SESSION 3

AI Limitations and Biases

Recognizing failure and developing calibrated trust

Preclinical M1/M2

Learning objectives

1. Fabrication	Identify unsupported claims and verify sources.
2. Algorithmic bias	Recognize biased proxies and examine equity effects.
3. Generalization and drift	Distinguish internal and external evaluation and recognize change over time.

Learning objectives

4. Automation bias	Recognize omission and commission errors and practice Read–Record–Reconcile.
5. Responsibility	Explain accountability, explainability, and the legal issues raised by AI-assisted care.

Characteristic failure profiles

Generative systems

Fabricated claims or citations
Persuasive but incorrect explanations
Contextual and cultural mismatch

Prediction and classification

Biased targets or labels
Poor transfer to new settings
Changing performance over time

Across both: evaluate the system, its workflow, and the clinical decision.

A historical example of fabricated references

69%

41 of 59 references were fabricated

20 medical questions
ChatGPT evaluated in February 2023

Median expert-rated answer quality: 60%

Gravel et al. Mayo Clin Proc Digit Health. 2023;1:226–234.

CONSTRUCTED TEACHING EXAMPLE

Persuasive fabrication

An LLM explains that hyperkalemia produces prominent U waves and a shortened PR interval.

The explanation sounds coherent.
Which physiological claims need checking?

Self-checking is not external verification

What self-checking can do

Prompt reconsideration or reveal inconsistency.
A repeated answer can still repeat the same error.

What verification requires

Find the source.
Confirm what it actually says.
Check relevance and current guidance.

Retrieval improves grounding, but still needs checking

What retrieval adds

External documents at answer time
Traceable source passages
Updates through the document collection

What can still fail

Wrong or outdated retrieval
Missing relevant material
Misreading a correct source
Unsupported synthesis

Lewis et al. NeurIPS. 2020; Amugongo et al. PLOS Digit Health. 2025.

CHECKPOINT 1

A real citation is only the first check

An approved RAG system summarizes a hypertension guideline. The citation resolves. What check remains?

A. No further check is needed
B. Ask the model to confirm itself
C. Read the guideline and compare the claim
D. Rely on the model's confidence

Proxy bias

The clinical goal	Allocate additional care according to health need.
The model target	Predict future healthcare spending.
The problem	Unequal access and spending can make equal predicted cost represent unequal illness burden.

Obermeyer et al. Science. 2019;366:447–453.

Labels and outcomes carry clinical history

Chest-radiograph classification

Underdiagnosis differed across underserved groups in evaluated datasets.
Overall performance concealed disparities.

Pneumonia prediction

Lower observed mortality among patients with asthma reflected intensive treatment.
The association was unsafe to interpret as lower untreated risk.

Seyyed-Kalantari et al. Nat Med. 2021; Caruana et al. KDD. 2015.

Sensitive information can remain in images

Models predicted self-reported race from several medical-image datasets.

The signal persisted under multiple image-degradation tests.

Removing a demographic field does not establish that a model is blind to it.

Gichoya et al. Lancet Digit Health. 2022;4:e406–e414.

Language models can reproduce bias

Clinical vignettes	Generated demographic patterns can reinforce stereotypes.
Assessment and recommendations	Output may change with identity cues that do not justify a clinical difference.
Verification	Review assumptions as well as factual accuracy.

Zack et al. Lancet Digit Health. 2024;6:e12–e22.

Four questions about equity

What is the target, and for whom is it a poor proxy?

Is performance reported across relevant patient groups?

Are demographic adjustments clinically justified and critically assessed?

Who audits performance and responds to disparities?

CHECKPOINT 2

Representative data do not settle fairness

A calibrated model predicts missed appointments and triggers double-booking. What is the clearest equity concern in this scenario?

A. Representative data rule out bias
B. The target may encode transport and caregiving barriers
C. The brand of hardware
D. Future demographic change

Internal and external validation

Internal validation

New observations from the development source or sampling framework.
Useful evidence within that context.

External validation

A meaningfully different site, population, time, device, or workflow.
Tests transfer beyond development.

Futoma et al. Lancet Digit Health. 2020;2:e489–e492.

Performance in real clinics

About 21%

Images rejected as ungradable in the reported field deployment

Retinopathy screening across 11 Thai clinics

Image quality, connectivity, and workflow affected how the system could be used.

Beede et al. CHI. 2020.

Drift after deployment

Patient mix	New case mix or demographic patterns
Equipment and assays	Different scanners, protocols, or laboratory methods
Documentation	New templates, codes, or input practices
Clinical context	Changes in disease prevalence or treatment

Monitoring makes drift visible

Evidence

AKI prediction models evaluated across nine years showed changing calibration and over-prediction.

Institutional response

Define monitoring ownership.
Track relevant outcomes and subgroups.
Re-evaluate after meaningful changes.
Recalibrate, update, or retire when justified.

Davis et al. J Am Med Inform Assoc. 2017;24:1052–1061.

Automation bias

Omission

The system misses a problem or remains quiet, and the user fails to act.

Commission

The system gives incorrect advice, and the user accepts it over a better assessment.

Tsai et al. JAMIA. 2003; Goddard et al. JAMIA. 2012.

Alert fatigue

49–96%

Drug-safety alert override rates in a 2006 review

High override rates call for investigation, not an assumption that every override was wrong.

Improve relevance, tier severity, suppress duplicates, and examine why alerts are dismissed.

van der Sijs et al. JAMIA. 2006;13:138–147.

Read–Record–Reconcile

Read	Review the patient and primary data before viewing the AI output when feasible.
Record	Commit your initial impression in writing.
Reconcile	Investigate discrepancies, including important omissions, and document the reasoning.

CHECKPOINT 3

A quiet model and a concerned nurse

On postoperative day 2, a nurse says a patient “seems off.” The model shows low risk and no alert. What governs the next move?

A. The score alone
B. The score if recently validated
C. Bedside assessment, reconciled with the score
D. A second model to break the tie

Clinical accountability and liability

Clinical duty persists	AI advice does not replace professional judgment or the standard of care.
Responsibility can be shared	Clinicians, institutions, and developers may have different duties.
Reasoning matters	Record what was reviewed, why the advice fit or did not fit, and how uncertainty was addressed.

Mello & Guha. N Engl J Med. 2024;390:271–278; Price et al. JAMA. 2019.

Explainability and professional responsibility

Explainability

A plausible explanation may not faithfully describe how a model produced its output. Consider validation, limitations, and clinical rationale.

Responsibility

The system cannot hold professional licensure or undertake a clinician's duty of care.
Keep clinical reasoning and oversight explicit.

Ghassemi et al. Lancet Digit Health. 2021; Mello & Guha. N Engl J Med. 2024.

Mitigations at a glance

Failure mode	Mitigation	Key source
Fabrication	Verify the source and what it supports.	Gravel 2023; Lewis 2020
Algorithmic bias	Examine the target and subgroup performance.	Obermeyer 2019; Seyyed-Kalantari 2021
Drift and generalization	Evaluate locally; monitor and reassess after change.	Beede 2020; Davis 2017
Automation and alerts	Read–Record–Reconcile; improve alert design.	Tsai 2003; van der Sijs 2006
Accountability	Keep clinical reasoning, review, and roles explicit.	Mello & Guha 2024; Ghassemi 2021

What to carry into clinical training

Match the check to the task and system.

Verify source existence, content, and applicability.

Examine targets and subgroup performance.

Expect validation to have a setting and a date.

Read, record, reconcile, and explain your decision.