MANCHESTER 1824

The University of Manchester

(What) can healthtech teach us about responsible AI?

DAVE WONG (@DRDAVECWONG)

DAVID.WONG@MANCHESTER.AC.UK

About Me

Lecturer in AI for healthcare (University of Manchester)

- Metrics from smartphone videos to assess hand tremor
- ECG disease classification
- Medical software development and deployment

Symptom Checkers

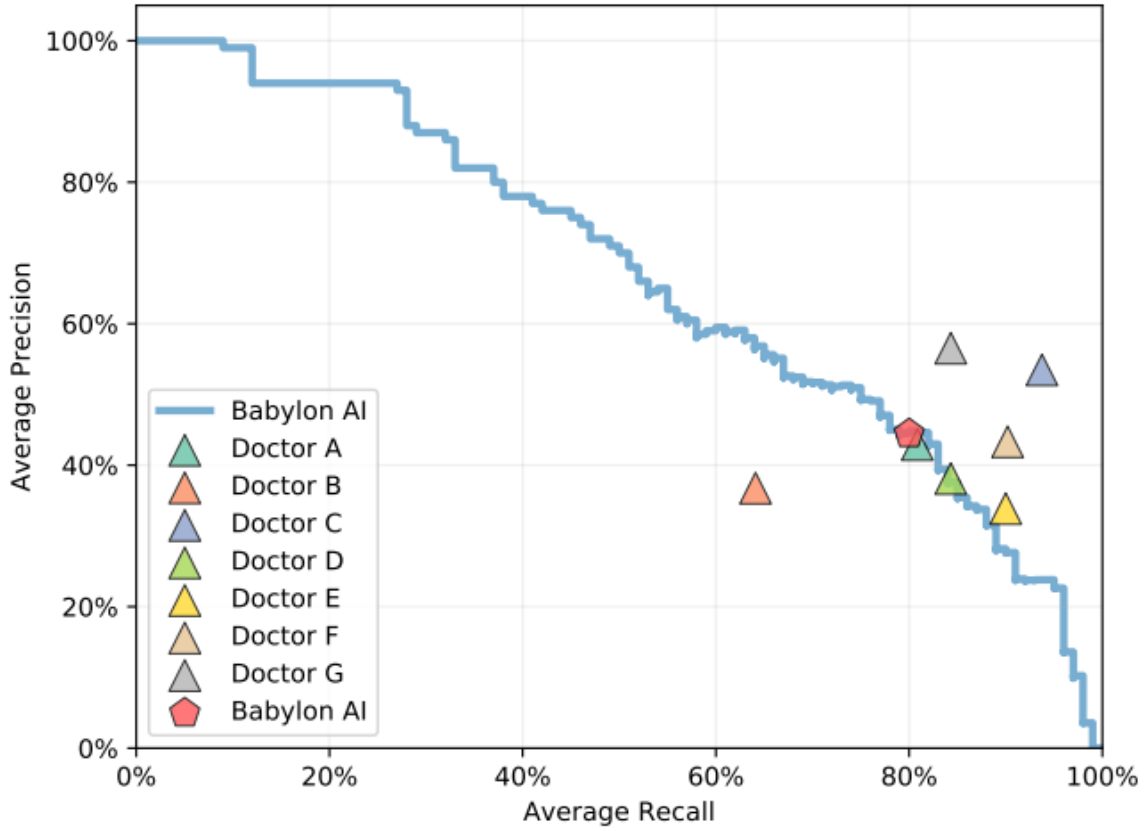
- Symptom checker that would ask questions to a patient and provide possible diagnoses

- In 2018, claims that it passed relevant parts of the Royal College of GPs exam

Symptom Checkers

- Safety of patient-facing digital symptom checkers, Lancet 2018, Fraser et al.

- Concerns around accuracy raised by multiple clinicians – example to follow



Symptom Checkers

NHS to sign up patients in tech revolution

Exclusive: Patients checking systems will be ... need to go to hospital

Denis Campbell
Health policy editor

Thu 23 May 2019 21.30 BST

TOPICS MAIN M

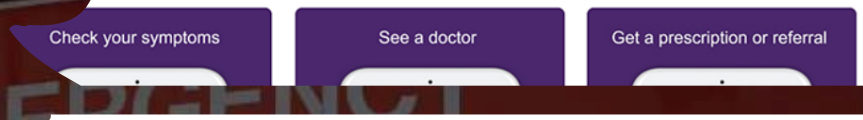
ANZ ASIA EMEA Globa

...n's triage chatbot launches in ...Columbia

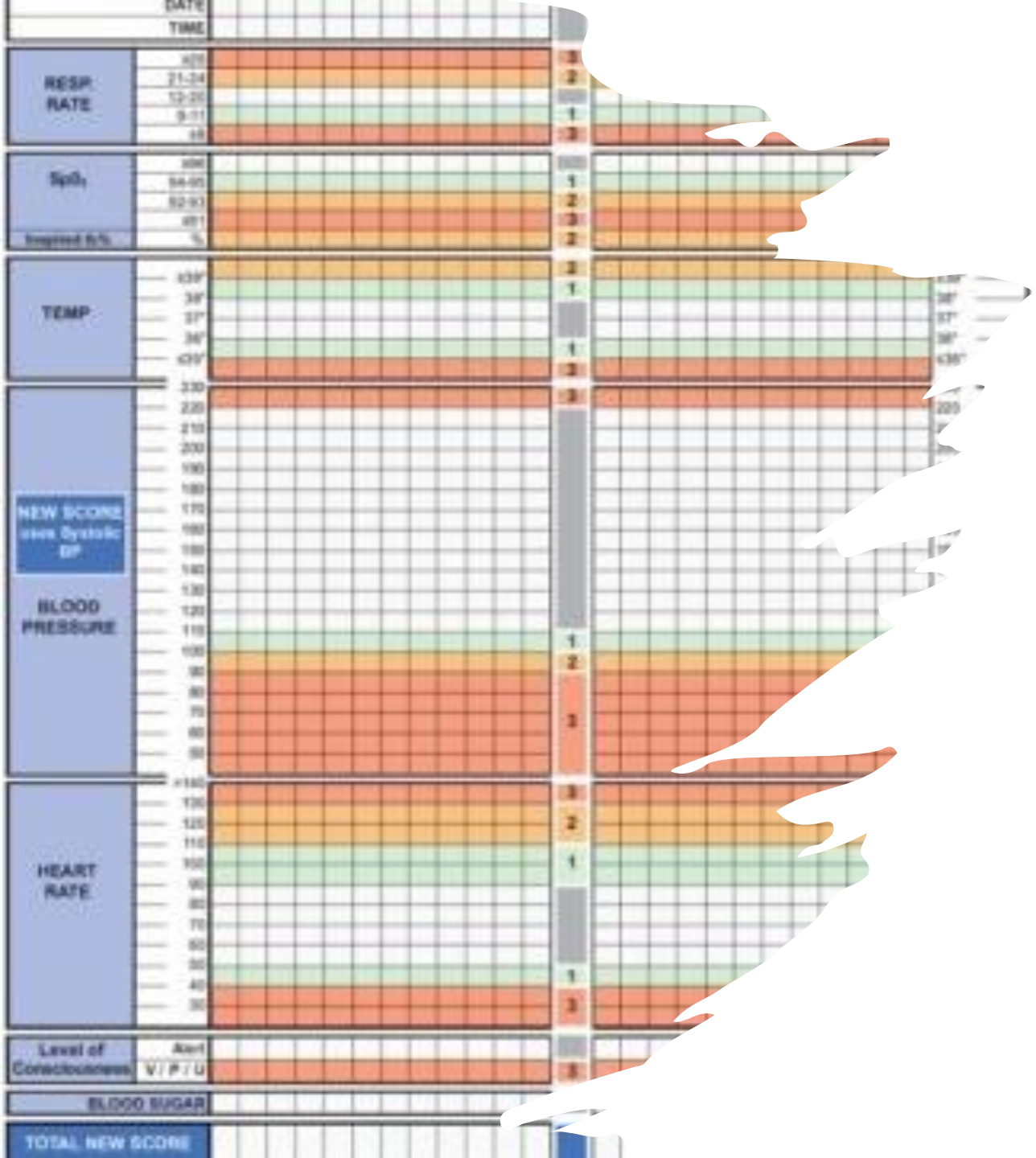
...ealth will allow patients in the province to input their sympto... ...deo visits with accredited physicians.

... 2019 | 01:27 pm

SHARE Share 49



...ing Babylon AI ...ider rollout



Vital Sign Monitoring

- Early Warning Scores – real-time prediction model for patient deterioration
- Traditionally calculated by hand, but increasingly calculated electronically

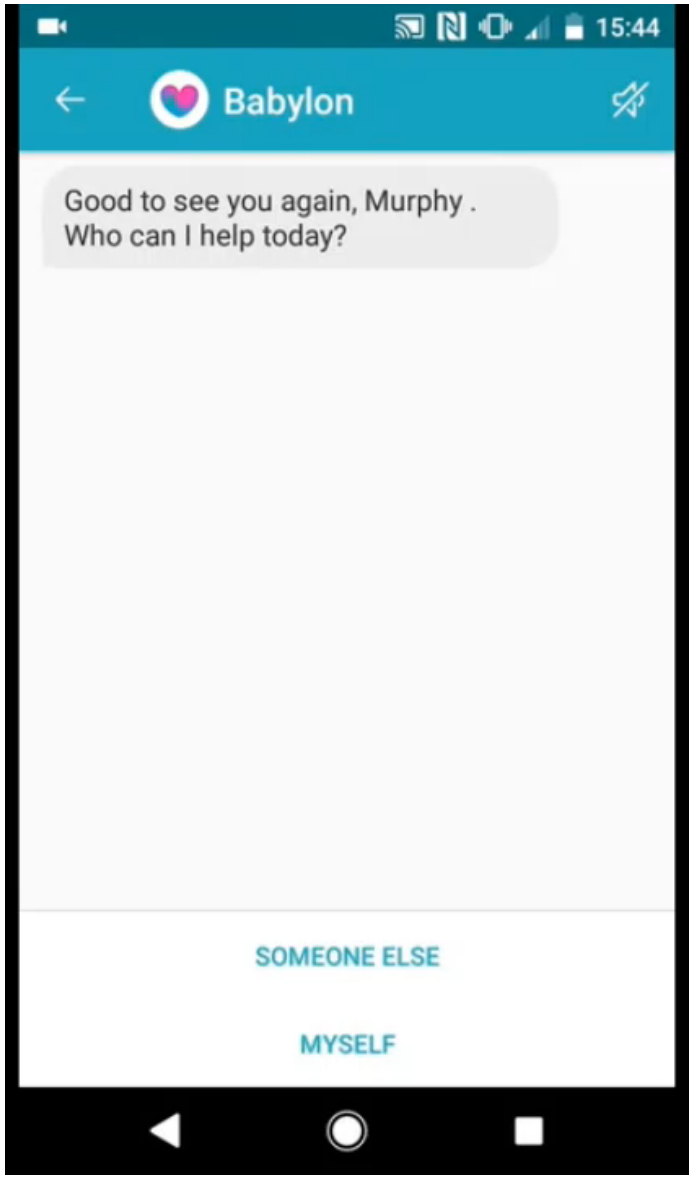
Vital Sign Monitoring

- SEND – electronic documentation of vital signs
- Used (and widely liked) by 5 hospitals in Oxfordshire and Warwickshire

Why focus on healthtech companies?

- Regulated
- Healthtech lets us see what **IS** going wrong or right in the real world
- Allow us to reflect on how responsible our own practices are.

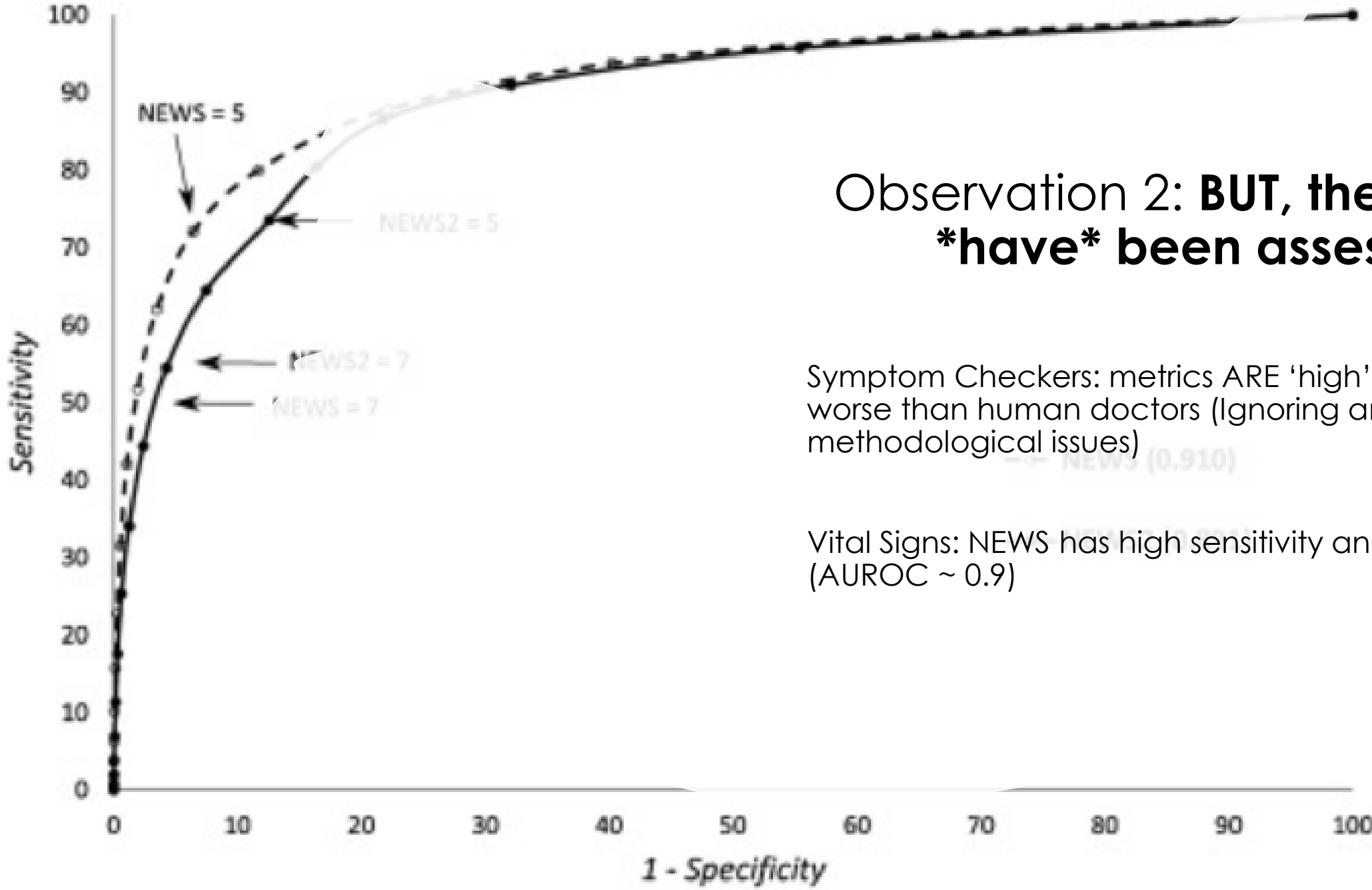
Observations



Observation 1: **The underlying Machine Learning models are not very good, by any clinically intuitive measure**

Mid-2018: symptom checker suggests ‘panic attack’ for classic heart attack symptoms

Early Warning Scores: senior nurses often think of it as a tool for junior staff



Observation 2: **BUT, the models *have* been assessed**

Symptom Checkers: metrics ARE ‘high’ and not much worse than human doctors (Ignoring any potential methodological issues)

Vital Signs: NEWS has high sensitivity and specificity (AUROC ~ 0.9)

Observation 3: **both systems have been successful**

Conjecture - people don't care if a Machine Learning model is 'good'

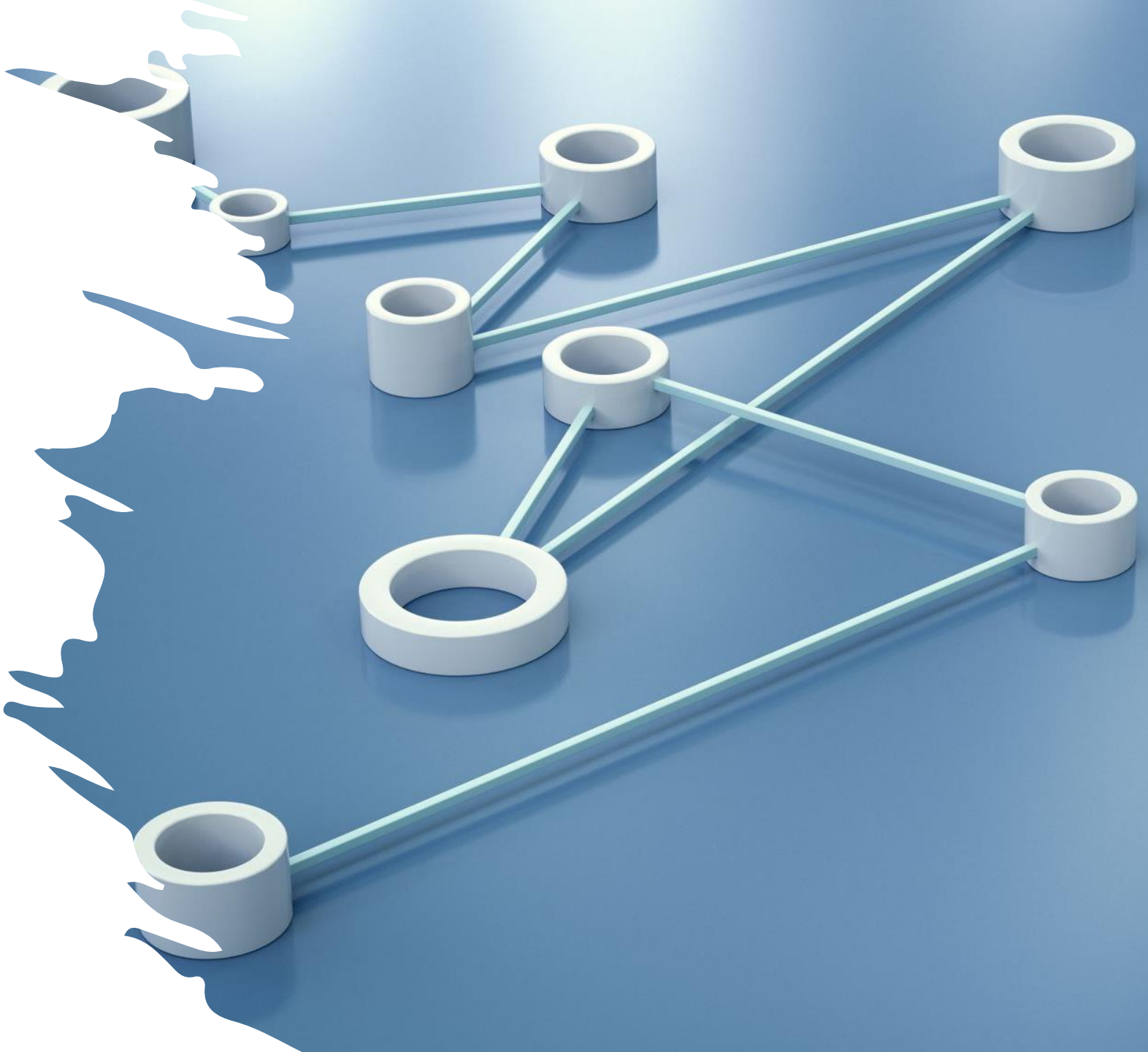
(but they do care if it 'works')

A Problem:

Machine learning models are often just one component of clinical software. That software is, in itself, part of a larger clinical process

A solution?

- Evaluation of systems, not just models
- Model development that takes into account the existing processes
  - Economic
  - Clinical constraints
  - Existing prediction models



1. System Evaluation

- DECIDE-AI: reporting guidelines for early stage clinical evaluation of decision support systems driven by AI – Vasey et al.

- ‘addressing the complexity of this collaboration between two independent forms of intelligence [human/AI], beyond measures of effectiveness alone.

- Covers: proof of clinical utility at small scale, safety, human factors evaluation, and preparation for larger scale summative trials.

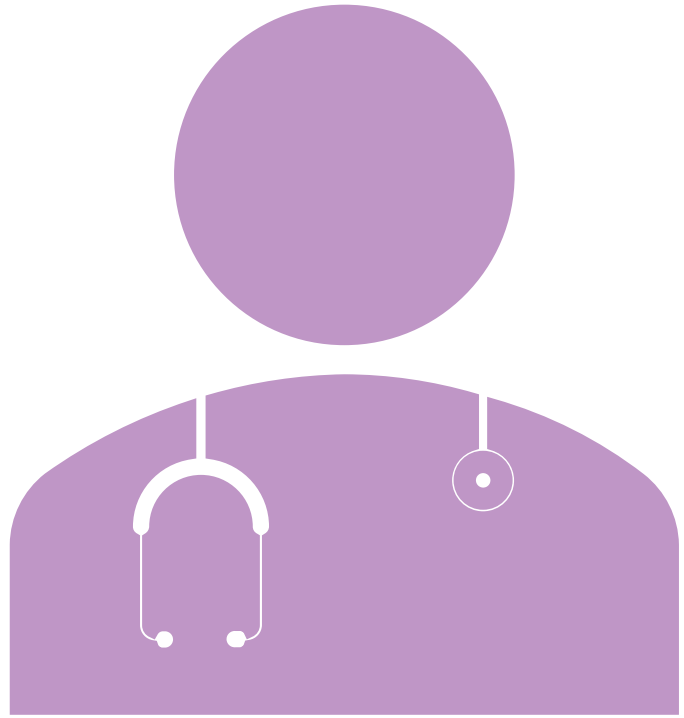
System Evaluation

Implementation	a) Describe the settings in which the AI system was evaluated
	b) Describe the clinical workflow/care pathway in which the AI system was evaluated, the timing of its use, and how the final supported decision was reached and by whom

Implementation	a) Report on the user exposure to the AI system, on the number of instances the AI system was used, and on the users' adherence to the intended implementation
	b) Report any significant changes to the clinical workflow or care pathway caused by the AI system

2. ML for systems

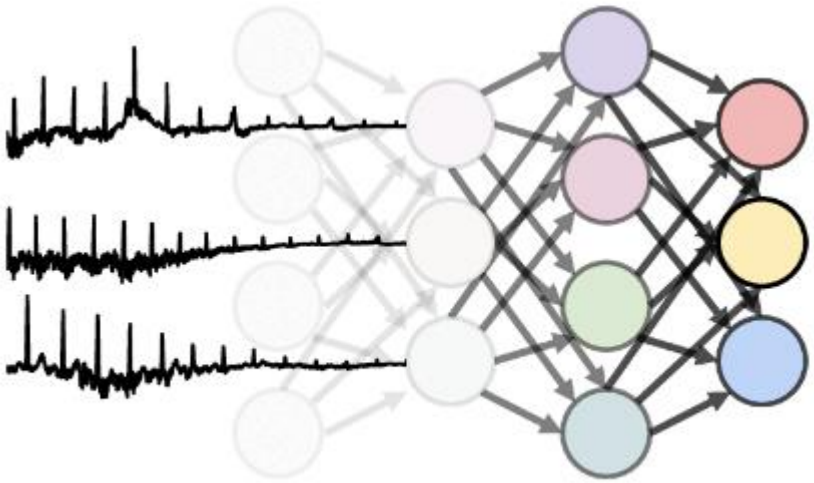
The disconnect between the metrics for algorithm performance and the realities of a clinician's workflow and decision making process is a fundamental but often overlooked issues. The inclusion of clinical context in the AI metrics for optimising and evaluating clinical algorithms could make AI tools more clinically relevant.



Reyna et al. Rethinking Algorithm Performance Metrics for Artificial Intelligence in Diagnostic Medicine, JAMA, 2022

2. ML for systems

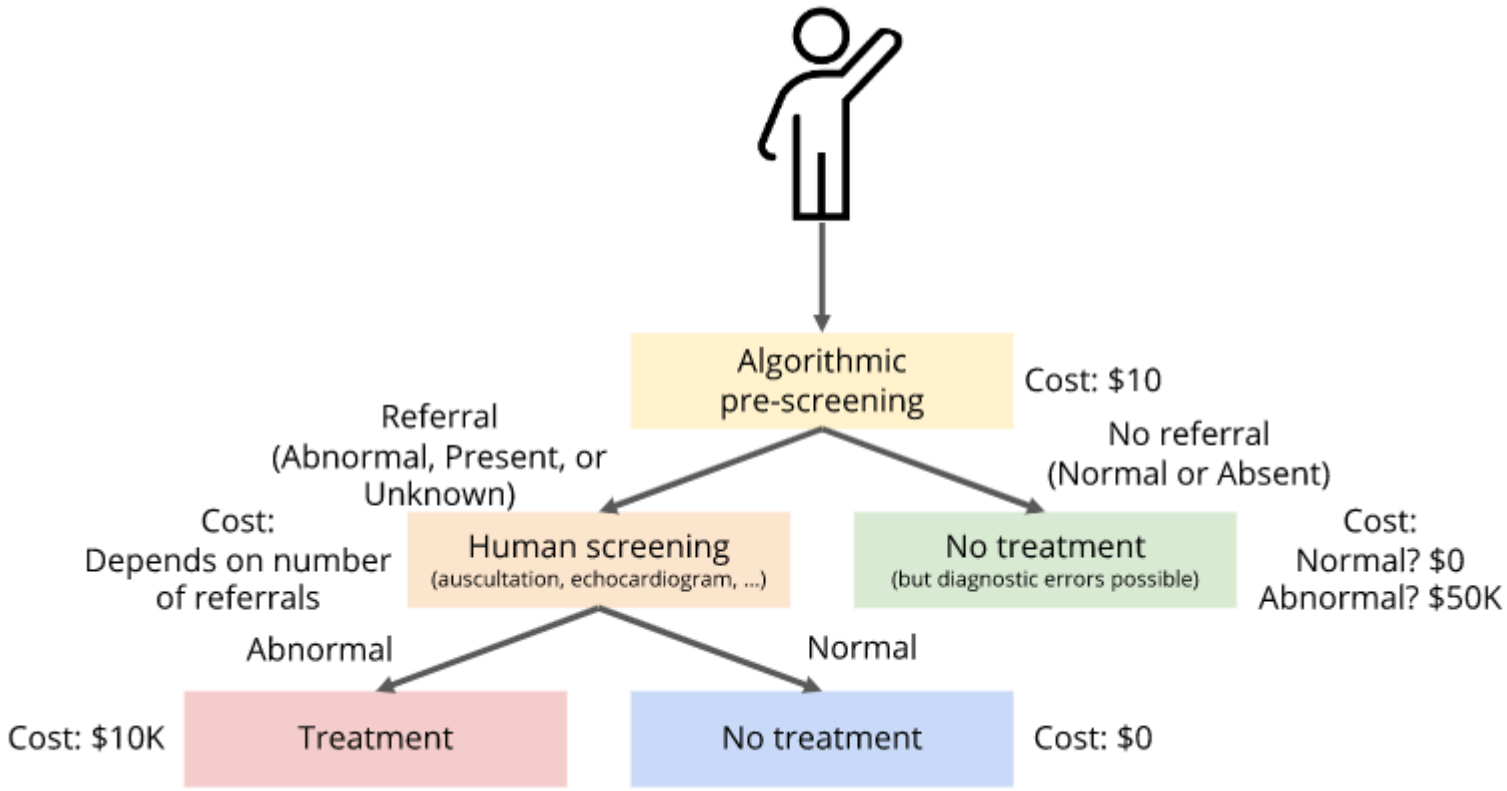
George B. Moody PhysioNet Challenge



- Aim: predict whether a patient is ‘normal’ or ‘abnormal’ based on audio recordings of heart sounds.
- Classical metrics: Sensitivity, Specificity, Accuracy

George B Moody Physionet Challenge 2022: https://moody-challenge.physionet.org/2022/

2. ML for systems



Metric: considers system cost in terms of money, clinical time, patient outcome

Idea: per-patient cost of human screening is dependent on number sent to for screening.

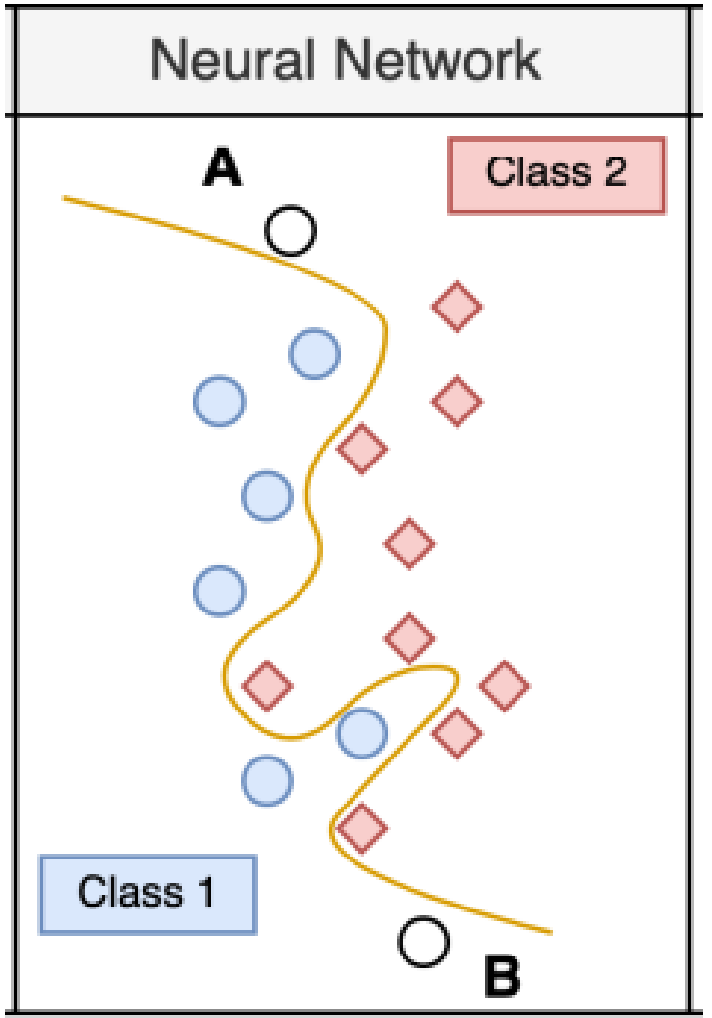
3. Previous Models

- Take into account existing workflow
- Default to existing models when unsure (graceful failover)
- Neural Network regularised against baseline model

A simple out-of-distribution regularisation approach for neural network classification using benchmark labels – Zhou, Yau, Wong. ML4H 2022

3. Previous Models

- Take into account existing workflow
- Default to existing models when unsure (graceful failover)
- Neural Network regularised against baseline model



A simple out-of-distribution regularisation approach for neural network classification using benchmark labels – Zhou, Yau, Wong. ML4H 2022

3. Previous Models

- Take into account existing workflow
- Default to existing models when unsure (graceful failover)
- Neural Network regularised against baseline model

A simple out-of-distribution regularisation approach for neural network classification using benchmark labels – Zhou, Yau, Wong. ML4H 2022

MANCHESTER 1824

The University of Manchester

(What) can healthtech teach us about responsible AI?

DAVE WONG (@DRDAVECWONG)

DAVID.WONG@MANCHESTER.AC.UK