

5.5a Trustworthy AI

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Trustworthy AI

Methodological concerns

Critical appraisal

Trustworthy AI

Education and training

Toward the future

Data quality and bias impacting ML

Dataset shifts (Finlayson, 2021; Knaus, 2022; Nelson, 2025) may lead to performance decline of algorithms over time (Ross, 2022; Vaid, 2023; Palmer, 2023)

Limitations for temporal EHR data in machine learning (Xie, 2021)

Clinician concerns about voice data collection include workflow disruptions, self-censorship, and errors that could impact patient eligibility for services (Wilcox, 2023)

Data leakage

Contamination of model input features with outcome information

Demonstrated to effect at least 294 papers across 17 disciplines for 8 reasons (Kapoor, 2023)

- Lack of clean separation of training and test dataset
- No test set
- Pre-processing on training and test set
- Feature selection on training and test set
- Duplicates in datasets
- Model uses features that are not legitimate
- Test set is not drawn from the distribution of scientific interest
- Temporal leakage
- Nonindependence between training and test samples
- Sampling bias in test distribution

Framework for assessing impact of “label leakage” into models (Davis, 2023)

Bias and questionable practices

Bias in medical AI (Cross, 2024)

- Training data
 - Imbalanced sample sizes
 - Nonrandomly missing patient data
 - Data not usually or easily captured
 - Biases in data labels and misclassification
- Model development and evaluation
 - Overreliance on whole-cohort performance metrics
- Publication
 - Where medical AI models are developed and published, and by whom
 - AI research biased towards positive results and certain medical domains
- Model implementation
 - Sample selection bias: real-world patients differ from training cohort
 - End-user biases

Questionable practices in ML (Leech, 2024)

- Contamination
 - Data, prompts, RAG, etc.
- Cherrypicking
 - Data, results, etc.
- Misreporting
 - From biasing results to outright fraud
- Amplifiers
 - “Handcrafting bias”

Explainability

Need to explain outputs, especially from neural networks (Hsieh, 2024)

- How to explain to patient when algorithm predicts what clinician cannot explain (Burt, 2018; Schiff, 2019)
- Methods evolving (Graziani, 2020; Hoyt, 2022)

Must maintain clinical perspective (Verghese, 2018)

Should healthcare accept “black box” medicine (Wang, 2020)?

- Lack of interpretability makes local testing imperative
- Clinicians and trainees must be educated on benefits, risks, and limitations

Explainability a false hope?

- Aim instead for trustworthiness through transparency (Ghassemi, 2021) or interpretability (Ning, 2025)?
- Algorithm could learn to say, “I don’t know” (Shashikumar, 2022)

Reproducibility

Predates ML: CDS studies rarely (0.3%) replicated (Coiera, 2021)

How can we reproduce results? (Van Calster, 2019)

- Algorithm must be publicly available to allow external validation
- For black-box algorithms, software should be available
- Need open data sets, code, and publications (Beam, 2020)

Clinical prediction models must be validated with many different data sets beyond those used in original papers (Adibi, 2020)

- “Living” models in cloud?
- Excess models may drown out most useful ones (Damen, 2016)

Poor methods leading to “reproducibility crisis” (Ball, 2023)

Algorithm bias

Algorithms may discriminate (“weapons of math destruction,” O’Neil, 2016), undermine civil rights (Barocas, 2017), or otherwise be mis-used (Angwin, 2016; Rainie, 2017; Brundage, 2018)

- Must ensure equity for representative data and ethical application (Rajkomar, 2018)

Failure to include race in models may propagate systemic inequities and discrimination but use may reward historical inequities (Basu, 2023)

Predicting race variables from medical images could be helpful for monitoring disparities and facilitating equity (Zou, 2023)

Deep learning helped performance in skin disease diagnosis, but improved more for light than dark skin (Groh, 2024)

Evidence mixed that algorithms can mitigate, perpetuate, or exacerbate racial and ethnic disparities (Siddique, 2024)

Examples of algorithm bias

Allocation of healthcare resources (Obermeyer, 2019)

- Risk score for allocation of resources
- At a given risk score, black patients considerably sicker than white patients, but historically received less care
- Bias arose because algorithm predicted healthcare costs rather than illness, but unequal access to care meant less money had been spent caring for black patients than for white patients
- Remedying disparity would increase the percentage of black patients receiving additional help from 17.7% to 46.5%

In behavioral healthcare (Walsh, 2020)

- Self and social stigma contribute to under-reported symptoms
 - Under-coding worsens ascertainment
- Health disparities contribute to algorithmic bias
- Lack of reliable biological and clinical markers hinders model development
- Model explainability challenges impede trust among users

Environmental concerns

Climate – computationally intensive ML/AI algorithms can generate notable carbon emissions (Ceurstemont, 2025)

- In US, data centers consumed 4.4% of all electricity in 2023, estimated to grow to 6.7-12% by 2028 (Shehabi, 2024)
- Electricity consumption of AI request 10-fold more than Google search (de Vries, 2023)
- Energy requirements increase from
 - Text to image to video generation (Delavande, 2025)
 - Use of larger and more accurate LLMs (Dauner, 2025)

Individual uses of generative AI small, but add up over worldwide use; details not all known due to lack of transparency (Kirkpatrick, 2023; O'Donnell, 2025)

Water (Li, 2025) – data centers consume water in three main areas

- Data center on-site for cooling
- Electricity generation
- Server manufacturing

Strategies for mitigating to sustainably advance health AI (Ramachandran, 2025)

Critical appraisal of AI research

As with all evidence-based medicine (EBM), major questions that guide patient care to ask about ML/AI include

- Diagnosis – “test” for a “disease”
 - Ideally from comparison with “gold standard”
 - Predictive models are type of diagnostic test
- Treatment – therapy or intervention to prevent or treat disease; ideally from randomized controlled trial (RCT)

Studies of both questions can be aggregated into systematic reviews that may (if data allow) be combined via meta-analysis

Many tools for reporting AI studies, critical appraisal, study quality, and risk of bias (Cabello, 2025; Shiferaw, 2025)

Instruments for assessing reporting of studies adapted for AI (Ibrahim, 2021)

Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD), adapted for AI (Collins, 2024) and LLMs (Galifant, 2025)

Standards for Reporting of Diagnostic Accuracy Study (STARD) (adapted for AI by Sounderajah, 2021)

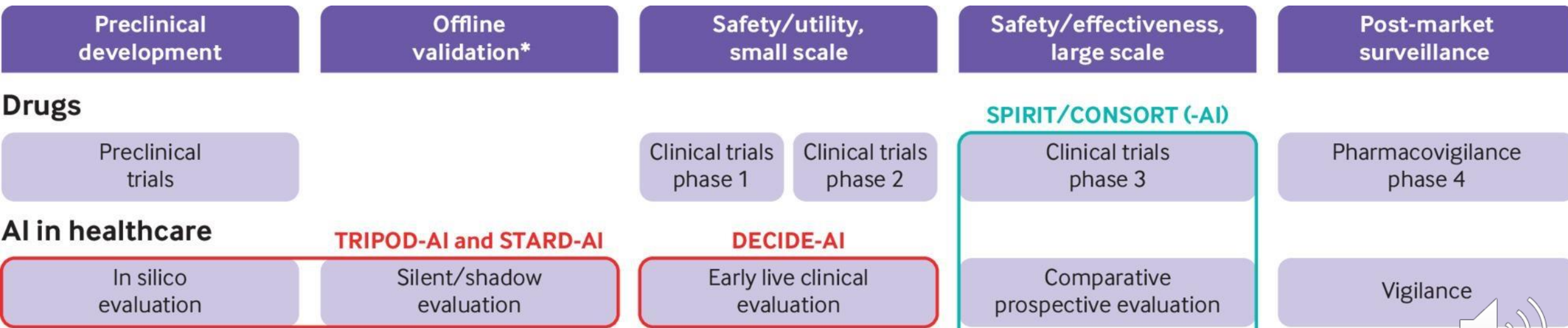
Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI) (Vasey, 2022)

Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) (adapted for AI by Rivera, 2020)

Consolidated Standards of Reporting Trials (CONSORT) (adapted for AI by Liu, 2020)

Prediction model Risk Of Bias ASsessment Tool (PROBAST) (Wolff, 2019)

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (adapted for AI by Cacciamani, 2023)



Key considerations for predictive model evaluation (Van Calster, 2025; Mercaldo, 2025)

To avoid overfitting of model, data sets should include

- Training set to build predictive model
- (Optional) validation set to assess performance and fine-tune model's parameters to selecting best-performing model
- Test set to assess likely future performance of model on unseen data

When only limited amount of data is available, use k-fold cross-validation

- In the k-fold cross-validation, divide the data into k subsets of equal size and each time leave out one of subsets from training to use in test set
- If k equals sample size, this is "leave-one-out" method

		Reference standard		
		Present	Absent	
Model classification	Positive	True Positive (TP)	False Positive (FP)	Positive Predictive Value $\frac{TP}{TP + FP}$
	Negative	False Negative (FN)	True Negative (TN)	Negative Predictive Value $\frac{TN}{TN + FN}$
		Sensitivity $\frac{TP}{TP + FN}$	Specificity $\frac{TN}{TN + FP}$	Accuracy $\frac{TP + TN}{TP + FN + FP + TN}$

Critical appraisal of ML/AI models

Many have been developed (Liu, 2019; Vollmer, 2020; Hernandez-Boussard, 2020; Vinny, 2021; van Smeden, 2022)

APPRAISE-AI tool provides standardized, quantitative tool for evaluating the methodological rigor and clinical utility of AI models (Kwong, 2023)

Recent extension of TRIPOD for AI (Collins, 2024) and for LLMs (Galifant, 2025)

Reporting guidelines for chatbot health advice studies: Chatbot Assessment Reporting Tool (CHART Collaborative, 2025)

Chatbot Assessment Reporting Tool (CHART Collaborative, 2025)

Title & Abstract

- Title
- Abstract/Summary

Introduction

- Background

Methods

- Model Identifiers
- Model Details
- Prompt Engineering
- Query Strategy
- Performance Evaluation
- Sample Size
- Data Analysis

Results

- Report the alignment between generative AI-driven chatbot output and ground truth or reference standard using quantitative or mixed methods approaches as applicable
- For responses deviating from the ground truth or reference standard, state the nature of the difference(s)
- Report the assessment for potentially harmful, biased, or misleading responses

Discussion

Open Science

- Disclosures
- Funding
- Ethics
- Protocol
- Data availability

Critical appraisal of ML/AI (and other) RCTs

Validity of any type of RCT (Straus, 2018)

- Did experimental and control groups begin study with similar prognosis?
 - Were patients randomized?
 - Was randomization concealed (blinded or masked)?
 - Were patients analyzed in groups to which they were randomized?
 - Were patients in treatment and control groups similar with respect to known prognosis?
- Did experimental and control groups retain a similar prognosis after the study started?
 - Were patients aware of group allocation?
 - Were clinicians aware of group allocation?
 - Were assessors aware of group allocation?
 - Was follow-up complete?

Risk of bias tool from Cochrane Collaboration indicates risk that RCT methods might lead to biased results (Sterne, 2019)

Additional questions to ask about RCTs of ML/AI interventions (Liu, 2020)

Title and abstract	• Indicate that the intervention involves AI/ML in the title and/or abstract and specify the type of model • State the intended use of the AI intervention within the trial in the title and/or abstract
Background and objectives	• Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g., healthcare professionals, patients, public)
Eligibility criteria for participants	• State the inclusion and exclusion criteria at the level of participants • State the inclusion and exclusion criteria at the level of the input data
Settings and locations	• Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements
Interventions for each group with sufficient details to allow replication, including how and when administered	• State which version of the AI algorithm was used • Describe how the input data were acquired and selected for the AI intervention • Describe how poor quality or unavailable input data were assessed and handled • Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users • Specify the output of the AI intervention • Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice
Harms	• Describe results of any analysis of performance errors and how errors were identified, where applicable – if no such analysis was planned or done, justify why not
Funding	• State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use

