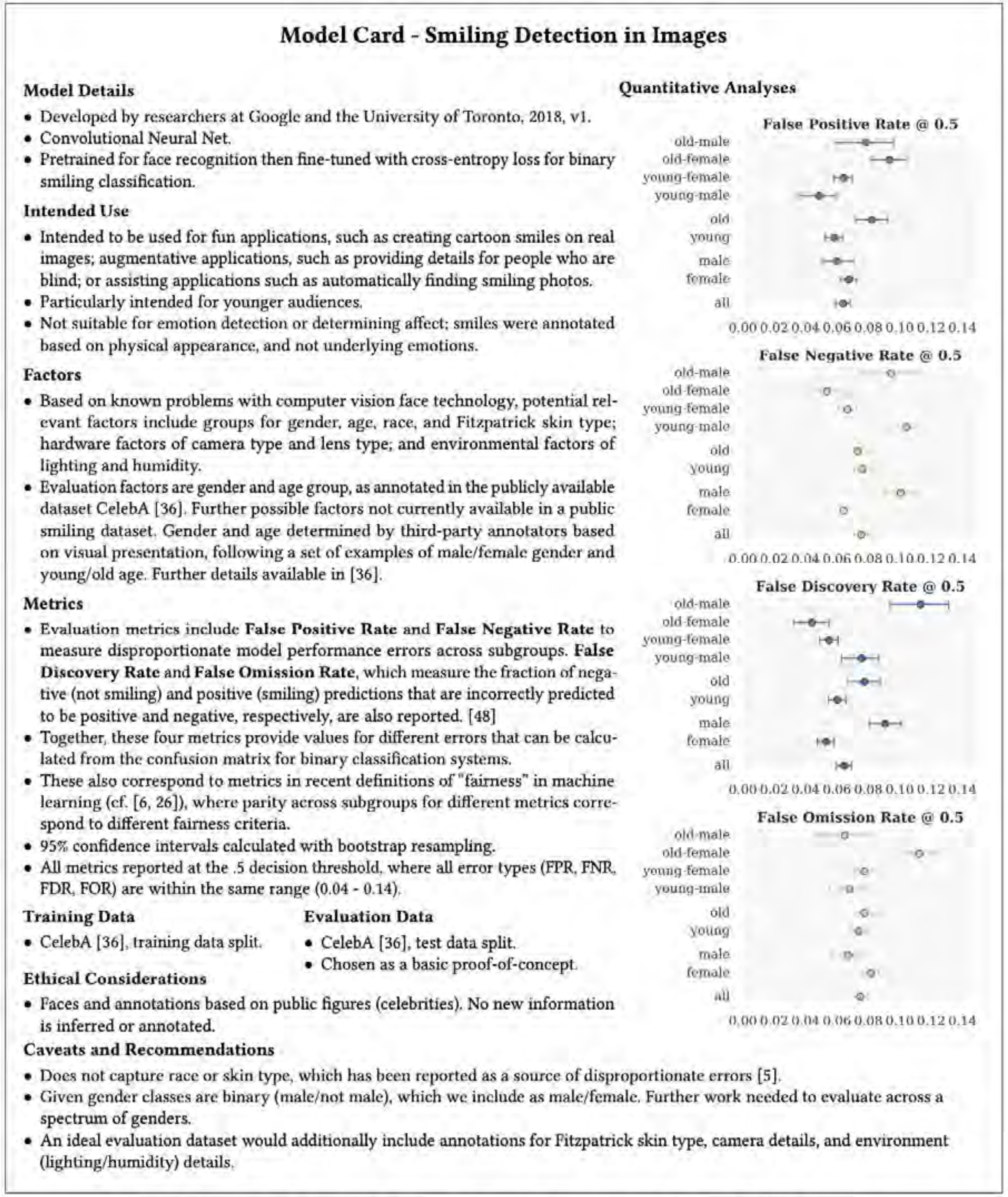


Model Cards for Model Reporting

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.

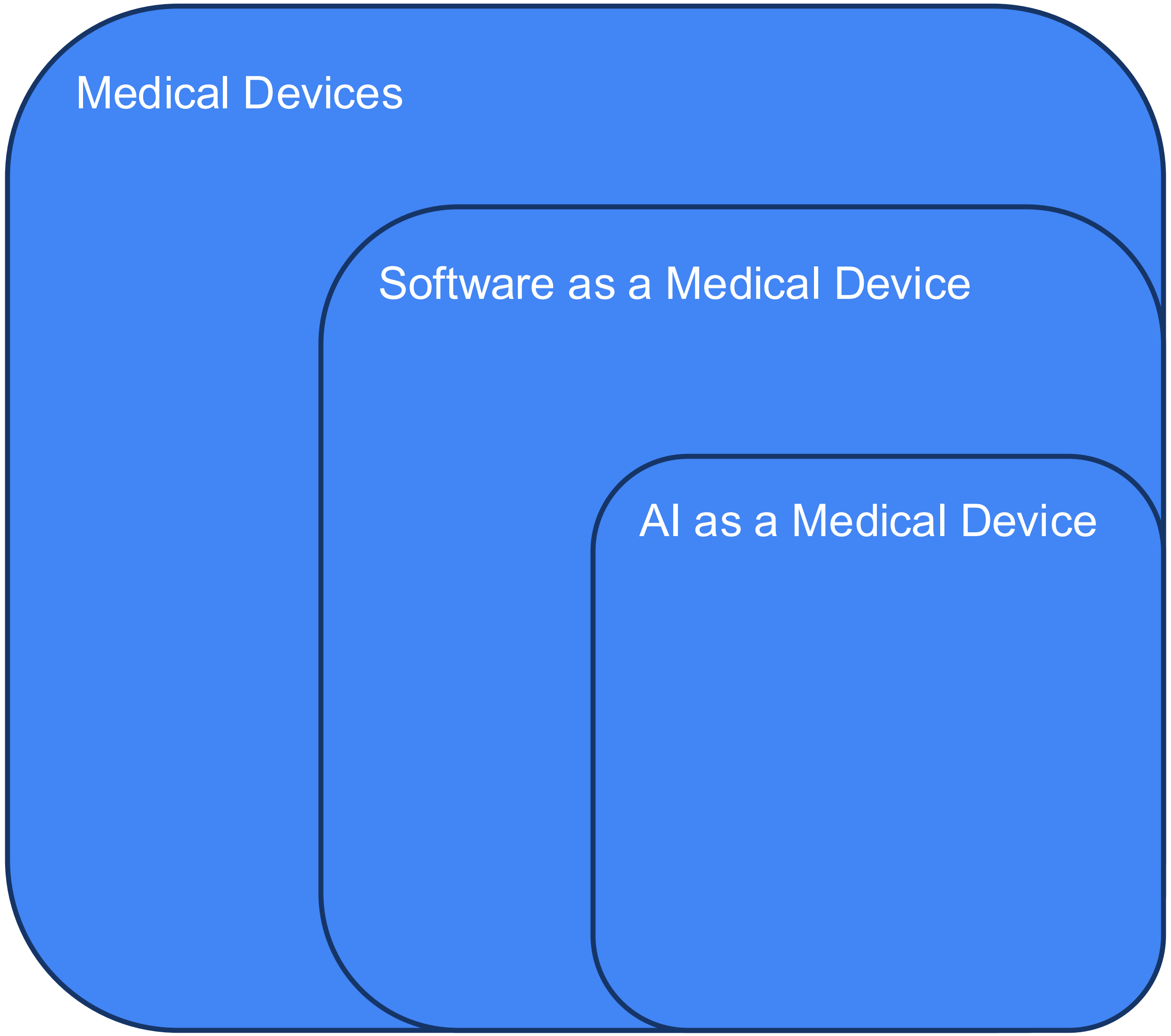
Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

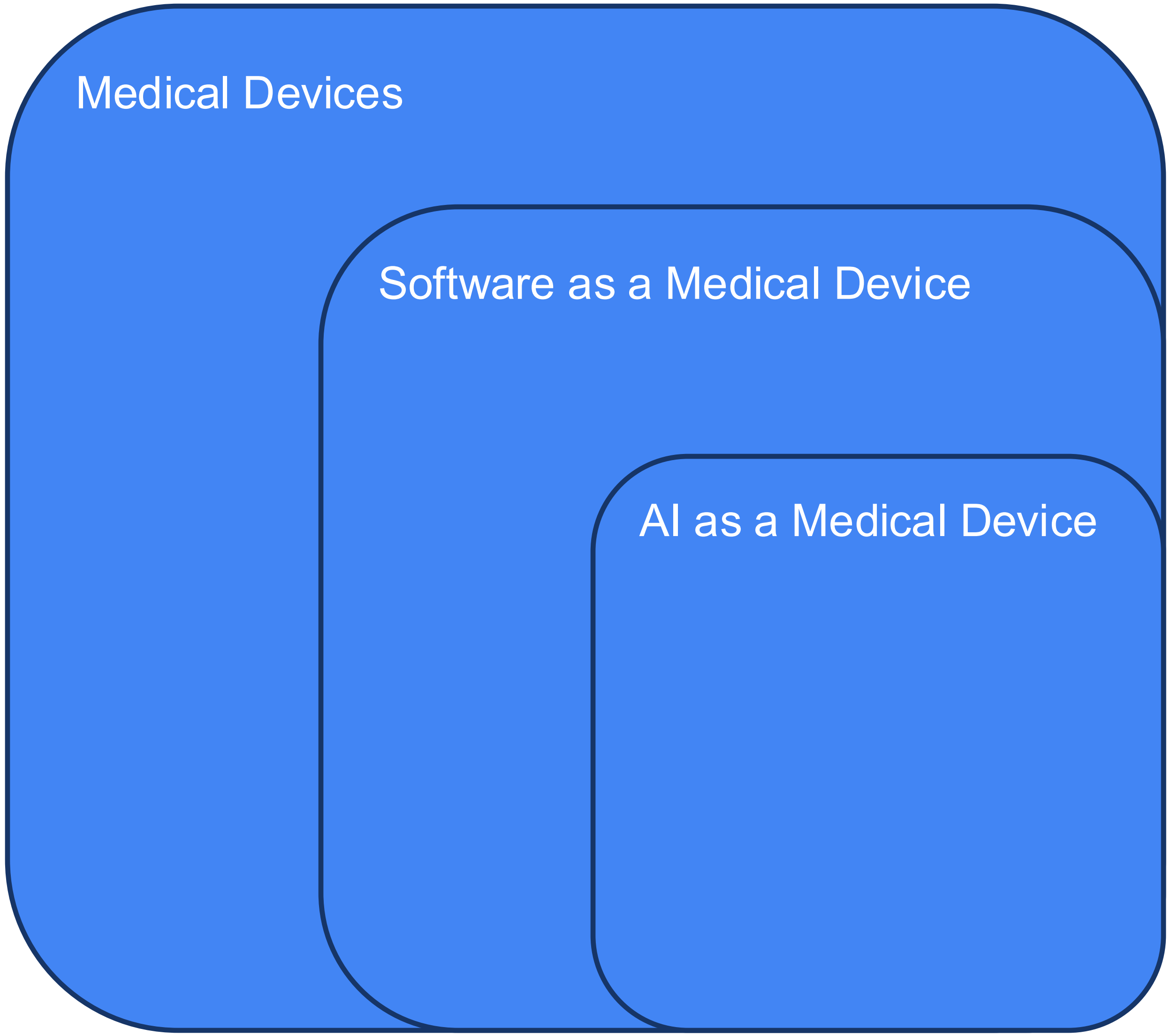
Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses

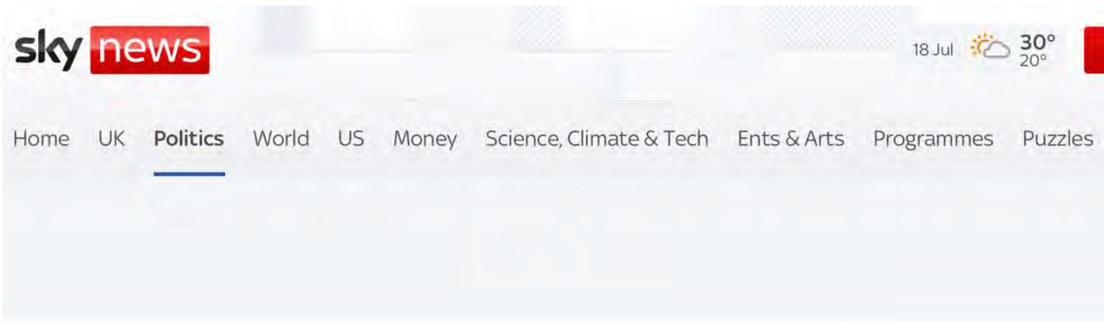
Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation 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.


Medical Devices
Software as a Medical Device
AI as a Medical Device

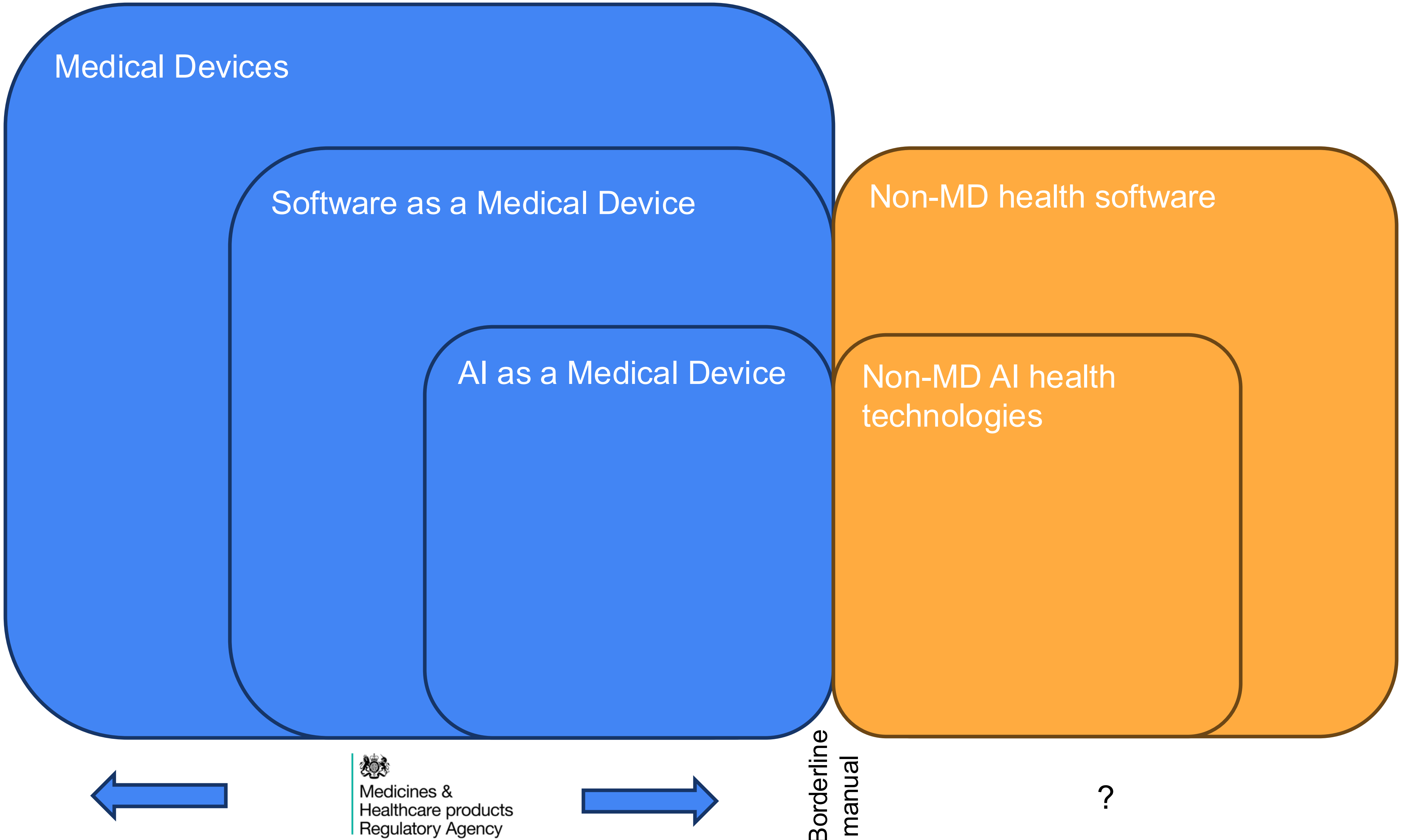


Medicines & Healthcare products Regulatory Agency

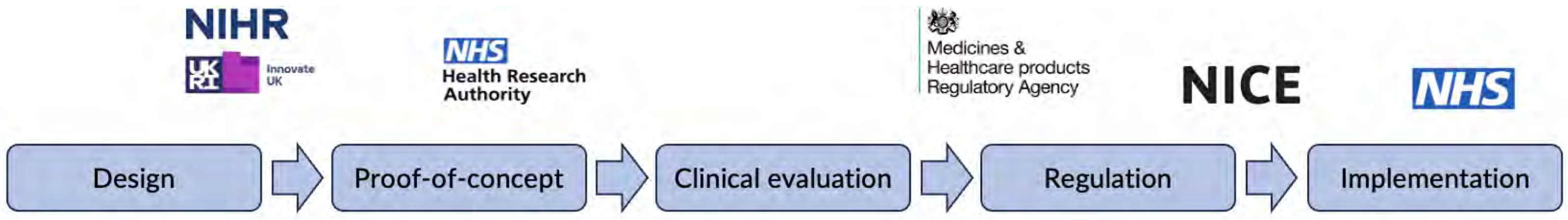


Doctors are using unapproved AI software to record patient meetings, investigation reveals

There is growing controversy around AI software that transcribes patient conversations, with GPs warned unauthorised tools could breach data rules.


Medical Devices
Software as a Medical Device
AI as a Medical Device
Non-MD health software
Non-MD AI health technologies
Medicines & Healthcare products Regulatory Agency
Borderline manual
?

Know the journey

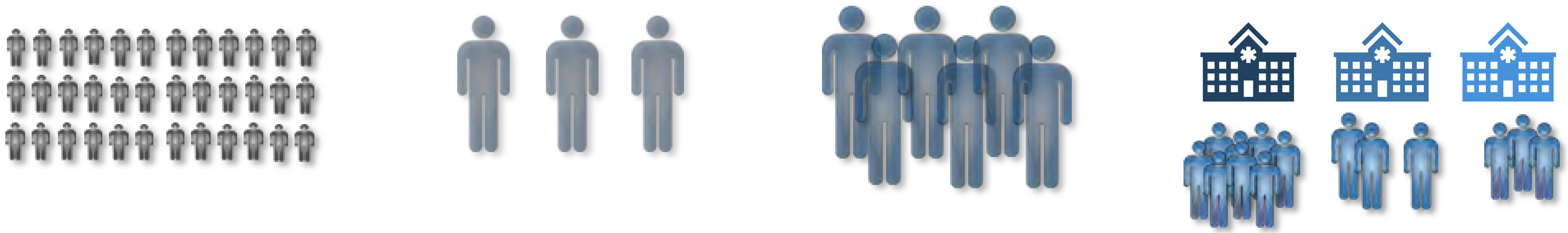


<u>Plan your evidence generation early, and with all ‘gate-keepers’ in mind</u>

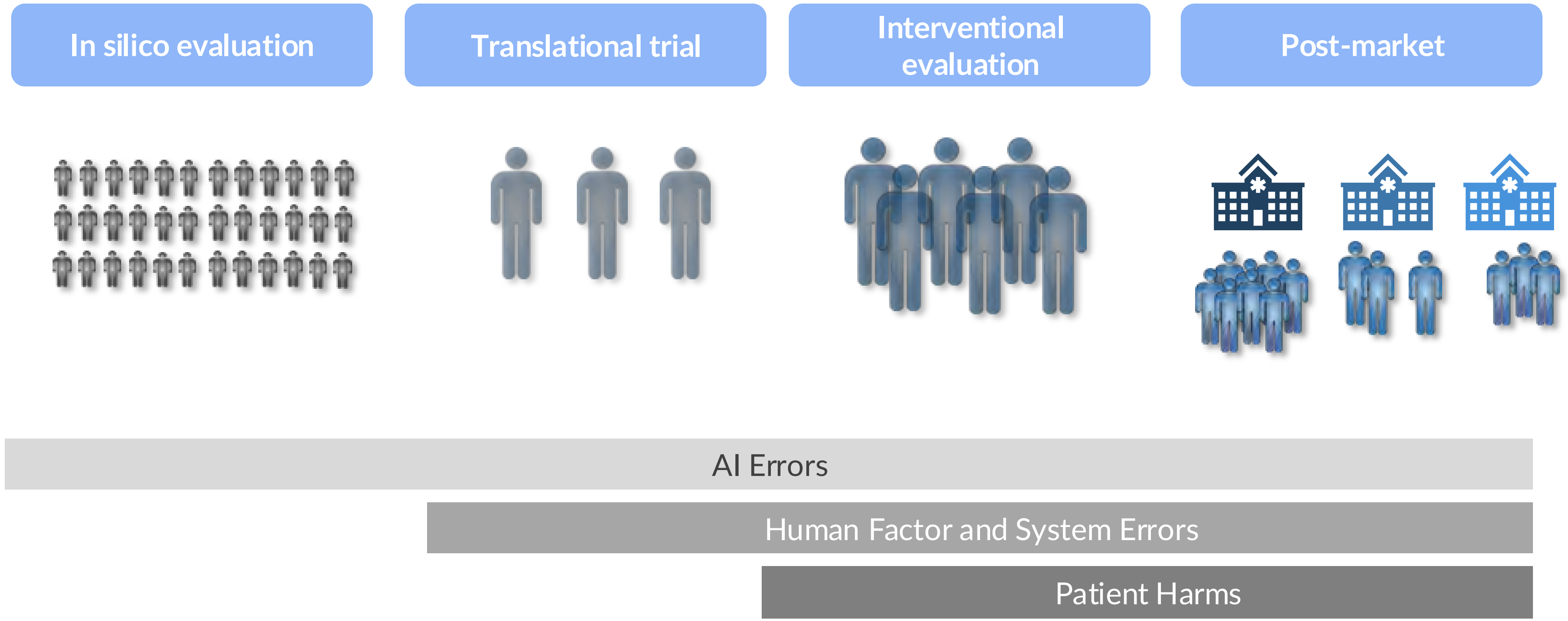
Getting there safely
with everyone on board

Evidence of performance in the real world

In silico evaluation | Translational trial | Interventional evaluation | Post-market



Evidence of performance in the real world



Safe AI implementation using Algorithmic Audit

Skin cancer

Scoping

Audit scope: Cases assessed between April 2022-April 2023

Intended use: screening, triage and assessment of skin lesions suspicious for skin cancer

Intended impact: improved access to dermatology care

Mapping

AI system: Deep Learning based AI system

Health-care task: Triage of suspicious lesions

Personnel and resources

Identify and prioritise risks: Risk Priority Numbers

Artefact Collection

Audit checklist:

- Intended use statement
- Intended impact statement
- FMEA clinical task risk analysis
- FMEA risk priority number document
- Datasets
- Data description
- Data, including explainability artefacts
- Data flow diagram
- AI model itself
- Model summary
- Previous evaluation materials

Testing

Overall Performance: 97.26% sensitivity for malignant lesions

Exploratory error analysis: 22 false negative cases audited

Subgroup testing: six subgroups included for analysis

Adversarial testing: no formal testing conducted

Reflection

Developer actions: Modifications to post-market surveillance

Updates to system thresholds

Clinical actions: Review of device autonomy

Education and training of staff (inc. sharing of audit report)

Post audit

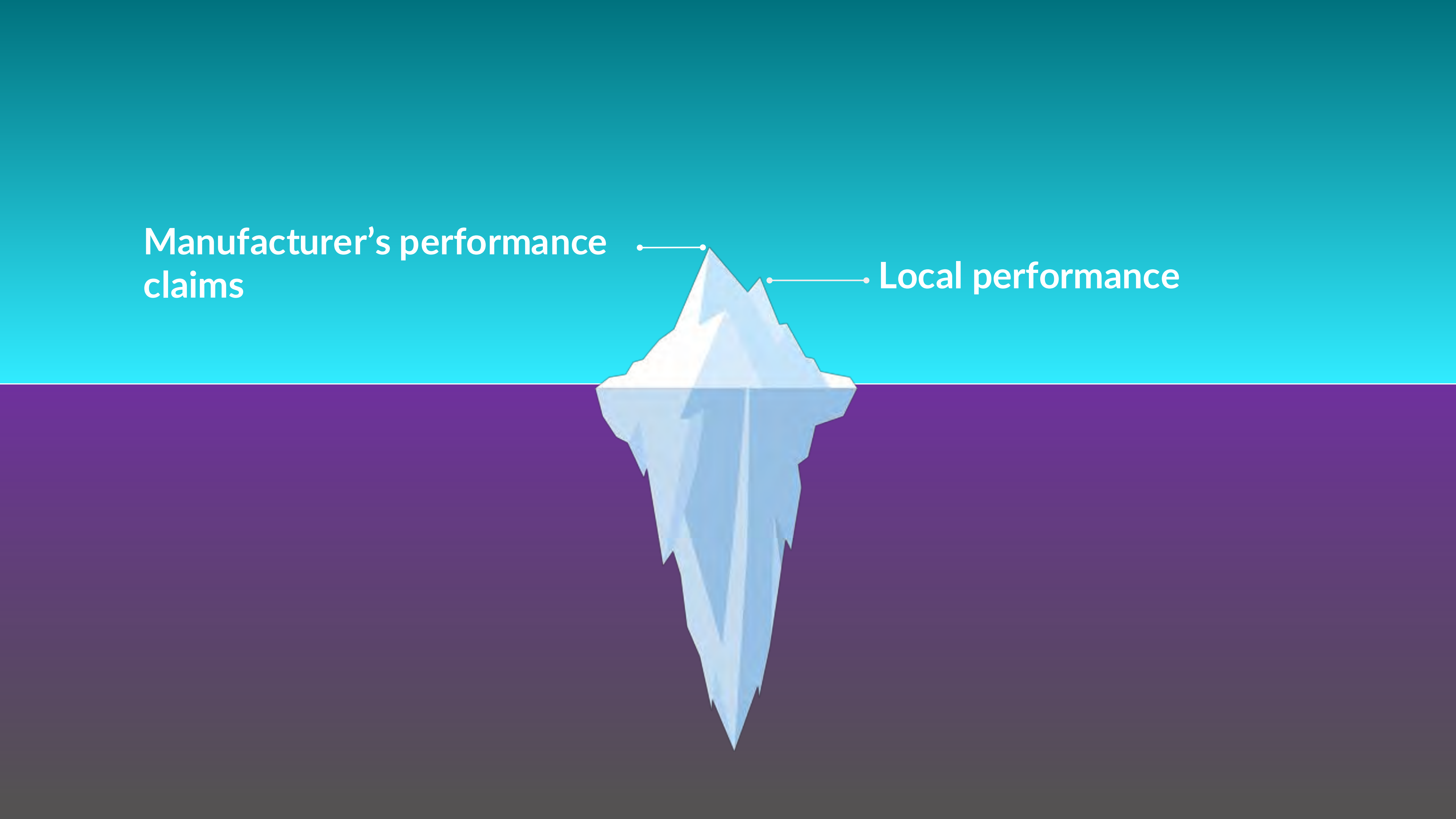
Algorithmic audit summary: Disseminated to key stakeholders relevant to local deployment including chief medical officers of both hospital and AI manufacturer.

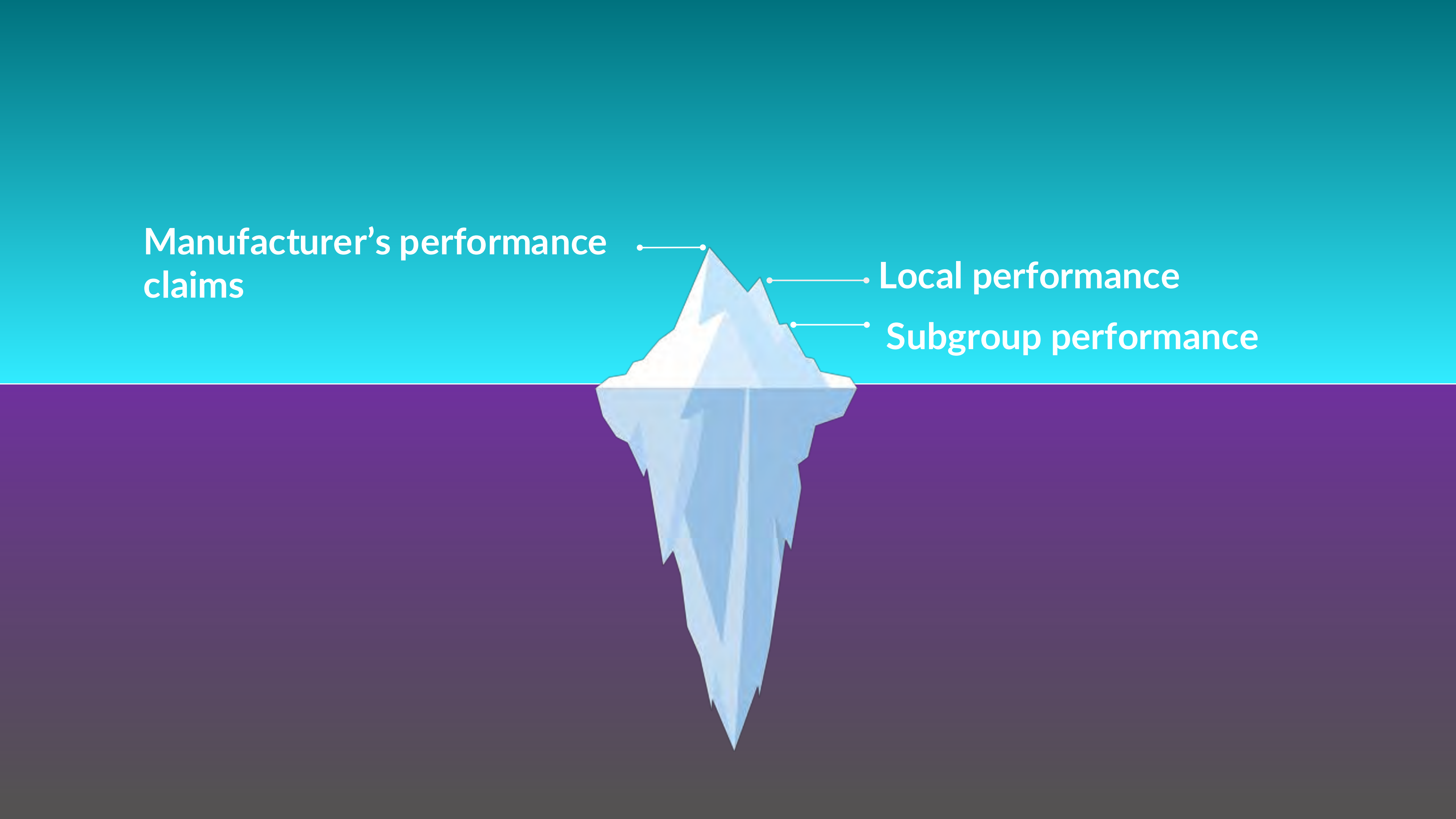
Plan re-audit: Medical Algorithmic Audit framework is being built into local governance processes

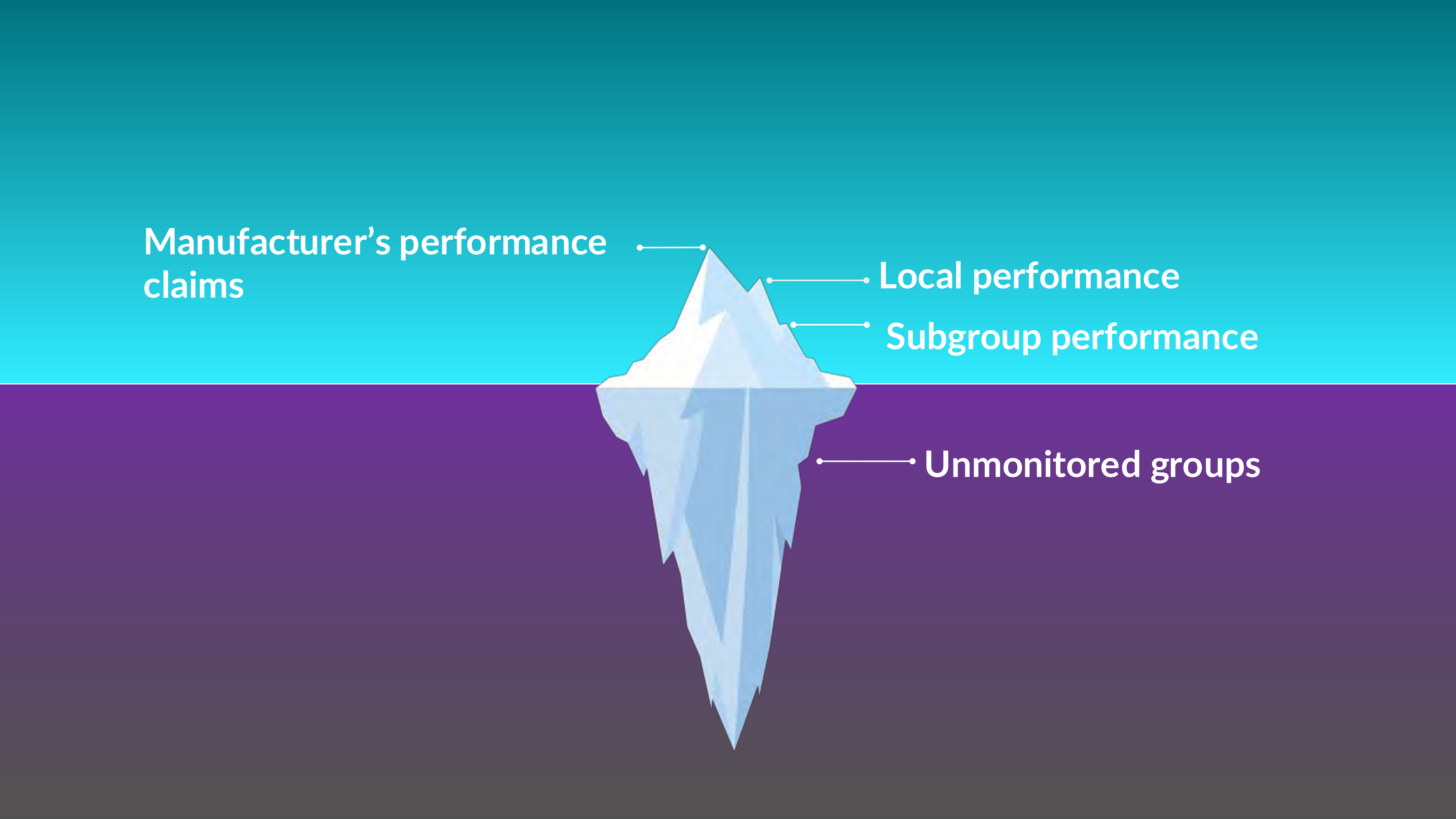
Failure Modes and Effects Analysis (FMEA)

Liu et al 2022
Kale et al

Manufacturer’s performance claims


Manufacturer’s performance claims
Local performance


Manufacturer’s performance claims
Local performance
Subgroup performance



Manufacturer's performance claims
Local performance
Subgroup performance
Unmonitored groups

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.

Science

RESEARCH ARTICLE

ECONOMICS

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

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

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

"**We show that a widely used algorithm**… affecting millions of patients, **exhibits significant racial bias**"

Obermeyer et al, Science (2019)

ARTICLES

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

nature medicine

Check for updates

OPEN

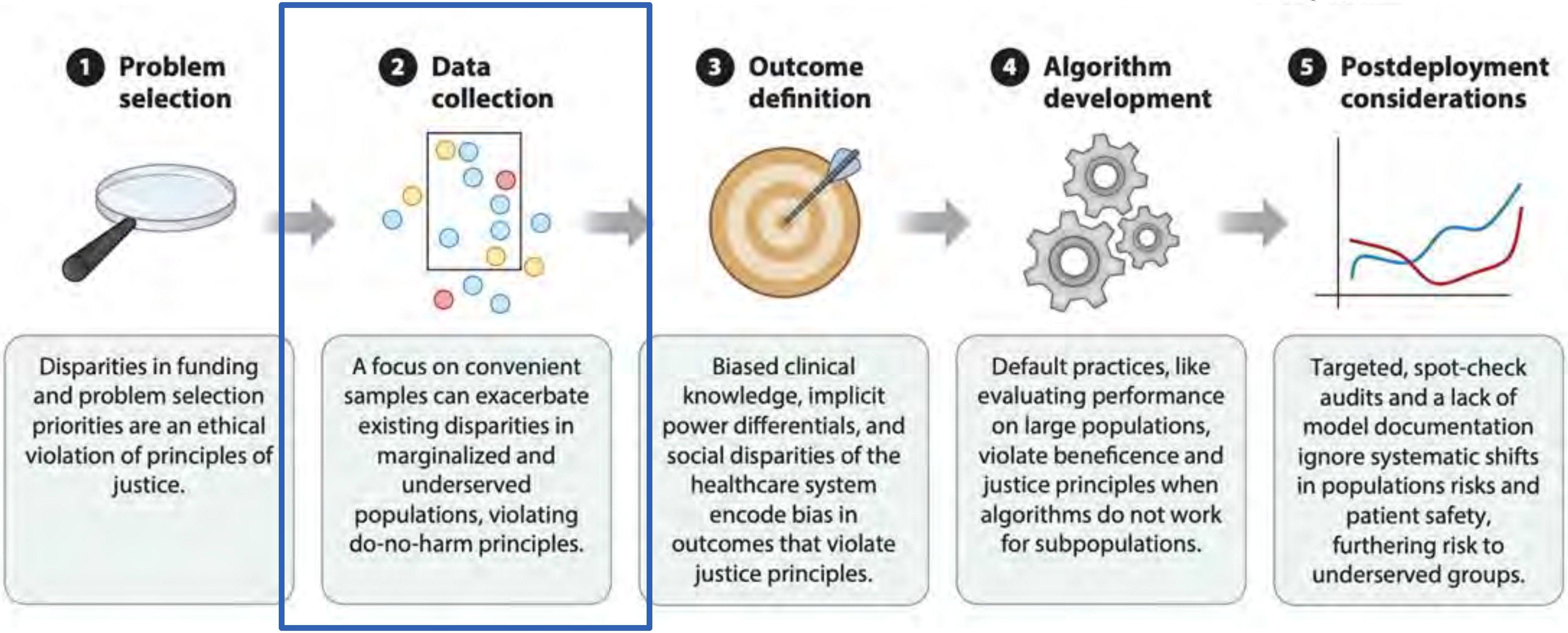
Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

"We find that **classifiers… consistently and selectively underdiagnosed under-served patient populations**"

Seyyed-Kalantari et al, Nature Medicine (2021)

Health data is one source of algorithmic bias

Annual Review of Biomedical Data Science
Ethical Machine Learning in Healthcare
Irene Y. Chen,[1] Emma Pierson,[2] Sherri Rose,[3] Shalmali Joshi,[4] Kadija Ferryman,[5] and Marzyeh Ghassemi[1,4]



Health Data Poverty

*The **inability** for*

individuals, groups, or populations to

benefit from

data-driven discoveries and innovations,

due to insufficient data

that are representative of them

THE LANCET
Digital Health

Health data poverty: an assailable barrier to equitable digital health care

Hussein Ibrahim, Xiaoxuan Liu, Nevine Zariffa, Andrew D Morris*, Alastair K Denniston*

Data-driven digital health technologies have the power to transform health care. If these tools could be sustainably delivered at scale, they might have the potential to provide everyone, everywhere, with equitable access to expert-level care, narrowing the global health and wellbeing gap. Conversely, it is highly possible that these transformative technologies could exacerbate existing health-care inequalities instead. In this Viewpoint, we describe the problem of health data poverty: the inability for individuals, groups, or populations to benefit from a discovery or innovation due to a scarcity of data that are adequately representative. We assert that health data poverty is a threat to global health that could prevent the benefits of data-driven digital health technologies from being more widely realised and might even lead to them causing harm. We argue that the time to act is now to avoid creating a digital health divide that exacerbates existing health-care inequalities and to ensure that no one is left behind in the digital era.

Lancet Digit Health 2021; 3: e260-65
Published Online March 4, 2021
https://doi.org/10.1016/S2589-7500(20)30317-4

*Joint senior authors
Centre for Regulatory Science and Innovation, Birmingham Health Partners, Birmingham,

NEJM AI

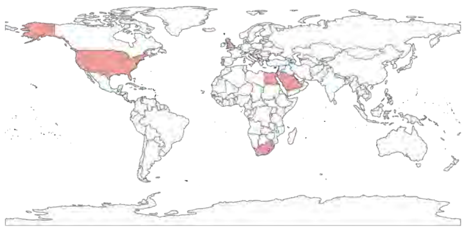
EDITORIAL

A Global Health Data Divide

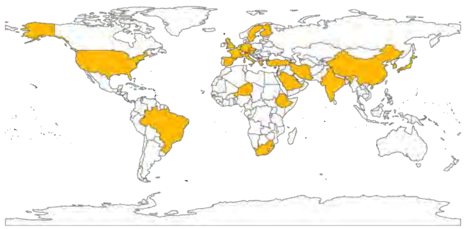
Authors: Xiaoxuan Liu, Ph.D., Joseph Alderman, M.B.Ch.B., and Elinor Laws, M.B.B.S.
Author Info & Affiliations

Published May 17, 2024 | DOI: 10.1056/AIe2400388

Mammography

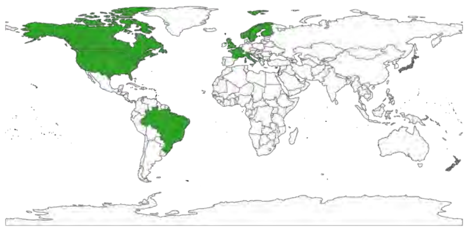


Ophthalmology

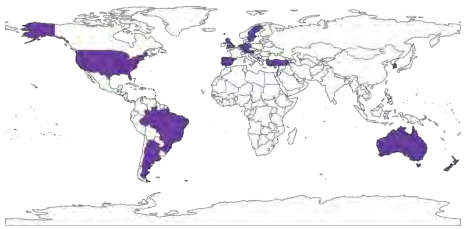


Covid-19

Heart Failure

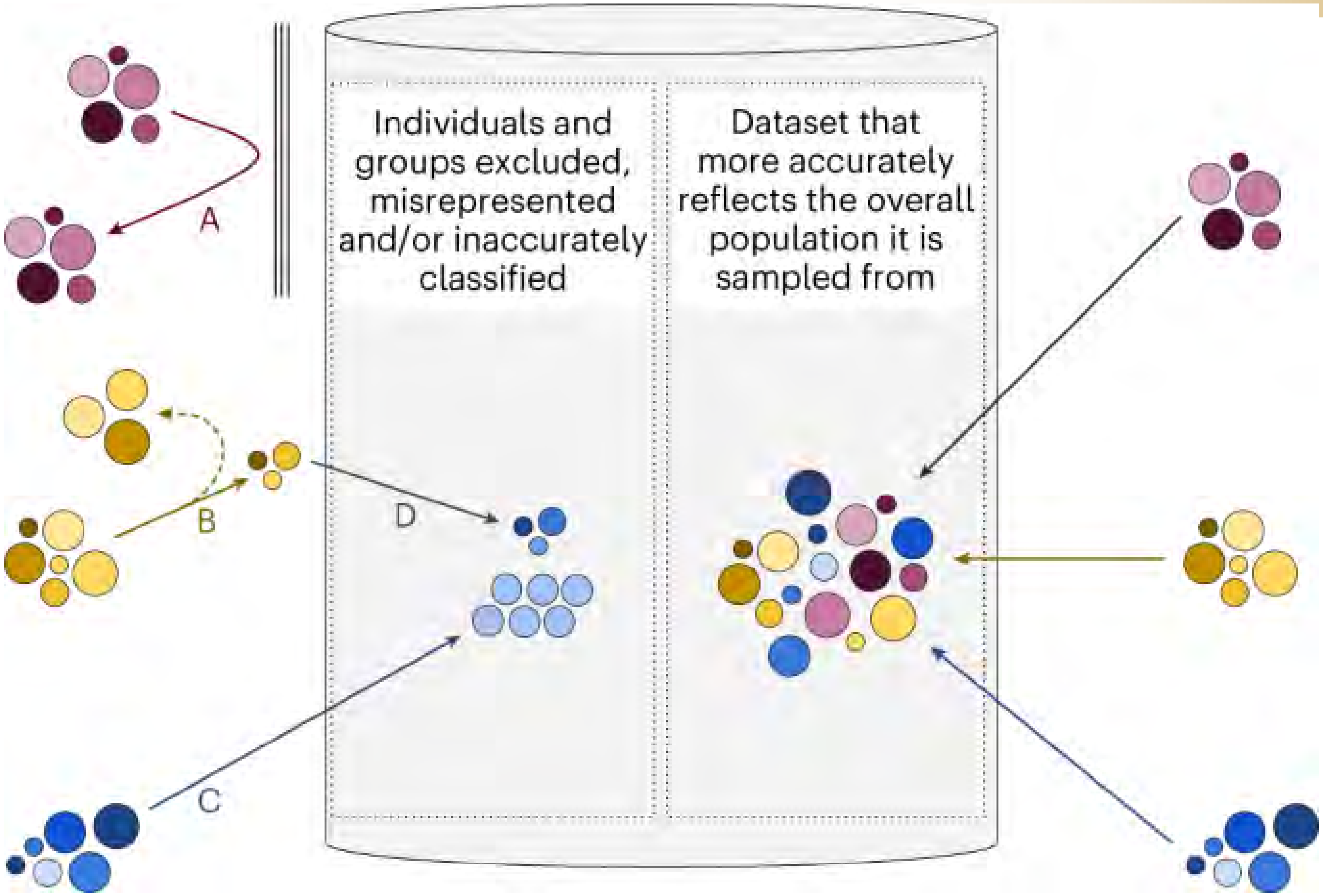


Skin cancer



Ibrahim H et al. Lancet Digital Health 2021

Liu et al. NEJM AI 2024



Arora, A., Alderman, J.E., Palmer, J…Liu X. The value of standards for health datasets in artificial intelligence-based applications.

STANDING Together

Australian Government
Department of Health and Aged Care
Therapeutic Goods Administration

Artificial Intelligence (AI) and medical device software

Information for software manufacturers about how we regulate AI medical devices.

Information about populations that this data is based on and justification for how this data would be appropriate for the Australian population and sub-populations for whom the AI is intended to be used. Independent global draft consensus standards have been developed for datasets used in health AI, which could provide a basis for structuring this information.

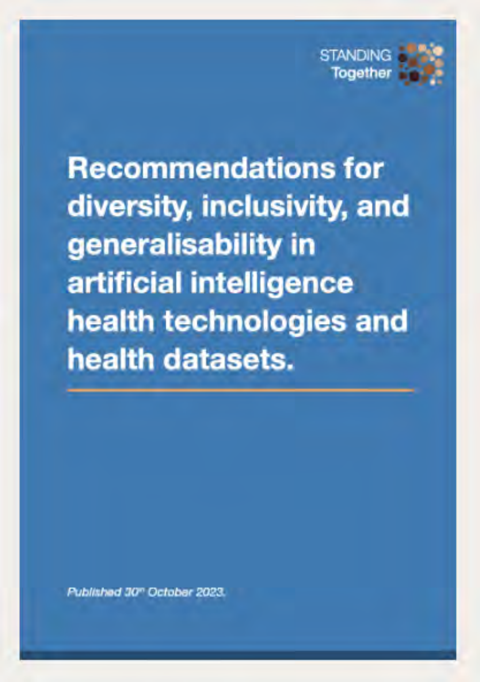
Medicines & Healthcare products Regulatory Agency

Guidance

Software and AI as a Medical Device

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.



STANDING Together

Recommendations for diversity, inclusivity, and generalisability in artificial intelligence health technologies and health datasets.

Published 30th October 2023.

Alderman, Palmer, Laws…Liu.
Lancet Digital Health 2025
NEJM-AI 2025
www.datadiversity.org

bsi.

AAMI

ISO International Organization for Standardization

World Health Organization

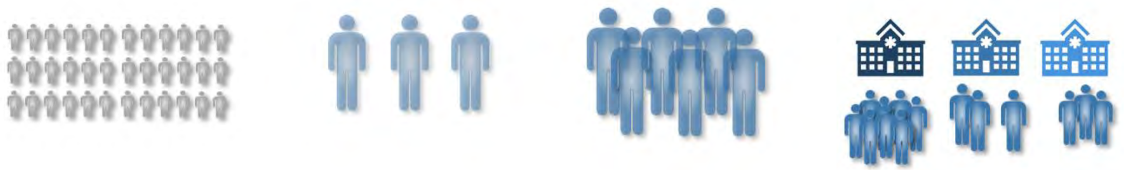
AI SAFETY SUMMIT

G20 BRASIL 2024

Equity in Medical Devices: Independent Review

Report 2023

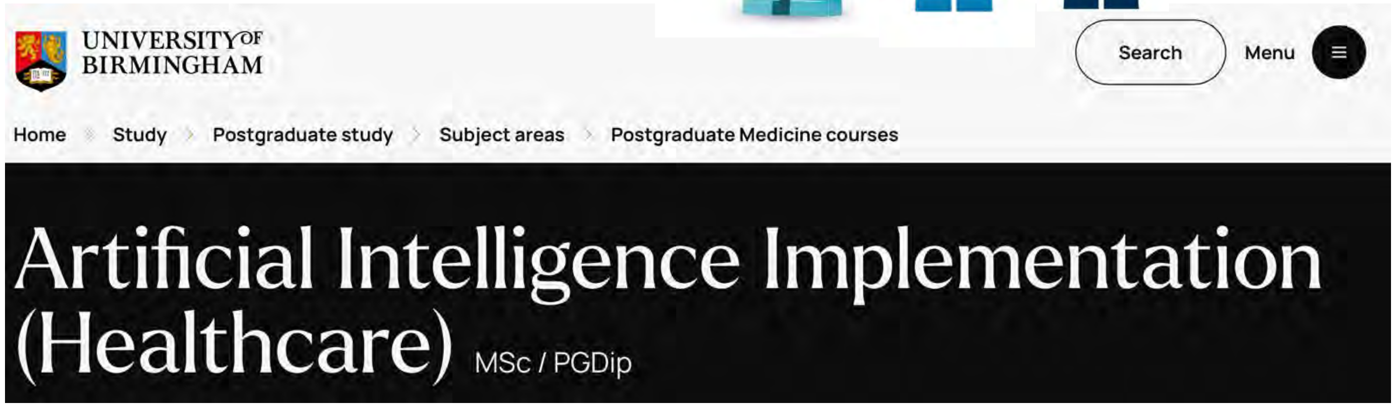
Know the journey



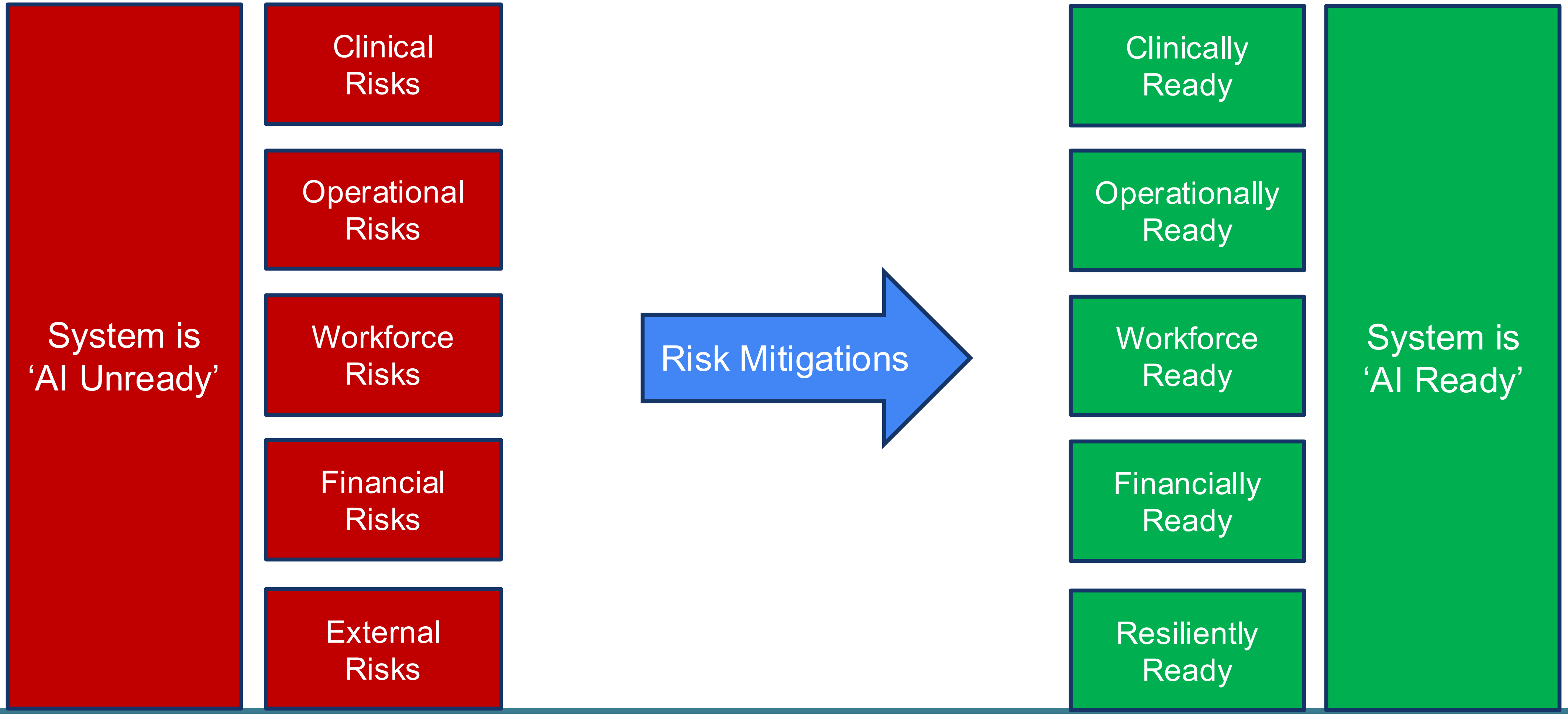
<u>Design with safety and equity in mind throughout the TPLC</u>

It’s not just about the product…

Training the workforce for digital/AI adoption



Enhancing NHS readiness for AI



Accelerating the pipeline

Design → Proof-of-concept → Clinical evaluation → Regulation → Implementation

This year, we will review regulations, and in 2026 we will publish a new regulatory framework for medical devices including AI. This will create faster, risk proportionate and more predictable routes to market. We will collaborate with AI developers, regulators and the Department for Science, Innovation and Technology through the AI Opportunities Action Plan.

UNIVERSITY OF BIRMINGHAM

NHS
University Hospitals Birmingham
NHS Foundation Trust

Thank you

Email: ai.incubator@uhb.nhs.uk

Community, seminars, workshops

AI Incubator for AI & Digital Healthcare

Courses in AI Implementation

Making regulation ‘AI-ready’

CERSI AI

www.cersi-ai.org