AI Ethics

Assessment Questions

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

Eight NBME-style items with answer key and explanations. Suggested use: pre/post administration for the Level 2 evaluation.

Question 1

Objective 1 — ethical principles applied to AI

A health system deploys a machine-learning model that estimates each discharged patient's 30-day readmission risk and routes high-risk patients to extra follow-up. External validation at the system's own hospitals shows good accuracy with equal performance across demographic groups. A two-year quality review finds reduced readmissions, no measurable harms, and benefits distributed evenly across the patient population. Patients are not informed that the model contributes to their discharge planning.

Which ethical concern remains unaddressed despite these findings?

A. Beneficence, because the model's clinical benefit has not been convincingly established

B. Non-maleficence, because rare harms cannot be excluded without a much longer observation period

C. Autonomy, because patients have had no opportunity to know of or question the model's role in their care

D. Justice, because equal measured performance does not guarantee that access to follow-up resources is equitable

E. None, because a tool that demonstrably helps and does not harm requires no further ethical justification

Question 2

Objective 2 — disclosure standards

A primary care clinic uses an ambient AI scribe that records visit audio and drafts clinical notes for physician review. The intake packet states: “This office uses electronic systems for documentation.” After a patient complains about learning of the recording from a nurse, the clinic asks an ethics consultant how to make its disclosure practice ethically adequate.

Which change best accomplishes this goal?

A. Add the software vendor's name and regulatory clearance status to the existing intake packet language

B. Tell patients in plain language that visits may be recorded and drafted by AI, and that they may decline

C. Post a waiting-room notice describing the clinic's electronic health record and documentation workflow

D. Obtain a signed general consent-to-treat that broadly covers all technologies the clinic may employ

E. Instruct clinicians to describe the scribe in detail whenever a patient asks about recording devices

Question 3

Objective 3 — patient autonomy

A 58-year-old woman with well-controlled type 2 diabetes (A1c 6.8% on metformin, BMI 24, no cardiovascular disease) brings a printout from a consumer AI chatbot recommending she start a GLP-1 receptor agonist, and asks her physician to prescribe it. The physician judges the medication not clinically indicated. The patient responds: “It's my body — I have the right to decide my treatment.”

Which response best reflects the ethical scope of patient autonomy?

A. Prescribe the requested medication, because autonomy gives patients final authority over the selection of their own treatments

B. Prescribe a time-limited trial while documenting the chatbot recommendation as the indication for therapy

C. Explore her reasons and explain the assessment, while declining to prescribe a treatment that is not indicated

D. Refer her to a colleague who may be more comfortable following the chatbot's recommendation

E. Advise her to stop consulting AI tools about her health, because they undermine the therapeutic relationship

Question 4

Objective 4 — weighing benefit against environmental cost

A hospital AI-governance committee is evaluating an imaging-AI platform expected to meaningfully reduce missed lung nodules on chest CT. The sustainability report notes the deployment would run in a vendor data center located in a drought-affected region, with substantial projected energy and water use. Committee members disagree about how environmental impact should enter the decision.

Which approach best reflects ethical reasoning grounded in justice and beneficence?

A. Exclude environmental factors from deliberation, since the committee's mandate concerns patient outcomes

B. Approve the deployment only on the strict condition that the vendor purchases verified carbon offsets covering all of the associated emissions

C. Reject the deployment, because environmental harms to communities outweigh benefits accruing to individual patients

D. Weigh the tool's documented benefit against its footprint, require vendor impact disclosure, and revisit the balance after deployment

E. Postpone any decision until national environmental standards for health care AI have been formally established

Question 5

Objectives 1 and 2 — justice and disclosure under a known performance gap

An externally validated deterioration-prediction model performs well overall, but the validation study shows meaningfully lower sensitivity for patients with sparse chart documentation — disproportionately patients with limited English proficiency. Retraining will take at least a year. The hospital must decide how to proceed in the interim.

Which approach best reflects the principles of justice and non-maleficence?

A. Withhold the tool from patients with sparse chart documentation until its measured performance is demonstrably equal across all patient groups

B. Deploy without restriction, since aggregate performance meets the accepted published standards for clinical use

C. Suspend all use of the model at every site until the sensitivity gap has been fully eliminated by retraining

D. Deploy with clinicians informed of the gap, reinforced routine monitoring for affected patients, and scheduled re-validation

E. Deploy the model only at hospitals whose patient populations closely resemble the original validation cohort

Question 6

New — generative AI vs. machine-learning limitations

In the same week, a medical student uses a secure large language model to summarize literature for a case write-up, and, on rounds, reviews an EHR-embedded model that displays each patient's deterioration-risk score.

Which statement correctly pairs one of these tools with its characteristic limitation and the appropriate safeguard?

A. The language model is prone to model drift, which the student should address by checking each of its citations against PubMed

B. The language model may fabricate plausible references, so its claims should be verified against primary sources

C. The risk model is prone to hallucination, which the hospital should address through prompt engineering

D. The risk model's main limitation is cultural gaps in its training text, addressed by rewording its input notes

E. The two tools share the same limitations, so one verification workflow is appropriate for both

Question 7

New — alert fatigue vs. automation bias

A hospital deploys a sepsis early-warning system with a positive predictive value of 12% at its chosen threshold. Nurses receive dozens of alerts per shift. Six months after deployment, a patient with true early sepsis deteriorates after an alert is dismissed without a bedside evaluation.

The root-cause review would most usefully conclude that the missed case primarily reflects:

A. Individual negligence by the dismissing nurse, best addressed through disciplinary referral

B. Automation bias, because the clinical staff had substituted the model's judgment for their own bedside assessments

C. Alert fatigue from a high false-alarm burden, best addressed by re-tuning the threshold and alert design

D. Model drift, because the system's predictive performance must have degraded in the months after deployment

E. Insufficient training, best addressed by a mandatory education module on responding to alerts

Question 8

New — AI as a tool, not a team member

A physician signs a clinic note drafted by an ambient AI scribe. The note states “no known drug allergies,” although the patient has a documented penicillin allergy that was not mentioned aloud during the visit. Relying on the signed note, a covering colleague later prescribes amoxicillin, and the patient has an allergic reaction. At the case review, the physician states that the scribe made the error.

Which statement best characterizes professional responsibility here?

A. Responsibility is shared equally between the signing physician and the software vendor whose product erred

B. The physician is responsible, because signing the note makes its content the physician's own

C. The vendor bears primary responsibility, because the documentation error originated in its product

D. The health system bears sole responsibility for choosing to deploy an imperfect documentation tool

E. No party is responsible, because the error represents an unforeseeable failure of an automated system

Answer Key and Explanations

Question 1 — Correct: C (Autonomy)

Why C is correct. The stem stipulates evidence that answers every other principle: demonstrated benefit (beneficence), no measurable harm (non-maleficence), equal performance and even distribution (justice). No outcome study can supply the patient's standing to know about and question their care — autonomy is not reducible to results.

Why the others are wrong. A and E are refuted by the stem's stipulations. B leans on the truism that harm can never be fully excluded — an argument that would forbid all of medicine. D invents an access problem the stem excludes.

Question 2 — Correct: B

Why B is correct. The materiality standard asks what a reasonable patient would want to know: that they are recorded, that AI drafts the note, and that they may decline. B supplies exactly those three elements.

Why the others are wrong. A adds detail less material than the fact of recording. C and D are transparency theater — generic notices and blanket consents inform no one of the material facts. E restricts honesty to patients who already suspect enough to ask, inverting the obligation.

Question 3 — Correct: C

Why C is correct. Autonomy is a right to be informed, to participate, and to refuse — not an entitlement to non-indicated treatment. Respecting this patient means engaging her reasons seriously while holding the clinical line. An AI recommendation does not create a clinical indication — the tool-not-team-member principle in patient-facing form.

Why the others are wrong. A and B capitulate (B doubly so — it documents an AI output as the indication). D off-loads the conflict rather than engaging it. E disparages the patient's information-seeking, which damages the relationship autonomy is meant to protect.

Question 4 — Correct: D

Why D is correct. Sound reasoning here is procedural, not categorical: credit the documented benefit (beneficence), take seriously a footprint borne by a community that does not share the benefit (justice), and weigh honestly over time — vendor disclosure plus post-deployment review, since both benefit and footprint are estimates that drift.

Why the others are wrong. A and C refuse to weigh a genuine trade-off. B converts an ethical deliberation into a purchasable indulgence and ignores water and siting. E defers a decision the institution is already making by default.

Question 5 — Correct: D

Why D is correct. Justice as remediation: deploy with eyes open — clinicians told where the model is weak, routine human monitoring reinforced exactly where the safety net is thinner, re-validation scheduled so the gap closes rather than fossilizes.

Why the others are wrong. The intuitive answers compound the injustice. A denies the affected group the substantial partial benefit the tool still provides — lower sensitivity is not zero benefit. C denies everyone a validated benefit to no one's gain. B pretends distribution doesn't matter; E simply exports the gap.

Question 6 — Correct: B

Why B is correct. Generative models characteristically fabricate — including realistic citations — because they compose plausible text rather than retrieve verified facts. The safeguard is external verification against primary sources.

Why the others are wrong. A and C swap the failure modes across tool families. D misattributes a text-corpus limitation to a structured-data model. E is the exact conflation this objective exists to prevent — different mechanisms require different safeguards.

Question 7 — Correct: C

Why C is correct. A 12% PPV means roughly seven of eight alerts are false; dismissal becomes the learned, locally rational response. That is alert fatigue, and its remedy is design — threshold tuning, alert tiering, workflow — not discipline or exhortation.

Why the others are wrong. A and E treat a systems failure as a character or knowledge failure; neither changes the signal-to-noise ratio. B names the opposite error — automation bias is over-trusting the system; this staff had learned to under-trust it. D is unsupported: the PPV was 12% by design from day one.

Question 8 — Correct: B

Why B is correct. Signature is the act by which a clinician adopts a document's content as their own professional statement. The note is the physician's, whoever — or whatever — drafted it. A tool cannot hold a duty of care, so accountability does not transfer to it.

Why the others are wrong. Vendor and system duties are real but arrive through different channels; they do not dilute the signing clinician's responsibility (A, C, D). E mistakes a foreseeable, known failure mode — scribes err, which is why review-before-signing is the standard — for an unforeseeable event.