What are the requirements for evidence generation of AI models to be guideline proof?

Rohan Khera, MD, MS
Assistant Professor, Yale University

Director, Cardiovascular Data Science (CarDS) Lab
Clinical Director, CORE Center for Health Informatics and Analytics
Associate Editor, JAMA
cards-lab.org | @rohan_khera

Funding and Disclosures

Funding

NIH
R01HL167858, R01AG089981, and K23HL153775

Foundations
Doris Duke Charitable Foundation, Blavatnik Fund for Innovation

Research funding through Yale
Bristol-Myers Squibb, Novo Nordisk, and BridgeBio

In-kind contribution from Apple Inc. through Investigator Support Program

Disclosures

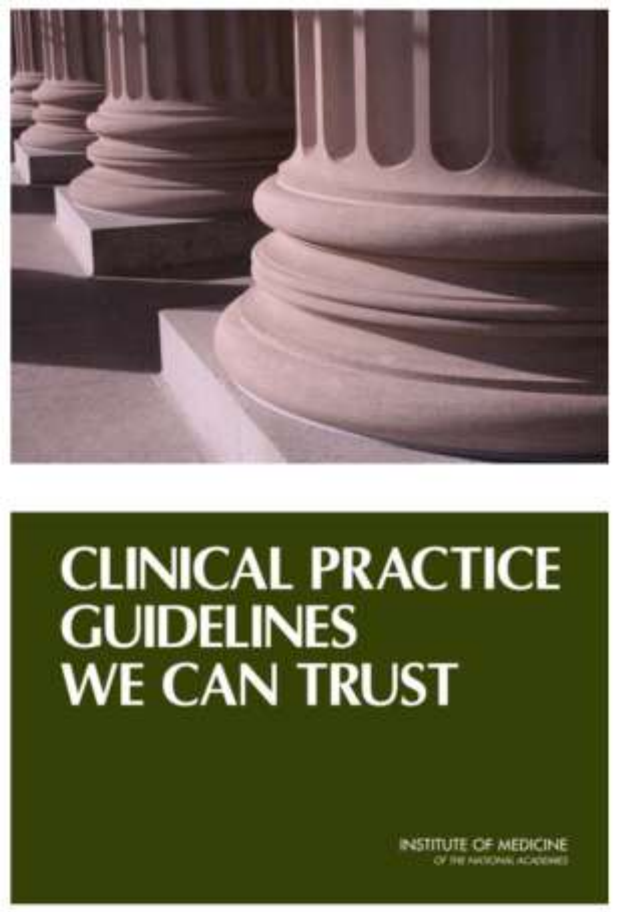
Pending patents related to the methods

Research tools presented

Co-founder of Ensight-AI and Evidence2Health

CarDS LAB

Framework for Trustable Clinical Practice Guidelines



- Detailed, explicit and publicly **accessible process**
- **Guideline writing group** with no conflicts of interest
- Writing group with **multidisciplinary expertise**
- Establish **evidence foundations** for each recommendation
- **Quality, quantity, and consistency** of the aggregated available evidence
- Level of strength based on the **certainty of evidence**
- All recommendations articulated in a **standard format**
- External reviewers from **full spectrum of relevant stakeholders**
- **Monitoring literature** for emerging relevant new evidence

Practice guidelines we can trust. Institute of Medicine. 2011

Guideline process for major cardiovascular societies

ESC

Systematic review and evidence identification

- Clinical question of interest
- Comprehensive search
- Selection of relevant studies

Evidence appraisal and quality assessment

- Nature + quality of studies
- Level of evidence
- Generalizability

Synthesis of evidence for recommendation

- Examining multiple studies
- Benefit/Harm balance
- Defining recommendation

CarDS LAB

ACC/AHA Task Force. Circulation. 2014

Reporting of guideline recommendations

Classes of Recommendation

Class I	Is recommended or is indicated
Class II	Conflicting evidence or divergence of opinion
Class IIa	Should be considered
Class IIb	May be considered
Class III	Is not recommended

Levels of evidence

Level of evidence A	Data derived from multiple RCTs or metanalysis
Level of evidence B	Data derived from single RCT
Level of evidence C	Consensus of opinion of experts and/or small, retrospective, registries

ESC Clinical practice guidelines: processes and Procedures. 2022

How does this framework apply to AI?

ESC

Transformative Role of AI in Cardiology

Optimizing Detection

Predicting Development

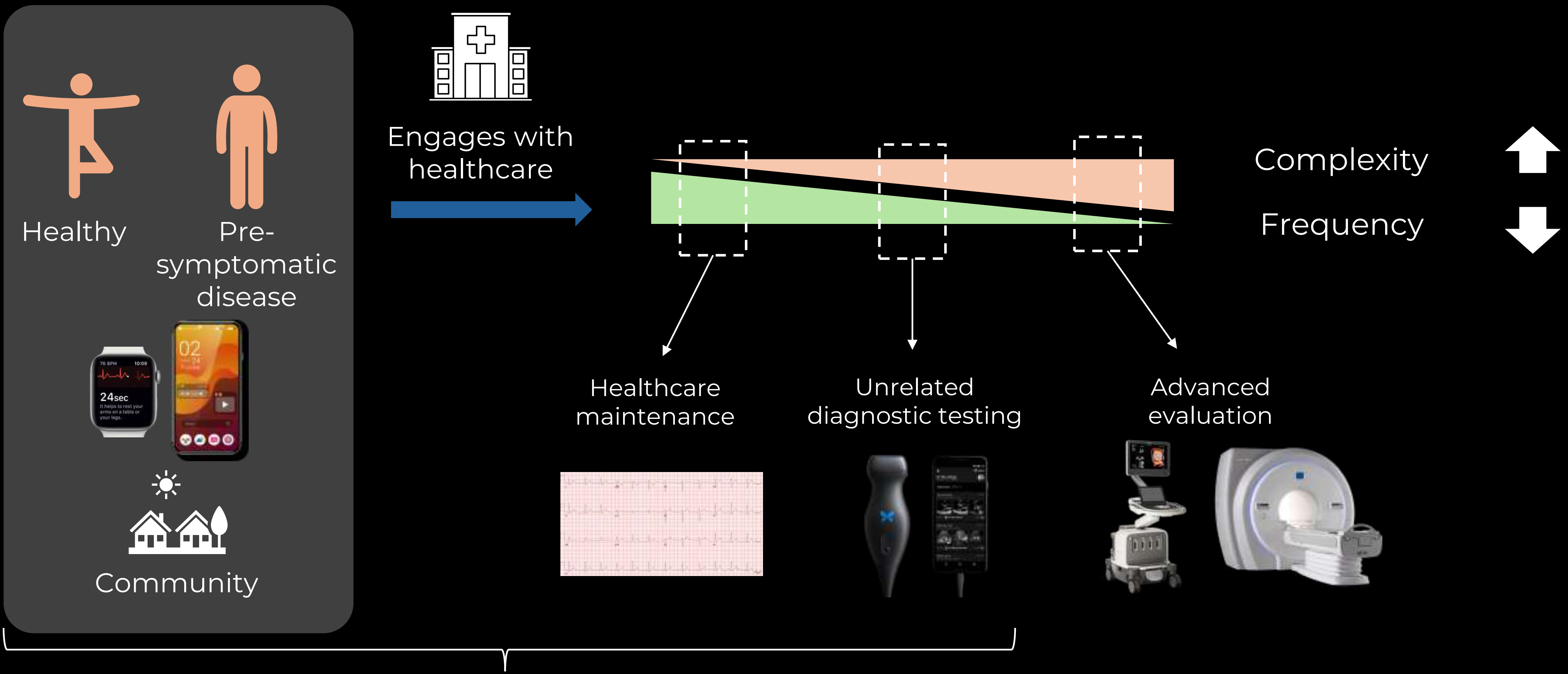
Precise Prognostication

Enabling Care Quality

Biological Discoveries

CarDS LAB

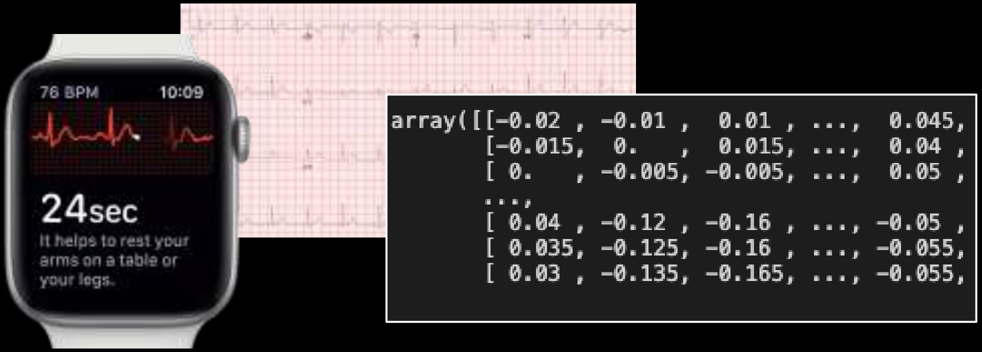
Khera et al. JACC. 2024



Experts **cannot consistently diagnose** structural heart disorders ("hidden disease")

Khera R. *JACC*. 2024 & *JAMA*. 2023

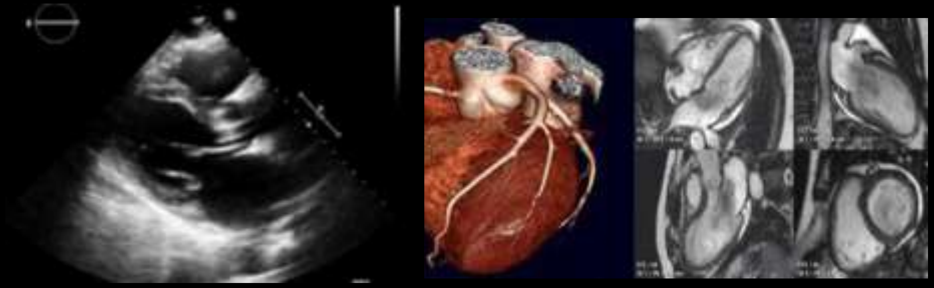
AI for augmenting cardiovascular diagnosis



Extensive AI-ECG Applications

- LV Systolic Dysfunction
- Hypertrophic CMP
- Cardiac amyloid

and many more…



Broad cardiac imaging applications

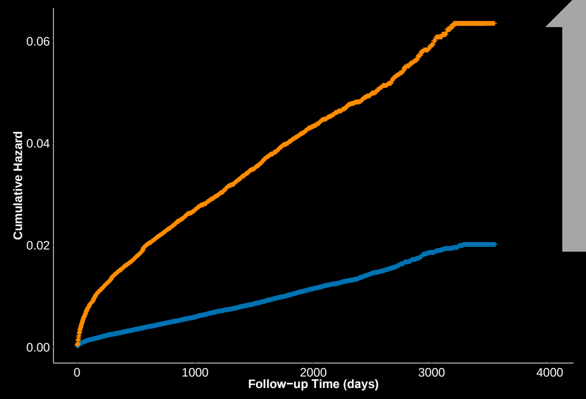
- Aortic stenosis
- Obstructive CAD
- Cardiac Amyloid

and many more…

Sangha V…Khera R. *Circulation*. 2023. Dhingra LD…Khera R. *JACC*. 2025.
Oikonomou E…Khera R. *JAMA Cardiol*. 2024. Aminorroaya A…Khera R. *EHJ Dig Health*. 2025 In print

AI for augmenting prediction and prognostication

Predicting HF risk

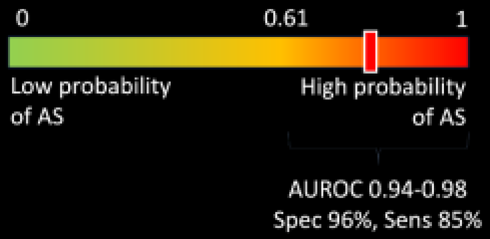


4-fold higher risk
HR 4.4 (4.1-4.7)

Performance vs Traditional Scores (c-statistic)
AI-ECG – 0.77 (0.74-0.79)
PCP-HF – 0.77 (0.75-0.79)
PREVENT-HF – 0.78 (0.76-0.80)

Dhingra LD...Khera R. *EHJ.* 2025
Dhingra LD, Aminorroaya A...Khera R. *JAMA Cardiology*. 2025 In print

Predicting AS Progression



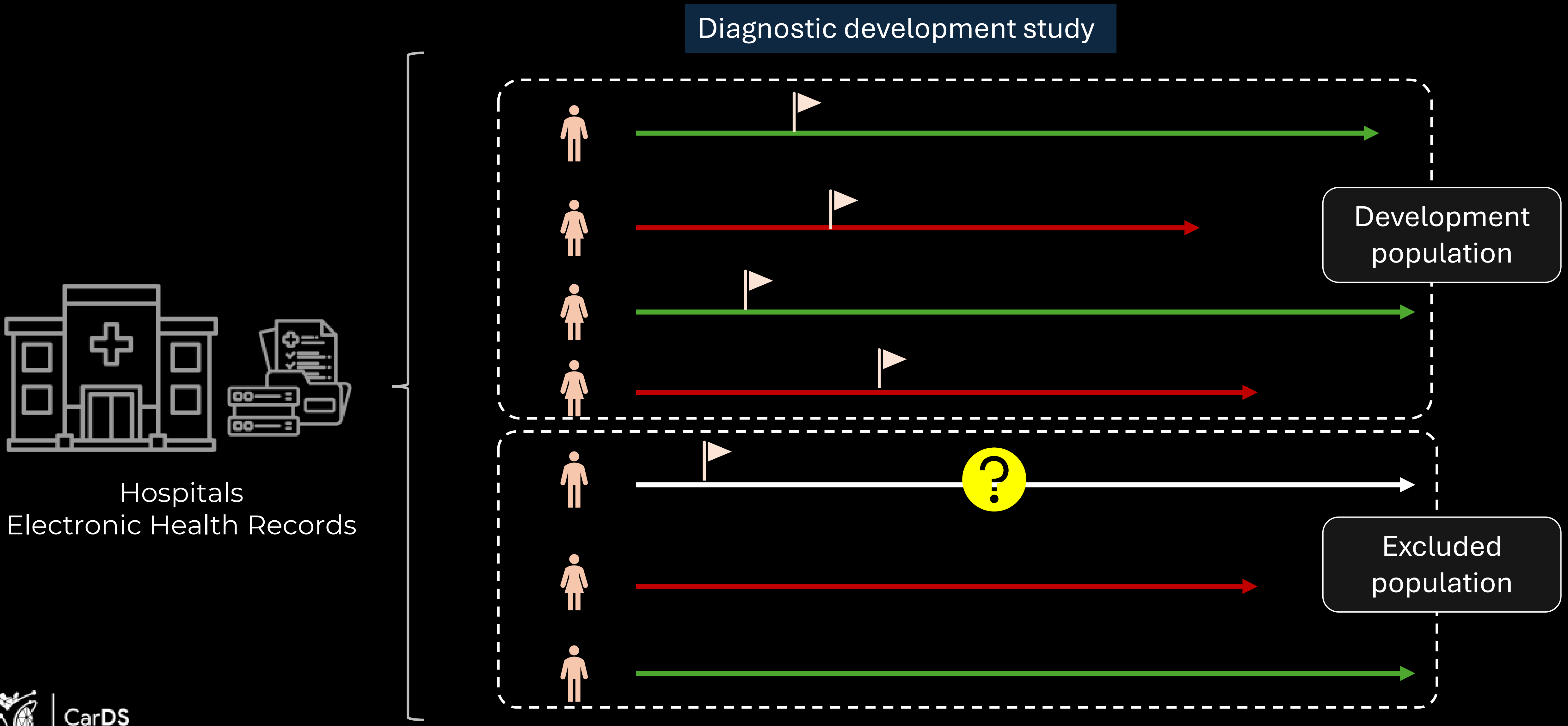
DASSi (Digital Aortic Stenosis Severity index)

Novel AI Biomarker
AS development
Progression from mild/moderate AS
Adverse outcomes

Holste G...Khera R. *EHJ.* 2023
Oikonomou E...Khera R. *JAMA Cardiology*. 2024

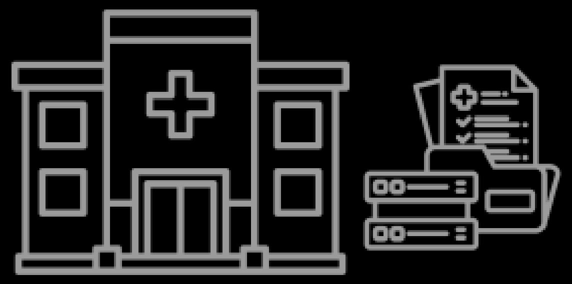
CarDS LAB

Common design of AI development/validation studies



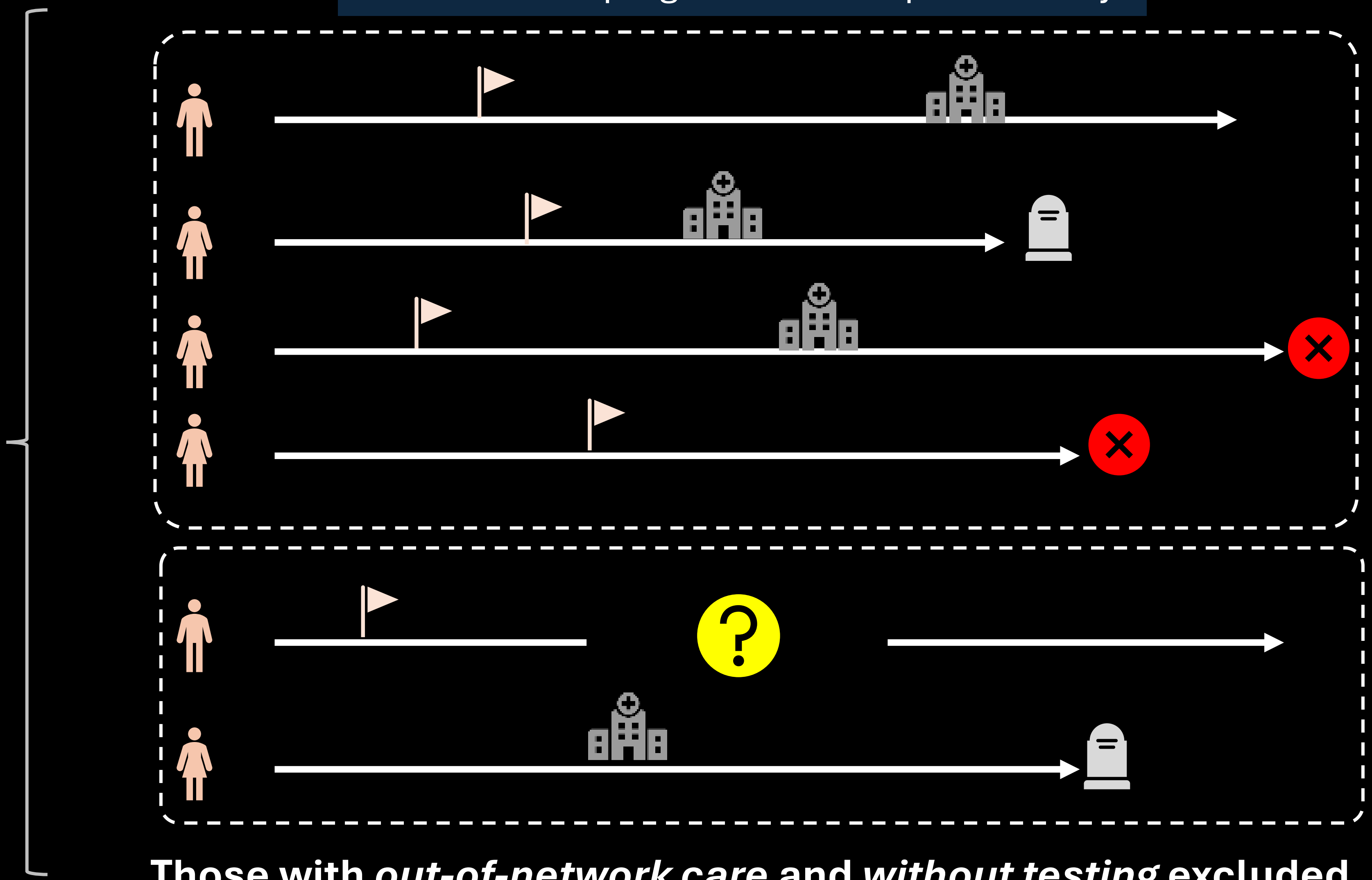
Those *not seeking care* are also excluded

Common design of AI development/validation studies



Hospitals
Electronic Health Records

Predictive and prognostic development study



Those with *out-of-network care* and *without testing* excluded

AI validation required for regulatory clearance

Retrospective study of the intervention at ≥3 sites

Case-control design acceptable

↓

Available for clinical use

↓

Would the ***evidence be sufficient*** to recommend use in clinical practice?

Randomized Clinical Trials of AI

ESC

Benefits	Challenges
Addresses unmeasured confounding	High resource need
Eliminates confounding by indication	Testing of tool vs mechanism
Enables blinding or masking	Unclear generalizability to diverse populations
Strong evidence for causality	Not suitable for complex and dynamic models
Less selection bias and comparable groups	Publication and reporting bias

CarDS LAB

Interventional Studies of AI applications

64 Studies
Prospective validation of AI interventions in cardiology

11 RCTs
Across 6 different clinical use case

All published after 2021

- Acute conditions
- Arrhythmias
- Coronary Artery Disease
- Ejection Fraction
- Heart Failure
- Patient Management

Most RCTs for operational interventions

3 RCTs of clinical interventions

None with hard clinical endpoints

Moosavi A et al. *JACC Adv*. 2024

What recommendation would we assign to contemporary AI technologies?

Classes of Recommendation

Class I	Is recommended or is indicated
Class II	Conflicting evidence or divergence of opinion
Class IIa	Should be considered
Class IIb	May be considered
Class III	Is not recommended

Levels of evidence

Level of evidence A	Data derived from multiple RCTs or metanalysis
Level of evidence B	Data derived from single RCT
Level of evidence C	Consensus of opinion of experts and/or small, retrospective, registries

The DETECT-AS Study

AI-Enabled AS Detection and Prognostication

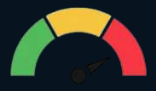
National Institute on Aging

Aim 1: Multicenter RCT

Older adults without advanced AS attending routine visit

(n=410)

Portable ECG + AI-ECG



vs. standard care

POCUS + AI-POCUS

Detection of advanced AS

Aim 2: Cohort Study

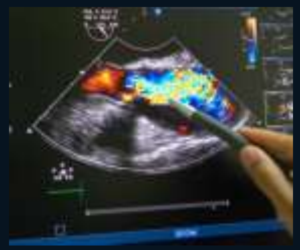
Older adults with baseline echo showing early stage AS based on MRR

(n=210)

Digital biomarker for precision risk stratification (DASSi)

+

Follow-up echo



Progression to advanced AS

Cedars Sinai

Yale

Mount Sinai

Path for guideline-proof AI development

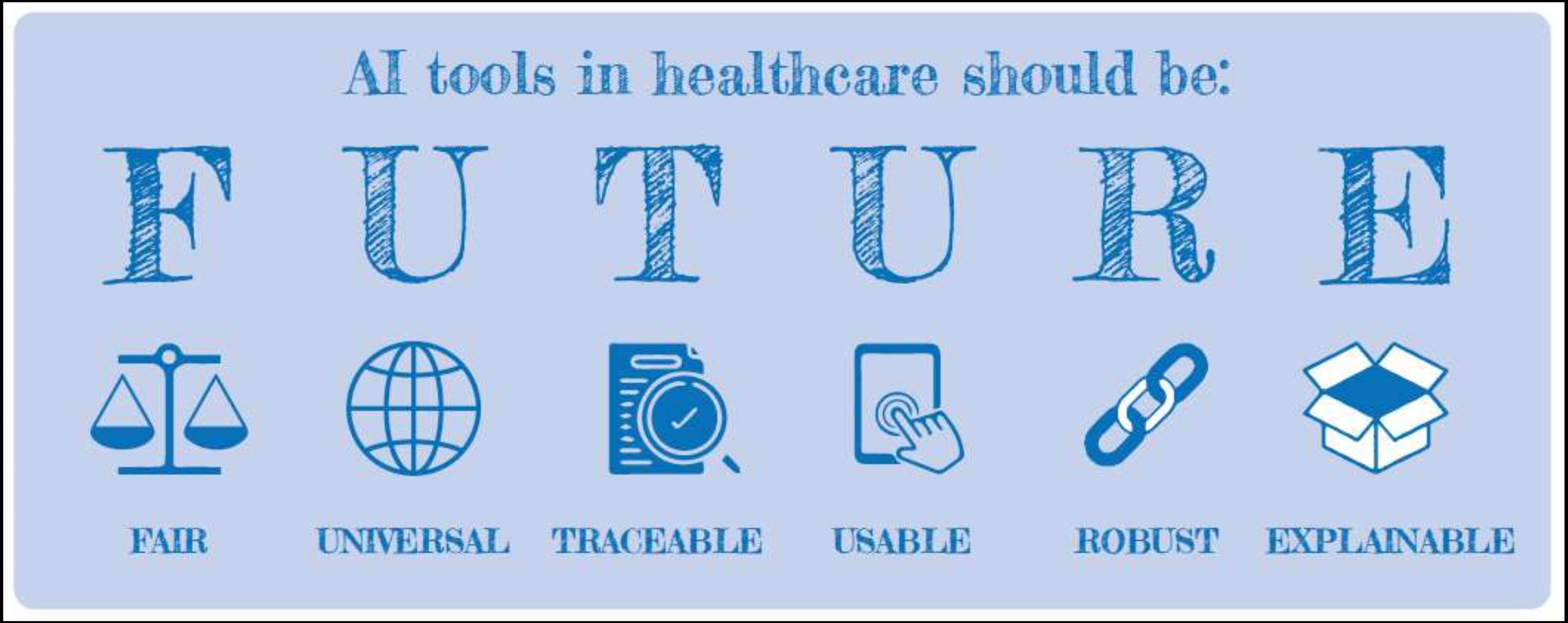
Best practices for AI development

- Development in diverse population
- Robustness check in population with intended use
- External validation in distinct populations and settings
- Enable explainability and algorithmic uncertainty
- Evaluation in an RCT, especially in new domains

Pragmatic guideline development process

- Broadening the pool of AI experts in guideline committees
- Evaluation of methodological rigor in observational designs
- Risk-benefit calculus in evidence evaluation
- Flexible, evolving recommendations for dynamic models
- Guidance on appropriate care strategy for algorithmic care

Khera R et al. *JAMA*. 2023



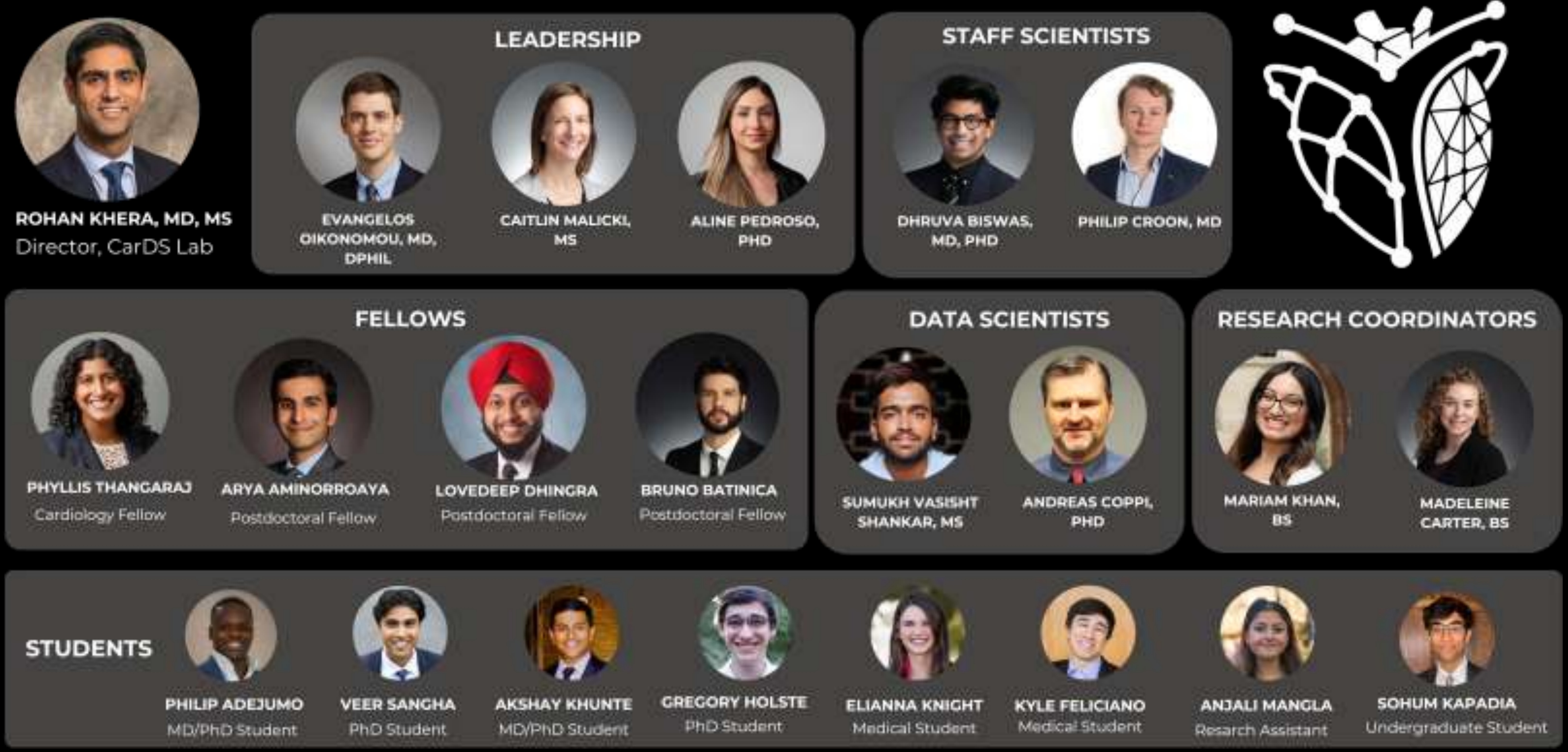
Lekadir K et al. BMJ. 2025

Conclusions

- AI can **augment human capacity** and improve care processes
- Evidence-base for AI intervention is evolving with **largely observational studies**
- To balance rapid adoption with intended impact, **need nimble guideline processes**
- Critical to guide clinical community to **incorporate algorithm-enabled care**

Follow our work: ***CarDS-Lab.org***

Team



Collaborators

Harlan Krumholz
Cynthia Brandt
Erica Spatz
Robert McNamara
Eric Velazquez
Lucila Ohno-Machado
Bobak Mortazavi
Norrisa Haynes
Marc Suchard
Folkert Asselbergs
Girish Nadkarni
Tom Ribeiro
James de Lemos
David Ouyang

Contact

Rohan Khera, MD, MS
rohan.khera@yale.edu
@rohan_khera

Thank you

National Institutes of Health
Doris Duke Charitable Foundation
Blavatnik Family Foundation
Bristol Myers Squibb
Novo Nordisk
BridgeBio