UNIVERSITAT DE BARCELONA

ESC European Society of Cardiology

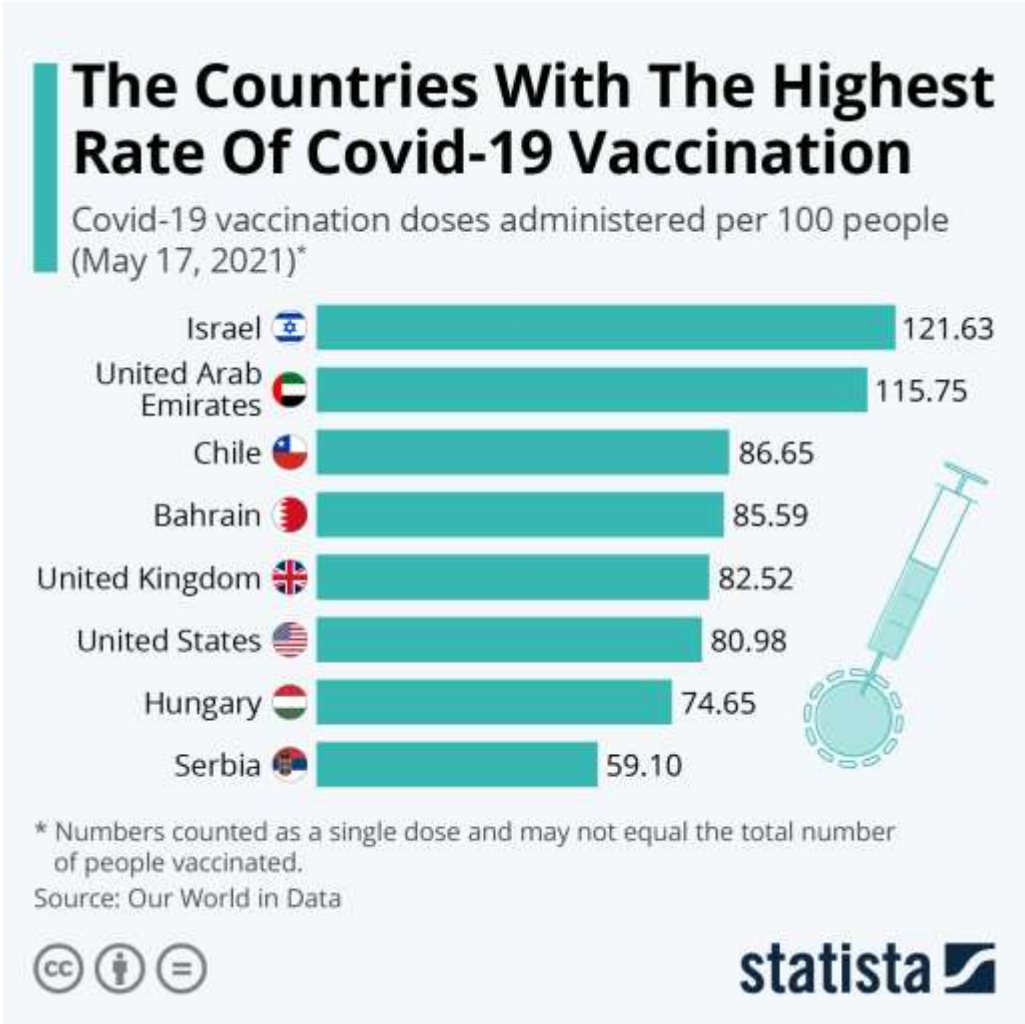
How to Ensure Trustworthiness in AI for Healthcare: The FUTURE-AI Guideline

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

Part 1 - What is trustworthy AI?

Part 2 - How do we achieve it?

Trustworthiness



Novak Djokovic: Tennis star deported after losing Australia visa battle

BBC

16 January 2022

AI in Cardiology

BBC Sign in Home News Sport Reel Worklife

NEWS

Home | War in Ukraine | Climate | Video | World | UK | Business | Tech | Science | Stories

Tech

NHS uses AI scan to detect hidden heart disease

29 March 2021 · Comments

The technology could help save "thousands of lives"

The Guardian

AI eye checks can predict heart disease risk in less than minute, finds study

Breakthrough opens door to a highly effective, non-invasive test that does not need to be done in a clinic

Ophthalmologists may soon be able to carry out cardiovascular screening by checking the retina - without the need for blood tests. Photograph: Zorica Nastasic/Getty Images

Robustness

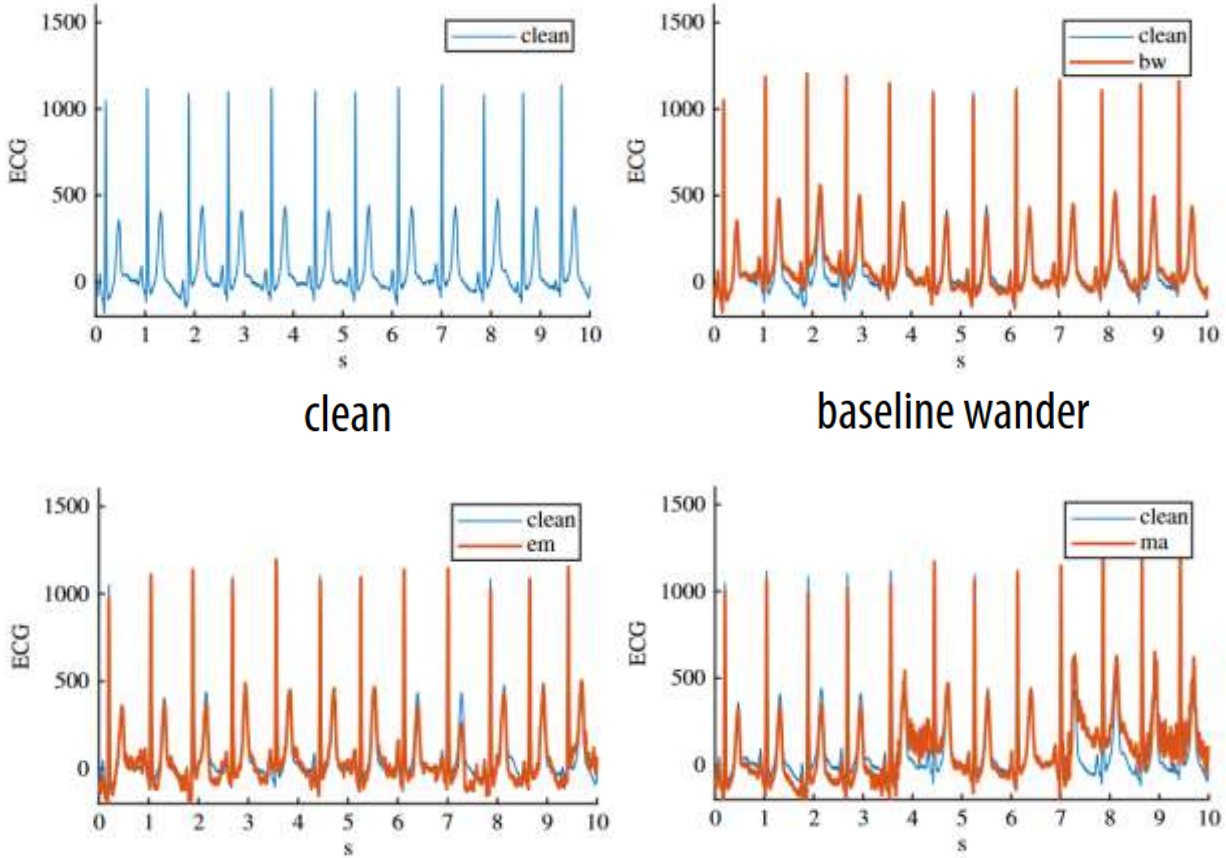
Robustness of convolutional neural networks to physiological electrocardiogram noise

J. Venton[1], P. M. Harris[1], A. Sundar[1], N. A. S. Smith[1] and P. J. Aston[1,2]

[1]Department of Data Science, National Physical Laboratory, Teddington, UK
[2]Department of Mathematics, University of Surrey, Guildford, UK

JV, 0000-0003-0547-1226

PHILOSOPHICAL TRANSACTIONS A



clean

baseline wander

electrode movement

motion artefact

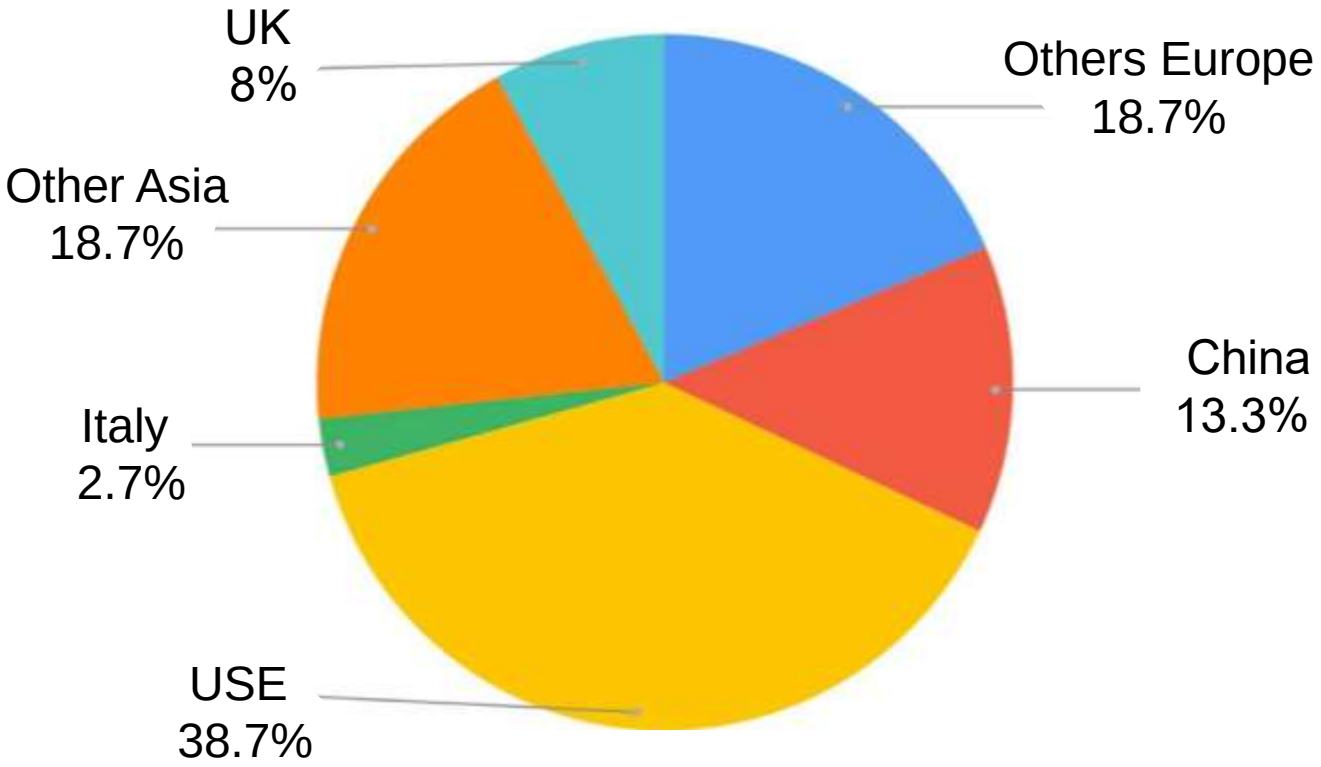
Universality

JACC: Advances
Volume 3, Issue 9, Part 2, September 2024, 101202

Original Research
Outcomes and Quality

Prospective Human Validation of Artificial Intelligence Interventions in Cardiology: A Scoping Review

Amirhossein Moosavi PhD [a b *], Steven Huang [b *], Maryam Vahabi MSc [a b], Bahar Motamedivafa BSc [a b], Nelly Tian MBAn [c], Rafid Mahmood PhD [a], Peter Liu MD [b], Christopher L.F. Sun PhD [a b]



Fairness

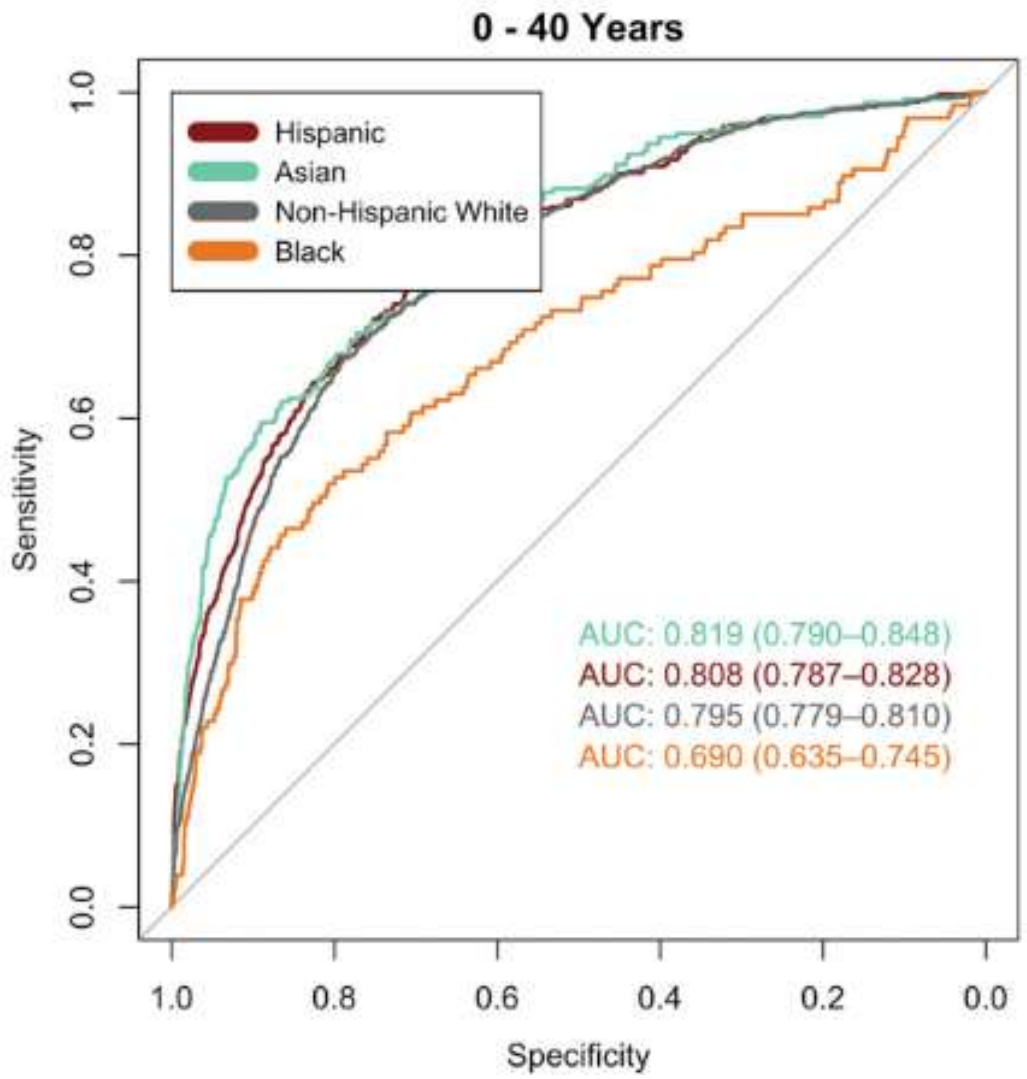
Circulation: Heart Failure
Volume 17, Issue 1, January 2024; Page e010879
https://doi.org/10.1161/CIRCHEARTFAILURE.123.010879

ORIGINAL ARTICLE

Race, Sex, and Age Disparities in the Performance of ECG Deep Learning Models Predicting Heart Failure

See Editorial by Rosenberg

Dhamanpreet Kaur, BS, J. Weston Hughes, BA, Albert J. Rogers, MD, MBA, Guson Kang, MD, Sanjiv M. Narayan, MD, PhD, Euan A. Ashley, DPhil, and Marco V. Perez, MD



Traceability

JMIRx Med. 2024 Jun 12:5:e45973. doi: 10.2196/45973.

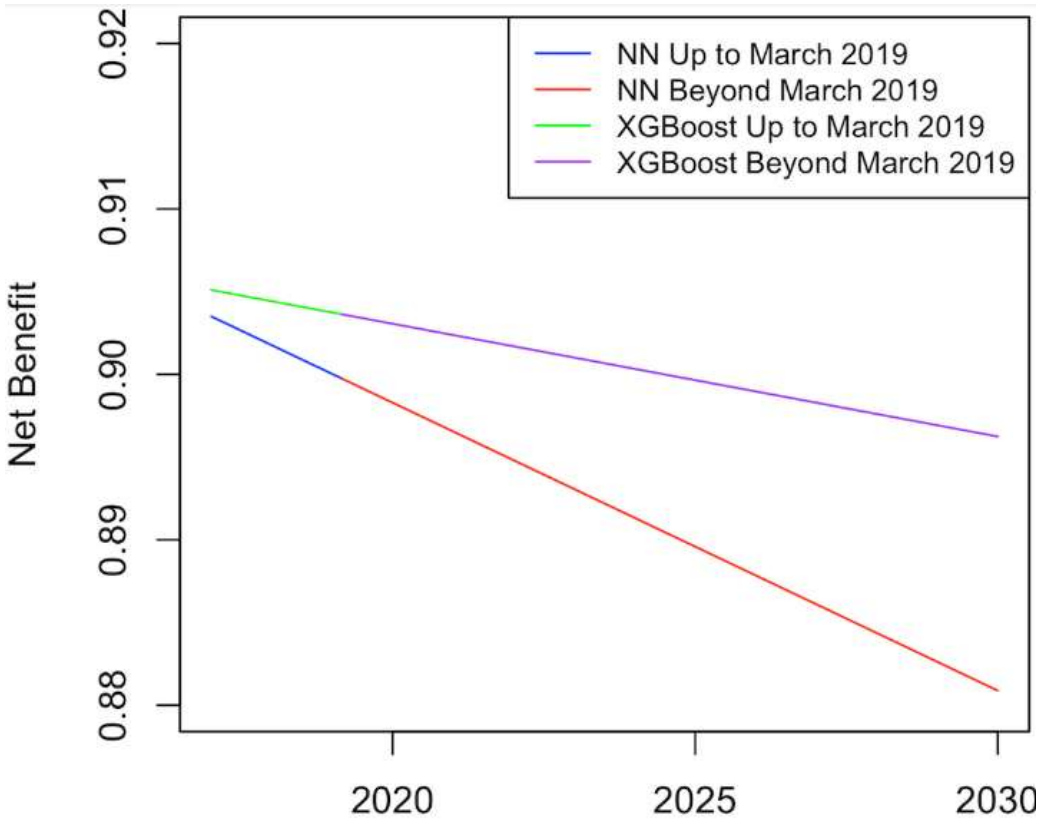
Performance Drift in Machine Learning Models for Cardiac Surgery Risk Prediction: Retrospective Analysis

Tim Dong [1], Shubhra Sinha [1], Ben Zhai [2], Daniel Fudulu [1], Jeremy Chan [1], Pradeep Narayan [3], Andy Judge [1], Massimo Caputo [1], Arnaldo Dimagli [1], Umberto Benedetto [1], Gianni D Angelini [1]

Affiliations — collapse

Affiliations

1 Bristol Heart Institute, Translational Health Sciences, University of Bristol, Bristol, United Kingdom.
2 School of Computing Science, Northumbria University, Newcastle upon Tyne, United Kingdom.
3 Department of Cardiac Surgery, Rabindranath Tagore International Institute of Cardiac Sciences, West Bengal, India.



Explainability

ELSEVIER

Check for updates

Canadian Journal of Cardiology 38 (2022) 204–213

Review

Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology

Jeremy Petch, PhD, MA, BA(H),[a,b,c,d] Shuang Di, MSc, BSc,[a,e,‡] and Walter Nelson, BSc(H)[a,f,‡]

[a] Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, Ontario, Canada
[b] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[c] Division of Cardiology, Department of Medicine, McMaster University, Hamilton, Ontario, Canada
[d] Population Health Research Institute, Hamilton, Ontario, Canada
[e] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[f] Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

ABSTRACT

Many clinicians remain wary of machine learning because of long-standing concerns about "black box" models. "Black box" is shorthand for models that are sufficiently complex that they are not straightforwardly interpretable to humans. Lack of interpretability in predictive models can undermine trust in those models, especially in health care, in which so many decisions are— literally—life and death issues. There has been a recent explosion of research in the field of explainable machine learning aimed at addressing these concerns. The promise of explainable machine learning is considerable, but it is

RÉSUMÉ

De nombreux cliniciens restent méfiants envers l'apprentissage automatique en raison de préoccupations de longue date concernant les modèles à « boîte noire ». Le terme « boîte noire » sert à désigner des modèles suffisamment complexes pour échapper à une interprétation simple par un humain. Le manque d'interprétabilité des modèles prédictifs peut miner la confiance en ces modèles, en particulier dans le domaine des soins de santé, où tant de décisions sont littéralement des questions de vie ou de mort. Il y a eu récemment une explosion de la recherche consacrée à l'apprentissage automatique explicable

Patient Name	Probability of Heart Disease
[redacted]	94%
[redacted]	79%
[redacted]	55%
[redacted]	49%
[redacted]	31%
[redacted]	26%

Probability of Heart Disease

Probability calculated: 7/21/2021 20:26

79%

High

Factors Contributing to Prediction

Thallium Stress Test	Normal
Number of Major Vessels	1
Exercise Induced Angina	No
Chest Pain Type	Typical Angina
Max Heart Rate Achieved	174
Sex	Male

Probability of Heart Disease

Probability calculated: 7/21/2021 20:26

26%

Low

Factors Contributing to Prediction

Thallium Stress Test	Normal
Max Heart Rate Achieved	131
Number of Major Vessels	1
Sex	Male
ST Depression	0.10
Age	69

Usability

ELSEVIER

Cardiovascular Digital Health Journal

Volume 4, Issue 3, June 2023, Pages 101-110

Original Article

Artificial intelligence–enabled tools in cardiovascular medicine: A survey of current use, perceptions, and challenges

Alexander Schepart PharmD, MBA *, Arianna Burton PharmD *, Larry Durkin MBA †, Allison Fuller BA †, Ellyn Charap MSc †, Rahul Bhambri PharmD, MBA *, Faraz S. Ahmad MD, MS ‡ §

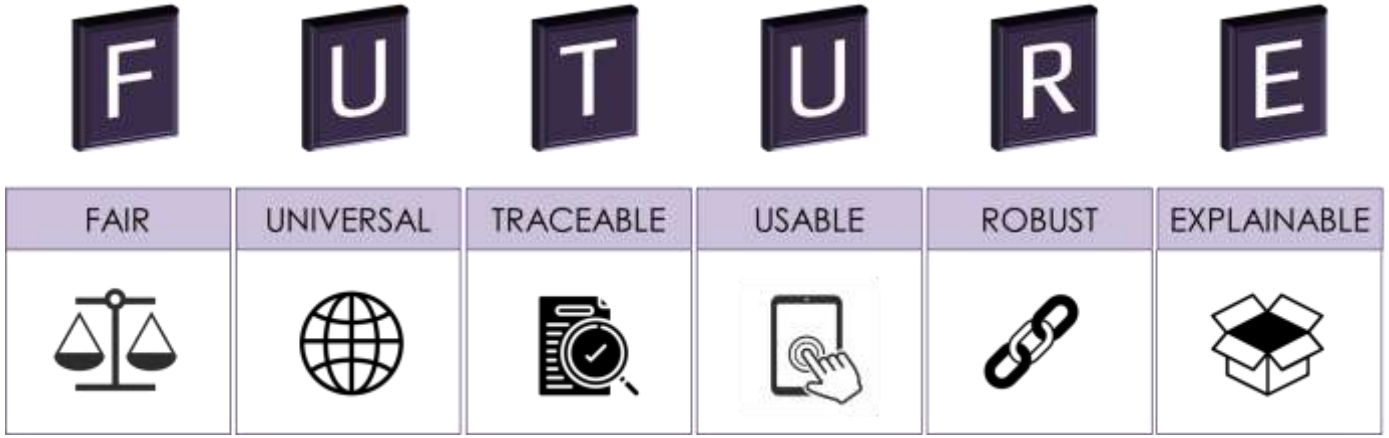
"I think it has to pull from the [health] record without you doing anything...you don't have to go and put the data into it yourself." –Cardiologist

"Sometimes when you getall of these risk calculators…and EPIC, it becomes more information, which is not better. We're pretty busy as clinicians. I think information that's not actionable or helpful, it just slows us down and we ignore it." –Cardiologist

"I don't really know what's going on as far as generating risk scores for something like [ATTR-CM]. The majority of the time, we look at the detailed data graph and use trends…I know there are scoring systems for drug withdraw though." –Cardiologist

Characteristics Trustworthy AI

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



Characteristics Trustworthy AI

	Clusters of requirements	Core principle
1	Diversity, Inclusiveness, Non-discrimination, Bias, Equity	Fairness
2	Generalisability, Adaptability, Interoperability, Applicability	Universality
3	Transparency, Monitoring, Auditing, Accountability	Traceability
4	Human-centred AI, User engagement, Accessibility, Efficiency	Usability
5	Reliability, Resilience, Safety, Security	Robustness
6	Interpretability, Understandability, Transparency	Explainability

Characteristics Trustworthy AI

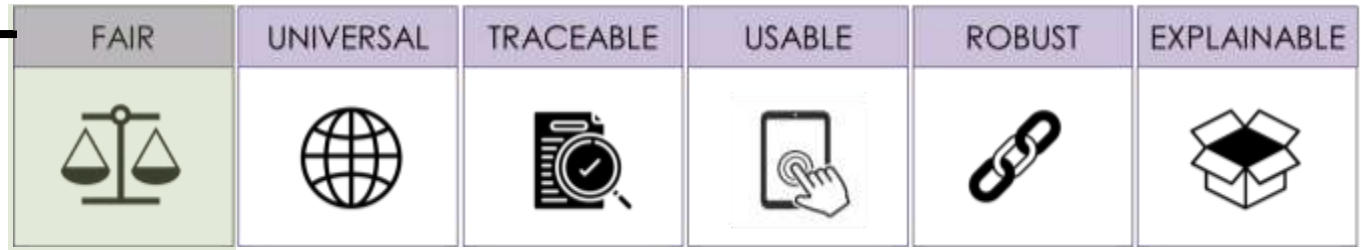
Based on ethical principles and fundamental rights:

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?

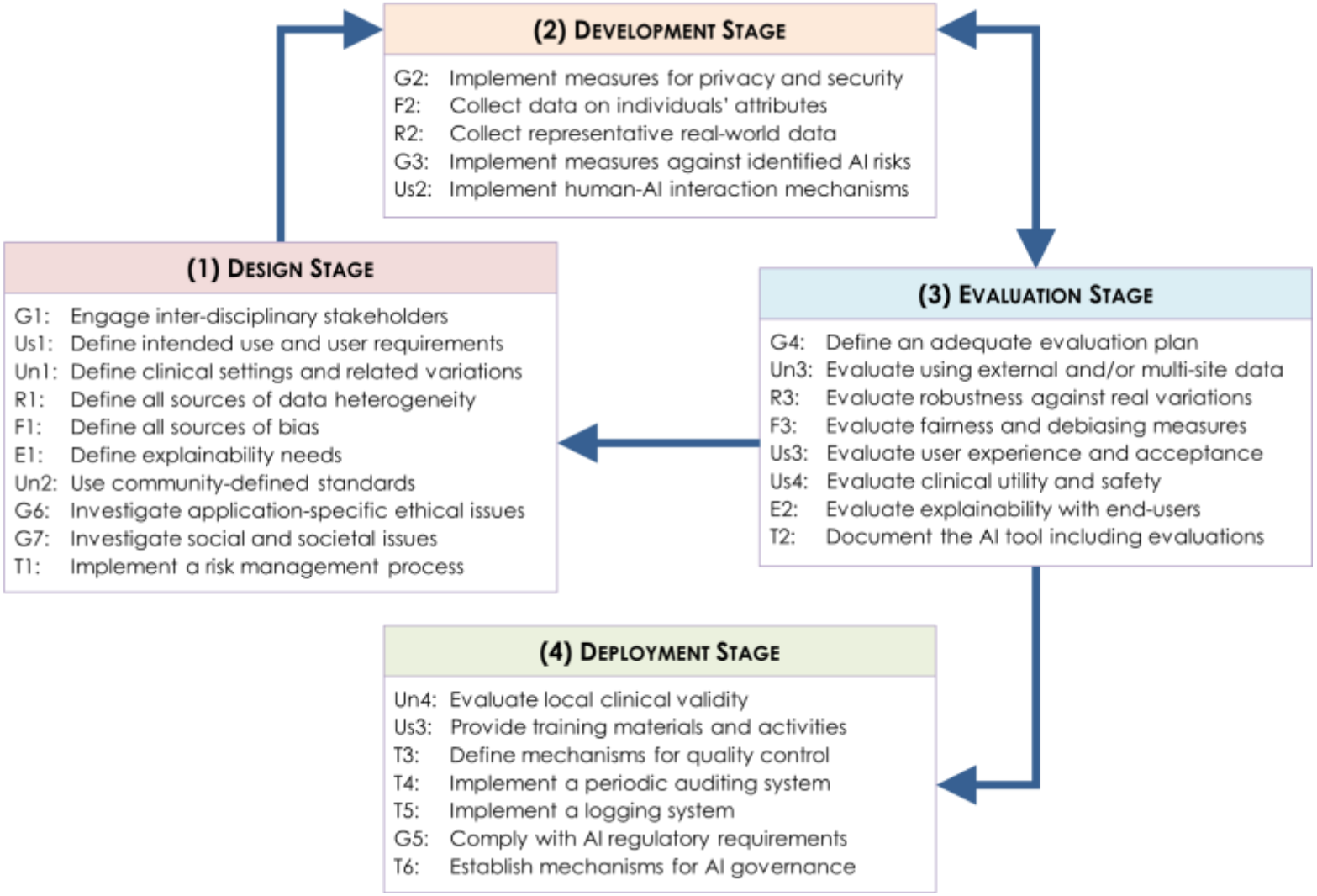
Operationalisation



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]

Operationalisation



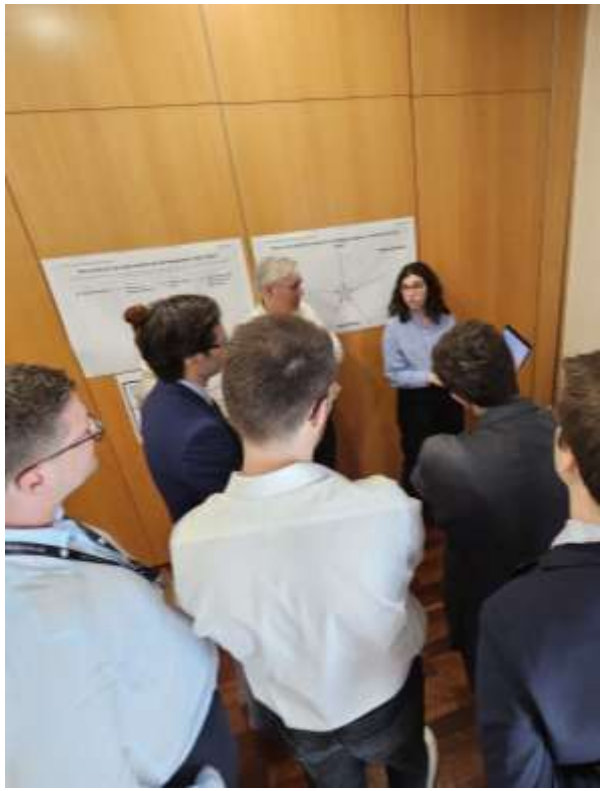
Best Practices – Design

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)

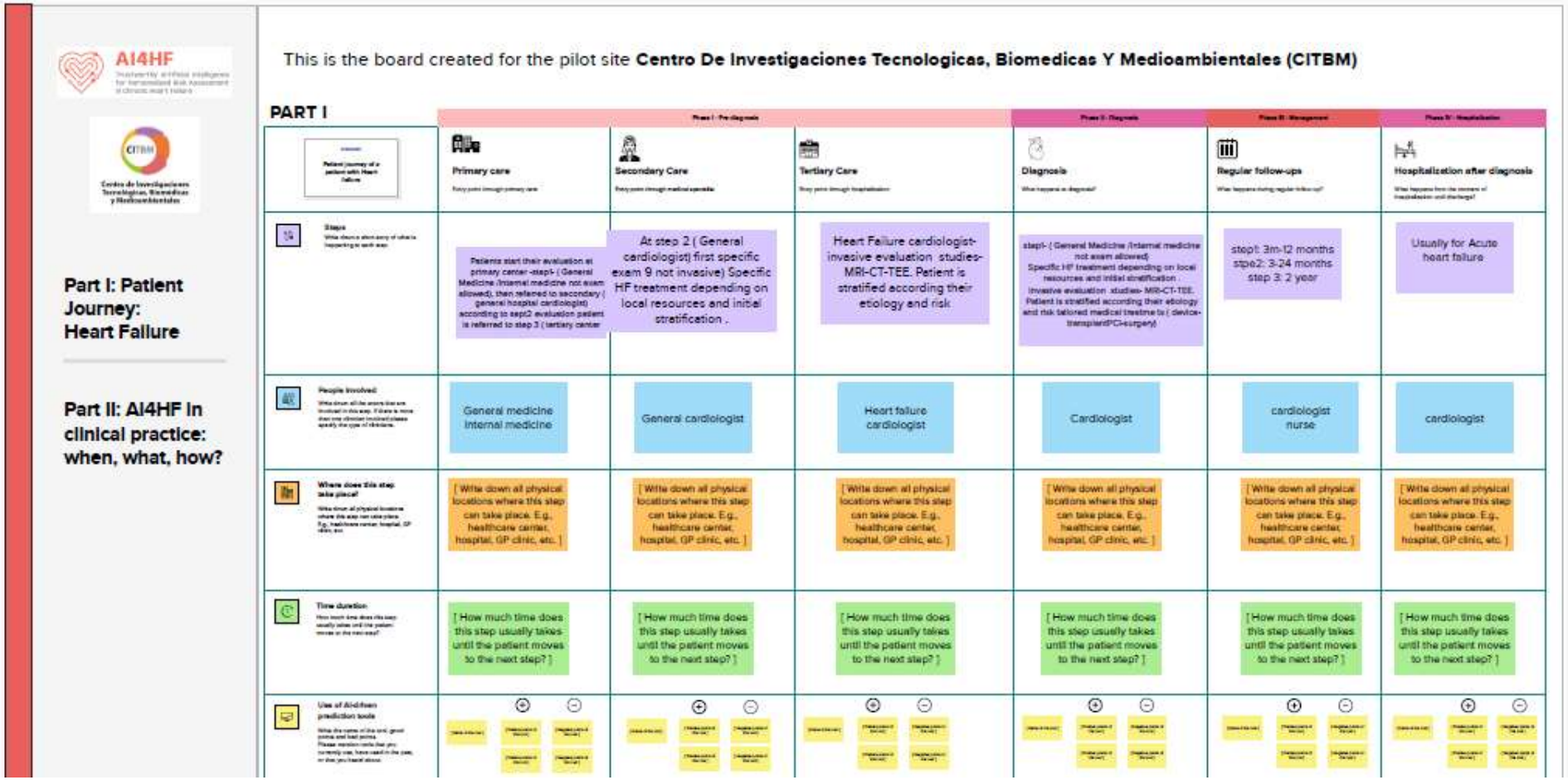
Stakeholder Engagement

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

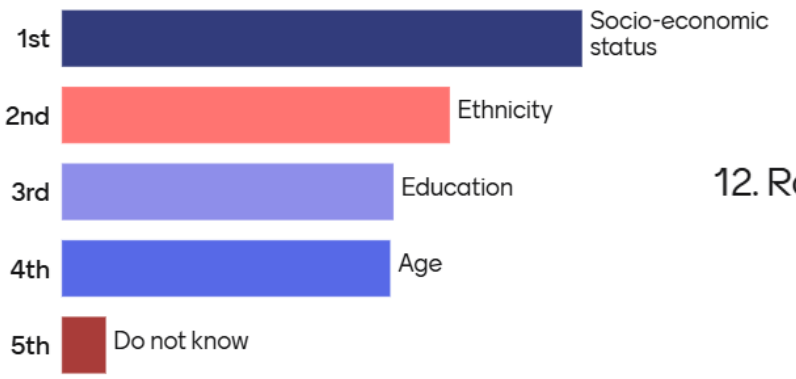


Stakeholder Engagement

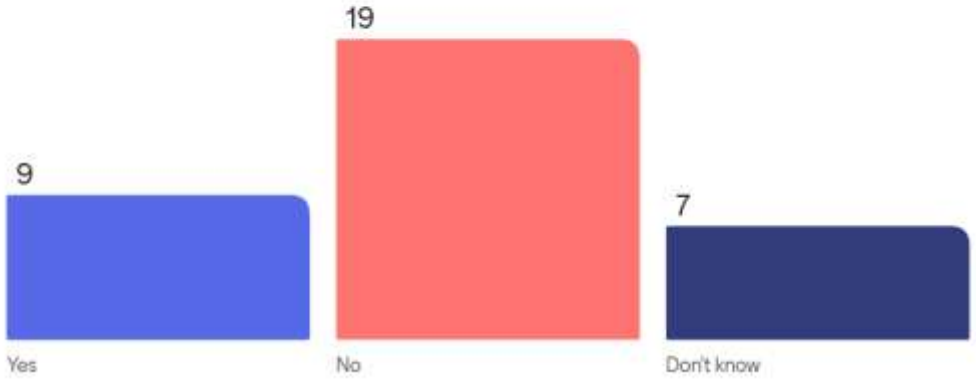


Stakeholder Engagement

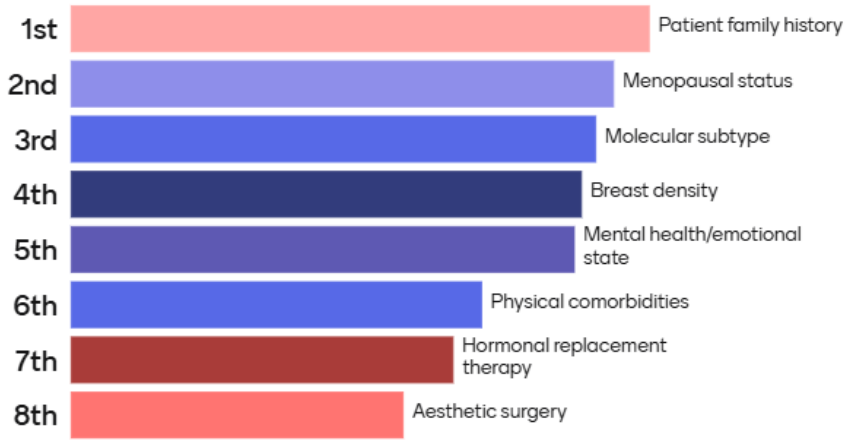
8. Rank the following variables based on their importance for bias estimation



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

Sources of Biases

AI4HF

Sources of Biases

Japanese

Quechua

European

Mestizo

Sources of Biases

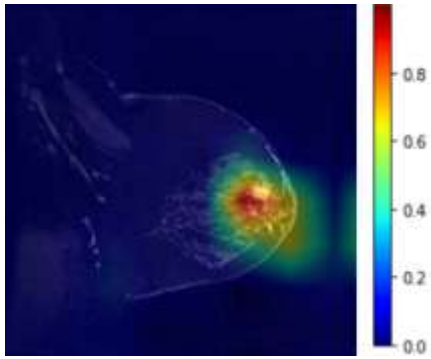
Lima: Sea level

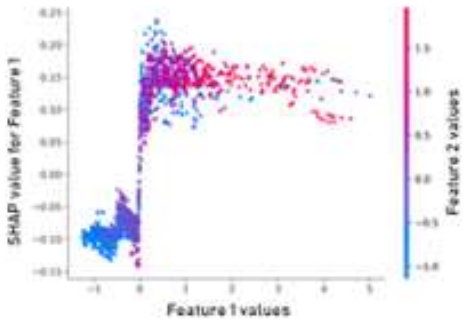
Arequipa: 2,335 m

Cusco: 3,400 m

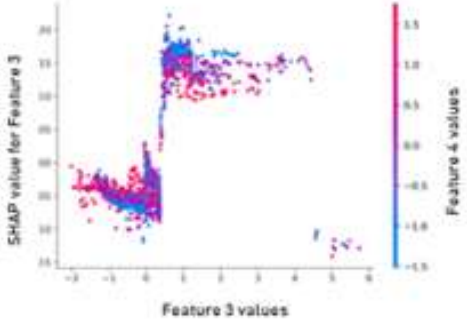
Rinconada: 5,100 m

Explainability Options

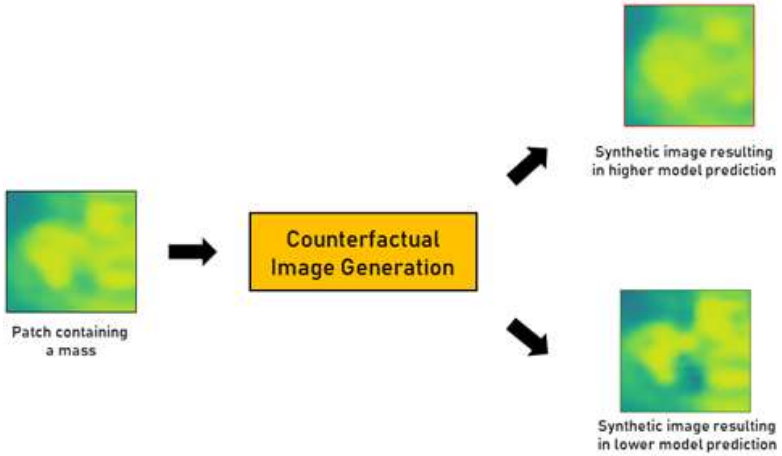


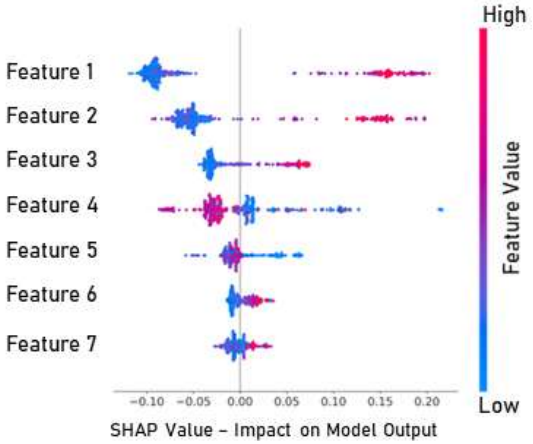


(a)



(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)

Data Collection

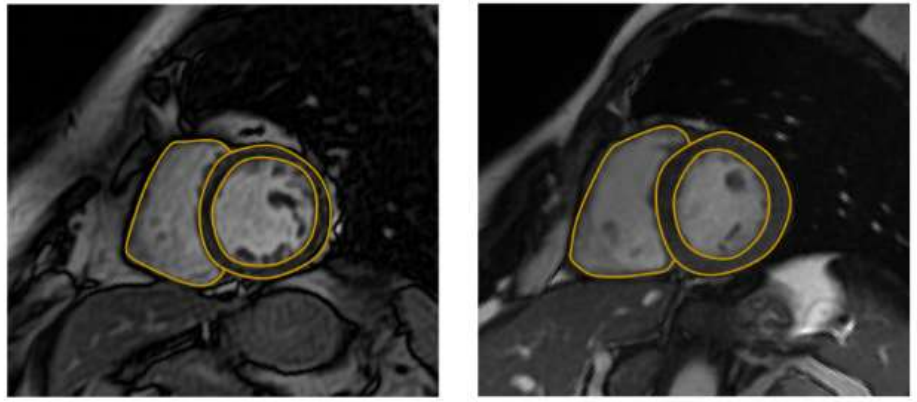
Patient	Age	Sex	Ethnicity	Neighbourhood	Altitude	Skin colour	Education	Menopause
Patient001	20	Male	E. Europe	Gracia	0	White	High-School	No
Patient002	20	Male	S. Europe	Gracia	0	White	University	No
Patient003	30	Male	N. Africa	Horta	500	White	University	No
Patient004	30	Male	N. Africa	Horta	500	White	University	No
Patient005	40	Male	S. Africa	Poblenou	1000	Black	High-School	No
Patient006	40	Female	S. Africa	Poblenou	1000	Black	High-School	Yes
Patient007	50	Female	S. Asia	Gervasi	1500	White	High-School	Yes
Patient008	50	Female	E. Asia	Gervasi	1500	White	University	Yes
Patient009	60	Female	L. America	Eixample	3000	White	University	No
Patient010	60	Female	L. America	Eixample	3000	Black	University	No

Data Representativeness

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 40, NO. 12, DECEMBER 2021 3543

Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge

Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreño, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, *Senior Member, IEEE*, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Viladés, Martín L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Çavuş, Steffen E. Petersen, Sergio Escalera, Santi Segui, José F. Rodríguez-Palomares, and Karim Lekadir



Centre	Vendor	Model	Field strength (T)
1	Siemens	MAGNETOM Avanto	1.5
2	Philips	Achieva	1.5
3	Philips	Achieva	1.5
4	GE	Signa Excite	1.5
5	Canon	Vantage Orian	1.5
6	Siemens	MAGNETOM Skyra	3.0

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

AI4HF

Spain

Vall d'Hebron
Institut de Recerca

Netherlands

Amsterdam UMC
University Medical Centers

UMC Utrecht

Czech Republic

FNUSA
ICRC
ST. ANNE'S UNIVERSITY HOSPITAL BRNO
INTERNATIONAL CLINICAL RESEARCH CENTER

Peru

Instituto Nacional
Cardiovascular

Tanzania

Muhimbili University of Health and
Allied Sciences

Usability Evaluation

Human evaluators in 5 sites:

- ✓ 2 GPs at each site
- ✓ 2 cardiologists at each site
- ✓ 7 patients for each clinician
- ✓ 2 IT/data manager
- ✓ 50% male + 50% female
- ✓ 50% early-career, 50% > 5-year experience

Explainability Evaluation

AI Explainability Score:

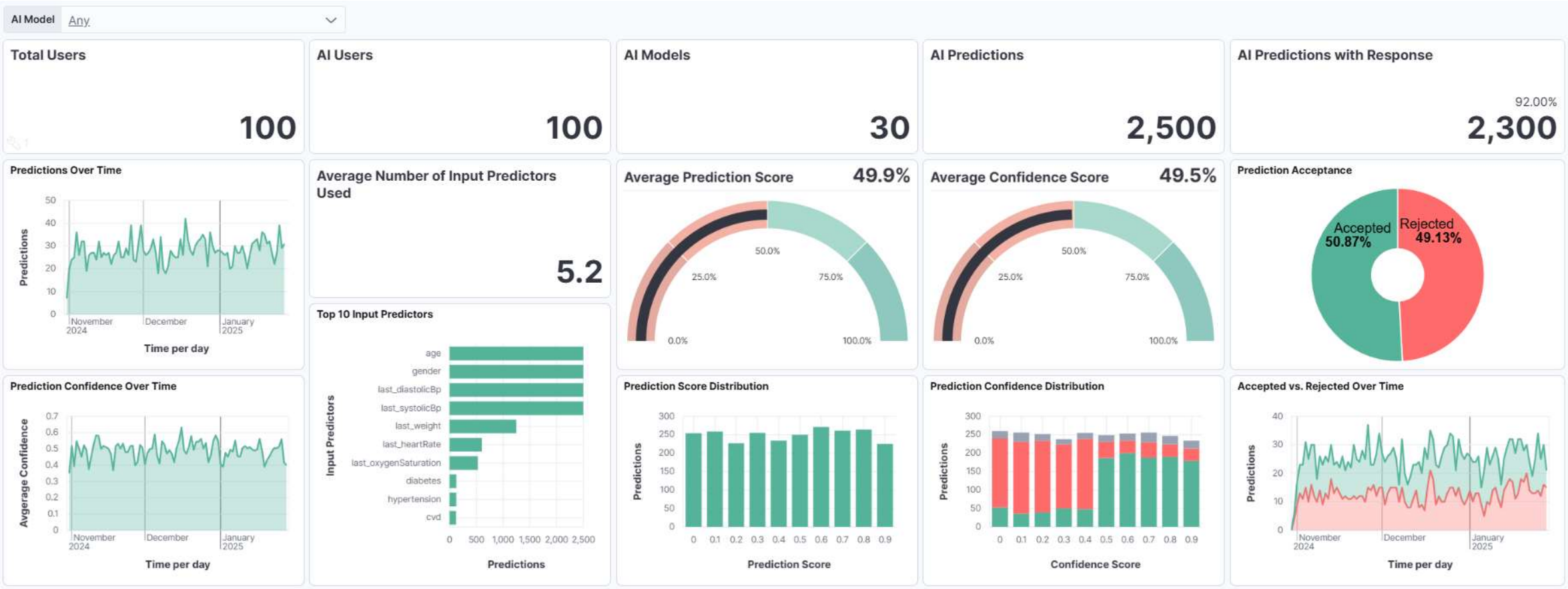
Questions to assess explainability with clinicians:

- Did you find the AI explanations clear?
- Did you find the AI consistent between cases?
- Did the AI explanations increase confidence in the decisions?
- Were the AI visualisations easy to use?

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)

Periodic Auditing



AI4HF

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

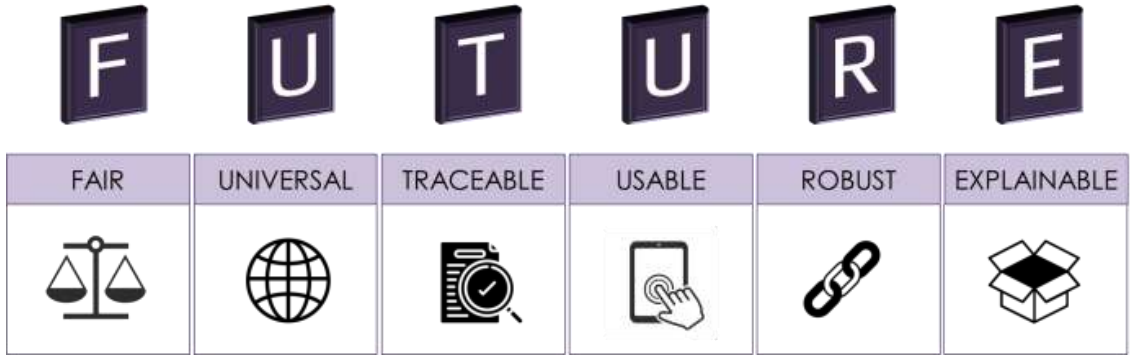
- Despite major advances in medical artificial intelligence (AI) research, clinical adoption of emerging AI solutions remains challenging owing to limited trust and ethical concerns
- The FUTURE-AI Consortium unites 117 experts from 50 countries to define international guidelines for trustworthy healthcare 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.[4-8]

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,

ESC European Society of Cardiology

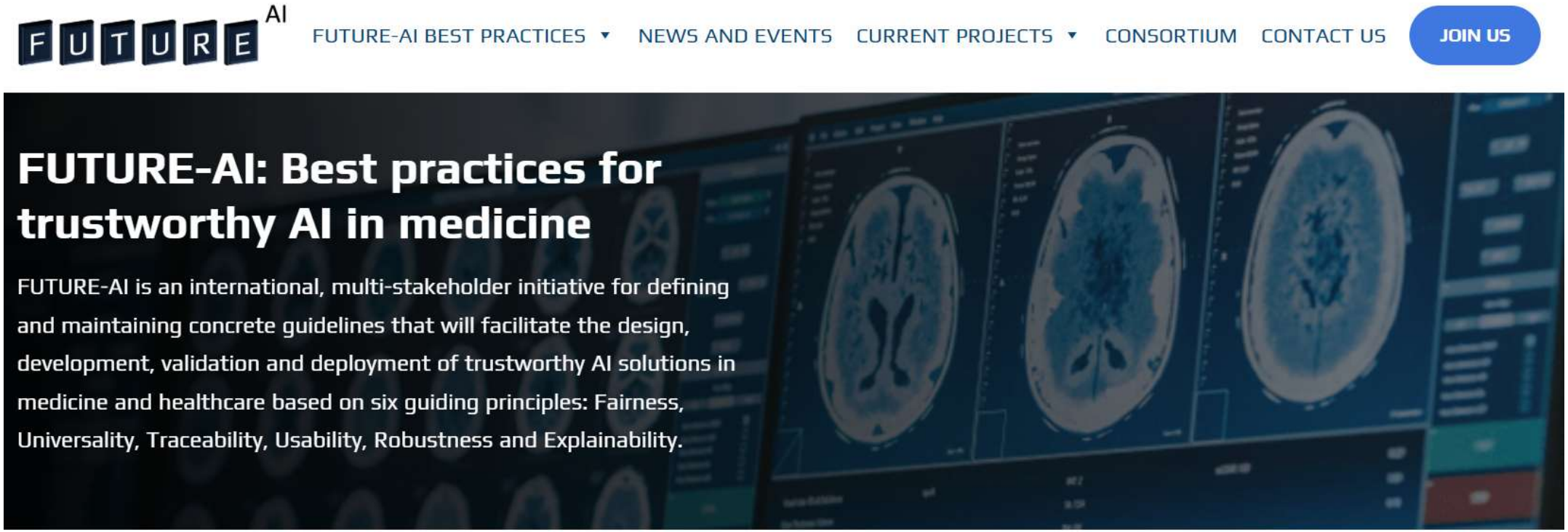


Next Steps

- Feedback gathering from over 10 EU projects
- Feedback gathering from external actors (e.g. ESC members)
- Next papers:
  - FUTURE-AI guideline: The patient perspective
  - FUTURE-AI guideline: Clinical validation methods
  - FUTURE-AI guideline: Implications for AI regulations
  - FUTURE-AI guideline: Adaptations for large language models

karim.lekadir@ub.edu

Join Us!



karim.lekadir@ub.edu

Many Thanks!

karim.lekadir@ub.edu