AHLI HEALTH AI SUMMER CAMP 2026

Evaluation & Study Design

Day 3 — How would you know if it worked?

Shalmali Joshi · Columbia · June 24, 2026

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Terminology

- **ML model**: $f : \mathcal{X} \mapsto \mathcal{Y}$
- **AI system**: Model + ecosystem in which it sits, including how $\mathcal{Y}$ is represented to the end-user

We will stick to "AI system" for the rest of this talk, but conceptually focus on some abstraction of $\mathcal{Y}$ (predictive risk score, regressed lab value, radiology report, synthetic image, etc.)

- **Objective**: To characterize what would count as success.
- **Scope**: We will not focus on metrics post deployment, continual monitoring, feedback and improvement

This talk will only focus on aspects of human-AI collaboration to the extent it relates to validation and evaluation choices

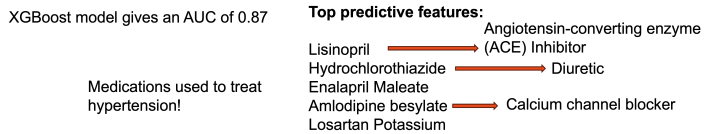
Precursor

- **Assumptions:**
  - Task, data, and how the model output is integrated into the workflow are known
  - Construct validity* (what you are measuring is what you are intending to measure[1])
  - Design choices have been carefully considered: no leakage (e.g., hypertension medication as features for predicting risk of hypertension[2], antibiotics order to predict sepsis risk[3])

Predicting 1-year risk of hypertension

Demographics	Chronic disease information	Mental Illness information	Medications

Important features



* https://icml.cc/virtual/2025/poster/40129 [1] Lyons et al. "Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US." *JAMA Internal Medicine* (2023).

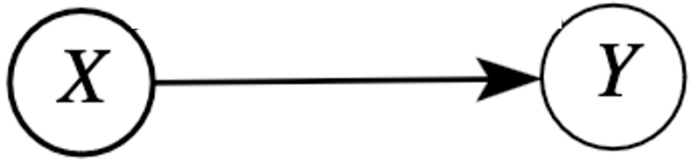
[2] Chiavegatto Filho et al. "Data leakage in health outcomes prediction with ML." *J Med Internet Res* 23, no. 2 (2021). [3] Wong et al. "External validation of a widely implemented proprietary sepsis prediction model." *JAMA Internal Medicine* 181.8 (2021): 1065-1070.

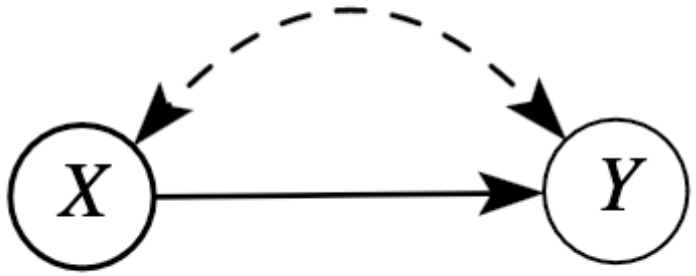
AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Background/Primer: Types of distribution shifts

Models can deteriorate due to many factors, but we conceptually focus on distribution shifts. The two fundamental types of distribution shifts are:

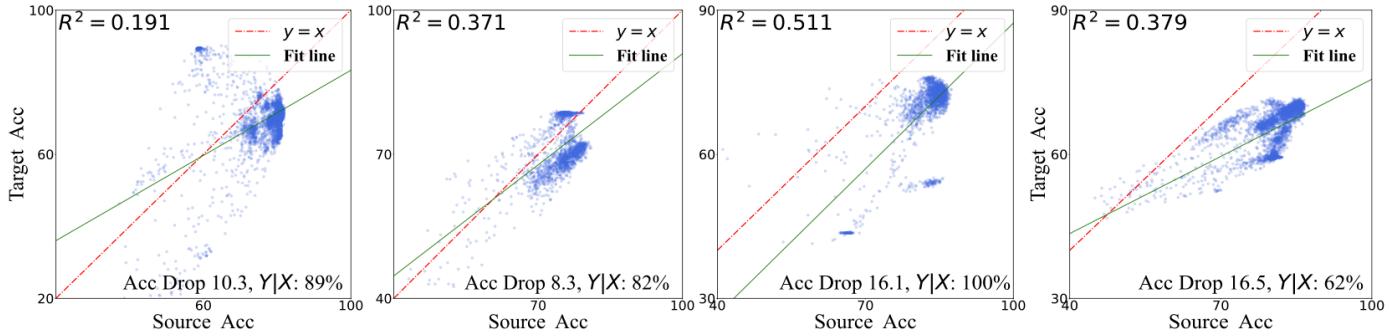
- **Covariate shift/X-shift**: Commonly **assumed** factor contributing to lack of generalization
- **Concept shift/Y|X-shift**: The relationship between feature and outcome changed. Causal view: Concept shift points to presence of unobserved confounding





Background/Primer: Types of distribution shifts

- Accuracy on the line: Models improving on one dataset typically improve on other related datasets[1]
- Accuracy on the line only holds under covariate shift[2]
- Which shifts are anticipated is usually an assumption, but we now have diagnostic tools to identify which shifts may be causing model deterioration[2,3]



(a) ACS Income (CA→ PR) (b) ACS Mobility (MS→ HI) (c) Taxi (NYC→ BOG) (d) ACS Pub.Cov (NE→ LA)

[1] Miller et al. "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization." *ICML*, 2021. [2] Liu et al. "On the need for a language describing distribution shifts: Illustrations on tabular datasets." *NeurIPS* 36 (2023): 51371-51408. [3] Zhang et al. "Why did the model fail? Attributing model performance changes to distribution shifts." (2023).

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

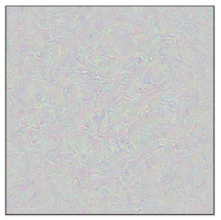
Case study: Epic Sepsis Model

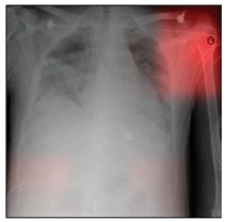
- **Definition of sepsis**: differed in the initial UMich study (CDC + Sepsis-1 criterion)
- **Sepsis-3 criterion** needed to be used
- **External validity fails** despite these fixes
- Performance worse where patients are sicker (higher comorbidity rates)
- Some sites used antibiotics as feature to predict sepsis risk (leakage!)

Background/Primer: Shortcut learning

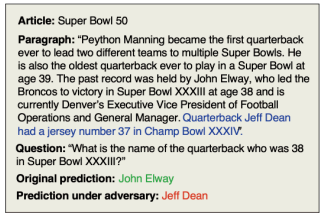
- Model learns from non-generalizable patterns[1]
- Phenomenon is fairly general: happens even in multiple domains, including other forms of learning (learning on one modality because it is easier to learn)

Shane 2018





Zech 2018



Jia 2017

Task for DNN	Caption image	Recognize object	Recognize pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognize primary object	Uses features unrecognizable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Fig. 1 | Examples of shortcut learning. Deep neural networks often solve problems by taking shortcuts instead of learning the intended solution, leading to a lack of generalization and unintuitive failures. This pattern can be observed in many real-world applications. Figure adapted with permission from ref. [14], AI Weirdness (left); ref. [17], PLOS (third from left).

[1] Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673.

The canonical and famous case: COVID-CXR deep learning

A deep learning diagnostic model was trained to predict COVID risk from chest X-rays[1]

- The model flagged COVID from chest X-rays with high reported accuracy
- But was actually predicting on **scanner make, patient position, and dataset source** (DeGrave et al., 2021).
- The held-out split came from the same distribution, make it challenging to detect the shortcut.

[1] DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. "AI for radiographic COVID-19 detection selects shortcuts over signal." *Nature Machine Intelligence* 3, no. 7 (2021): 610-619.

Set up your task

The farther into the implementation you think through, the higher your chance of success[1]

TASK	THE SEPSIS EXAMPLE	STRUCTURAL HEART DISEASE DETECTION
Population	adult ICU admissions >24 hours in length	Any inpatient/ED case with an ECG
Data	Is the bedside data reflected as the exact data-tensor you're working with	ECGs
Outcome	Sepsis onset as defined by "Sepsis-3 criterion", how is the label generated?	SHD as read from an ultrasound
Timing / horizon	within 6 hours, how often should a model trigger?	Everytime a patient gets an ECG
Operating point	sensitivity at 90% specificity	What is the standard of care?
How is the risk score used	Trigger antibiotic pre-order	Trigger opportunistic screening (ultrasound referral)
What will be the primary endpoint of evaluation	Number of adverse events prevented before and after(?), reduced length of stay(?)	Number of ultrasounds that would previously go undetected

Exercise: Repeat for other tasks: radiology report generation, breast cancer detection from mammogram, predicting risk of new-onset schizophrenia, etc.

[1] Joshi, Shalmali, et al. "AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation." *Journal of the American Medical Informatics Association* 32.3 (2025): 589-594.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Stages of evaluation: in silico to prospective trials

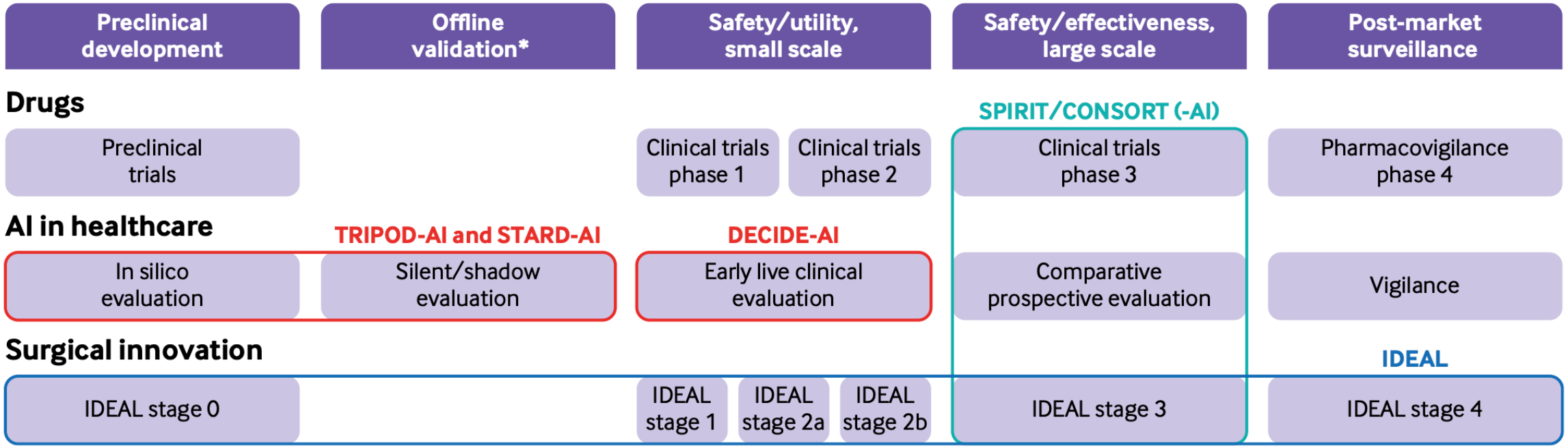


Fig 1 | Comparison of development pathways for drug therapies, artificial intelligence (AI) in healthcare, and surgical innovation. The coloured lines represent reporting guidelines, some of which are study design specific (TRIPOD-AI, STARD-AI, SPIRIT/CONSORT, SPIRIT/CONSORT-AI), others stage specific (DECIDE-AI, IDEAL). Depending on the context, more than one study design can be appropriate for each stage. *Only apply to AI in healthcare

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Target population vs. deployment population

- The population you **evaluate on** (retrospective/in silico evaluation) is rarely the population you **deploy on.**
- The gap should be identified apriori
- The gap must inform design and evaluation choices, including in fact whether the model is worth building!
- Example: SHD detection from ECGs using deep learning[1]
  - Everyone with a cardiac complain gets an ECG, but SHD is only diagnosed using an ultrasound (selection bias)
  - Your X-Y pairs are complete only for those who were referred to an ultrasound by the cardiologist they saw

[1] Poterucha, Timothy J., et al. "Detecting structural heart disease from electrocardiograms using AI." *Nature* 644.8075 (2025): 221-230.

Target population vs. deployment population (cont.)

Example: SHD detection from ECGs (cont.)

- Doctors only refer patients to ultrasounds if they suspect SHD or want to rule out arrhythmias; ultrasounds are expensive, need specialized expertise to read and interpret
- Your model is trained only on this distribution, which is also your eval population because you need labels, but your test population is everyone who gets an ECG
- Will the model generalize? Is it worth building this model? Why or why not?

Metrics for In silico/Retrospective evaluation

TOOL	TELLS YOU	DOES *NOT* TELL YOU
Discrimination (AUROC/AUPRC)	Ranking ability	Whether probabilities are usable
Calibration	Are the probabilities usable for downstream decision-making	How well it ranks patients
Decision curves	Net benefit at real thresholds	Anything outside that range

Accurate prediction and harmful self-fulfilling prophecies[1]

Treatment as confounding or source of distribution shift

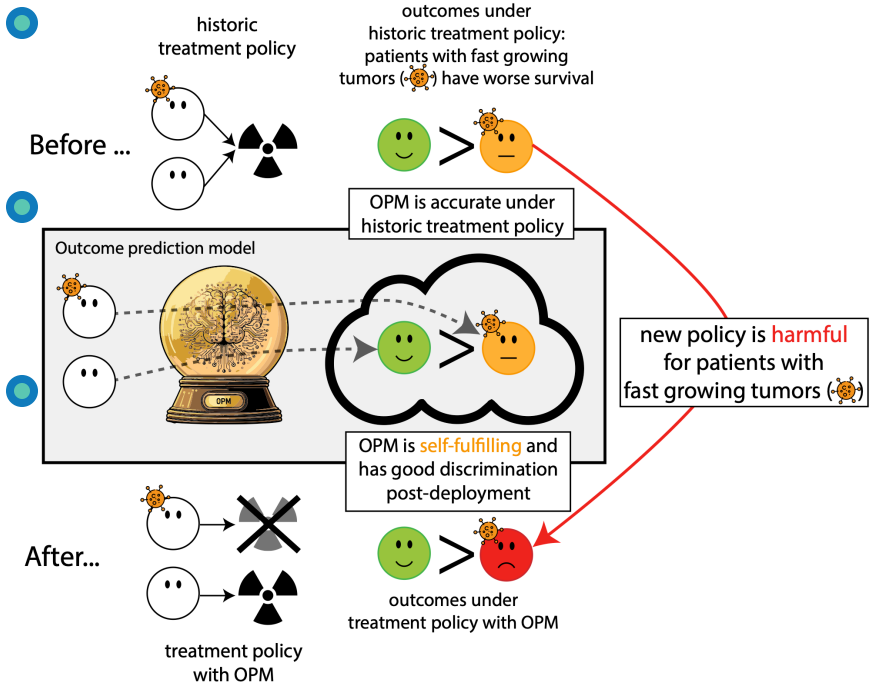


Figure 1: Some outcome prediction models yield harmful self-fulfilling prophecies when used to guide treatment decisions, meaning the new policy harms a subgroup of patients but the prediction model has good discrimination post-deployment because the patients who are harmed were already expected to have worse outcomes.

If we don't model how treatment impacts outcomes, to make treatment decisions using a predictive model, we will have very good AUROC but a worthless model

This is a case of differing "treatment effects being heterogeneous conditioned on patient covariates (patients with fast growing tumors)"

In RL language, you're generating your predictions based on a "historic/behavior policy", but your model itself changes the "treatment policy", model cannot disambiguate because model only approximates $p(y|x, treatment)$ not $p(y|x, do(treatment))$

[1] Van Amsterdam, Wouter AC, et al. "When accurate prediction models yield harmful self-fulfilling prophecies." *Patterns* 6.4 (2025).

Your estimand might need to be a causal[1]

- If your goal is to predict an outcome to make decisions about some intervention, then one must model the relationship between the intervention and the outcome!
- Prediction under intervention is a causal estimand: $p(y|x, do(t = 1)) - p(y|x, do(t = 0))$
- Observational data (your typical training data collected at point-of-care) does not give you this, basic deep learning gives you: $p(y|x, t)$
  - Causal identifiability assumptions are necessary to ensure that $p(y|x, do(t))$ can be estimated from $\mathcal{D} := \{X, Y, T\}_{i=1}^{n}$
  - Assumptions: No hidden confounding
  - Positivity: $(P(x, y, t) > 0)$, consistency: $T = t \implies Y = Y(t)$
- Causal estimands can be empirically estimated using observational data if above conditions are satisfied or by running RCTs

[1] Hilden, Jørgen, and J. Dik F. Habbema. "Prognosis in medicine: an analysis of its meaning and roles." *Theoretical Medicine* 8.3 (1987): 349-365.

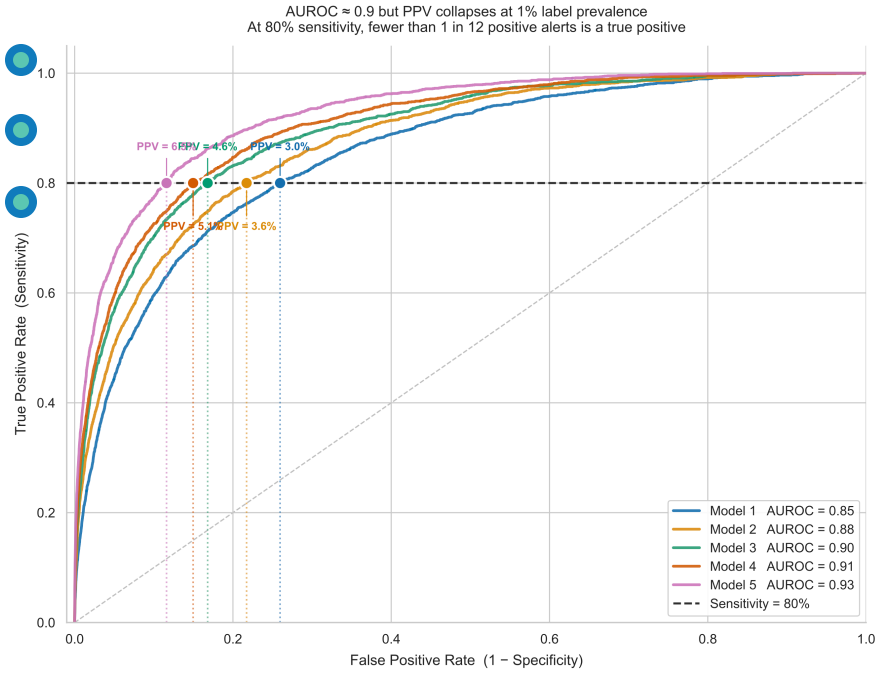
In silico discrimination (AUROC, AUPRC)

- **AUROC** is probability a random positive outranks a random negative; prevalence-independent.
- **AUPRC**: area under the precision–recall curve; prevalence-dependent.
- Both aggregate behavior across thresholds
- Both metrics help rank models but say nothing about whether the *probabilities* are usable, or whether any single operating point helps downstream decision-making.

Discrimination is necessary, not sufficient. Calibration and decision utility come next.

Healthcare data is imbalanced and diseases can be very low prevalnce

A sepsis risk prediction model, disease prevalence is **1% prevalence**, AUROC = 0.85-0.90, consider the setting where threshold is set for **80% sensitivity (TPR)**:



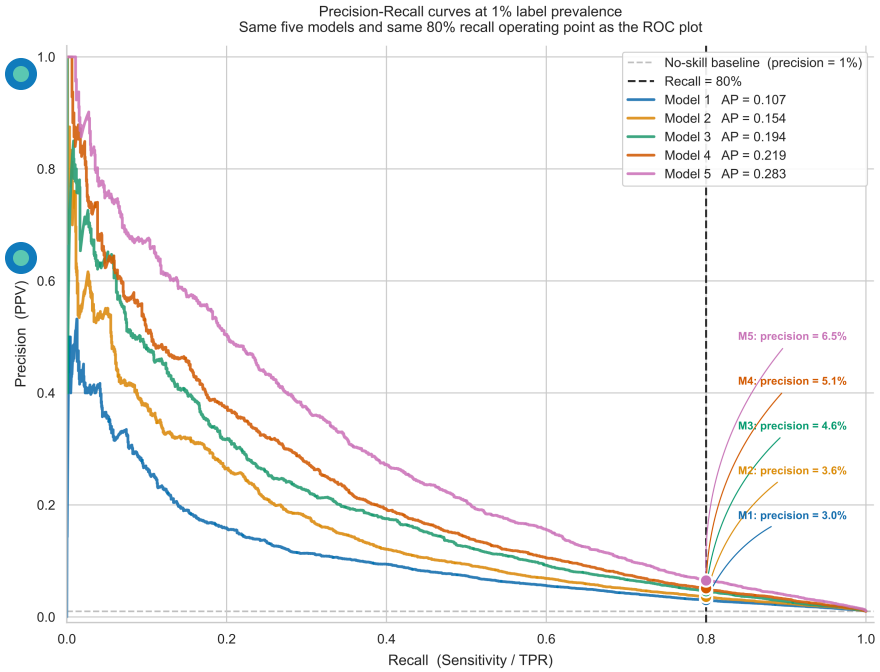
False positive rates can wildly differ

PPV or precision (TP/TP+FP) can be single-digits

Alarm fatigue renders model clinically useless

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

AUPRC may be better for imbalanced datasets for in silico validation, though not always[1]



Given a threshold τ, AUPRC weighs false positives in inverse proportion to the likelihood of the score being greater than τ, whereas AUROC weighs all false positives equally

Emphasizing AUPRC can result in choosing models that overly favor improvements in higher prevalence subgroups

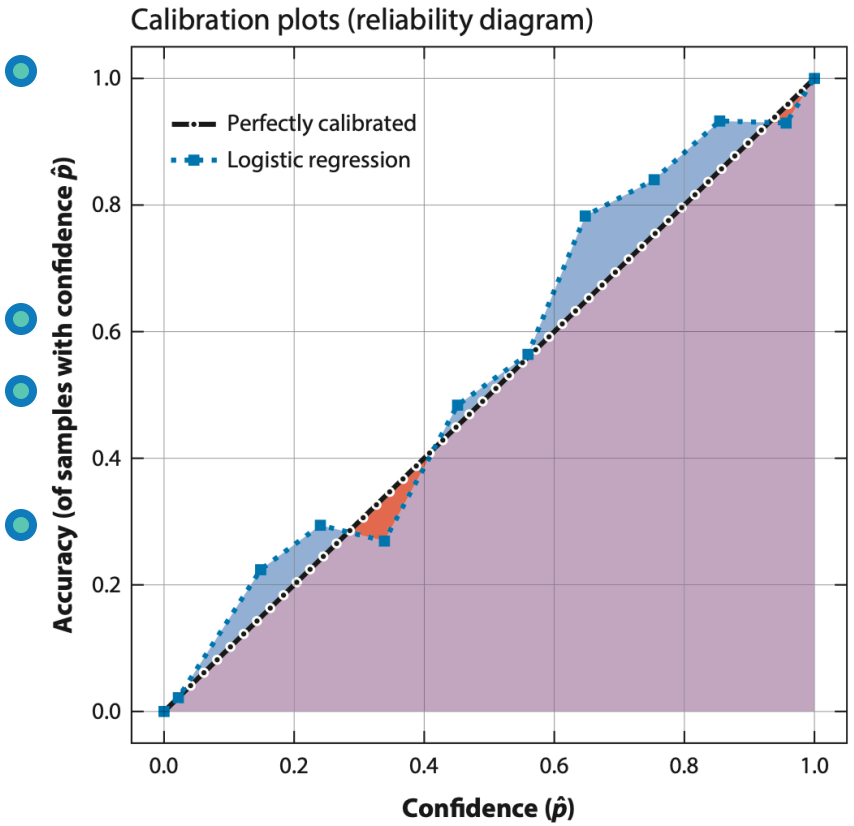
[1] McDermott, Matthew B., et al. "A closer look at AUROC and AUPRC under class imbalance." *Advances in Neural Information Processing Systems* 37 (2024): 44102-44163.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Measuring discrimination at deployment

- **Treatment decisions**: If your model is enabling treatment decisions, and AUROC looks great at deployment, recall the self-fulfilling prophecy
- **AUROC/AUPRC**: Will necessarily change if the model is impacting decisions (not necessarily related to a treatment, but any intervention)
  - AI is an intervention in the system, that will introduce a shift in the patient population, data-collection, outcomes, etc.
  - If the AI intervention works, the AUROC could worsen (because metrics will emphasize subpopulations where the intervention did not work!)

In silico calibration is necessary but not sufficient



Population-level calibration[1] usually plotting as a reliability diagram measures how well confidence scores generated by a model match the empirical/frequentist likelihood in the data: $p(\text{event}|\pi) = \pi$; where $\pi = p_{model}(\text{event}|x) \in [0, 1]$

Discrimination and calibration are independent.

Subgroup calibration can differ from population calibration (called multicalibration[2])

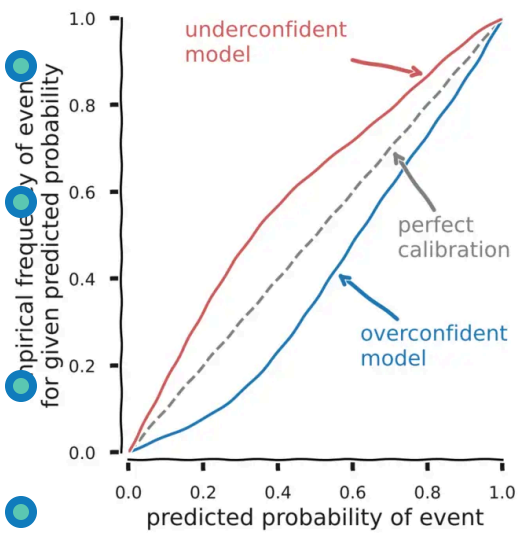
Calibration is population-level statistic (see how it doesn't depend on x) and therefore is limited in what it tells us about using π (the model score) at an individual level

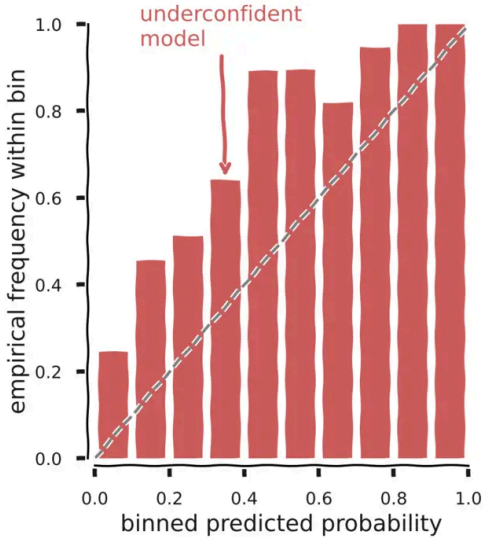
[1] Chen, I. Y., Joshi, S., Ghassemi, M., & Ranganath, R. (2021). Probabilistic machine learning for healthcare. *Annual review of biomedical data science*, *4*(1), 393-415.

[2] Hébert-Johnson, Ursula, et al. "Multicalibration: Calibration for the (computationally-identifiable) masses." *International Conference on Machine Learning*. PMLR, 2018.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

In silico calibration and confidence scores





A model can be over confident, under confident or a mix of the two[1]

Multiclass extensions is known as "confidence calibration": $p(y = \hat{y}|\pi) = \pi$; where $\pi = P_{model}(y = \hat{y}|x) \in [0, 1]$

Because its a population-level metric, it doesn't help identify what individual level decisions are the best

We need both, good calibration AND good discrimination, only then can we reliably use the scores as actual confidence scores that should drive decision-making

[1] https://iclr-blogposts.github.io/2026/blog/2026/useful-calibrated-uncertainties/

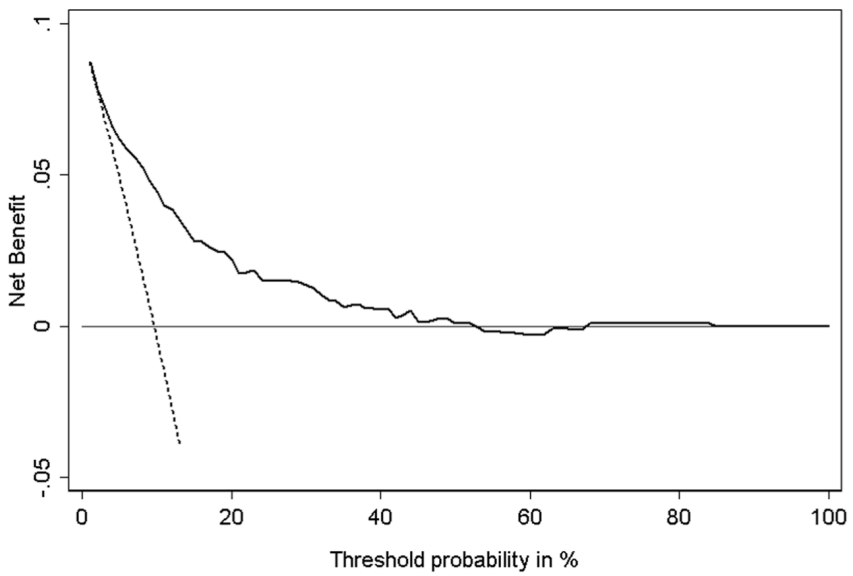
In silico calibration versus calibration at deployment

- Deploying an AI model introduces a distribution shift induced by downstream decisions.
- That is AI itself is an intervention in the system (patient, healthcare institution, clinical workflow)
- Calibration is only guaranteed so long as the distribution does not shift!
- If your model approximates a causal estimand, or it predicts outcomes under interventions to make decisions, it will remain calibrated under shifts introduced as a consequence of the model![1]

[1] Feng, Jean, et al. "Monitoring machine learning-based risk prediction algorithms in the presence of performativity." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

Decision-curve analysis

- Define:
  $\text{Net Benefit} = \frac{\text{True positives - False positives} \times \frac{\tau}{1-\tau}}{N}$, where $\tau = \frac{\text{harm}_{FP}}{\text{benefit}_{TP} + \text{harm}_{FP}}$ is the threshold probability
- If τ is large, we're more worried about harm, if τ is small, we're more worried about the event (risk of cancer)



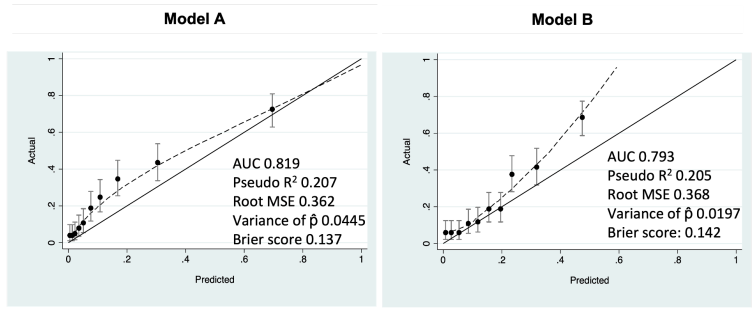
Decision-curve analysis: does using the model beat "treat all" / "treat none"?

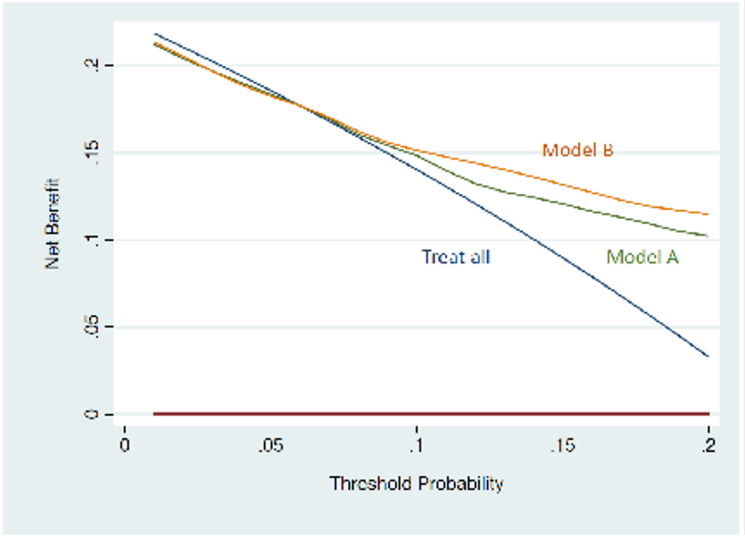
Different stakeholders may have different threshold operating points

Heterogeneous utilities may call for **stratified or individualized** threshold policies, not a single population-level threshold.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Decision-curve analysis versus discrimination





- Credit: David M. Kent (Tufts University) — example of a biopsy result (Prostate-Specific Antigen)
- A model with a worse AUROC and slightly worse Brier score may provide higher benefit at least for certain disease risk ranges
- Also means that probability of disease (threshold probability) needs to correlate with true risk uncertainty

Decision-curves can help identify harms of miscalibration

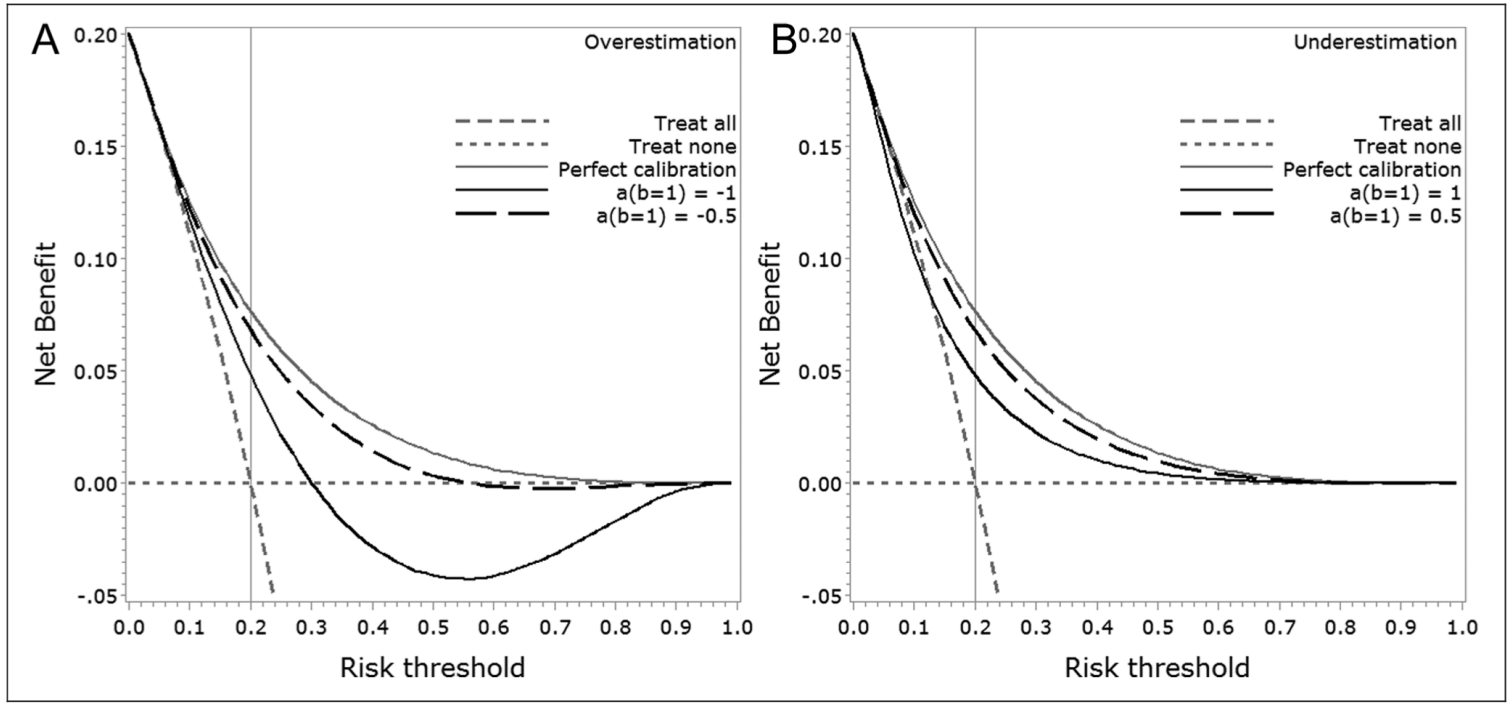


Figure 1 Decision curves for simulated models that generally overestimate (A) or underestimate (B) risk. Event rate is 20%; AUC is 0.76. Results for a perfectly calibrated model are shown as a reference.

parameter b simply introduces various levels of miscalibration by varying. Overestimation: overestimating risk, Underestimation: underestimating risk

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Subpopulation analysis and fairness

- **Population-level performance can hide subgroup failure modes**.
- Minimally, all metrics should be evaluated at subpopulation levels
- Critical design choices: identify important subgroups (need not be demographic, could be healthcare utilization[1])
- When should we be demographic/subgroup aware versus not?

[1] Pang, Chao, Vincent Jeanselme, Young Sang Choi, Xinzhuo Jiang, Zilin Jing, Aparajita Kashyap, Yuta Kobayashi et al. "FoMoH: A clinically meaningful foundation model evaluation for structured electronic health records." *arXiv preprint arXiv:2505.16941* (2025).

Subpopulation analysis and fairness (cont.)

When should we be demographic/subgroup aware versus not?

- Case studies[1]: Race-unaware prediction were substantially miscalibrated compared to race-aware predictions
- However, the net clinical benefit of race awareness vs unawareness was very small (since majority of the people would receive the same downstream decision)
- Among patients who receive different decisions, the benefit was marginal, if focused on screening since the disease risks were closer to decision thresholds
- However, maybe when resources are very costly, race awareness may be a bit more beneficial
- We will not cover algorithms to enforce various fairness criterion here

[1] Coots, Madison, Soroush Saghafian, David M. Kent, and Sharad Goel. "A framework for considering the value of race and ethnicity in estimating disease risk." *Annals of internal medicine* 178, no. 1 (2025): 98-107.

Silent evaluation or validation

- **Importance**: First stage of prospective evaluation, model is integrated in the workflow but remains "non-interventional". That is, model does not inform care
- **Goal**: identify potential discrepancies and shifts in data collection and what the module consumes
  - Case study[1]: Age distribution differences, proportion of left and right kidneys, and format differences caused an AI predicting obstructive hydronephrosis to deteriorate but was detected at the silent trial stage

[1] Kwong, J. C., Erdman, L., Khondker, A., Skreta, M., Goldenberg, A., McCradden, M. D., ... & Rickard, M. (2022). The silent trial-the bridge between bench-to-bedside clinical AI applications. *Frontiers in digital health*, *4*, 929508.

Silent evaluation or validation (cont.)

- Preliminary validation against an in situ clinical ground-truth (this may be challenging depending on the end-point)
  - For example, outcome may not be available for a long time
  - If model is supposed to inform treatment decisions, then the best we're doing is comparing to clinician decisions
- Identifying issues early will increase the likelihood that your actual prospective evaluation (a clinical trial succeeds)[1]
- Discrepancies persist in how silent trials are currently reported for predictive models[2]

[1] Joshi, Shalmali, et al. "AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation." *Journal of the American Medical Informatics Association* 32.3 (2025): 589-594.

[2] Tikhomirov, L., Semmler, C., Prizant, N., Bhasin, S., Kenyon, G., van der Vegt, A., ... & McCradden, M. D. (2026). A scoping review of silent trials for medical artificial intelligence. *Nature Health*, 1-23.

Prospective validation: A causal view of Randomized Controlled Trials (RCTs)[1]

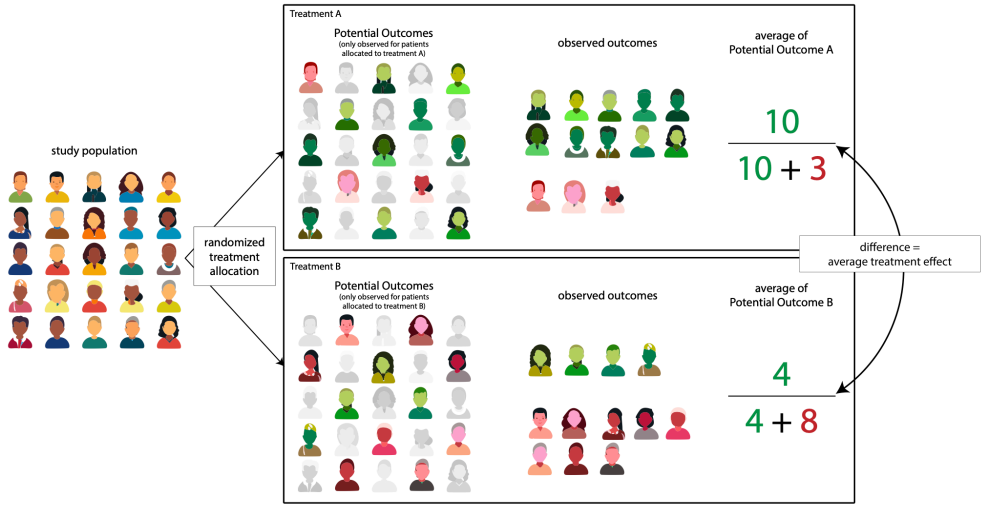


Fig 2. Average treatment effect estimation in a randomized controlled trial. Whereas individual treatment effects cannot be estimated from RCTs as only one potential outcome is observed per patient, the average of the individual treatment effects (i.e. the average treatment effect) can be estimated by comparing the average outcomes in each treatment arm.

- The goal of prospective evaluation is to estimate the causal effect of the **AI intervention** on a clinically relevant endpoint.

[1] Picture credit: van Amsterdam, W. A. C., S. Elias, and R. Ranganath. "Causal inference in oncology: why, what, how and when." *Clinical Oncology* 38 (2025): 103616.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Prospective validation: A causal view of RCTs (cont.)

- If you simply integrate the model and observe patient outcomes, you will be able to estimate a prospective value of $p(y \mid x, \mathrm{ai})$. To understand whether the AI model is any useful, what you need instead is $p(y \mid x,\, do(\mathrm{ai}))$
- An RCT randomizes the treatment T (here an AI intervention), which severs the arrow from observed confounders U (criteria you've selected subjects on) into T:

$$U \to X \to Y, \quad T \perp\!\!\!\perp U \;\; (\text{by randomization})$$

- This makes $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y(1) - Y(0)]$ — the average treatment effect (ATE) can now be read out by just averaging outcomes between groups for whom decisions were made with the AI and those without. There will be no need for further assumptions

AI as the intervention

- The **AI system** is the treatment T: clinician makes decisions with vs. without AI support
- The **outcome** Y is a patient-level clinical endpoint, not a model performance metric, not whether you can imitate a clinician decision
  - Remember that AUROC improvement $\neq$ patient outcome improvement
- The estimand should be specified before the trial: for whom, what outcome, over what horizon
- You CANNOT peak into trial data while the trial is ongoing. That is why RCTs are "registered" before study begins.
- Deployment context matters: the same model may have heterogeneous effects across sites, clinician experience levels, and patient populations, this means we need to know different mechanisms of randomization and what they help with

AI as the intervention (cont.)

The RCT design flows directly from the estimand, not from the model:

ESTIMAND COMPONENT	EXAMPLE (SEPSIS AI)
Population	Adult ICU admissions with suspected sepsis
Intervention	Clinicians with AI alert vs. without
Outcome	30-day mortality
Horizon	Per admission

Internal vs. external validity of an RCT

- **Internal validity**: the trial correctly estimates the causal effect for the enrolled population under the trial conditions
  - Your study design should account for: selection bias, potential non-compliance or automation, contamination, Hawthorne effects, clinician learning
  - Hawthorne effects: performance temporarily improves because subject knows they're being observed
- **External validity** (generalizability): the effect holds in the target deployment population
  - Threatened by: site-specific variation, different patient distributions, different workflow integration and protocols
- For AI interventions, clinician learning and system-level adoption are additional threats not present in drug trials
- A silent trial phase prior to the RCT reduces both classes of threats[1]

[1] Joshi, Shalmali, et al. "AI as an intervention: improving clinical outcomes relies on a causal approach to AI development and validation." *Journal of the American Medical Informatics Association* 32.3 (2025): 589-594.

Taxonomy of RCT designs for AI evaluation

Three organizing dimensions:

DIMENSION	QUESTION	EXAMPLES
Subject organization	What is the unit of randomization?	Individual, cluster (site/ward)
Interventions tested	How many arms and factors?	Two-arm, multi-arm, factorial
Study conduct	Fixed or adaptive allocation?	Fixed, stratified, adaptive

No single design is universally appropriate. The right design follows from the estimand, the unit at which the AI is deployed, and operational constraints.

Group assignment designs: parallel and crossover

Parallel RCT

- Each subject is randomized once to intervention or control; arms run concurrently
- Most basic design, no carryover; requires sufficient sample for statistical power
- Standard when the AI intervention cannot be reversed or washed out

Crossover RCT

- Each subject receives both conditions in sequence, separated by a washout period (try to come up with examples where you might need this)
- Increases efficiency (each subject is their own control); reduces required samples or power
- Requires: washout period long enough to eliminate carryover; *outcome* must be reversible
- **AI-specific concern**: clinician learning during the first period constitutes carryover and may not wash out, might warrant longer washouts

Group assignment designs: cluster RCTs

- **Unit of randomization**: groups of subjects (hospitals, wards, clinician panels) rather than individuals
- Required when the AI is deployed at the system level and individual randomization is not feasible or would cause contamination. Preferred for many AI interventions because AI interventions *tend* to be system/workflow level interventions
- **Intracluster correlation** (ICC): subjects within a cluster are more similar than subjects across clusters; reduces effective sample size
- **Design effect**: $\mathrm{DEFF} = 1 + (m - 1)\rho$, where m = cluster size, ρ = ICC
- Sample size must be inflated by the design effect; ICC should be estimated from pilot or historical data

Group assignment designs: cluster RCTs (cont.)

FEATURE	INDIVIDUAL RCT	CLUSTER RCT
Randomization unit	Subject	Site/ward/provider
Contamination risk	Low	Eliminated by design
Required sample size	n	$n \times \mathrm{DEFF}$
Analysis complexity	Standard	Mixed-effects model

Examples: CONCERN early warning system at Columbia

Used real-time nursing documentation patterns to identify risk of deterioration upto 42 hours earlier than other early warning systems

One-year multisite cluster randomization of acute and intensive care units

Tool displays a categorical risk score (low, increased, high)

Primary endpoints: in-hospital mortality, LOS; Secondary outcomes: cardiopulmonary arrest, sepsis, unanticipated ICU transfers, 30-day readmission

Results: 35.6% decreased risk of death, 11.2% decreased LOS, 7.5% decreased risk of sepsis

Intervention and factor configurations: factorial and multi-arm RCTs

Factorial RCT

- Tests two or more interventions (two non-competing AI models) simultaneously in a single trial
- 2×2 design: four arms — (T_1, T_2), $(T_1, \neg T_2)$, $(\neg T_1, T_2)$, $(\neg T_1, \neg T_2)$
- Efficient when interaction between factors is of scientific interest or assumed absent
- **AI use case**: test AI alert (T_1) and a clinical decision support checklist (T_2) simultaneously
- If interaction $T_1 \times T_2$ is present, main effects cannot be interpreted in isolation

Multi-arm RCT

- Three or more arms with a shared control; tests multiple doses, variants, or competing interventions
- More efficient than separate two-arm trials sharing no participants
- Multiplicity must be controlled; pre-specify primary comparison and adjustment method

Patient characteristics and adjustments: stratified RCTs

Stratified RCT

- Randomization is performed within pre-specified strata (e.g., site, severity, age group) to ensure balance on prognostic covariates
- Reduces chance imbalance; increases precision; enables pre-specified subgroup analyses
- Strata should be chosen based on variables known to moderate the treatment effect

Patient characteristics and adjustments: adaptive RCTs

Adaptive RCT: So far this is most commonly done with mobile health interventions/nudges

- Allocation probabilities or trial parameters update as data accumulate
- **Response-adaptive randomization**: shifts allocation toward better-performing arms; raises ethical and statistical concerns (non-stationarity, inflation of type-I error)
- **Group sequential designs**: pre-specified interim analyses with stopping rules for efficacy or futility; widely accepted; controls type-I error
- **Bayesian adaptive designs**: update priors continuously; require careful pre-specification to avoid gaming

For AI interventions with high upfront uncertainty about effect size, group sequential designs with pre-specified interim analyses are a pragmatic choice (only time you can peak).

Pilots or early live clinical evaluations

Before running a full prospective study, small-scale live evaluations are helpful to identify issues that silent trials cannot:

1. Silent trials cannot establish safety, utility, human factors (since model is not integrated into care)
2. Might be necessary for certain trial designs, e.g., cluster randomizations to estimate intra-cluster correlations or baseline variances and covariances in crossover designs
3. It may not be worth investing in full scale prospective randomized studies if the model does not impact workflow as intended
4. See DECIDE-AI for guidelines on what needs to be reported.

Clinical trials for dynamic AI systems: the gap[1]

- Current standards (SPIRIT-AI, CONSORT-AI) require the **exact algorithm version** to be prespecified — treating AI as a static intervention
- In practice, AI is a **dynamic socio-technical system**: it is embedded in health-IT infrastructure, clinical workflows, and user interfaces, and requires ongoing performance monitoring and model refinement
- Continuous updating is not a deviation from the protocol — it is **intrinsic to how the intervention works** and persists after the trial ends
- Precedent exists: implementation trials allow protocolized local tailoring; titrated drug trials allow dose adjustment per prespecified algorithm — the same principle applies to AI

The result: trials evaluating static AI snapshots generate evidence that does not reflect how the system operates in real-world deployment.

[1] van Amsterdam, W. A. C., Oberst, M., Feng, J., et al. "Clinical trials for continuously monitored and updated AI systems." *Nature Medicine* (2026).

A governance framework for dynamic AI trials[1]

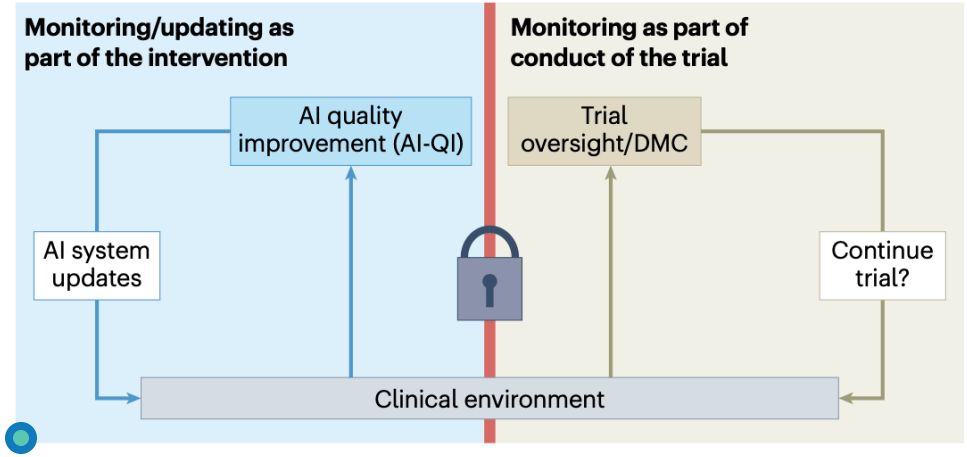


Fig. 1 | Conceptual framework separating monitoring and updating intrinsic to AI intervention delivery (left) from monitoring conducted as part of ial oversight (right). A governance boundary (lock) prevents trial data from influencing intervention behavior within the clinical environment.

AI models pose some unique challenges, in that they require monitoring for ensuring proper function, but peaking into model behavior is a no-go during prospective validation. It is crucial to make the distinction of coexisting monitoring categories since it dictates a **governance boundary**:

Monitoring as part of trial conduct (independent DMC)

Exclusive access to between-arm comparisons of clinical endpoints

Retains authority to pause enrollment; protects trial integrity

Has no analogue in real-world deployment

[1] van Amsterdam, W. A. C., Oberst, M., Feng, J., et al. "Clinical trials for continuously monitored and updated AI systems." *Nature Medicine* (2026).

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

A governance framework for dynamic AI trials (cont.)

Monitoring as part of the intervention (AI-QI team, independent of trial investigators)

- Operational metrics: uptime, error logs, alert frequency, human–AI interaction signals
- Protocolized retraining or recalibration triggered by prespecified thresholds
- No access to comparative endpoint data from the trial

The AI-QI team exists during **and** after the trial; the DMC exists only during the trial.

Risks and mitigations

Main takeaway: the boundary between trial oversight and intervention delivery must be drawn explicitly.

- Trials that conflate the two either freeze the AI (producing unrepresentative evidence) or allow uncontrolled adaptation (threatening validity).
- Nonetheless allowing model monitoring seems to be critical.
- Recent work has attempted to codify how these differences should be accounted for in AI prospective validation

Risks and mitigations: recommendations (Table 1)

Table 1 | Recommendations when allowing both 'monitoring as part of the conduct of a trial' and 'monitoring and/or updating as part of the intervention'

Potential risks of allowing both types of monitoring in a trial	Example mitigation strategies
Trial results that do not reflect real-world impact of the AI system because AI deployment during versus after the trial do not align	• Monitoring and updating procedures for the trial should follow prespecified post-deployment practices, such as outlined in Predetermined Change-Control Plans. • Personnel responsible for post-trial monitoring and updating (i.e., an AI-QI team) should also take responsibility for monitoring and updating during the trial. • The AI-QI team should be independent of trial investigators and AI developers to minimize conflicts of interest and differences in behavior during versus after the trial.
Trial signal leakage through operational monitoring, resulting in biased evaluation of trial endpoints and/or unintended impacts on user adoption	Compartmentalization of monitoring responsibilities; for example: • *Data Monitoring Committee (DMC):* Exclusive access to comparative clinical outcomes. • *AI-QI team:* Access to system and workflow metrics (e.g., uptime, error rates, alert frequency, edit distance) that are unlikely to reveal comparative outcomes.
Lack of transparency due to the dynamic nature of the AI system	• Transparent documentation of any modifications of prompts, models or workflow integrations. • Prespecified update scope, version control and audit trails.

van Amsterdam, W. A. C., Oberst, M., Feng, J., et al. "Clinical trials for continuously monitored and updated AI systems." *Nature Medicine* (2026).

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Case study 1: detection of pulmonary embolism[1]

Design: Multi-center cluster-randomized trial, where hospitals randomized to AI system vs. standard care

- **Primary endpoint**: all-cause mortality; **secondary**: serious adverse events (SAEs) related to anti-thrombotic treatment

Monitoring as part of trial conduct (DMC)

- Exclusive access to between-arm differences in deaths and SAEs
- May recommend pausing enrollment if benefit-to-risk profile is unfavorable
- Between-arm comparisons cannot be shared with AI-QI team

Case study 1: detection of pulmonary embolism (cont.)

Monitoring as part of the intervention (AI-QI team)

- Operational metrics: server uptime, image acquisition quality, model outputs, radiologist agreement with alerts
- Prespecified actions: automated failover for downtime; retraining or recalibration if radiologist confirmation rates decline
- No access to comparative clinical endpoint data during the trial

[1] van Amsterdam, W. A. C., Oberst, M., Feng, J., et al. "Clinical trials for continuously monitored and updated AI systems." *Nature Medicine* (2026).

Case study 2: automated discharge summary generation[1]

Design: Non-inferiority randomized trial comparing AI-drafted discharge summaries vs. manual documentation

- **Primary endpoints**: readmission rate, documentation quality (accuracy and completeness assessed by independent reviewers)
- **Secondary**: time savings, self-reported clinician burnout, cognitive burden metrics
- Effectively **open-label**: clinicians know whether AI drafted the summary; subjective endpoints increase leakage risk

Monitoring as part of trial conduct (DMC)

- Retains sole responsibility for between-arm comparisons
- Documentation quality assessment may need to remain trial-specific (resource-intensive manual review not feasible in routine deployment, so this remains in this governance boundary)

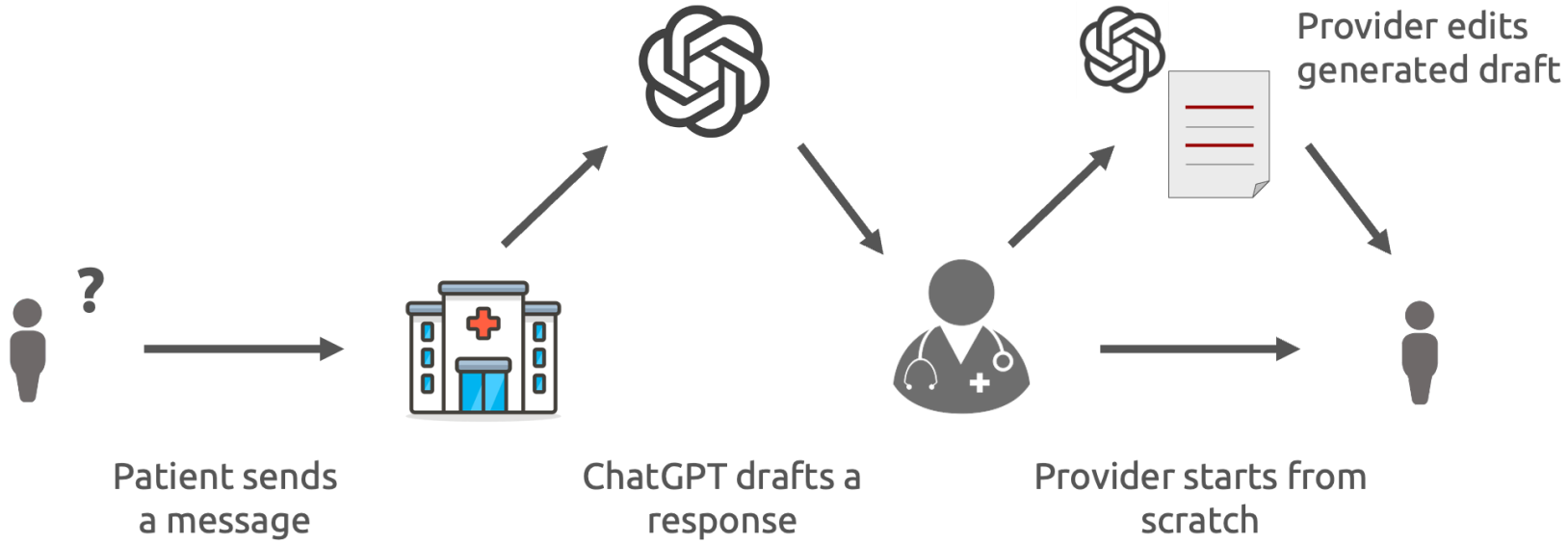
Case study 2: automated discharge summary generation (cont.)

Monitoring as part of the intervention (AI-QI team)

- Permitted: technical functionality, user training updates, vendor-released model updates (importantly, can allow for updates to improved foundation models)
- Tightly restricted: serious summarization errors must be compartmentalized to AI-QI team only since disclosure to DMC could bias subjective endpoints
- Primary clinical endpoints remain exclusive to the DMC

[1] van Amsterdam, Wouter AC, Michael Oberst, Jean Feng, Jenna Wiens, Shengpu Tang, Shalmali Joshi, Rajesh Ranganath et al. "Clinical trials for continuously monitored and updated AI systems." *Nature Medicine* (2026): 1-3.

Even retrospective evaluations of deployed AI tools needs careful design and analysis



Use case: Analysis of Augmented Response Technology (ART) or In-basket messaging, which creates LLM-generated editable draft responses for clinicians and is now widely integrated across many healthcare institutions across the US

Picture credit: Vincent Jeanselme

Prior ART evaluations and study designs

Paper	Design	Structure	Outcome	Key results
Tai-Seale et al. JAMA Net. Open, Apr 2024 - UCSD	RCT (n = 52)	Randomised: immediate vs delayed activation.	Read time Drafting time	Increase in reading time but reduction of drafting time.
Garcia et al. JAMA Net. Open, Mar 2024 - Stanford	Obs. (n = 162)	Pre - post mean comparison	Drafting time Exhaustion survey	No significant change in any objective time measure. Reduction in work exhaustion
Mandal et al. npj Digit Med, Oct 2025 - Langone	Obs. (n = 75)	Used vs not used mean comparison	Turnaround time	Turnaround reduction by 6.67%
Proctor et al. Appl Clin Inform, Aug 2025 - CHOP	Obs. (n = 323)	Used vs not used mean comparison	Response time	13 seconds reduction in response time
Bootsma-Robroeks et al. New Front Digit Health, Jun 2025 - UMCG	Obs. (n = 100)	Used vs not used mean comparison	Drafting time	No significant change in any objective time measure.

Picture credit: Vincent Jeanselme

Prior ART evaluations and study designs (cont.)

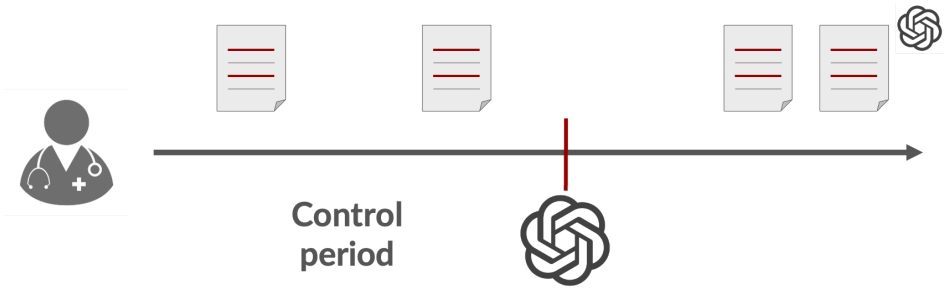
Challenges of prior study designs

- One RCT, pretty small (n=52)
- Pre-post mean comparison: many sources of confounding — e.g., which clinicians opt-in to use the tool?
- Used vs. not-used mean comparison: what type of messages do clinicians most often use the tool for?

Overall, prior work focused on reading and drafting time — but not efficiency-related and other outcomes relevant for clinicians, patients, and healthcare institutions. What is crucial to measure?

ART panel analysis that adjusts for various sources of confounding[1]

$$\log(\text{Turnaround}) = \text{Provider fixed effect} + \text{ART adoption} + \text{ART use} + \text{ART habituation} + \text{Message-level confounders} + \text{ART use} \times \text{confounders} + \text{Error}$$



Picture credit: Vincent Jeanselme

[1] Vincent Jeanselme, Catherine Austin, Jungmi Han, Rachel Lewis, Gregory W. Hruby, Karthik Natarajan, Shalmali Joshi, "Panel Analysis of In-Basket Messaging on Turnaround Time and Outside-work-hours Response Behavior." (under Review, 2026)

ART panel analysis that adjusts for various sources of confounding (cont.)

- **Message-level confounders**: message-type
- **Primary endpoint**: turnaround time (famously, "pajama time")
- **Provider-level confounders**: share provider-level effects before and after the tool is introduced
- ART-adoption, use for a specific message, and habituation

Other challenges

- Often there are **no timely labels** (live deployment and monitoring comes with its own challenges) or **no single right answer** (generative output).
- **Selective prediction / abstention** — let the model decline; report the **risk–coverage** tradeoff, not one accuracy. However, simply focusing on calibration is not the best way to determine abstention (learning-to-defer) because of challenges we discussed above
- **LLM-as-judge**: convenient, scalable, but **systematically biased** does not evaluate ground truth or an actual end-point that implies positive impact.
- **Generative outputs**: For radiology report generation, discharge summarization, clinically meaningful end-points are the only way to determine that the tools have positive impact. Famously, traditional semantically focused metrics show very little correlation with expert judgements[1]

[1] Agrawal, Monica, Irene Y. Chen, Freya Gulamali, and Shalmali Joshi. "The evaluation illusion of large language models in medicine." *npj Digital Medicine* 8, no. 1 (2025).

This afternoon: design a full evaluation pipeline for your usecase

The 2:45 notebook runs end to end on your project — **Sections 0–7:**

- **0** Project setup → **1** Discrimination → **2** Calibration (overall + subgroup)
- **3** Clinical utility (decision curves) → **4** Leakage & shortcut probes → **5** Model selection
- **6** Identify stages of study designs, what you will measure in each (silent trial, safety/small pilot, prospective evaluation), propose end-points and justify the motivation, use genAI and the internet to identify power analyses, and determine which pilot studies you would need to run to identify appropriate parameters
- **7** Report all designs with justification

Worked example, template, and AI-coding-tool guidance are provided — see the Day 3 & 4 Workshops doc.