AI Concepts in Healthcare

Assessment Questions

Macy Foundation / Millennium Conference • Preclinical M1/M2 • November 2026

Five NBME-style items calibrated for preclinical (M1/M2) students, each mapped to a Session 1 learning objective. Suggested use: pre/post administration for the Level 2 evaluation (see the Alignment Matrix & Evaluation Plan document). Answer key with explanations follows the items.

Question 1

Maps to learning objective 1 — foundational AI taxonomy: generative versus predictive systems, applied

A hospital deploys an artificial intelligence system that analyzes chest radiographs and reports the probability that a pneumothorax is present. A first-year medical student who has been warned that chatbots can fabricate plausible-looking citations asks whether this system's output should be doubted for the same reason. Which response is most accurate?

A. Yes — every AI system produces generated content, so any of them can fabricate references and explanations when it is uncertain about its answer.

B. Yes — the system was trained on historical data, and any system trained on historical data fabricates outputs at a rate proportional to the size of its training set.

C. No — because the system is regulated as a medical device, its outputs have been verified to be error-free before marketing clearance was granted.

D. No — it is a pattern-recognition model that outputs a probability, not generated text; its characteristic risks are miscalibration, bias, and drift.

E. It cannot be determined without knowing whether the system's internal architecture uses a neural network or a statistical regression model.

Question 2

Maps to learning objective 2 — mechanics of large language models and hallucination

A first-year medical student asks a publicly available large language model to summarize three studies on beta-blockers in heart failure with reduced ejection fraction. The model returns a fluent paragraph naming three studies; one of the cited studies does not exist, although its title and author list look entirely plausible. Which of the following best explains why this occurred?

A. The model predicts statistically likely word sequences, and can generate a plausible citation pattern with no check that the study exists.

B. The model retrieved the study from a private subscription database that the student's own institutional credentials cannot access.

C. The model deliberately fabricated the study because it was trained to prioritize user satisfaction over factual correctness.

D. The model's training dataset included only outdated cardiology literature, so it substituted an older trial that has since been renamed.

E. The prompt failed to specify a citation format, so the model defaulted to inventing references in its own internal style.

Question 3

Maps to learning objective 3 — model performance metrics and the accuracy trap

An artificial intelligence screening tool for a disease with a 1% prevalence in the screened population reports 99% accuracy. Which of the following statements is most appropriate to make on the basis of this information alone?

A. The tool is ready for clinical deployment, since 99% exceeds the accuracy benchmark generally accepted for screening instruments of this kind.

B. The tool's specificity must be at least 99%, since overall accuracy places a mathematical floor beneath both specificity and sensitivity.

C. The tool's positive predictive value must also be approximately 99%, since predictive values track overall accuracy when prevalence is stable.

D. The accuracy figure rules out overfitting, because an overfit model would be unable to achieve high accuracy in the screened population.

E. No conclusion about clinical usefulness can be drawn — a model labeling every patient negative would also be 99% accurate here.

Question 4

Maps to learning objectives 1 and 2 — private versus public systems; data privacy in prompting

A medical student on a clinical preceptorship wants to use an artificial intelligence assistant to help draft a differential diagnosis for a patient she has just interviewed. Which of the following is the most appropriate next step?

A. Paste the patient's history into a publicly available consumer chatbot after deleting the patient's name, medical record number, and date of birth.

B. Paste the patient's history into the institution's HIPAA-compliant, sanctioned AI tool.

C. Paste the patient's identified history into a publicly available consumer chatbot, since use for the student's own education is not a clinical disclosure.

D. Use only her own clinical reasoning, since consulting an AI tool about a real patient is inappropriate at every stage of medical training.

E. Send a screenshot of the patient's chart to a classmate at another institution whose school provides a sandboxed AI tool for student use.

Question 5

Maps to learning objective 3 — algorithmic drift and the model lifecycle

A clinical artificial intelligence tool that performed well at deployment shows a steady decline in positive predictive value over 18 months. The software has not been updated during this period. Which of the following best describes this phenomenon?

A. Overfitting — the model memorized idiosyncrasies of its training data and is now being exposed on genuinely novel patients.

B. Underfitting — the model was always too simple for the task, and accumulating patient volume has finally made the deficit statistically visible.

C. Algorithmic drift — the population, practice patterns, or data-capture conventions have shifted away from the training distribution.

D. Hallucination — the model is generating plausible but incorrect predictions in place of the validated outputs it produced at launch.

E. Adversarial attack — an external actor is deliberately submitting manipulated inputs designed to degrade the model's performance.

Answer Key & Explanations

Question 1 — Correct: D

The pneumothorax detector is a predictive, pattern-recognition (computer-vision) system: image in, probability out. It generates no text, so citation fabrication — a failure mode of generative systems — does not apply. It has its own characteristic risks instead: miscalibration, bias inherited from training data, poor generalization, and drift. A anthropomorphizes all AI as generative. B invents a fabrication rate tied to training-set size. C mistakes regulatory clearance for a guarantee of error-free output — clearance is evidence review, not infallibility. E is a dodge: the family distinction, not the internal architecture, determines the relevant failure modes, and that distinction is given in the stem.

Question 2 — Correct: A

Large language models predict the next token; citation formats are highly regular patterns, so a model can generate a plausible-looking citation with no underlying referent. B is unsupported: the scenario provides no evidence that a real inaccessible study was retrieved. C anthropomorphizes — the model has no intent. D fails because the fabricated study does not exist under any name. E confuses formatting instructions with a verification mechanism; no citation style prevents fabrication.

Question 3 — Correct: E

At 1% prevalence, a model that predicts “negative” for everyone achieves 99% accuracy and catches zero cases — so the headline figure, alone, licenses no conclusion about usefulness. Sensitivity, specificity, positive predictive value, and calibration are required. A is the marketing conclusion the item exists to reject. B and C are mathematically false — accuracy places no such floor on specificity, and PPV at low prevalence can be tiny despite high accuracy. D is unrelated: accuracy on one population says nothing about overfitting.

Question 4 — Correct: B

Identifiable patient information may be entered only into HIPAA-compliant, institutionally sanctioned tools. A fails because removing three identifiers is not de-identification — quasi-identifiers (age, dates, rare diagnoses, locations) leave re-identification risk, and institutional policy generally bars patient data from public tools regardless. C is a privacy violation regardless of educational intent. D overcorrects: sanctioned AI use on real patient data is widely permitted and often valuable — the curriculum's stance is verified use, not abstinence. E discloses protected health information to an unauthorized recipient and system.

Question 5 — Correct: C

Performance decline in an unchanged model is consistent with dataset shift: the world the model reads has moved away from the world it learned. Overfitting and underfitting describe model fit during development; the progressive decline after initially good deployment performance more directly suggests a change in the population or data-generating process. Declining PPV can also reflect changing prevalence and warrants investigation. Hallucination (D) is a generative-model failure mode, not a property of a predictive tool's declining PPV. Adversarial attack (E) requires a deliberate actor and a specific manipulated-input pattern, neither of which is in evidence — and drift needs no adversary, only time.