AI in Clinical Skills and Assessment

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 (see the Alignment Matrix & Evaluation Plan).

Question 1

Objective 1 — recognizing tools, the tasks they support, and their characteristic limitations

During a single clinic day, a medical student encounters three AI tools: an EHR banner displaying each patient's 30-day readmission risk score; an ambient scribe that drafts visit notes from recorded conversation; and a chart summarizer that condenses two hundred pages of outside records into one screen.

Which statement correctly pairs one of these tools with the clinical task it supports and a characteristic limitation the student should anticipate?

A. The readmission score supports discharge planning, and because it is computed from structured data rather than generated text, a validated score can be integrated into planning without further verification against the patient in front of you.

B. The ambient scribe supports diagnostic reasoning, and its principal limitation is algorithmic bias inherited from the demographics of its training population.

C. The chart summarizer supports information retrieval, and it may omit or distort source content — so claims that would change management should be checked against the original records.

D. The readmission score supports diagnosis, and a high score should prompt the team to treat the patient as actively deteriorating.

E. The ambient scribe supports documentation, and because it transcribes the recorded audio verbatim, its errors are limited to misheard words.

Question 2

Objective 1 — de-skilling versus never-skilling, and why the distinction changes the remedy

An attending physician who trained before decision-support tools existed notices that after a decade of relying on an EHR-integrated differential generator, she struggles to construct a differential when the system is down. Her M3 student, who has used the same tool since his first clinical day, produces excellent AI-assisted differentials but cannot begin one unassisted — and, unlike the attending, does not experience this as a loss.

Which characterization of the two clinicians is most accurate?

A. Both exhibit deskilling, differing only in severity; the same remediation applies to each.

B. The attending exhibits deskilling; the student exhibits never-skilling — a skill never built because the tool absorbed the cognitive work first.

C. The attending exhibits automation bias; the student exhibits deskilling accelerated by early exposure to the tool.

D. Both exhibit never-skilling, since neither can currently perform the task without technological assistance.

E. The attending exhibits mis-skilling; the student exhibits automation bias, since he accepts the tool's output in place of his own assessment.

Question 3

Objective 1 — compliant use of sanctioned versus publicly available tools

After a clinic session, a medical student wants to learn more about a memorable patient: a 62-year-old recent immigrant, seen that afternoon, whose fatigue and orthopnea suggest new heart failure. The institution provides a secure, sanctioned clinical LLM approved for patient data; institutional policy permits publicly available AI tools for general educational use only.

Which of the following uses complies with the policy?

A. Pasting the patient's history into a public chatbot after deleting her name, medical record number, and date of birth.

B. Asking a public chatbot to explain the pathophysiology of orthopnea in heart failure, including no details about the patient.

C. Describing the patient to a public chatbot as “a 62-year-old woman who just immigrated, seen in my Tuesday afternoon clinic,” since no formal identifiers are included.

D. Uploading a photograph of the patient's leg edema to a public image-analysis tool, reasoning that a photograph of skin contains no identifying information.

E. Sending the case summary to a friend at another institution so the friend can run it through that institution's sanctioned tool.

Question 4

Objective 2 — input quality, training data, and reliability in a specific patient case

An urgent care clinic uses a validated machine-learning tool that estimates the probability of serious pathology from structured EHR data: vital signs, laboratory results, problem lists, and prior documentation. A 62-year-old woman who immigrated eight months ago presents with a month of fatigue; this is her first visit in this health system, her chart contains no laboratory results or problem list, and tonight's intake note is three lines long. The tool displays a reassuringly low risk estimate.

Which consideration should most influence how the clinician weighs this estimate?

A. The tool's published external-validation performance, which established good discrimination and calibration and indicates that the estimate can be relied upon wherever the software is deployed.

B. The direction of the estimate — low-risk predictions are inherently safer to trust than high-risk predictions, which require confirmation.

C. The sparseness of her record: the model computed its estimate from near-absent input, so the low score largely reflects missing data.

D. The tool's regulatory clearance status, since cleared devices have been shown to perform equivalently across patient populations and record types.

E. The consistency of the estimate on recalculation, since a stable output across repeated runs indicates the underlying prediction is trustworthy.

Question 5

Objective 2 and Objective 4 — independent reasoning first, then iterative use of AI output

An M3 completes a full history and physical on a patient with dyspnea, writes her own prioritized differential, and then — per her clerkship's AI policy — enters her complete H&P into the sanctioned clinical LLM. The tool returns a differential that overlaps hers but adds two diagnoses she had not considered, one of which fits the patient's presentation well.

Which next step makes the best use of the tool for both this patient and her own development?

A. Replace her differential with the AI's, since the tool had access to her complete H&P and produced the more comprehensive list.

B. Retain her original differential unchanged, since integrating AI-generated diagnoses at this stage of training would undermine the independence of her clinical reasoning.

C. Evaluate each discrepancy against the patient's findings, incorporate what survives the check, and record the missed diagnoses as study targets.

D. Re-enter the H&P with adjusted wording until the tool's differential matches her own, confirming her original reasoning.

E. Present the AI's additions to her attending as her own analysis, since she has verified that they are clinically plausible.

Question 6

Objective 3 — validating AI-generated documentation before it becomes the record

At the end of an urgent care shift, a resident reviews an ambient-scribe draft for a patient she evaluated with a student. The draft reads: “Review of systems: negative for chest pain, palpitations, or dyspnea. All other systems reviewed and negative. Medications: hydrochlorothiazide 25 mg daily.” The resident recalls that no review of systems was performed, cardiac symptoms were never asked about, and the patient said only that she takes “a pill for blood pressure.”

Which action is most appropriate before signing?

A. Sign the draft as written: the negatives are clinically plausible for this presentation, and the note can be amended later if any prove wrong.

B. Retain the ROS language because “reviewed and negative” is a recognized template convention, but correct the medication entry to “antihypertensive, name unverified.”

C. Remove the unperformed ROS and unverified medication details, then return to the patient to obtain the missing information before completing the note.

D. Regenerate the draft with the scribe and sign the new version if the fabricated content no longer appears.

E. File a report on the scribe's error and write this note manually, resuming use of the tool once the vendor confirms the defect is fixed.

Question 7

Objective 3 — documentation accuracy, clinician attention, and patient-centered communication

Six weeks after adopting an ambient scribe, a physician notices two things: she now maintains eye contact throughout visits and takes no notes of any kind, the scribe's drafts repeatedly miss symptoms that patients mention in passing, conversational language. She wants to preserve the relational gains without signing incomplete notes.

Which adjustment to her in-visit behavior best serves both goals?

A. Return to typing into the EHR during visits so that significant findings are captured contemporaneously by her own hand.

B. Keep the visit unchanged and compensate by reviewing each draft more carefully afterward, relying on her memory of the conversation to supply what the scribe missed.

C. Restate and explore significant symptoms aloud when patients mention them — signaling their importance to the patient and giving the scribe its salience cues.

D. Limit the scribe to visually straightforward visits and document complex patients manually.

E. Coach patients at the start of each visit to describe symptoms in precise clinical terminology so the scribe can recognize them.

Question 8

Objective 4 — AI-generated output within shared decision-making

A 58-year-old woman with knee osteoarthritis arrives with a printout from a consumer AI chatbot recommending topical NSAIDs, structured physical therapy, and weight management before considering injections or surgical referral. The recommendations are consistent with current guidelines. The physician had been planning to offer an intra-articular corticosteroid injection today. The patient says: “I want to do this right — what do you think of this plan?”

Which response best integrates the AI-generated recommendation into shared decision-making?

A. Explain that consumer AI tools are unreliable sources of medical advice and that treatment decisions should rest on the physician's clinical assessment instead.

B. Adopt the chatbot's plan in full, since its guideline concordance combined with the patient's evident preference makes further clinical discussion unnecessary.

C. Explore what “doing this right” means to her, assess the suggestions on their clinical merits, and build a shared plan incorporating the reasonable elements.

D. Refer her to orthopedic surgery so that a specialist can adjudicate between the chatbot's recommendations and the injection.

E. Proceed with the planned injection while documenting that the patient consulted outside sources, revisiting the printout only if the injection fails.

Answer Key and Explanations

Question 1 — Correct: C

Why C is correct. Each tool family maps to a task and a characteristic failure: the summarizer supports retrieval and can silently omit or distort what the source records say, so management-changing claims get checked against the originals. This is the session's core evaluative habit — AI output is a claim to verify, not a fact to inherit.

Why the others are wrong. A grants a validated score exemption from patient-level verification — validation is a population property, not a guarantee about this patient. B assigns the scribe the wrong task and the wrong signature limitation. D converts a probabilistic estimate into a diagnosis. E misdescribes the mechanism: scribes summarize with a language model rather than transcribing verbatim, which is exactly why fabrication — not just mishearing — is on the error list.

Question 2 — Correct: B

Why B is correct. The attending built the skill and lost it through disuse — deskilling. The student never built it — never-skilling — and the stem's final clause (“does not experience this as a loss”) is its diagnostic signature: never-skilling is invisible to the person affected, which is why it is the more dangerous condition for a learner and why the remedies differ (the attending can rehearse a skill she once had; the student must build one from nothing while a faster alternative sits open in another window).

Why the others are wrong. A erases the distinction the remedies depend on. C and E mislabel states of skill loss as automation bias, which is an error of over-trusting output in the moment, not a developmental condition. D stretches never-skilling to cover a clinician who demonstrably had the skill.

Question 3 — Correct: B

Why B is correct. A general educational question containing no patient information is precisely what public tools are permitted for under the stated policy. Note what this item rewards: not abstinence, but discrimination — knowing which uses are open, which are restricted, and why.

Why the others are wrong. A relies on informal de-identification; a narrative HPI can remain re-identifiable after obvious identifiers are stripped. C is the same failure in disguise — “62-year-old recent immigrant, Tuesday clinic” is a quasi-identifier set that could narrow to one person. D moves protected clinical information (a photograph of the patient's body) into a public tool. E launders the violation through a friend: the data still leaves the institution without authorization, and the friend's tool is sanctioned for their patients, not this one.

Question 4 — Correct: C

Why C is correct. A model can only assess the data it receives. This patient's record is nearly empty, so the “low risk” output is substantially an artifact of absent input — and the clinical response is to treat the estimate as uninformative and weight the primary history and examination accordingly. The equity dimension is deliberate: patients with fragmented records (recent immigrants, the uninsured, the recently arrived) are systematically the patients for whom such scores are least reliable.

Why the others are wrong. A confuses population-level validation with patient-level reliability. B inverts the safety logic — a false negative is precisely the dangerous direction here. D overstates what regulatory clearance establishes. E mistakes precision for accuracy: a model recomputing on the same absent data will stably produce the same uninformative answer.

Question 5 — Correct: C

Why C is correct. This is the iterative workflow the curriculum teaches: reason independently, then use the AI twice over — once as a check on the patient's differential (integrating only what survives verification against actual findings), and once as a mirror for the learner (the missed diagnoses become the week's study list). The tool improves the assessment and the student.

Why the others are wrong. A and B are the two purisms: wholesale adoption abandons verification, wholesale rejection abandons the benefit — and both waste the learning signal in the discrepancy. D is confirmation-seeking dressed as diligence. E converts verified-but-borrowed reasoning into a misrepresentation to a supervisor.

Question 6 — Correct: C

Why C is correct. A signature attests that the documented care occurred. An unperformed ROS and an invented medication dose cannot be attested — they must be corrected, and the underlying information must be actually obtained, which takes two minutes with the patient still in the building. The scribe's fabrications are foreseeable behavior of a summarization system; the checked, corrected note is the clinician's product.

Why the others are wrong. A signs false statements because they are plausible — plausibility is the property that makes fabrication dangerous, not a defense. B is the seductive one: template conventions do not convert an unperformed examination into a performed one, and “name unverified” documents a gap the resident could simply close. D re-rolls the dice without obtaining the missing information; a cleaner-looking draft would still describe care that never happened. E treats expected system behavior as a defect to escalate and forfeits the tool's benefits while waiting for a fix that isn't coming.

Question 7 — Correct: C

Why C is correct. Restating and exploring a significant symptom aloud is simultaneously excellent communication (the patient hears that they were heard) and the mechanism by which a summarization system learns what matters — its salience follows the spoken emphasis of the encounter. One behavior serves the relationship and the record, which is why it beats every option that trades one for the other.

Why the others are wrong. A recovers accuracy by re-purchasing the attention problem the scribe solved. B fails on mechanism: with no contemporaneous anchor, her post-hoc review checks the draft against a memory biased toward the conversational moments — the very content the scribe already captures best. D shrinks the tool's benefit instead of fixing the workflow, and complexity is not always visible at scheduling. E transfers the clinician's job to the patient.

Question 8 — Correct: C

Why C is correct. Shared decision-making asks the physician to bring clinical judgment, the patient to bring values, and any recommendation — human or algorithmic — to be weighed on its merits. Here the AI's advice is reasonable and guideline-concordant; the physician's task is to explore what “doing this right” means to her and integrate honestly, which may well mean deferring the injection. The item's point: balanced integration, not reflexive dismissal or reflexive deference.

Why the others are wrong. A rejects content because of its source — and in this case, rejects a defensible plan. B abandons clinical assessment entirely; concordant-in-general is not the same as right-for-her. D outsources a primary-care conversation. E dismisses the patient's explicit invitation to deliberate, which is the shared-decision-making failure the objective names.