tions and Biases

rriculum Alignment and Evaluation

undation / Millennium Conference • Preclinical M1/M2 • November 2026

ument maps Session 3 learning objectives to teaching, practice, and assessment. It also provides an evaluation plan across Ki –4, a session survey, and an annotated reading list.

n at a glance

ent	Time	Delivery
29 light slides, 3 ed checkpoints)	50 min	Whole class
Five Failure Modes" d)	75 min	Small groups, 1 facilitator per ~30 students
e 1 — The Confident	30–35 min	Small groups of 5–7
e 2 — The Algorithm pped Catching Sepsis	35–40 min	Small groups of 5–7
st-test + session	20 min	Individual (pre-test as pre-work)

ing options. Full implementation: lecture + Jigsaw in one 130-minute block (or two meetings), with the PBL cases in existing s me. Minimum implementation: lecture + Case 2 + MCQ pre/post. The lecture and the small-group materials use disjoint exampl out, so components can run in any order without spoiling one another.

culum Alignment Matrix

slide numbers refer to the Session 3 deck. PBL, Jigsaw, and MCQ/SAQ references refer to the Session 3 deliverables. The ma ective to teaching, practice, and assessment. Proxy bias is taught on slides 10–11 and reinforced in PBL Case 2, Primer B, MC Q 1.

e (abbreviated)	Bloom's	Lecture	PBL	Jigsaw	MCQ /SAQ	Evaluation
fy ions/fabrications in ontent	Understand/Apply	5–9	Case 1 Parts 1–3	Primer A	MCQ 2	L2: MCQ pre/post
AI-generated citations elines against primary	Apply	7–9	Case 1 Part 2 (live verification)	Primer A	MCQ 1	L2: MCQ; L3: OSCE verificatio station
ze persuasive fabrication but wrong)	Analyze	6	Case 1 Q1–Q2	Primer A mini-case	MCQ 1, 2	L2: MCQ pre/post
e proxy bias; wrong target → discriminatory	Understand/Apply	10–11	Case 2 Q6	Primer B	MCQ 3	L2: MCQ pre/post
ze how historical data nequities	Analyze	11–12	Case 2 Q6	Primer B	MCQ 3	L2: MCQ pre/post
ate an algorithm for ioeconomic bias before t	Evaluate	14–15	Case 2 Q6	Primer B mini-case; application round	MCQ 4; SAQ 1	L2: MCQ + SAQ; L3: observat card
ss cultural gaps that bias	Understand	13	Case 1 equity extension	Primer B	—	L1 survey; L2 teach-out rubric B)
al vs. external (and local)	Understand/Apply	16–17	Case 2 Q1	Primer C	MCQ 5; SAQ 1	L2: MCQ pre/post
fy dataset-shift / drift ms	Analyze	18–19	Case 2 Q2	Primer C mini-case	MCQ 6; SAQ 1	L2: MCQ pre/post

e (abbreviated)	Bloom's	Lecture	PBL	Jigsaw	MCQ /SAQ	Evaluation
ue the Epic Sepsis Model	Evaluate	PBL Case 2	Case 2 entire	Primer C	MCQ 5 (novel analog)	L2: MCQ; L1 survey
nation bias: errors of vs. commission	Analyze	20, 23	Case 2 Q3 + drill	Primer D	MCQ 7	L2: MCQ pre/post
nstrate the human-in-the-flow (Read–Record–e)	Apply	22–23	Case 2 Part 2 drill	Primer D mini-case; application round	SAQ 2	L2: SAQ; L3: OSCE + observa card
gnize alert fatigue and its pact	Understand	21	Case 2 Q4	Primer D; application round	SAQ 1 (operational bullet)	L2: teach-out rubric (Group D)
ed intermediary; liability on AI advice	Understand/Apply	24	Case 2 Q5	Primer E	MCQ 8	L2: MCQ pre/post
inability as an ethical	Evaluate	25	—	Primer E mini-case; application round	—	L1 survey; L2 teach-out rubric E)
a tool, not a team member	Understand	25–27	Case 2 Q5	Primer E	MCQ 8 (in part)	L2: MCQ pre/post

Teach-Out Rubric (Level 2 artifact)

ch expert's 2-minute teach-out 0–2 per criterion; 8 points total. Scored live by the facilitator (feasible at one facilitator per five h y sampling one expert per group), or from the expert's written teach-out notes collected at the end.

	0 — Absent	1 — Named	2 — Specified
definition	Missing or wrong	Correct but recited from the primer	Correct, in the expert's own words
example	None	Primer's example, partially garbled	Primer's example accurately — or new one
e mitigation	None or aspirational ("be careful")	A real behavior, vaguely stated	A behavior a shadowing classmat observe
stitution	None	Generic ("is it safe?")	Specific and answerable (names t evidence sought)

ted competence threshold: 6/8 with no zero. The home group's Phase 4 synthesis sentence is checked pass/revise against : it must name a mechanism, not a mood.

ation Plan — Kirkpatrick Levels 1–4

y tiers accompany the higher levels so all eleven consortium schools can implement at least the core plan. the mapping table a is document states explicitly the relationship between each evaluation instrument and session component.

— Reaction

ession survey (below), about five minutes. Core signals: perceived ability on each objective cluster, the value of the extended J e in-lecture checkpoints, and pacing.

2 — Learning

pre/post (new; core). The eight MCQs, administered before the session (or asynchronously as pre-work) and after. Paired sc the objective learning measure per objective cluster for each objective cluster.
hand-in (optional). The two short-answer items, scored with their point guides (6 + 4 points), as an individual graded hand-in o product. Recommended where schools want assessment 'built into' the session beyond recognition items.
aw teach-out rubric. The 0–8 rubric above — an artifact-based learning measure for the active-learning component specificall
cture checkpoints (formative). The three checkpoint votes embedded in the lecture (slides 9, 15, 23) give the lecturer an imm on each objective cluster and give students retrieval practice; tallies can be recorded (audience-response or hand-count) as a ative signal but are not graded.

3 — Behavior

E verification station. 3–6 months post-session. The student receives an AI-generated literature summary for a clinical questi
aining one fully fabricated citation and one real citation whose content is misstated, with PubMed access and five minutes. Che
mpts verification unprompted; distinguishes existence-checking from content-checking; catches both plants; states conclusions
ces. Feasibility tier: schools without OSCE capacity run it as a 15-minute proctored classroom exercise.

ct-observation card (Read–Record–Reconcile). A brief EPA-style card for preceptors, completed once per student per releva
on: commits a written impression before consulting a predictive tool or AI read; verifies AI-supplied citations or content before
enting; can state why they accepted or overrode an AI output. Aggregated, this is the session's primary behavioral endpoint an
ed with Session 2's card where both sessions run.

ent self-audit (low-cost tier). At clerkship midpoint, students log one real encounter with an AI tool or alert in three sentences
read first, what they recorded, how they reconciled. Reviewed in existing mentoring structures.

4 — Results

t-governance signal. Where deploying institutions expose them, alert override rates trended by unit — the population-level tra
sion-by-fatigue failure this session teaches, and a metric students should learn to read as a design signal.

ion-integrity signal. Where schools audit student scholarly work or notes, tag citation errors traceable to unverified AI output,
us after the curriculum. The ratio proxies verification behavior at scale.

ernance-adoption signal. Count of consortium institutions whose AI-procurement or governance processes require subgroup
rmance reporting and local validation — the institutional behavior this session's Objective 2–3 content argues for.

on caveat. Level 4 outcomes are influenced by far more than one preclinical session; these indicators are monitored as conso
text, not claimed as session effects. A separate consortium evaluation working group should own this layer.

on Evaluation Survey

tems (1–5, strongly disagree to strongly agree)

explain why fabricated AI citations are hard to spot by inspection, and describe the verification steps that actually settle the qu
explain how the choice of a target variable can encode bias into an algorithm, even when training data are representative and
r an input.
state what I would ask about subgroup performance and local validation before an algorithm is used on patients like mine.
explain why a model that performed well at deployment may degrade silently over time.
describe the Read–Record–Reconcile procedure and say when I would use it.
erstand that I — not the algorithm vendor — am the responsible party when I act on, or am reassured by, AI output.
hing my failure mode to my home group, and applying all five to the shared vignette, was an effective way to learn this materia
checkpoint questions during the lecture helped me consolidate the material as it was taught.
ecture content was clearly presented and appropriately paced.

ence self-assessment (retrospective pre/post, 1–5)

ally appraising a clinical AI tool's published performance claims, including what additional evidence to demand.
ing with a predictive tool or AI interpretation without letting it replace my own read of the primary data.

ended

t was the single most valuable thing you learned in this session?
t is one thing about this session that could be improved before it is taught again?

instrument speaks to which component

component	Evaluation instruments
incl. checkpoints)	Survey items 1–6, 8–9; MCQ pre/post; checkpoint tallies (formativ
	Survey item 7; teach-out rubric; synthesis-sentence check
e 1	Survey item 1; MCQ 1–2; OSCE verification station (L3)
e 2 (incl. drill)	Survey items 2–6; MCQ 3–8; SAQ 1–2; observation card and self- (L3)
ession (behavioral)	Observation card; self-audit; Level 4 signals

ed Reading List (revised)

owing readings support the session. Distinguish empirical findings from conceptual frameworks, and check current local poli guidance.

, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized p tern Med. 2021;181(8):1065–1070. — The primary source for Case 2 and the clearest published example of the generalization l reading.

yer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. 6(6464):447–453. — The foundational proxy-bias paper — cost as a proxy for need. Required reading; grounds Objective 2.

M, Guha N. Understanding liability risk from using health care artificial intelligence tools. N Engl J Med. 2024;390(3):271–278. clinician-facing analysis of where liability sits and concentrates; the primary legal reading, superseding (but not replacing) Price

, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT fo
questions. Mayo Clin Proc Digit Health. 2023;1(3):226–234. — The measured fabrication-rate study behind Objective 1; pairs w

JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Hea
):e406–e414. — Models can read race from images humans cannot — why bias can be uninspectable. Pairs with the equity cl

Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model
n study. Lancet Digit Health. 2024;6(1):e12–e22. — The LLM cultural-bias evidence for Objective 2's language-model arm.

go LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare
tic review. PLOS Digit Health. 2025;4(6):e0000877. — Fills the reading list's previously unfilled RAG slot: what retrieval demon
what it does not.

d K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inf
012;19(1):121–127. — Foundational automation-bias review. Technically demanding — assign with a reading guide.

N 2nd, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA. 2019;322(18):1765–1766. — Co
atement of the learned-intermediary framing; a two-page alternative for time-pressed readers.

R. The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age. McGraw-Hill, 2015. (Excerpts.) — Th
narrative treatment of alert fatigue and unintended CDS consequences for preclinical learners.