AI Evaluation in Real World

Xiaoxi Yao, PhD MPH MS FACC FAHA
(First name “Shao-She” meaning “sunlight in the morning” in Chinese, she/her pronouns)
Robert D. and Patricia E. Kern Scientific Director of Pragmatic Trials and Evaluation Professor of Health Services Research, Medicine

January 7th 2025

THREE STAGES OF AI EVALUATION IN ROUTINE PRACTICE

Stage	Primary Expertise
1. Technical performance	Data science & informatics
2. Usability & Acceptability	AI-human interaction & mixed methods studies
3. Health & Economic Impact	Health services research

1 TECHNICAL PERFORMANCE

TWO TYPES OF ML MODELS

Do we have ground truth for model outcomes?

1. Ground Truth Yes – examples: models predicting an adverse outcome
2. Ground Truth No – examples: causal machine learning and LLM

TECHNICAL PERFORMANCE EVALUATION

- Outcomes with ground truth
  - Discrimination
  - Calibration
  - Sensitivity, specificity, negative predictive value, positive predictive value
- Outcomes without ground truth
  - Causal ML
  - LLM

COUNTERFACTUAL OUTCOME

Universe A
Receive Control

Universe B
Receive Intervention

Compare
Outcomes

LLM TECHNICAL PERFORMANCE #1

- NLP Metrics
  - BLEU
  - ROUGE-L
  - BERT Score
  - MEDCON

Van Veen D et al. Nat Med. 2024

LLM TECHNICAL PERFORMANCE #2

Attribute	Score	Ideal note
1. Correctness	1 Not at all; 2 Slightly; 3 Moderately; 4 Very; 5 Extremely	The note is generally free of mistakes and incorrect information.
2. Completeness	1 Not at all; 2 Slightly; 3 Moderately; 4 Very; 5 Extremely	The note is complete and documents all of the issues of importance to the provider and patient.
3. Conciseness	1 Not at all; 2 Slightly; 3 Moderately; 4 Very; 5 Extremely	The note keeps to important topics without adding extraneous or unrelated information.
4. Appropriateness	1 Not at all; 2 Slightly; 3 Moderately; 4 Very; 5 Extremely	The note demonstrates a style consistent with medical care and professionalism.
5. Acceptability	1 Not at all; 2 Slightly; 3 Moderately; 4 Very; 5 Extremely	The note is acceptable *as is* for release to consulting clinicians.

Blinded clinical reviewers

2 USABILITY & ACCEPTABILITY

USABILITY & ACCEPTABILITY

Technology Acceptance Model

- Perceived ease of use
- Perceived usefulness
- Intention to use

XAI scale

- Trust in AI

Holden et al. Journal of biomedical informatics 2010;43(1):159-172.
Hoffman et al. 2021

3 HEALTH IMPACT

HOW TO EVALUATE IMPACT

- Randomized controlled trial
- Non-randomized studies
  - Non-randomized clinical trial
  - Prospective observational study
  - Retrospective cohort study

EXAMPLE 1 RCT

ECG AI Guided Screening for Low Ejection Fraction (EAGLE) Trial

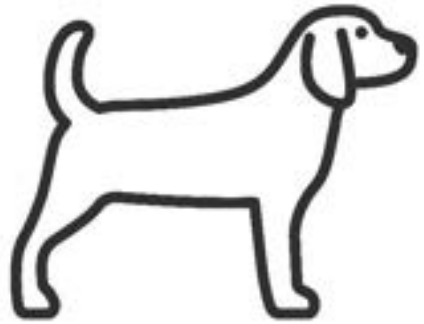
Yao & Noseworthy et al. Nature medicine. 2021;27(5):815-819.

ECG AI GUIDED SCREENING FOR LOW EJECTION FRACTION (EAGLE)

- A pragmatic, cluster-randomized controlled trial (NCT04000087)
- 358 clinicians and 22,641 patients from 45 primary care clinics in Minnesota and Wisconsin
- One of the first large RCTs to assess whether AI improves patient care in everyday practice

Yao & Noseworthy et al. Nature medicine. 2021;27(5):815-819.

EXAMPLE 2 NONRANDOMIZED TRIAL - BEAGLE

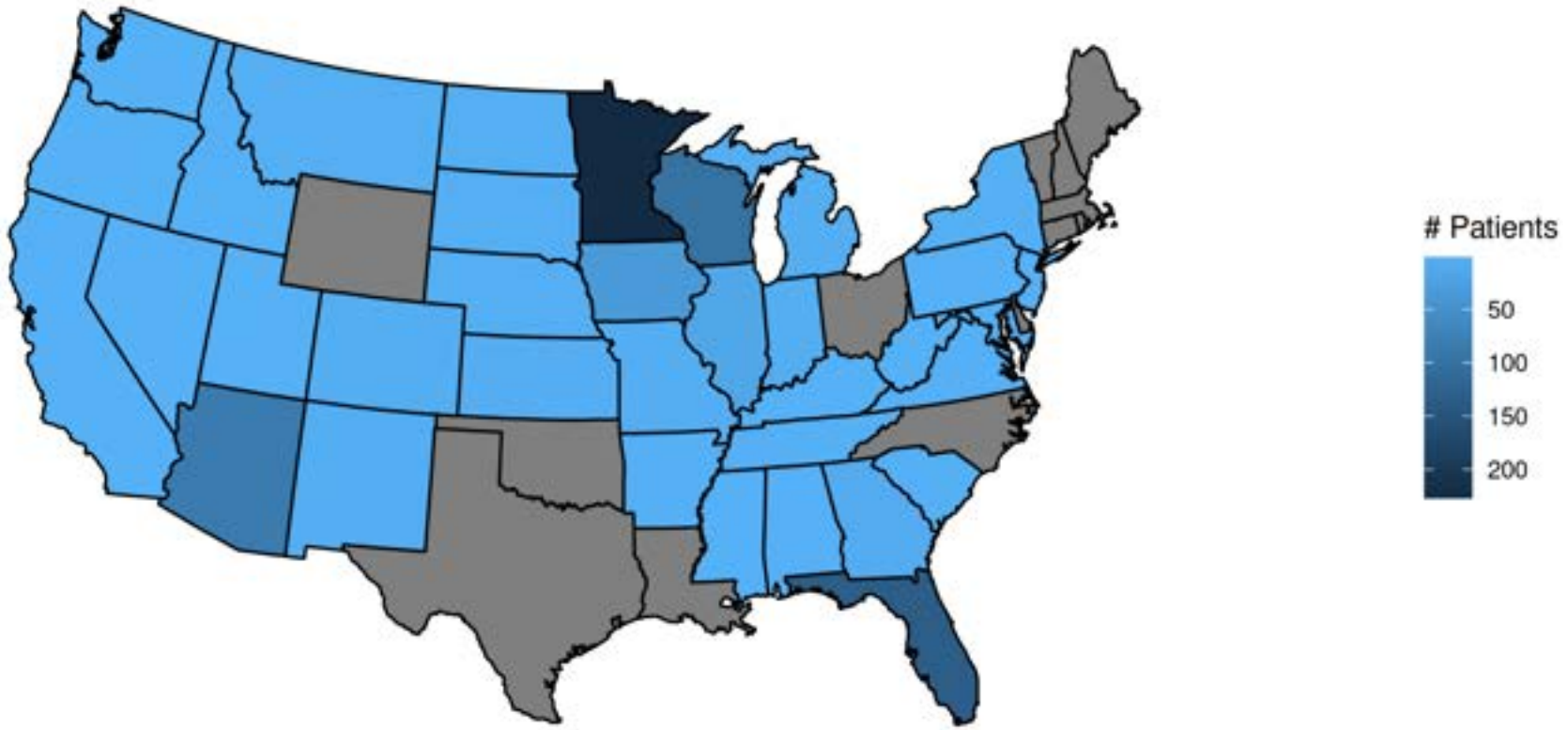


Batch Enrollment for AI-Guided Intervention to Lower Neurologic Events in Patients with AF

Borrow real-world controls

Noseworthy & Yao et al. Lancet 2022; 400 (10359): 1206-1212

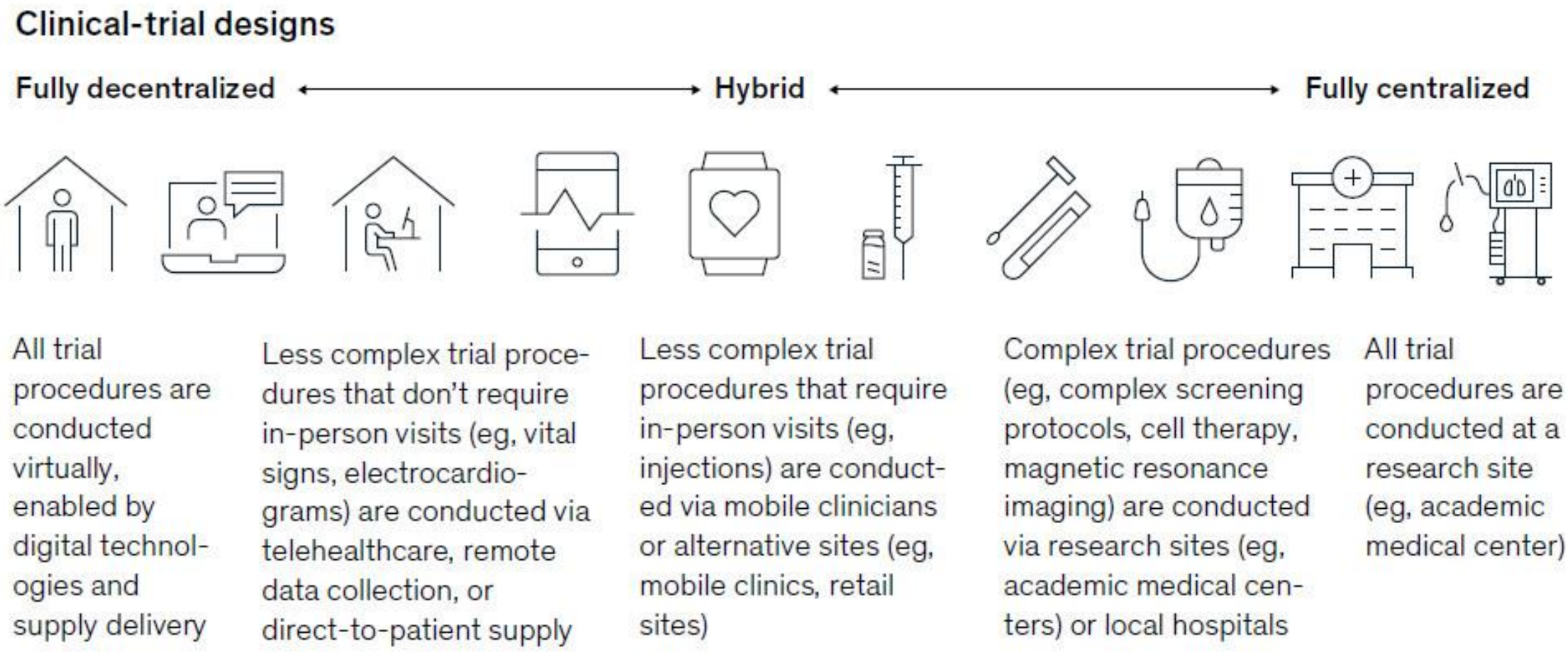
BEAGLE ENROLLMENT – DECENTRALIZED TRIAL



Meet patients where they are

Improve trial generalizability

DECENTRALIZED CLINICAL TRIALS



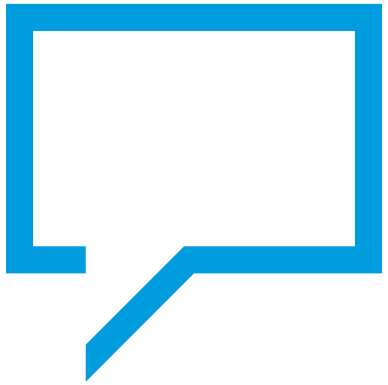
https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/no-place-like-home-stepping-up-the-decentralization-of-clinical-trials

CONCLUSIONS

AI Evaluation Framework

- Technical performance
- Usability & Acceptability
- Health impact

QUESTIONS & ANSWERS



yao.xiaoxi@mayo.edu

