Como desmostrar que la IA que uso es robusta?

Prof. Karim Lekadir
ICREA Research Professor
Universitat de Barcelona
Artificial Intelligence in Medicine Lab

EU Projects

Project	Dates	Role	Field
euCanSHare	2018-2023	Coordinator	Cardiology
EarlyCause	2020-2024	Coordinator	Mental health
EuCanImage	2020-2025	Coordinator	Oncology
RadioVal	2022-2026	Coordinator	Oncology
AIMIX (ERC)	2023-2027	PI	Obstetrics
DataTools4Heart	2022-2026	Coordinator	Cardiology
AFRICAI-RI	2025-2029	Coordinator	Pulmonology

Content

1.3 Objectives

⚠ *If you plan to use, develop and/or deploy artificial intelligence (AI) based systems and/or techniques you must demonstrate their technical robustness. AI-based systems or techniques should be, or be developed to become:*

- *technically robust, accurate and reproducible, and able to deal with and inform about possible failures, inaccuracies and errors, proportionate to the assessed risk they pose*
- *socially robust, in that they duly consider the context and environment in which they operate*
- *reliable and function as intended, minimizing unintentional and unexpected harm, preventing unacceptable harm and safeguarding the physical and mental integrity of humans*
- *able to provide a suitable explanation of their decision-making processes, whenever they can have a significant impact on people's lives.*

Terminology

- Technically & societally robust AI
- Trustworthy AI
- Ethical AI
- Responsible AI
- Human-centred AI

Part 1 - What is trustworthy AI?

Part 2 - How do we achieve it?

Part 3 - Trustworthy AI in an EIC

Part 1 - What is trustworthy AI?

Part 2 - How do we achieve it?

Part 3 - Trustworthy AI in an EIC

AI in Healthcare

BBC Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

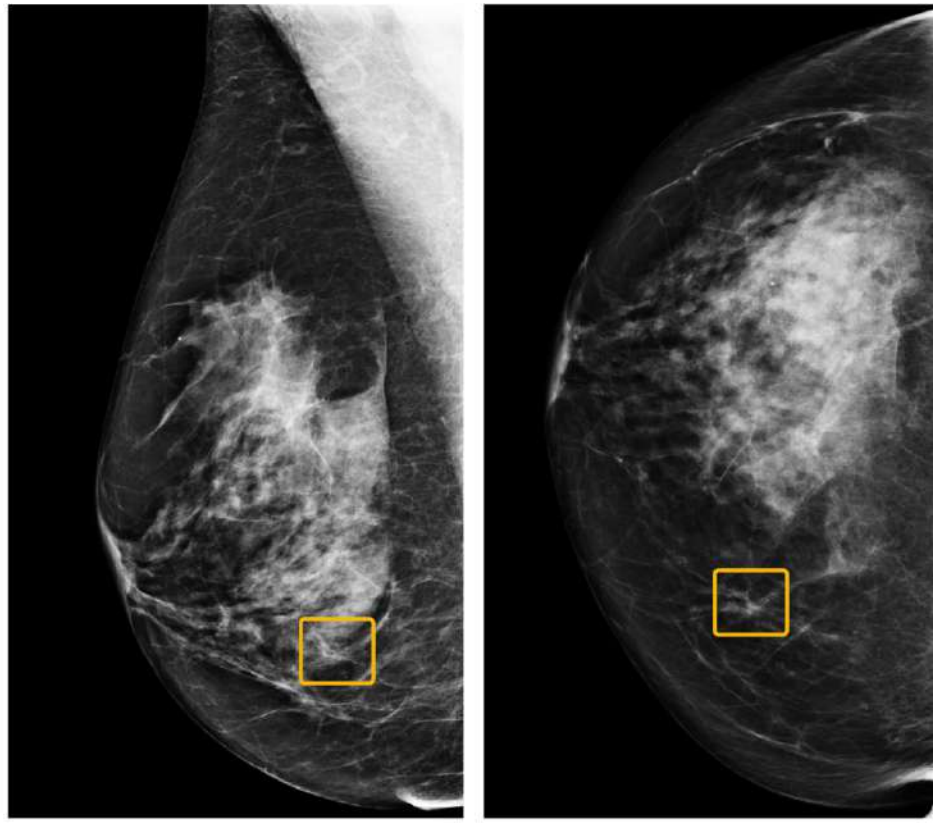
Home | Coronavirus | Climate | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts

Health | Coronavirus

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh

2 January 2020



Tumours missed by 6 clinicians

Detected by the AI tool

Robustness

nature

Explore content ▾ About the journal ▾ Publish with us ▾

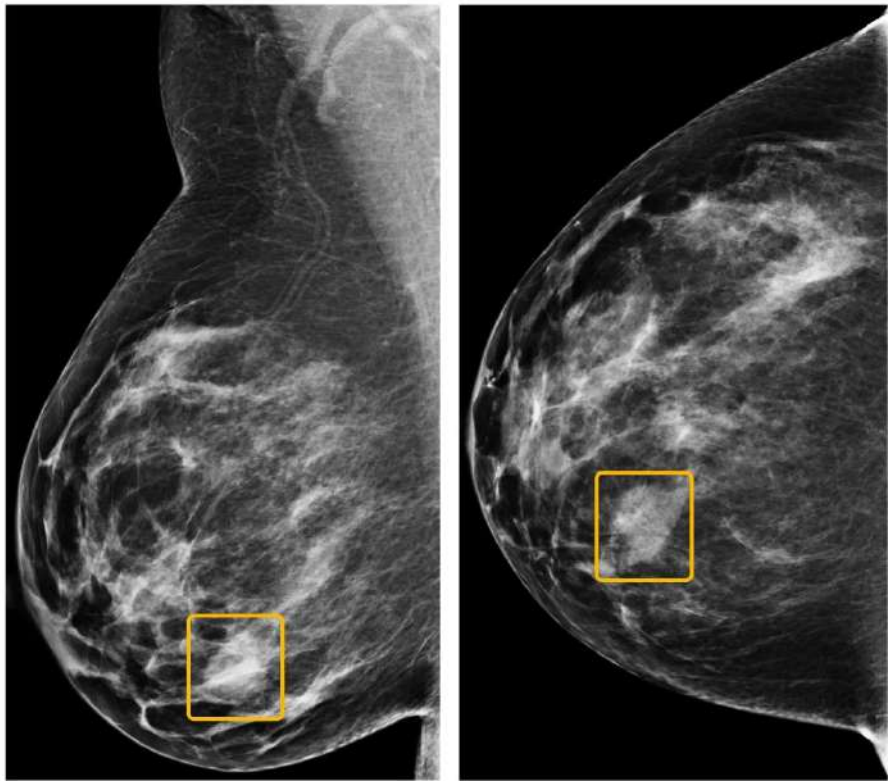
nature › articles › article

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney ✉, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse ✉, Kenneth C. Young, Jeffrey De Fauw & Shravya Shetty ✉ — Show fewer authors

Nature **577**, 89–94 (2020) | Cite this article



Tumours detected by 6 clinicians

Missed by the AI tool

Transparency

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature › matters arising › article

Matters Arising | Published: 14 October 2020

Transparency and reproducibility in artificial intelligence

Benjamin Haibe-Kains ✉, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush & Hugo J. W. L. Aerts

Nature **586**, E14–E16 (2020) | Cite this article

"The lack of details of the methods and algorithm code undermines its scientific value"

Universality

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature › articles › article

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney ✉, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona J. Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher J. Kelly, Dominic King, Joseph R. Ledsam, David Melnick, Hormuz Mostofi, Lily Peng, Joshua Jay Reicher, Bernardino Romera-Paredes, Richard Sidebottom, Mustafa Suleyman, Daniel Tse ✉, Kenneth C. Young, Jeffrey De Fauw & Shravya Shetty ✉ — Show fewer authors

Nature **577**, 89–94 (2020) | Cite this article

Test datasets

		
Number of women	25,856	3,097
Interpretation	Double reading	Single reading
Screening interval	3 years	1 or 2 years
Cancer follow-up	39 months	27 months
Number of cancers	414 (1.6%)	686 (22.2%)
AI vs. Doctor(s)	**Equivalent**	**Superior**

Fairness

ARTICLES
https://doi.org/10.1038/s41591-021-01595-0
nature medicine
Check for updates

OPEN

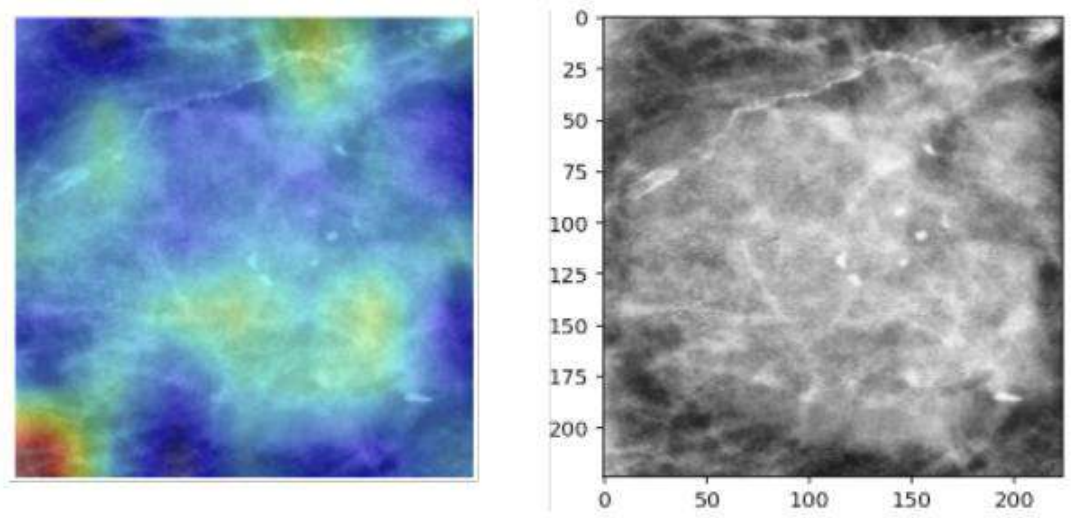
Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

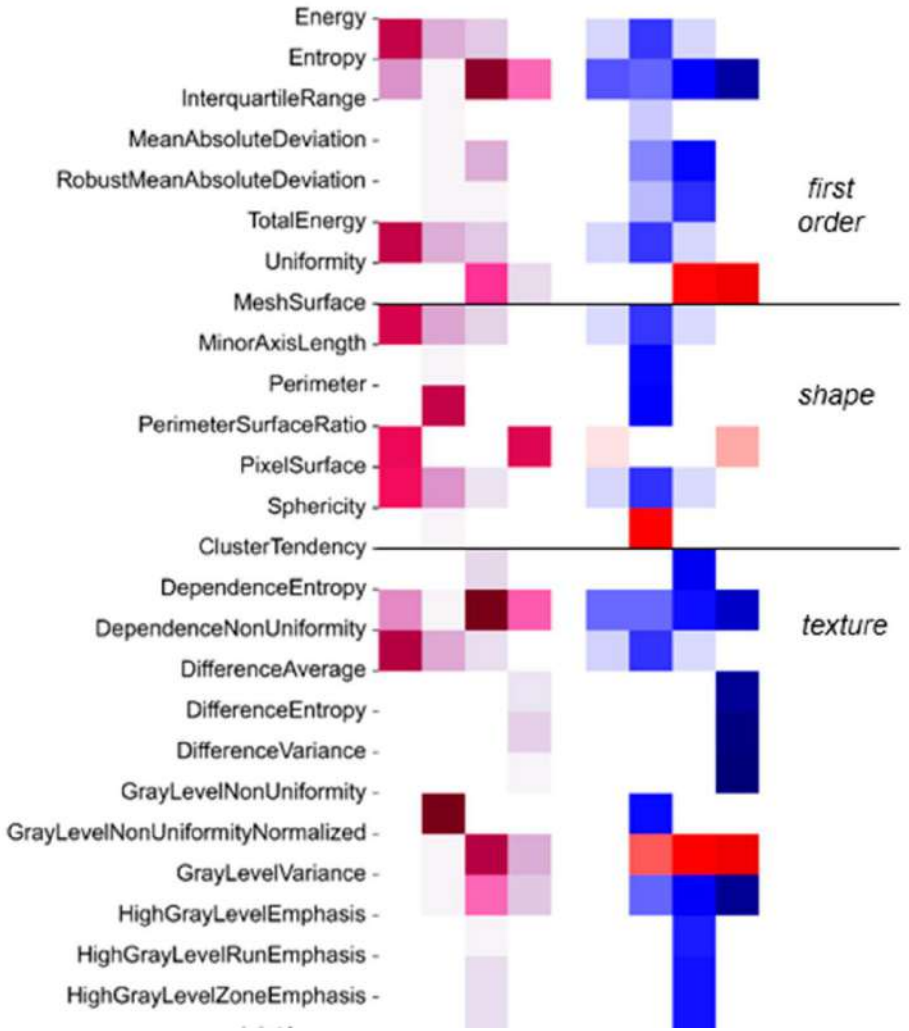
Laleh Seyyed-Kalantari[1,2], Haoran Zhang[3], Matthew B. A. McDermott[3], Irene Y. Chen[3] and Marzyeh Ghassemi[2,3]

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 inaccurately 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.

Attribute	Dataset	Average Cross-Label Gap	Lowest	Greatest	Unfavorable	Favorable
Sex	ALL	**0.045**	Ef:0.001	Pa:0.105	Female (4/7)	Male (4/7)
	CXP	0.062	Ed:0.000	Co:0.139	Female (7/13)	Male (7/13)
	CXR	0.072	Ed:0.011	EC:0.151	Female (10/13)	Male (10/13)
	NIH	0.190	M:0.001	Cd:0.393	Female (8/14)	Male (8/14)
Age	ALL	**0.215**	Ef:0.115	NF:0.444	0-20 (5/7)	40-60,60-80(5/7)
	CXR	0.245	SD:0.091	Cd:0.440	0-20, 20-40 (7/13)	60-80 (10/13)
	CXP	0.270	SD:0.084	NF:0.604	0-20, 20-40, 80- (7/13)	40-60 (8/13)
	NIH	0.413	In:0.188	Em:1.00	60-80 (7/14)	20-40 (9/14)
Race	CXR	0.226	NF:0.119	Pa:0.440	Hispanic (9/13)	White (9/13)
Insurance	CXR	0.100	SD:0.021	PO:0.190	Medicaid (10/13)	Other (10/13)

Explainability





Nowakowska, S., et al. 2024. Bioengineering, 11(6), p.556.

Traceability

Article

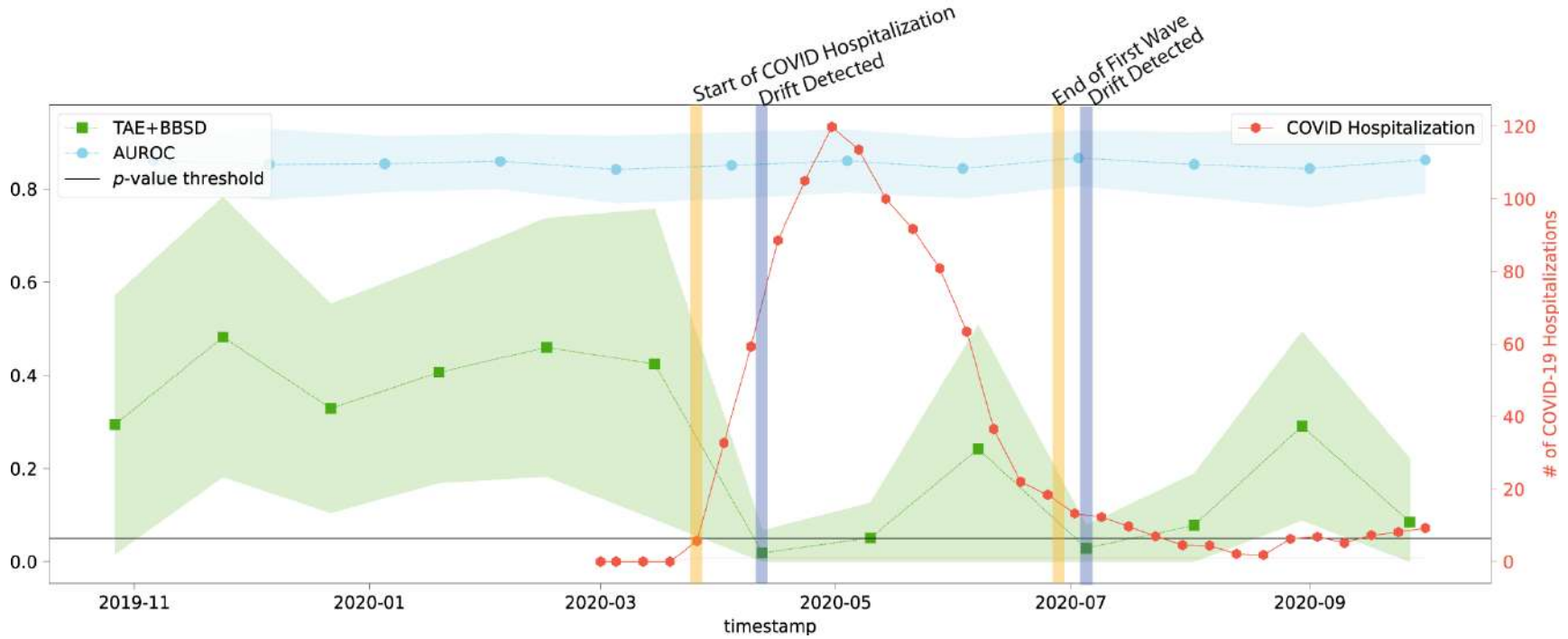
https://doi.org/10.1038/s41467-024-46142-w

Empirical data drift detection experiments on real-world medical imaging data

Received: 31 July 2023

Accepted: 14 February 2024

Ali Kore[1], Elyar Abbasi Bavil[2], Vallijah Subasri[3], Moustafa Abdalla[4], Benjamin Fine[5,6], Elham Dolatabadi[1,7] & Mohamed Abdalla[5]



Usability

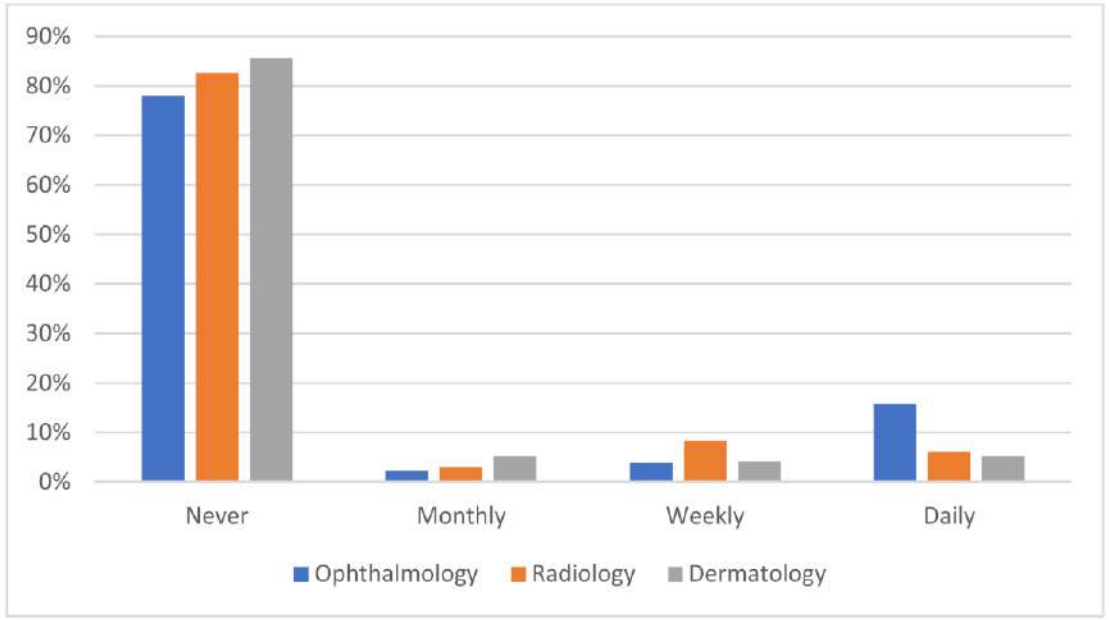
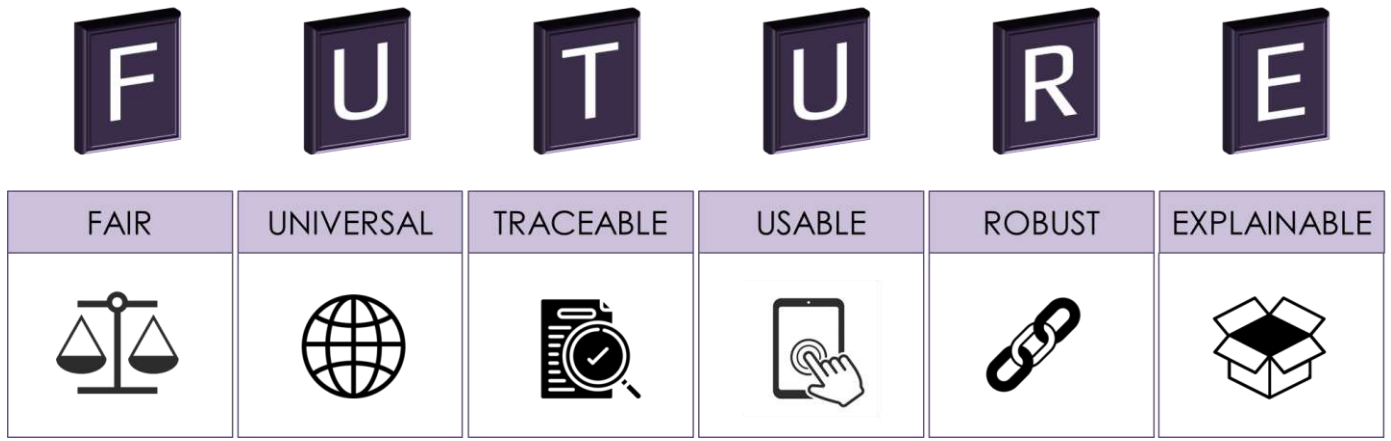


Figure 2. Current frequency of artificial intelligence use in clinical practice.

Scheetz, J. et al. 2021, A survey of clinicians on the use of AI, Scientific reports, 11(1), p.5193.

Characteristics Trustworthy AI

- Robustness
- Universality
- Fairness
- Explainability
- Traceability
- Usability



Characteristics Trustworthy AI

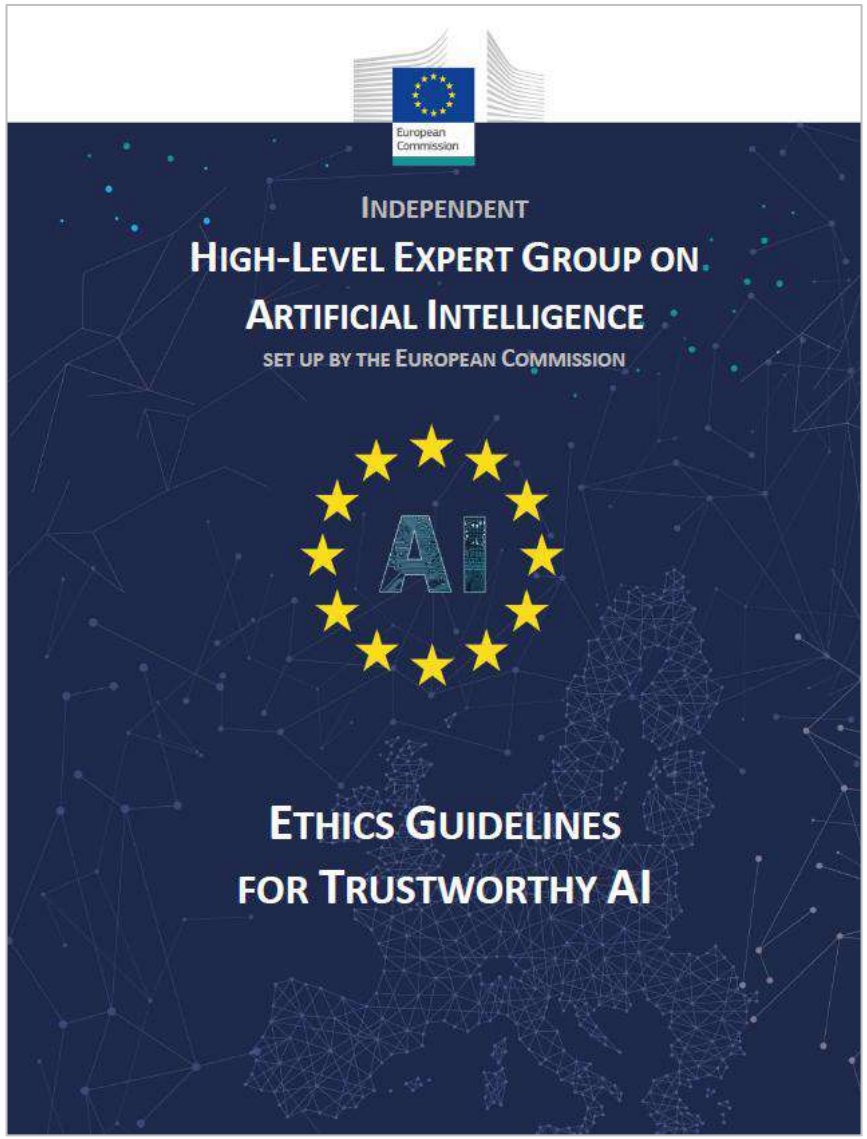
Based on ethical principles and fundamental rights:

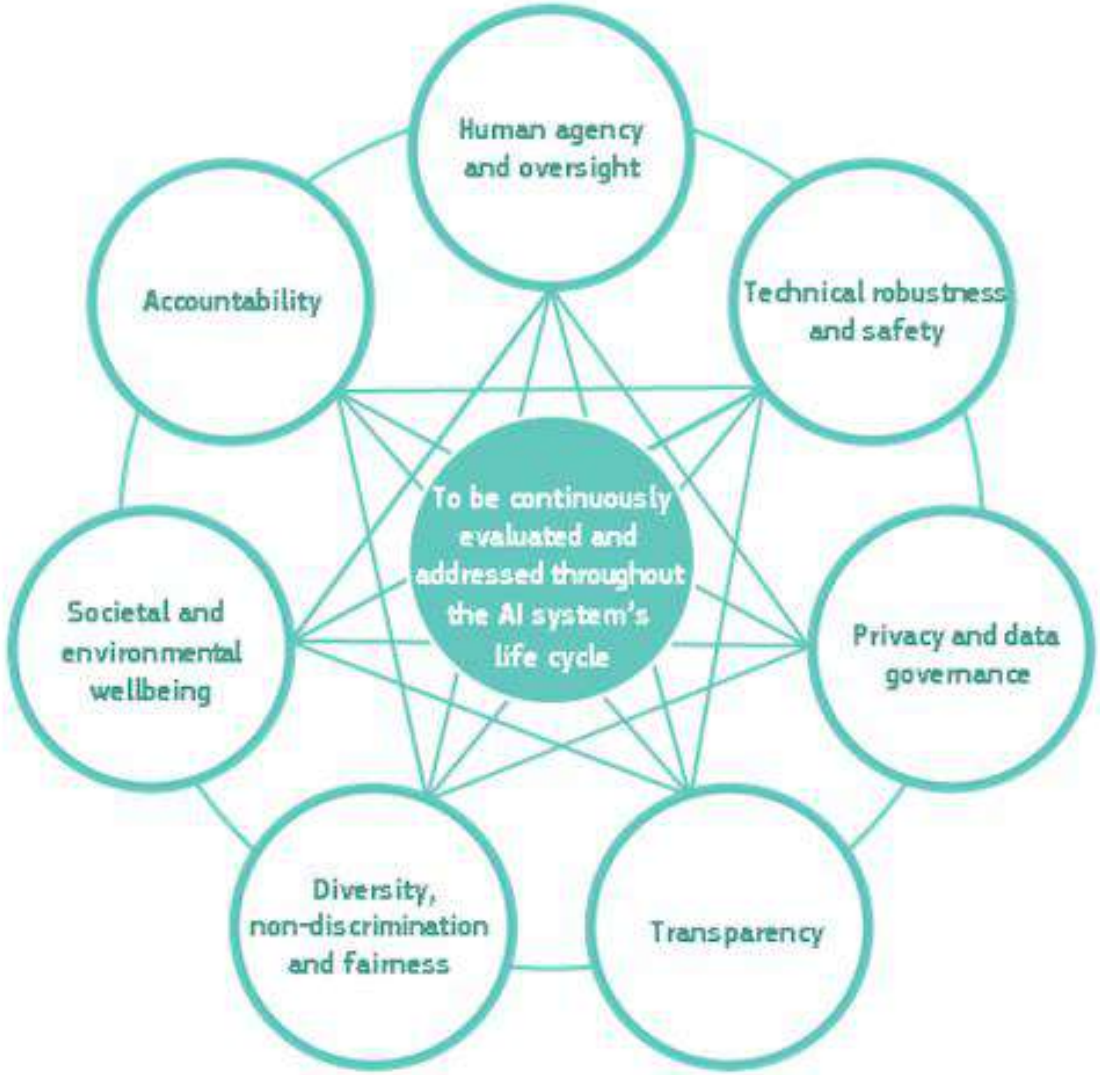
FAIRNESS		Right to non-discrimination
UNIVERSALITY		Right to equity
TRACEABILITY		Right to accountability
USABILITY		Right to autonomy
ROBUSTNESS		Right to safety
EXPLAINABILITY		Right to transparency

Part 1 - What is trustworthy AI?

Part 2 - How do we achieve it?

Part 3 - Trustworthy AI in an EIC


European Commission
INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION
AI
ETHICS GUIDELINES FOR TRUSTWORTHY AI


Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Karim Lekadir,[1,2] Alejandro F Frangi,[3,4] Antonio R Porras,[5] Ben Glocker,[6] Celia Cintas,[7] Curtis P Langlotz,[8] Eva Weicken,[9] Folkert W Asselbergs,[10,11] Fred Prior,[12] Gary S Collins,[13] Georgios Kaissis,[14] Gianna Tsakou,[15] Irène Buvat,[16] Jayashree Kalpathy-Cramer,[17] John Mongan,[18] Julia A Schnabel,[19] Kaisar Kushibar,[1] Katrine Riklund,[20] Kostas Marias,[21] Lameck M Amugongo,[22] Lauren A Fromont,[23] Lena Maier-Hein,[24] Leonor Cerdá-Alberich,[25] Luis Martí-Bonmatí,[26] M Jorge Cardoso,[27] Maciej Bobowicz,[28] Mahsa Shabani,[29] Manolis Tsiknakis,[21] Maria A Zuluaga,[30] Marie-Christine Fritzsche,[31] Marina Camacho,[1] Marius George Linguraru,[32] Markus Wenzel,[9] Marleen De Bruijne,[33] Martin G Tolsgaard,[34] Melanie Goisauf,[35] Mónica Cano Abadía,[35] Nikolaos Papanikolaou,[36] Noussair Lazrak,[1] Oriol Pujol,[1] Richard Osuala,[1] Sandy Napel,[37] Sara Colantonio,[38] Smriti Joshi,[1] Stefan Klein,[33] Susanna Aussó,[39] Wendy A Rogers,[40] Zohaib Salahuddin,[41] Martijn P A Starmans[33]; on behalf of the FUTURE-AI Consortium

For numbered affiliations see end of the article

Correspondence to: K Lekadir karim.lekadir@ub.edu (ORCID 0000-0002-9456-1612)

Additional material is published online only. To view please visit the journal online.

Cite this as: BMJ 2025;388:e081554 http://dx.doi.org/10.1136/bmj-2024-081554

Accepted: 10 January 2025

Despite major advances in artificial intelligence (AI) research for healthcare, the deployment and adoption of AI technologies remain limited in clinical practice. This paper describes the FUTURE-AI framework, which provides guidance for the development and deployment of trustworthy AI tools in healthcare. The FUTURE-AI Consortium was founded in 2021 and comprises 117 interdisciplinary experts from 50 countries representing all continents, including AI scientists, clinical researchers, biomedical ethicists, and social scientists. Over a two year period, the FUTURE-AI guideline was established through consensus based on six guiding principles—fairness, universality, traceability, usability, robustness, and explainability. To operationalise trustworthy AI in healthcare, a set of 30 best practices were defined, addressing technical, clinical, socioethical, and legal dimensions. The recommendations cover the entire lifecycle of healthcare AI, from design, development, and validation to regulation, deployment, and monitoring.

SUMMARY POINTS

- Despite major advances in medical artificial intelligence (AI) research, clinical adoption of emerging AI solutions remains challenging owing to limited trust and ethical concerns
- The FUTURE-AI Consortium unites 117 experts from 50 countries to define international guidelines for trustworthy healthcare AI
- The FUTURE-AI framework is structured around six guiding principles: fairness, universality, traceability, usability, robustness, and explainability
- The guideline addresses the entire AI lifecycle, from design and development to validation and deployment, ensuring alignment with real world needs and ethical requirements
- The framework includes 30 detailed recommendations for building trustworthy and deployable AI systems, emphasising multistakeholder collaboration
- Continuous risk assessment and mitigation are fundamental, addressing biases, data variations, and evolving challenges during the AI lifecycle
- FUTURE-AI is designed as a dynamic framework, which will evolve with technological advancements and stakeholder feedback

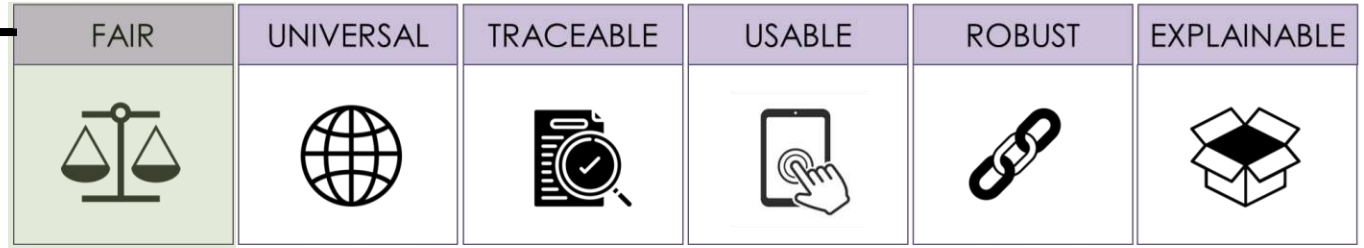
Introduction

In the field of healthcare, artificial intelligence (AI)—that is, algorithms with the ability to self-learn logic—and data interactions have been increasingly used to develop computer aided models, for example, disease diagnosis, prognosis, prediction of therapy response or survival, and patient stratification.[1] Despite major advances, the deployment and adoption of AI technologies remain limited in real world clinical practice. In recent years, concerns have been raised about the technical, clinical, ethical, and societal risks associated with healthcare AI.[2 3] In particular, existing research has shown that AI tools in healthcare can be prone to errors and patient harm, biases and increased health inequalities, lack of transparency and accountability, as well as data privacy and security breaches.[4-6]

To increase adoption in the real world, it is essential that AI tools are trusted and accepted by patients, clinicians, health organisations, and authorities. However, there is an absence of clear, widely accepted guidelines on how healthcare AI tools should be designed, developed, evaluated, and deployed to be trustworthy—that is, technically robust, clinically safe,

FAIR	UNIVERSAL	TRACEABLE	USABLE	ROBUST	EXPLAINABLE
					

Operationalisation



Recommendations	Research	Deployable
Fairness		
1. Define any potential sources of bias from an early stage	++	++
2. Collect information on individuals' and data attributes	+	+
3. Evaluate potential biases and, when needed, bias correction measures	+	++

Recommendations	Operations	Examples
Define any potential sources of bias (fairness 1)	Engage relevant stakeholders to define the sources of bias	Patients, clinicians, epidemiologists, ethicists, social carers[97 98]
	Define standard attributes that might affect the AI tool's fairness	Sex, age, socioeconomic status[99]
	Identify application specific sources of bias beyond standard attributes	Skin colour for skin cancer detection,[100 101] breast density for breast cancer detection[34]
	Identify all possible human biases	Data labelling, data curation[99]

Methodology

- Engage stakeholders (e.g. patients, clinicians, ethicists, …)
- Understand all potential issues, risks, needs, ethical issues, etc:
  1. User requirements (incl. intended use)
  2. Intended clinical settings
  3. Sources of errors, data variations
  4. Sources of bias
  5. Explainability options
  6. Sources of performance degradation
  7. Application-specific ethical risks (e.g. in children)
- Implement suitable appropriate AI methods/mitigation measures
- Evaluate all dimensions of trustworthy AI, then iterate if needed
- Report everything, including benefits and limitations

Usability

- Understand intended use and user requirements, preferences, etc
- Accordingly, implement user-centred solutions
- Examples: User-friendly human-AI interfaces
- Then, evaluate usability measures
- Examples: User satisfaction through questionnaires

Universality

- Understand intended clinical settings and variations.
- Examples: High-income country, low-income country, big hospital, rural clinic.
- Evaluate applicability across the intended settings.
- Then, apply mitigation measures
- Examples: Train and test with data from multiple clinical centres.

Robustness

- Understand potential sources of errors or failures
- Examples: Noise, motion of the patient during scanning, low- quality equipment
- Then evaluate robustness and identify potential issues.
- Then apply mitigation measures, if necessary.
- Examples: Noise removal, data harmonisation, human oversight.

Fairness

- Understand sources of bias
- Examples: Sex, age, ethnicity, socio-economics
- Then evaluate potential AI biases
- If necessary, apply mitigation measures
- Examples: Data re-sampling, equalised odds techniques

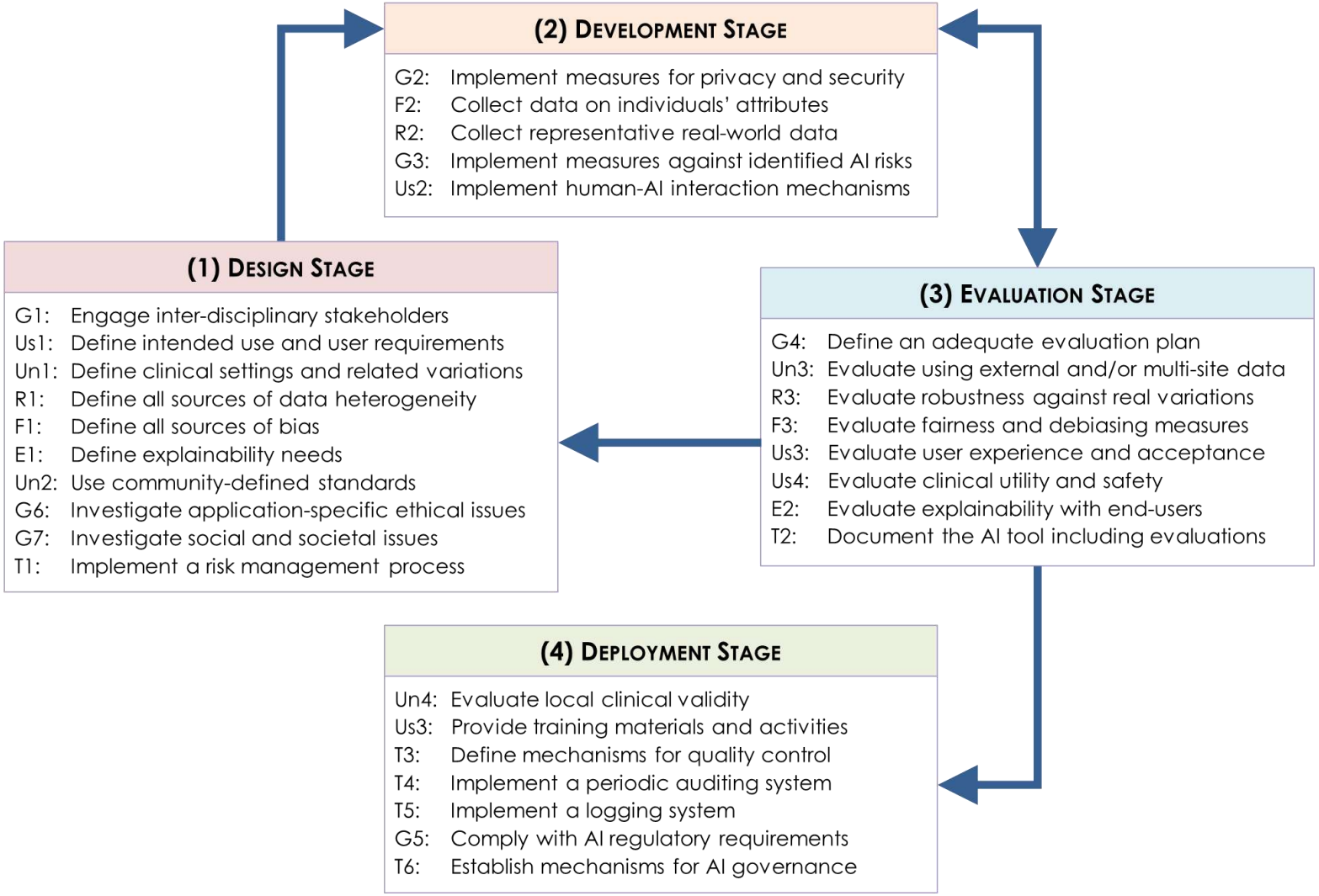
Explainability

- Understand more suitable explainability options
- Example: Image heatmaps, feature importance, counterfactuals
- Evaluate the explanations with end-users
- If necessary, apply mitigation measures
- Examples: Use a different explainable AI method.

Traceability

- Understand sources of performance degradation over time
- Examples: Change in imaging protocols
- Accordingly, implement specific mechanisms for monitoring the AI tool over time
- Examples: Quality control of input data, yearly evaluation, model recalibration

Through AI Lifecycle



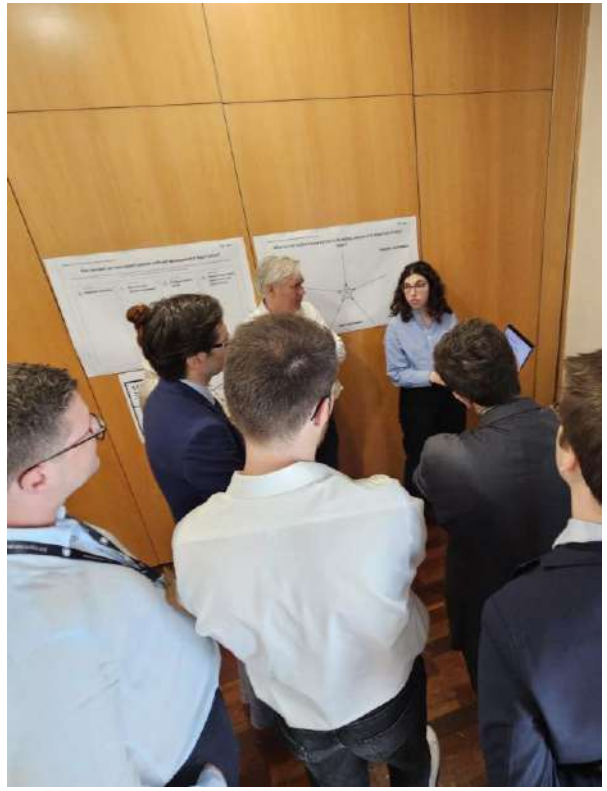
Best Practices – Design

G1:	Engage inter-disciplinary stakeholders	(1)
Us1:	Define intended use and user requirements	(2)
Un1:	Define clinical settings and related variations	(3)
R1:	Define all sources of data heterogeneity	(4)
F1:	Define all sources of bias	(5)
E1:	Define explainability needs	(6)
Un2:	Use community-defined standards	(7)
G6:	Investigate application-specific ethical issues	(8)
G7:	Investigate social and societal issues	(9)
T1:	Define a risk management process	(10)

Stakeholder Engagement

Best practice (What)	Practical steps (How)	Examples (References)
Engage inter-disciplinary stakeholders (General 1)	Identify all relevant stakeholders	Patients, GPs, nurses, ethicists, data managers (78,79)
	Provide information on the AI tool and AI	Educational seminars, training materials, webinars (80)
	Set up communication channels with stakeholders	Regular group meetings, one-to-one interviews, virtual platform (81)
	Organise co-creation consensus meetings	One-day co-creation workshop with *n*=15 multi-disciplinary stakeholders (82)
	Use qualitative methods to gather feedback	Online surveys, focus groups, narrative interviews (83)

Stakeholder Engagement

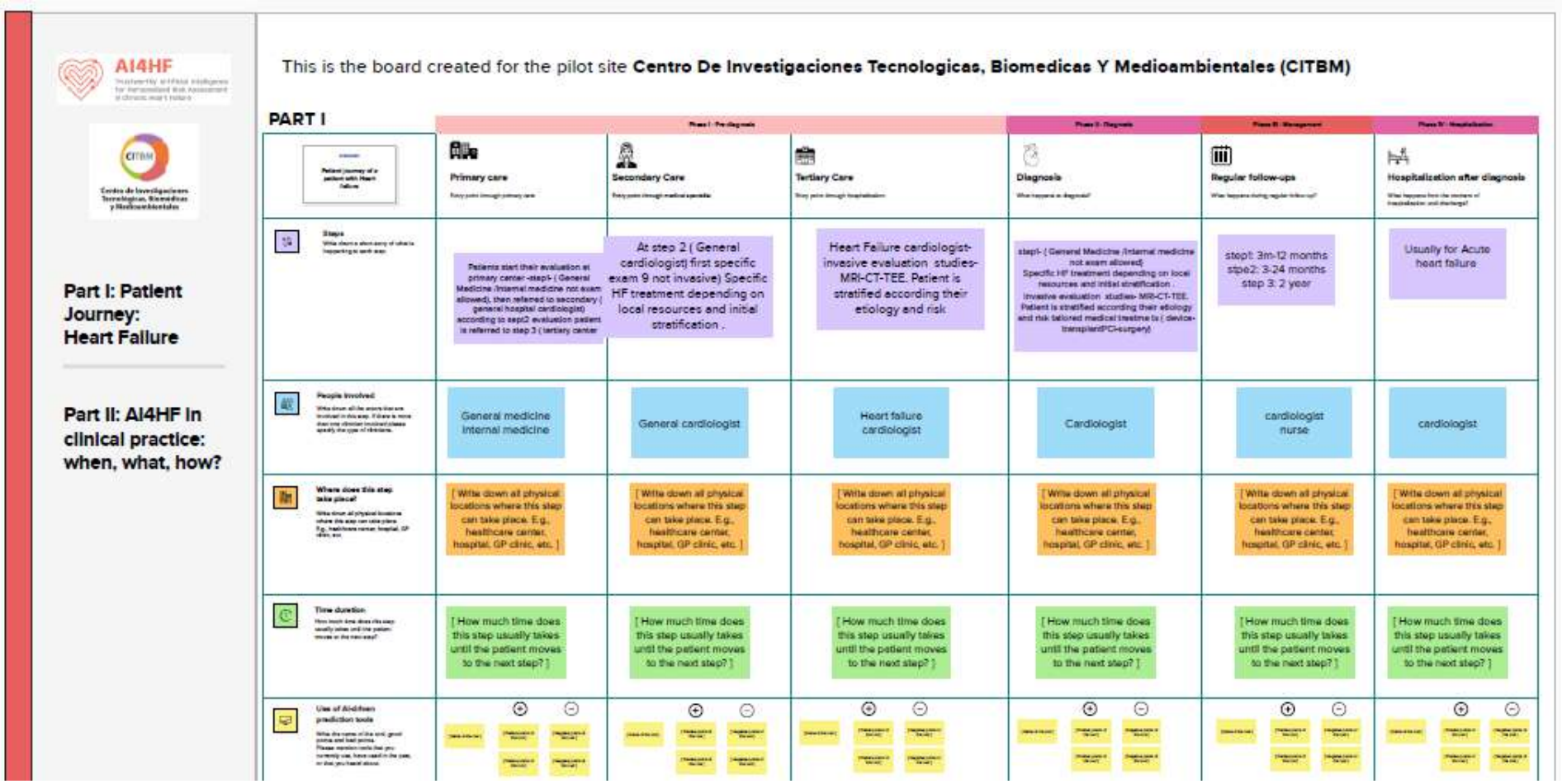


AI4HF

Stakeholder Engagement

AIMIX

Stakeholder Engagement



AI4HF

Part I: Patient Journey: Heart Failure

Part II: AI4HF in clinical practice: when, what, how?

This is the board created for the pilot site **Centro De Investigaciones Tecnologicas, Biomedicas Y Medioambientales (CITBM)**

PART I

	Primary care	Secondary Care	Tertiary Care	Diagnosis	Regular follow-ups	Hospitalization after diagnosis
Steps	Patients start their evaluation at primary center -step1- (General Medicine /Internal medicine not exam allowed), then referred to secondary (general hospital cardiologist) according to sept2 evaluation patient is referred to step 3 (tertiary center	At step 2 (General cardiologist) first specific exam 9 not invasive) Specific HF treatment depending on local resources and initial stratification .	Heart Failure cardiologist- invasive evaluation studies- MRI-CT-TEE. Patient is stratified according their etiology and risk	step1- (General Medicine /Internal medicine not exam allowed) Specific HF treatment depending on local resources and initial stratification . Invasive evaluation studies- MRI-CT-TEE. Patient is stratified according their etiology and risk tailored medical treatme ts (device-transplantPCI-surgery)	step1: 3m-12 months step2: 3-24 months step 3: 2 year	Usually for Acute heart failure
People involved	General medicine internal medicine	General cardiologist	Heart failure cardiologist	Cardiologist	cardiologist nurse	cardiologist
Where does this step take place?	[Write down all physical locations where this step can take place. E.g., healthcare center, hospital, GP clinic, etc.]	[Write down all physical locations where this step can take place. E.g., healthcare center, hospital, GP clinic, etc.]	[Write down all physical locations where this step can take place. E.g., healthcare center, hospital, GP clinic, etc.]	[Write down all physical locations where this step can take place. E.g., healthcare center, hospital, GP clinic, etc.]	[Write down all physical locations where this step can take place. E.g., healthcare center, hospital, GP clinic, etc.]	[Write down all physical locations where this step can take place. E.g., healthcare center, hospital, GP clinic, etc.]
Time duration	[How much time does this step usually takes until the patient moves to the next step?]	[How much time does this step usually takes until the patient moves to the next step?]	[How much time does this step usually takes until the patient moves to the next step?]	[How much time does this step usually takes until the patient moves to the next step?]	[How much time does this step usually takes until the patient moves to the next step?]	[How much time does this step usually takes until the patient moves to the next step?]
Use of AI-driven prediction tools						

Stakeholder Engagement

RadioVal FUTURE AI Survey

Time required: 45 minutes

Goal

We are conducting a survey on trustworthy artificial intelligence (AI) for breast cancer treatment planning. Our aim is to reach healthcare experts across various institutions to get an insight into their day-to-day practice: issues encountered, agreements and more importantly, disagreements regarding clinical practices, definitions, and protocols to set requirements for our tool. We also want to get their perspective on AI tools and ways to make them more reliable and applicable in clinical settings.

4. What is your profession? *

- ○ Radiologist
- ○ Oncologist
- ○ Radiation Oncologist
- ○ Surgeon
- ○ General practitioner
- ○ Pathologist
- ○ Other

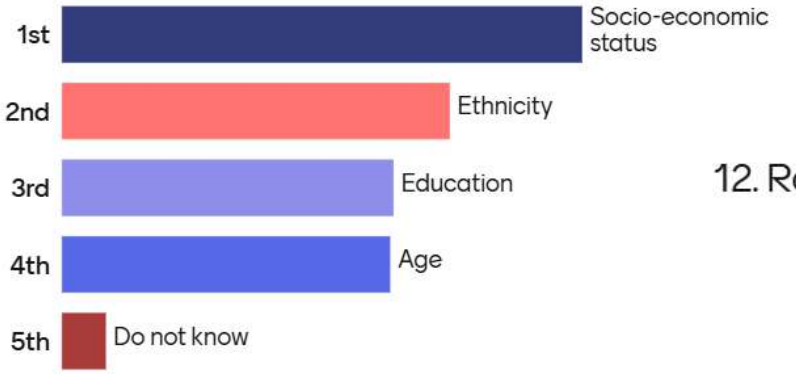
3. What sources of bias do you think we should possibly take into account? *

3. What information should be monitored over time? *

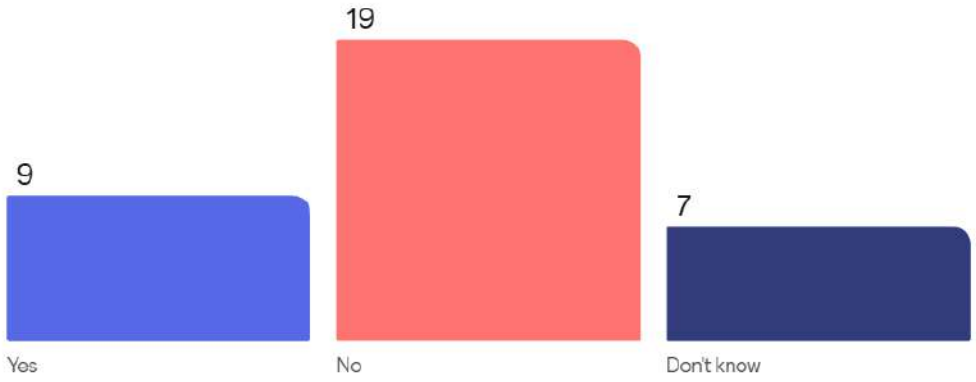
	Definitely	Maybe	Not needed
Usage	○	○	○
Errors	○	○	○
Limitations	○	○	○

Stakeholder Engagement

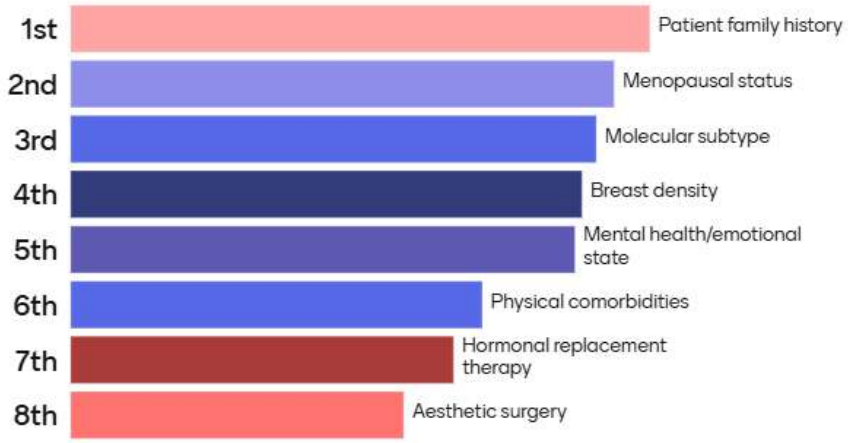
8. Rank the following variables based on their importance for bias estimation



9. Do you have information on ethnicity in your centre/country?



12. Rank the following variables based on their importance for bias estimation



Best Practices – Design

G1:	Engage inter-disciplinary stakeholders	(1)
Us1:	Define intended use and user requirements	(2)
Un1:	Define clinical settings and related variations	(3)
R1:	Define all sources of data heterogeneity	(4)
F1:	Define all sources of bias	(5)
E1:	Define explainability needs	(6)
Un2:	Use community-defined standards	(7)
G6:	Investigate application-specific ethical issues	(8)
G7:	Investigate social and societal issues	(9)
T1:	Define a risk management process	(10)

Intended Use

AI4HF

Secondary risk prevention in heart failure

Clinicians

What should the AI tool predict ?

- Change in cardiac function
- Risk of myocardial infarction
- Risk of mortality

Patients

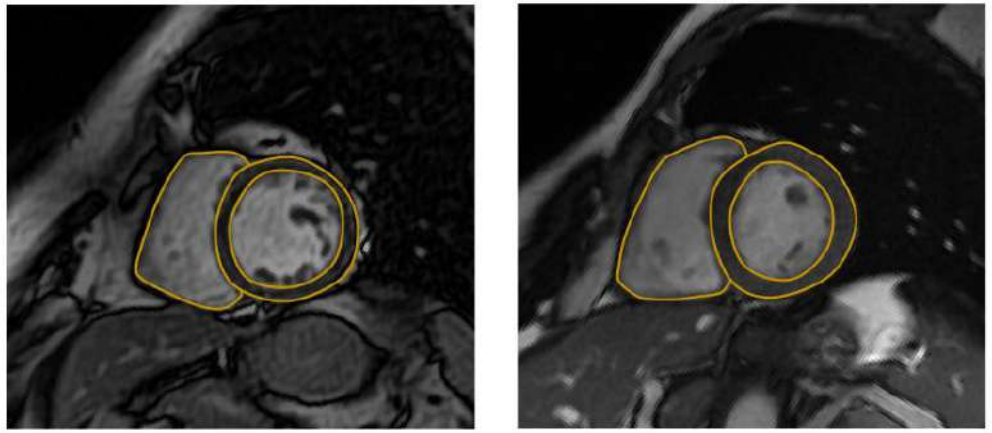
What should the AI tool predict ?

- Risk of fatigue
- Risk of backpain
- Risk of hospital re-admission

Best Practices – Design

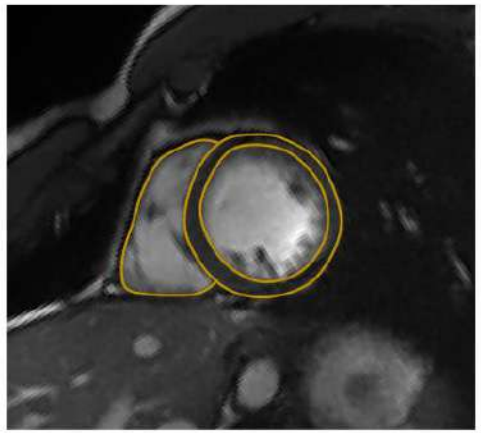
G1:	Engage inter-disciplinary stakeholders	(1)
Us1:	Define intended use and user requirements	(2)
Un1:	Define clinical settings and related variations	(3)
R1:	Define all sources of data heterogeneity	(4)
F1:	Define all sources of bias	(5)
E1:	Define explainability needs	(6)
Un2:	Use community-defined standards	(7)
G6:	Investigate application-specific ethical issues	(8)
G7:	Investigate social and societal issues	(9)
T1:	Define a risk management process	(10)

Data Heterogeneity



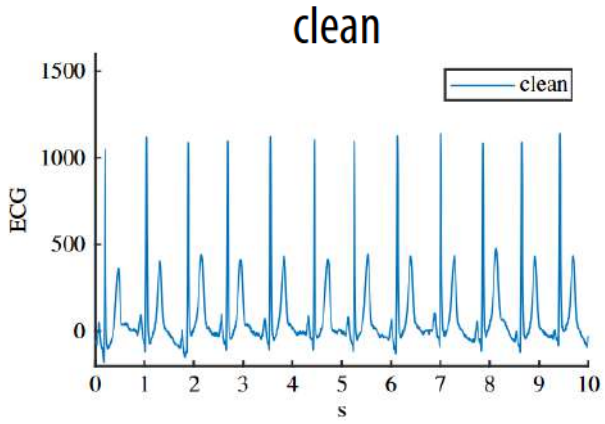
(1) Canon

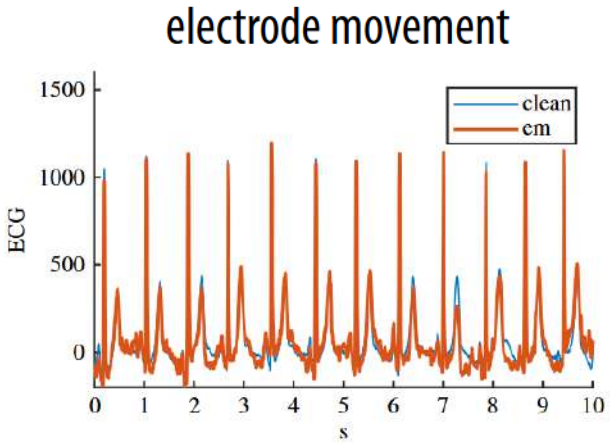
(2) General Electric

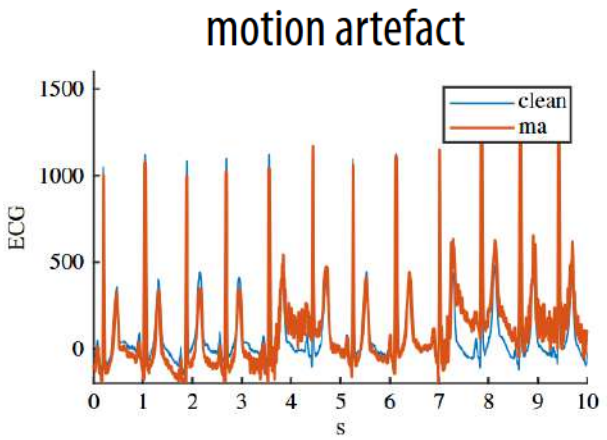


(3) Philips

(4) Siemens







Venton *et al.* 2021. Robustness CNNs to physiological ECG noise.

Best Practices – Design

G1:	Engage inter-disciplinary stakeholders	(1)
Us1:	Define intended use and user requirements	(2)
Un1:	Define clinical settings and related variations	(3)
R1:	Define all sources of data heterogeneity	(4)
F1:	Define all sources of bias	(5)
E1:	Define explainability needs	(6)
Un2:	Use community-defined standards	(7)
G6:	Investigate application-specific ethical issues	(8)
G7:	Investigate social and societal issues	(9)
T1:	Define a risk management process	(10)

Sources of Biases

AI4HF

Sources of Biases

Lima: Sea level

Arequipa: 2,335 m

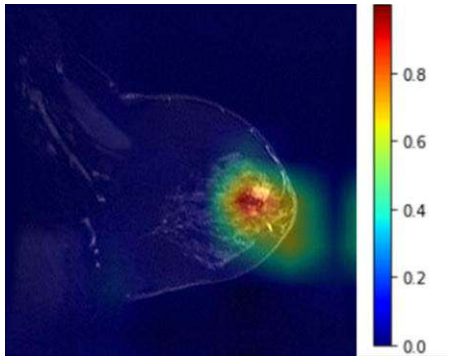
Cusco: 3,400 m

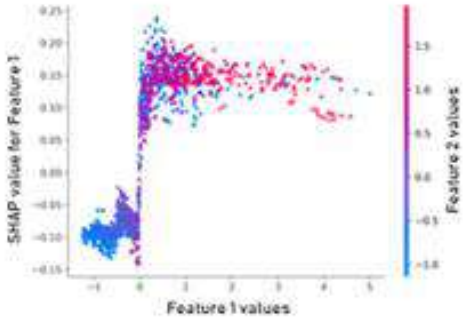
Rinconada: 5,100 m

Best Practices – Design

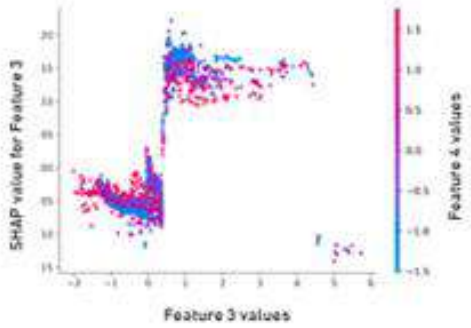
G1:	Engage inter-disciplinary stakeholders	(1)
Us1:	Define intended use and user requirements	(2)
Un1:	Define clinical settings and related variations	(3)
R1:	Define all sources of data heterogeneity	(4)
F1:	Define all sources of bias	(5)
E1:	Define explainability needs	(6)
Un2:	Use community-defined standards	(7)
G6:	Investigate application-specific ethical issues	(8)
G7:	Investigate social and societal issues	(9)
T1:	Define a risk management process	(10)

Explainability

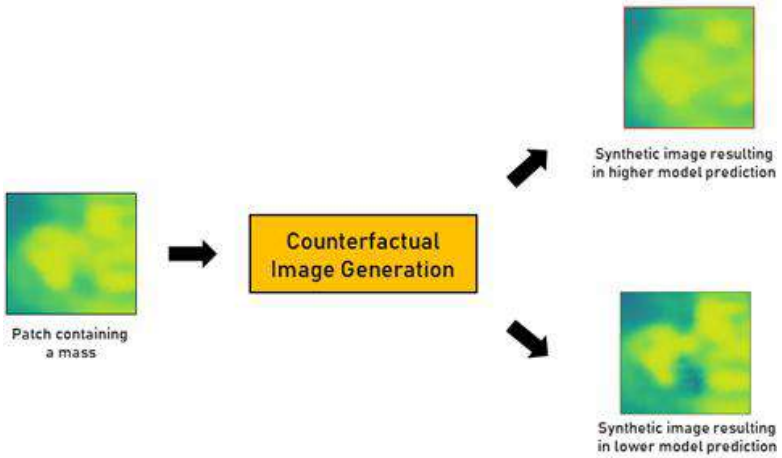


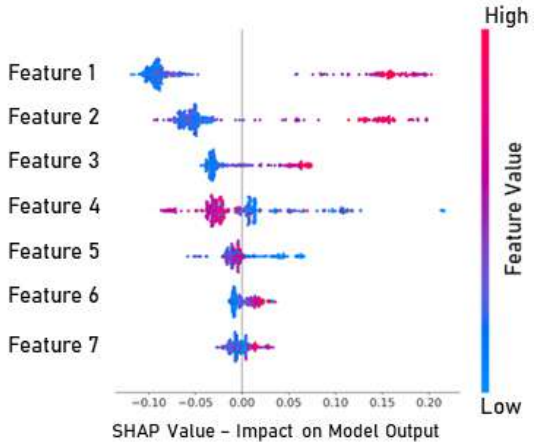


(a)



(b)





Best Practices – Development

G2:	Define measures for privacy and security	(11)
F2:	Collect data on individuals' attributes	(12)
R2:	Collect representative real-world data	(13)
G3:	Implement measures against identified AI risks	(14)
Us2:	Implement human-AI interaction mechanisms	(15)

Individuals' Attributes

Patient	Age	Sex	Ethnicity	Neighbourhood	Altitude	Skin colour	Education
Patient001	20	Male	E. Europe	Gracia	0	White	High-School
Patient002	20	Male	S. Europe	Gracia	0	White	University
Patient003	30	Male	N. Africa	Horta	500	White	University
Patient004	30	Male	N. Africa	Horta	500	White	University
Patient005	40	Male	S. Africa	Poblenou	1000	Black	High-School
Patient006	40	Female	S. Africa	Poblenou	1000	Black	High-School
Patient007	50	Female	S. Asia	Gervasi	1500	White	High-School
Patient008	50	Female	E. Asia	Gervasi	1500	White	University
Patient009	60	Female	L. America	Eixample	3000	White	University
Patient010	60	Female	L. America	Eixample	3000	Black	University

Training Data

High-quality data

Heterogeneous data

Private hospital

Public hospital

Best Practices – Development

G2:	Define measures for privacy and security	(11)
F2:	Collect data on individuals' attributes	(12)
R2:	Collect representative real-world data	(13)
G3:	Implement measures against identified AI risks	(14)
Us2:	Implement human-AI interaction mechanisms	(15)

Universality

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

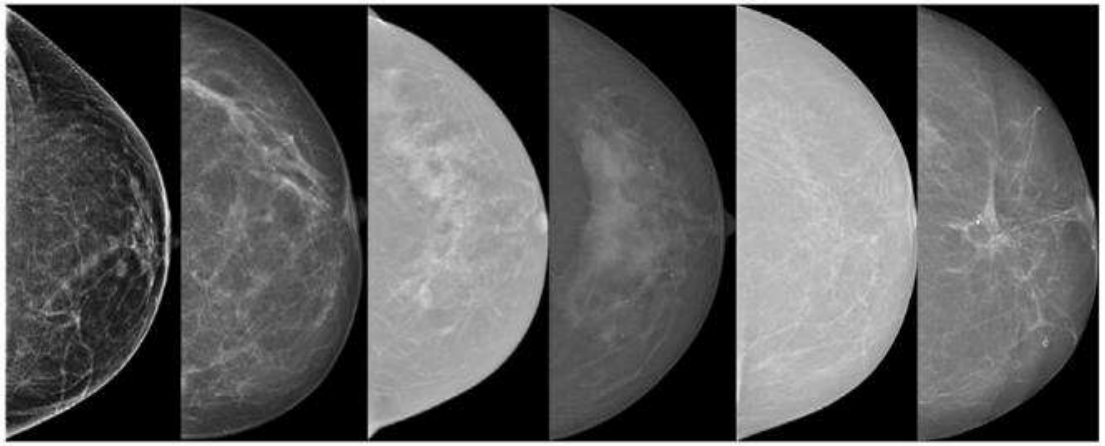
journal homepage: www.elsevier.com/locate/artmed

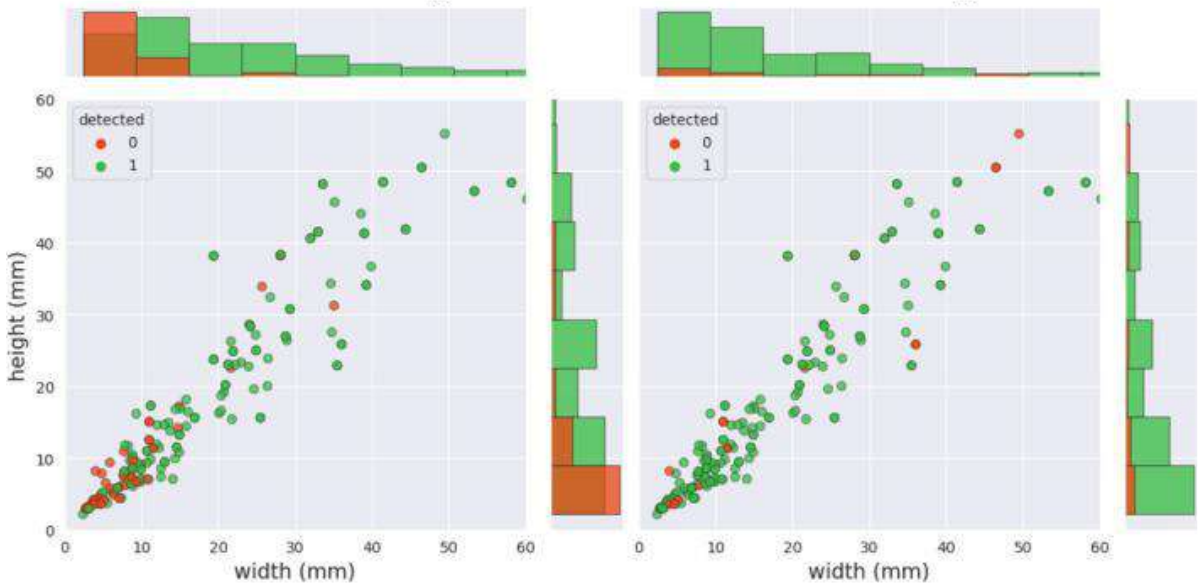
Research paper

Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study

Lidia Garrucho *, Kaisar Kushibar, Socayna Jouide, Oliver Diaz, Laura Igual, Karim Lekadir

Artificial Intelligence in Medicine Lab (BCN-AIM), Faculty of Mathematics and Computer Science, University of Barcelona, Gran Via de les Corts Catalanes 585, Barcelona, 08007, Barcelona, Spain





Baseline model **Domain adaptation**

Fairness

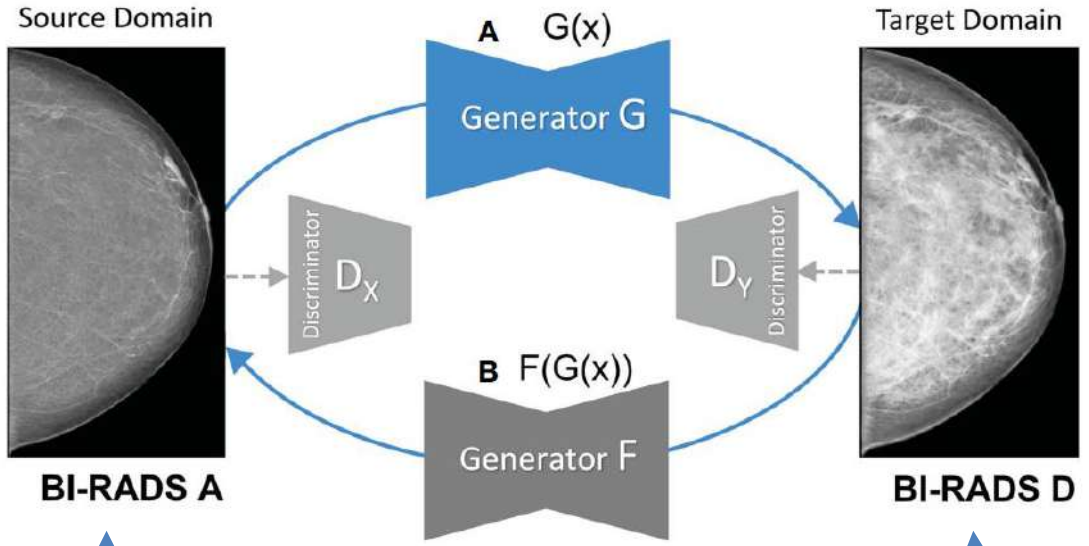
frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 23 January 2023
DOI 10.3389/fonc.2022.1044496

High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection

Lidia Garrucho[1*], Kaisar Kushibar[1], Richard Osuala[1], Oliver Diaz[1], Alessandro Catanese[2], Javier del Riego[3], Maciej Bobowicz[4], Fredrik Strand[5], Laura Igual[1] and Karim Lekadir[1]

[1]Barcelona Artificial Intelligence in Medicine Lab, Facultat de Matemàtques i Informàtica, Universitat de Barcelona, Barcelona, Spain, [2]Unitat de Diagnòstic per la Imatge de la Mama (UDIM), Hospital Germans Trias i Pujol, Badalona, Spain, [3]Área de Radiología Mamaria y Ginecológica (UDIAT CD), Parc Taulí Hospital Universitari, Sabadell, Spain, [4]2nd Department of Radiology, Medical University of Gdansk, Gdansk, Poland, [5]Breast Radiology, Karolinska University Hospital and Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden



Low-density breasts

High-density breasts (~5%)

Fairness

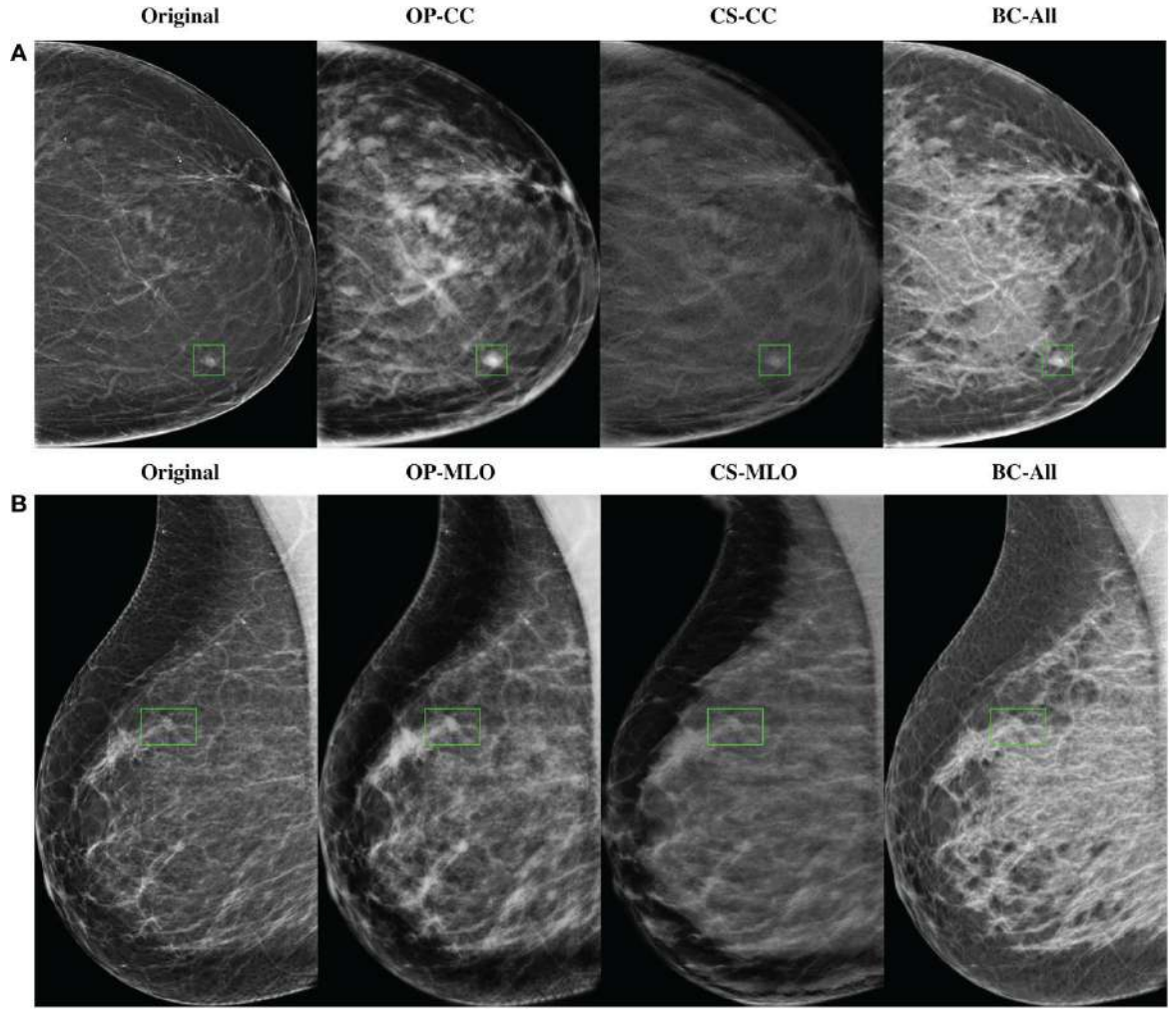
frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 23 January 2023
DOI 10.3389/fonc.2022.1044496

High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection

Lidia Garrucho[1*], Kaisar Kushibar[1], Richard Osuala[1], Oliver Diaz[1], Alessandro Catanese[2], Javier del Riego[3], Maciej Bobowicz[4], Fredrik Strand[5], Laura Igual[1] and Karim Lekadir[1]

[1]Barcelona Artificial Intelligence in Medicine Lab, Facultat de Matemàtques i Informàtica, Universitat de Barcelona, Barcelona, Spain, [2]Unitat de Diagnòstic per la Imatge de la Mama (UDIM), Hospital Germans Trias i Pujol, Badalona, Spain, [3]Área de Radiología Mamaria y Ginecológica (UDIAT CD), Parc Taulí Hospital Universitari, Sabadell, Spain, [4]2nd Department of Radiology, Medical University of Gdansk, Gdansk, Poland, [5]Breast Radiology, Karolinska University Hospital and Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden



Best Practices – Validation

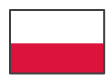
G4:	Define an adequate evaluation plan	(16)
Un3:	Evaluate using external and/or multi-site data	(17)
R3:	Evaluate robustness against real variations	(18)
F3:	Evaluate fairness and debiasing measures	(19)
Us3:	Evaluate user experience and acceptance	(20)
Us4:	Evaluate clinical utility and safety	(21)
E2:	Evaluate explainability with end-users	(22)
T2:	Document the AI tool including evaluations	(23)

Universality Evaluation

Instituto de Investigación Sanitaria La Fe

KAROLINSKA UNIVERSITY HOSPITAL

MEDIZINISCHE UNIVERSITÄT WIEN



GDAŃSKI UNIWERSYTET MEDYCZNY

HACETTEPE ÜNİVERSİTESİ

CentroMamario IAF INSTITUTO ALEXANDER FLEMING

Faculty of Medicine Ain Shams University

SVEUČILIŠTE U ZAGREBU MEDICINSKI FAKULTET

radioval

Fairness Evaluation

scientific reports

OPEN

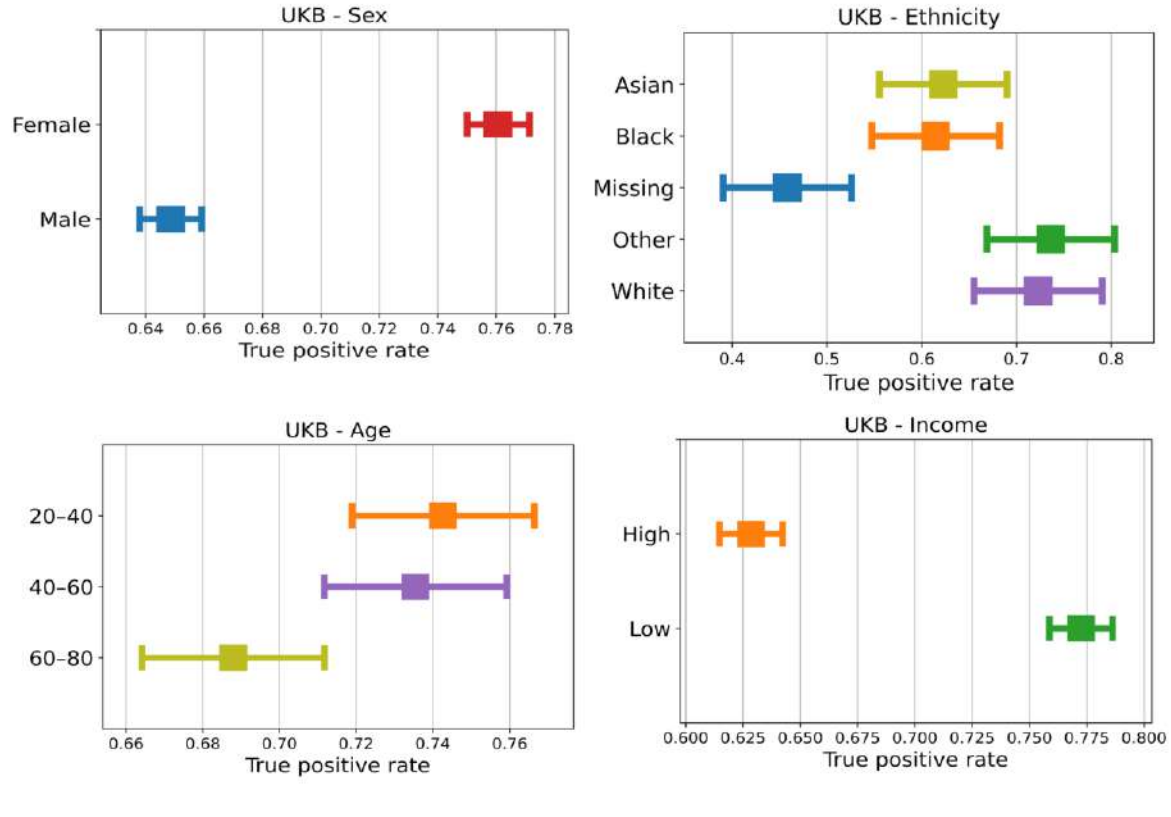
Fairness and bias correction in machine learning for depression prediction across four study populations

Vien Ngoc Dang[1✉], Anna Cascarano[1], Rosa H. Mulder[2,3], Charlotte Cecil[2,4,5], Maria A. Zuluaga[6], Jerónimo Hernández-González[7] & Karim Lekadir[1,8]

A significant level of stigma and inequality exists in mental healthcare, especially in under-served populations. Inequalities are reflected in the data collected for scientific purposes. When not properly accounted for, machine learning (ML) models learned from data can reinforce these structural inequalities or biases. Here, we present a systematic study of bias in ML models designed to predict depression in four different case studies covering different countries and populations. We find that standard ML approaches regularly present biased behaviors. We also show that mitigation techniques, both standard and our own post-hoc method, can be effective in reducing the level of unfair bias. There is no one best ML model for depression prediction that provides equality of outcomes. This emphasizes the importance of analyzing fairness during model selection and transparent reporting about the impact of debiasing interventions. Finally, we also identify positive habits and open challenges that practitioners could follow to enhance fairness in their models.

Demographic factors

Socioeconomic status



Standard AI models

Usability Evaluation

Human evaluators in 5 sites:

- ✓ 2 GPs at each site
- ✓ 2 cardiologists at each site
- ✓ 2 nurses at each site
- ✓ 7 patients for each clinician
- ✓ 2 IT/data manager

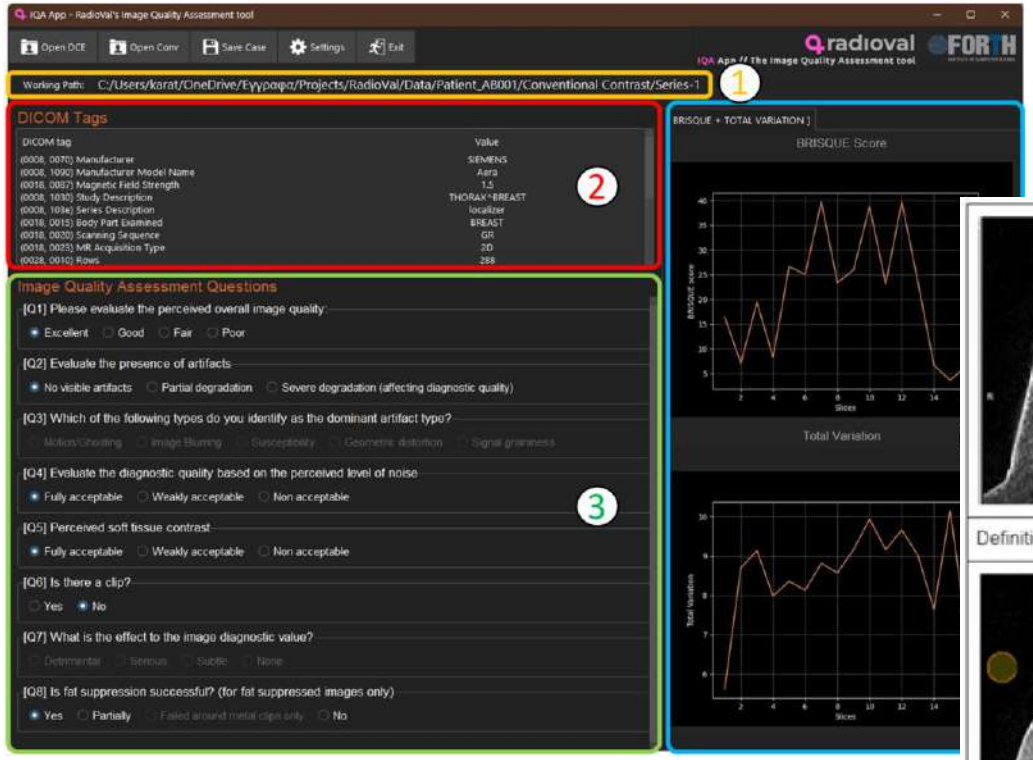
- 50% male + 50% female
- 50% early-career, 50% > 5-year experience

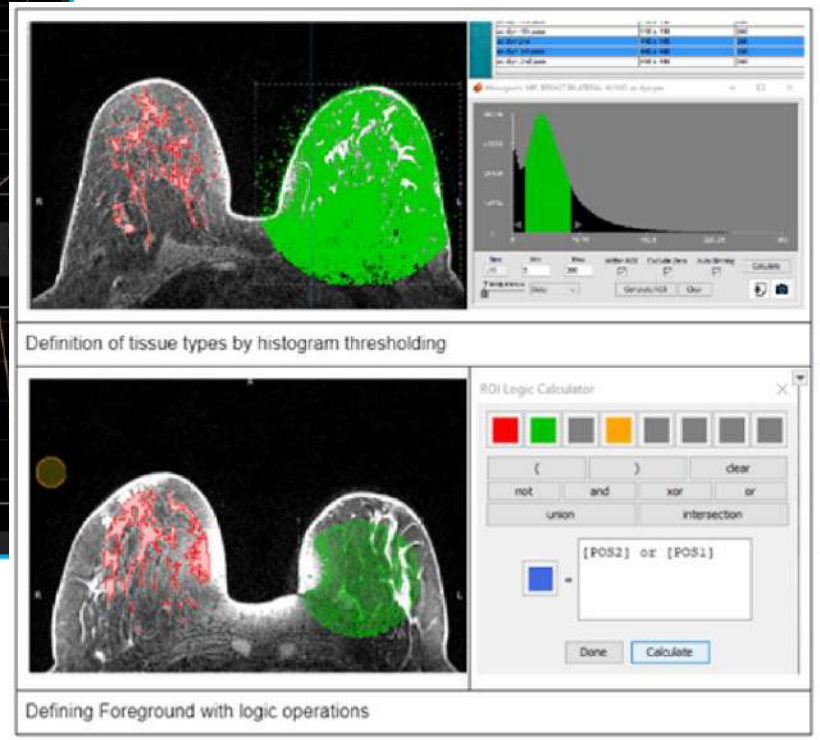
AI4HF

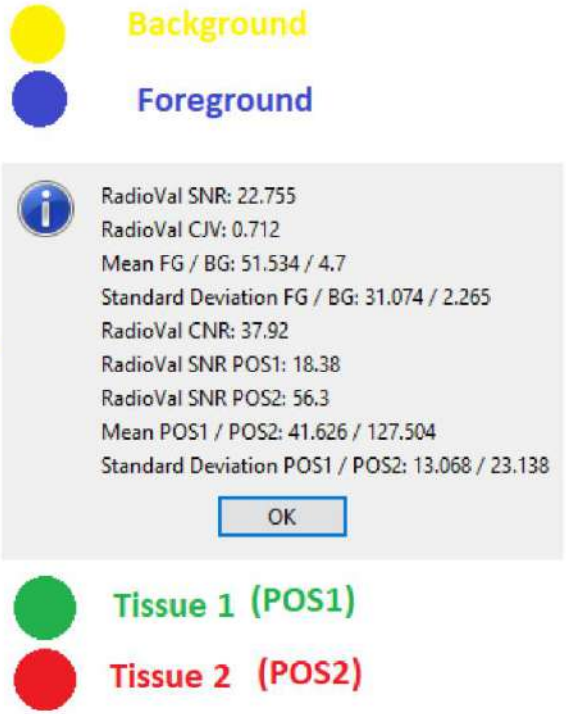
Best Practices – Deployment

Un4:	Evaluate local clinical validity	(24)
Us3:	Provide training materials and activities	(25)
T3:	Define mechanisms for quality control	(26)
T4:	Implement a periodic auditing system	(27)
T5:	Implement a logging system	(28)
G5:	Comply with AI regulatory requirements	(29)
T6:	Establish mechanisms for AI governance	(30)

Quality Control







Nikiforaki et al. 2024. Image Quality Assessment Tool for Conventional and Dynamic Magnetic Resonance Imaging Acquisitions. Journal of Imaging, 10(5), p.115.

Part 1 - What is trustworthy AI?

Part 2 - How do we achieve it?

Part 3 - Trustworthy AI in an EIC

Objectives

Objective 1: Design, develop and demonstrate the first trustworthy AI technology for personalised risk assessment and improved management of HF patients in clinical cardiology.

Objective 2: Implement a human-centred, multi-stakeholder, inclusive approach to improve awareness, acceptance and promotion of trustworthy AI in cardiovascular risk assessment.

Objective 4: Implement the very first international, multi-faceted clinical validation study for AI-powered risk assessment in cardiology in multiple developed and developing countries.

Objective 1: Design, develop and evaluate a trustworthy and ethical AI tool for early cancer detection

Method Section

1. Stakeholder engagement
2. Training and testing datasets
3. AI methods implementation
4. AI validation studies
5. Reporting, dissemination, exploitation

Stakeholder Engagement

Understand clinical needs, potential barriers and clinical features of the AI tools.	• 1-2 radiologists and 1-2 physicians per SSA site (Total ~20). • 1 healthcare manager per SSA site (Total 10). • 10 patients / advocates (*e.g.* patients at local sites, StopTB local groups).
Define pathways towards implementation and adoption of the AI tools in real-world practice.	• 15 local policy makers (*e.g.* health ministries, regional health agencies). • 7 policy / advocacy organisations (*e.g.* StopTB, WHO-Africa, National Tuberculosis Control Programs, Digital Health Africa, etc).

Participatory action research (PAR): We will employ various PAR methods, such as qualitative interviews, focus groups and co-creation workshops, to gather insights on the needs, views and concerns of SSA stakeholders. At the start of the project, CISM, KUHES and ISGLOBAL will elaborate a qualitative master protocol with all SSA site, which will be then adapted by each SSA site to ensure relevance to the local contexts. The local social scientists will then collect data through direct interactions with the local stakeholders, employing narrative interviews or focus groups as needed. For instance, each SSA partner will organise focus groups with n=7-10 local patients to discuss ethical concerns and the patient's views on (i) the clinical sites waiving consent on their behalf for the public good, (ii) the application of AI to diagnose their family members or children, (iii) the monetisation of their data to gather extra resources for the healthcare facilities. All qualitative data will be transcribed verbatim then structured through a qualitative data management software (Nvivo).

Datasets

Table 2 – List of selected cardiovascular cohorts to be leveraged in AI4HF

Name	Country	Partner	Type	Number of HF patients	Harmonised	Clinical reports	Cardiac imaging	ECG signals	Cardiac biomarkers	Clinical end-points
Dataset 1 (Available at the start of AI4HF)				**>715,000**						
CALIBER	UK	NLHI	EHR	502,536	**Y**	S	N	N	Y	Y
UK Biobank	UK	UOXF	Cohort	8,000	**Y**	S	I+R	S	Y	Y
SwedeHF	SE	UMCU	Registry	150,000	**Y**	S	R	R	Y	Y
ABUCASIS	ES	NLHI	EHR	48,000	**Y**	S	R	S	Y	Y
UPOD	NL	UMCU	EHR	8,000	**Y**	S	R	S+R	Y	Y

AI Methods / Validation

Fairness: To assess and address AI biases (*e.g.* with respect to sex or ethnicity), we will include dedicated metrics (*e.g.* Statistical Parity, Group Fairness), and bias mitigation techniques (*e.g.* adaptive data re-sampling, equalised odds post-processing).
Universality: To ensure wide applicability of the AI solutions, the toolkit will comprise transfer learning and knowledge distillation techniques to optimise and calibrate pre-existing AI models for each new clinical site across different SSA countries.
Traceability: We will implement statistical tests (*e.g.* population stability index, performance variability) to identify drifts in the input images or AI decisions over time. We will include continuous learning methods to re-calibrate the AI models.
Usability: Existing usability questionnaires, such as the System Usability Scale, will be updated and adapted to local contexts of African clinicians, so they can be utilised to assess acceptance and applicability of newly developed AI solutions.
Robustness: Methods for assessing robustness (*e.g.* against variations in data quality or image scanners across SSA sites) will be implemented, together with image harmonisation and domain adaptation techniques to enhance robustness.
Explainability: Explainable AI methods will be provided in the open toolkit, such as heatmaps, depending on the needs and preferences of the local clinicians for interpreting visually the decisions of the AI tools on the input images.

AI Validation

In-silico validation of AI-based TB diagnosis from adult CUS: This study will implement the same logical steps as above across five different SSA partners, in Gabon, Malawi, Gambia, Ghana, and Uganda. We will evaluate the AI tool's performance in adult TB diagnosis with CUS images, with a focus on its adaptability and consistency across various CUS machines and setups that are prevalent in SSA settings. Given the operator-dependent nature of CUS, we will specifically assess how variations in image acquisition by different sonographers influence the AI's diagnostic accuracy. Five radiologists, sonographers, and general practitioners per site will participate to evaluate the AI tool's accessibility and practicality in typical clinical environments. The trial will also explore the effectiveness of explainable AI features in aiding clinicians to identify TB-related abnormalities in CUS such as pleural effusion and echogenic patterns indicative of fibrosis, calcification, or fluid accumulation. Importantly, we will thoroughly compare AI-aided diagnosis of TB in adults with CUS with standard of care based on CXR expert interpretation.

Gender Dimensions

1.2.15 Gender Dimension

It is established that there are sex differences in respiratory disease, both in terms of pathophysiology and disease manifestation[39]. Hence, in AFRICAI-RI, sex differences will be taken into consideration from day one, as follows:

- First, during the requirements analysis in WP1, our social scientists will ensure a balanced representation in terms of sex and gender in the stakeholder engagement activities, to obtain unbiased requirements.
- During data preparation in WP2-3, we will carefully estimate the distribution of men and women for all data sources, thus ensuring potential bias is reported or corrected (*e.g.* using sampling methods).
- Bias estimation will be applied in WP4-5 to assess if the AI tools are biased or fair across male and female subjects. Correction measures will be systematically applied to correct for any identified biases,
- The selection of patients and clinicians for the multi-site in silico trials in WP6 will ensure a balanced representation and the analyses will be stratified for potential sex differences.
- In WP7, the early career researchers and students will be recruited ensuring a perfect gender balance.
- Finally, findings and recommendations will be disseminated in WP8-9 to promote accessible AI-powered imaging diagnostics of respiratory diseases across all subgroups, including male and female patients.

WPs / Tasks

T1.2. Qualitative studies (CISM, ISGLOBAL, ALL) [M1-M48]: At M1-M3, CISM and ISGLOBAL will coordinate elaborate a qualitative master protocol with all SSA sites. At M3-M12, data will be continuously collected through site workshops and individual interviews with stakeholders in all SSA sites (as described in Section 1.2.4). Content and thematic analyses will be used to analyse data (Nvivo software) and extract requirements. The process will be repeated throughout the project depending on emerging questions and needs.

T1.3. Co-creation activities and consensus requirements (CISM, ISGLOBAL, ALL) [M9-M48]: Starting from M9, the results will be shared through local workshops to foster feedback. MRCG will organise an in-person co-creation workshop in The Gambia during the first annual meeting to resolve any misalignments (*e.g.* a technical requirement does not align with ethical concerns). Follow-up online co-creation meeting will be convened regularly to refine all requirements. Requirements documents and scientific publications will be elaborated by all sites, coordinated by CISM.

WPs / Tasks

T3.3. Robust AI methods for cross-centre generalisability (UOXF, UB, NLHI, UMCU, VHIR, ICRC, MUHAS, INCOR, BSC, SRDC) [M12-M36]:

- At M12-M18, the baseline models will be assessed with respect to the sources of heterogeneity identified during requirement analysis in WP1 (*e.g.* variations in biomarkers, ECG machines, image scanners, data quality), by performing "leave one centre out" tests based on the 5 cohorts of Dataset 1.
- At M18-M24, UOXF and UB will implement and assess several strategies for enhanced AI robustness across datasets and centres, such as data synthesis, transfer learning, domain adaptation and knowledge distillation.

T3.4. Bias detection and mitigation methods (UB, NLHI, UOXF, UMCU, VHIR, ICRC, MUHAS, INCOR, CERTH) [M12-M36]:

- At M12-M18, the baseline models will be tested for their fairness depending on the sources of bias identified in WP1 (*e.g. sex, comorbidity)*, by implementing metrics such as Statistical Parity and Group Fairness.
- At M18-M27, based on Dataset 1, UB will implement bias mitigation measures, such as 1) pre-processing approaches (*e.g.* re-sampling or data augmentation), (2) in-processing approaches that explicitly remove discrimination during model training, and (3) post-processing *e.g.* the equalised odds technique.

Final Point

- Space is limited (20 pages): How much for AI (1 or 10 pages)?
- Reviewers could be AI experts or/and domain experts
- Depends on the EIC grant:
  - AI has a major role: It's an AI-focused EIC grant proposal
  - AI is prominent, but it's not the core of the EIC proposal
  - AI has a minor role: It's just a method amongst others
  - There is no AI at all

FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Karim Lekadir,[1,2] Alejandro F Frangi,[3,4] Antonio R Porras,[5] Ben Glocker,[6] Celia Cintas,[7] Curtis P Langlotz,[8] Eva Weicken,[9] Folkert W Asselbergs,[10,11] Fred Prior,[12] Gary S Collins,[13] Georgios Kaissis,[14] Gianna Tsakou,[15] Irène Buvat,[16] Jayashree Kalpathy-Cramer,[17] John Mongan,[18] Julia A Schnabel,[19] Kaisar Kushibar,[1] Katrine Riklund,[20] Kostas Marias,[21] Lameck M Amugongo,[22] Lauren A Fromont,[23] Lena Maier-Hein,[24] Leonor Cerdá-Alberich,[25] Luis Martí-Bonmatí,[26] M Jorge Cardoso,[27] Maciej Bobowicz,[28] Mahsa Shabani,[29] Manolis Tsiknakis,[21] Maria A Zuluaga,[30] Marie-Christine Fritzsche,[31] Marina Camacho,[1] Marius George Linguraru,[32] Markus Wenzel,[9] Marleen De Bruijne,[33] Martin G Tolsgaard,[34] Melanie Goisauf,[35] Mónica Cano Abadía,[35] Nikolaos Papanikolaou,[36] Noussair Lazrak,[1] Oriol Pujol,[1] Richard Osuala,[1] Sandy Napel,[37] Sara Colantonio,[38] Smriti Joshi,[1] Stefan Klein,[33] Susanna Aussó,[39] Wendy A Rogers,[40] Zohaib Salahuddin,[41] Martijn P A Starmans[33]; on behalf of the FUTURE-AI Consortium

For numbered affiliations see end of the article

Correspondence to: K Lekadir karim.lekadir@ub.edu (ORCID 0000-0002-9456-1612)

Additional material is published online only. To view please visit the journal online.

Cite this as: BMJ 2025;388:e081554 http://dx.doi.org/10.1136/bmj-2024-081554

Accepted: 10 January 2025

Despite major advances in artificial intelligence (AI) research for healthcare, the deployment and adoption of AI technologies remain limited in clinical practice. This paper describes the FUTURE-AI framework, which provides guidance for the development and deployment of trustworthy AI tools in healthcare. The FUTURE-AI Consortium was founded in 2021 and comprises 117 interdisciplinary experts from 50 countries representing all continents, including AI scientists, clinical researchers, biomedical ethicists, and social scientists. Over a two year period, the FUTURE-AI guideline was established through consensus based on six guiding principles—fairness, universality, traceability, usability, robustness, and explainability. To operationalise trustworthy AI in healthcare, a set of 30 best practices were defined, addressing technical, clinical, socioethical, and legal dimensions. The recommendations cover the entire lifecycle of healthcare AI, from design, development, and validation to regulation, deployment, and monitoring.

SUMMARY POINTS

- Despite major advances in medical artificial intelligence (AI) research, clinical adoption of emerging AI solutions remains challenging owing to limited trust and ethical concerns
- The FUTURE-AI Consortium unites 117 experts from 50 countries to define international guidelines for trustworthy healthcare AI
- The FUTURE-AI framework is structured around six guiding principles: fairness, universality, traceability, usability, robustness, and explainability
- The guideline addresses the entire AI lifecycle, from design and development to validation and deployment, ensuring alignment with real world needs and ethical requirements
- The framework includes 30 detailed recommendations for building trustworthy and deployable AI systems, emphasising multistakeholder collaboration
- Continuous risk assessment and mitigation are fundamental, addressing biases, data variations, and evolving challenges during the AI lifecycle
- FUTURE-AI is designed as a dynamic framework, which will evolve with technological advancements and stakeholder feedback

Introduction

In the field of healthcare, artificial intelligence (AI)—that is, algorithms with the ability to self-learn logic—and data interactions have been increasingly used to develop computer aided models, for example, disease diagnosis, prognosis, prediction of therapy response or survival, and patient stratification.[1] Despite major advances, the deployment and adoption of AI technologies remain limited in real world clinical practice. In recent years, concerns have been raised about the technical, clinical, ethical, and societal risks associated with healthcare AI.[2 3] In particular, existing research has shown that AI tools in healthcare can be prone to errors and patient harm, biases and increased health inequalities, lack of transparency and accountability, as well as data privacy and security breaches.[4-6]

To increase adoption in the real world, it is essential that AI tools are trusted and accepted by patients, clinicians, health organisations, and authorities. However, there is an absence of clear, widely accepted guidelines on how healthcare AI tools should be designed, developed, evaluated, and deployed to be trustworthy—that is, technically robust, clinically safe,

FAIR	UNIVERSAL	TRACEABLE	USABLE	ROBUST	EXPLAINABLE
					

Many Thanks!

karim.lekadir@ub.edu