AI in Clinical Skills and Assessment

Problem Based Learning Cases

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

Two problem-based learning cases, restructured so that learners actively use AI tools and appraise real AI output rather than critiquing a character who used them. Case 1 places learners in an outpatient general-medicine clinic: they commit to their own clinical reasoning first, then generate and appraise an AI differential. Case 2 places learners in an urgent care setting: they review verbatim encounter transcripts against AI-generated draft notes, find what the scribe missed, altered, and invented, and produce a corrected note they would be willing to sign.

Every technical term is defined in the concept boxes in Section C; no prior AI coursework is assumed. Section D is a faculty-preparation one-pager on the operational mechanics of ambient AI scribes, for facilitators who have not used one. Definitions and case mechanics are referenced to the literature listed at the end of the document; per consortium policy, every citation requires verification before publication.

Running the cases: two implementation tiers

- **Tier A (live AI).** Groups use an institutionally sanctioned, secure clinical AI tool (a sandboxed LLM or CDS interface) to generate the outputs in real time, using the standardized prompts provided. Tier A adds a powerful input-quality demonstration in Case 1, Part 3.
- **Tier B (no technology).** Groups use the pre-generated AI outputs printed in this document. The appraisal work — the heart of both cases — is identical. Pre-generated outputs were deliberately written to contain the error types learners must find; Tier A facilitators should review the live output and, if it is cleaner than the printed version, run the appraisal against the printed exhibit as well.

Timing: each case runs in 30–35 minutes with groups of 5–7. Programs running a single discussion should pick one case; both together need 60–70 minutes.

Case 1 — “The Differential, Twice”

Objectives 1 and 2, and the self-reflection objective • Groups of 5–7 • 30–35 minutes • Outpatient general medicine

The setting

You are an M2 on a longitudinal preceptorship in a community general-medicine clinic. Your preceptor, Dr. Osei, asks you to see the next patient and present your assessment. The clinic's EHR includes a sanctioned clinical decision support (CDS) AI that can generate differentials and suggested workups from entered clinical data. Dr. Osei tells you: “Use whatever you like — but I want to hear your reasoning.”

The patient

Jordan Reyes is 24 years old, presenting with a sore throat for three days. You take the history and examine the patient. Here is what you find:

History. Sore throat ×3 days, worse on swallowing. Subjective fevers at home; measured 38.3 °C in clinic. No cough. No rhinorrhea. Fatigue “more than a cold usually gives me.” No sick contacts at work; roommate had “a cold” two weeks ago. Not vaccinated against influenza this season; childhood vaccinations complete. A new sexual partner within the past two months; on further nonjudgmental questioning, reports receptive oral sex; inconsistent barrier use. No injection drug use. No medications, no allergies.

Examination. Temp 38.3 °C, HR 92, BP 118/74. Erythematous posterior pharynx with tonsillar exudates. Tender anterior cervical lymphadenopathy; posterior cervical chain also mildly tender. No trismus, no muffled voice, no drooling; uvula midline. Abdomen soft; spleen not palpable. No rash.

Part 1 — Your reasoning first (10 minutes)

Working as a group, and before any AI tool is opened, commit the following to paper:

- A one-sentence problem representation (age, key features, tempo).
- A differential diagnosis organized into three bins: most likely, less likely, and do-not-miss.
- For each diagnosis: the piece of history or exam that raises or lowers it.
- Your initial diagnostic plan (tests, and what each would change).

Write it down. The comparison in Part 3 only works against a committed answer — a differential you would have defended before seeing the AI's.

Part 2 — Now ask the AI (5 minutes)

Tier A: enter the case into the sanctioned tool using this standardized prompt: “24-year-old with 3 days of sore throat, fever 38.3, tonsillar exudates, tender anterior and posterior cervical nodes, fatigue, no cough. New sexual partner, receptive oral sex, inconsistent barrier use. Generate a prioritized differential diagnosis with recommended initial workup.”

Tier B: use the printed output below, generated from the same prompt.

AI OUTPUT — EXHIBIT 1 (appraise; do not assume it is correct)

- Streptococcal pharyngitis — most likely given exudates, fever, adenopathy, absent cough. Recommend rapid antigen test; begin empiric penicillin while awaiting results given high clinical probability.
- Infectious mononucleosis — posterior adenopathy and fatigue are suggestive. Recommend heterophile antibody testing.
- Viral pharyngitis — common, but exudates and absent cough argue against.
- Gonococcal pharyngitis — indicated by sexual history. Recommend pharyngeal NAAT.
- Acute HIV seroconversion — sore throat, fever, and fatigue after a new exposure warrant consideration. Recommend 4th-generation antigen/antibody testing.

Summary: presentation is most consistent with streptococcal pharyngitis; empiric treatment is reasonable pending testing.

Part 3 — Appraise the output (10–12 minutes)

Work through the appraisal worksheet as a group. The facilitator will mediate; the point is not to catch the AI being wrong (much of this output is reasonable) but to practice the habit of verification before integration.

Appraisal question	Your group's answer
1. Coverage. What did the AI include that your group missed? What did your group include that the AI missed?	
2. Verification. Pick the two claims that would most change management. How exactly would you verify each — against what source (guideline, primary literature, the patient)?	

Appraisal question	Your group's answer
3. Guideline check. The output recommends empiric antibiotics before test results. Is that guideline-concordant? How would you find out?	
4. Do-not-miss scan. Does the output address the red flags you should actively exclude in severe sore throat (e.g., signs of peritonsillar abscess or airway compromise)? What does its silence teach you?	
5. Input quality. Which single line of the prompt, if omitted, would most change the output — and what does that imply about whose differentials get impoverished when history-taking is rushed?	

Tier A extension (input quality, live). Re-run the prompt with the sexual history deleted. Compare the two differentials. This is the fastest demonstration in the curriculum of why the quality of AI output is bounded by the quality of the history that feeds it.

Part 4 — The skipper in the mirror (5 minutes)

You now know the tool produces, in seconds, a differential resembling the one your group built in ten minutes. On a real clerkship — tired, behind, and graded — the temptation to reverse the order will be real, and it will be yours, not a hypothetical classmate's.

- Name the situation in which you, personally, would be most tempted to ask the AI first (end of a long shift? a complaint that feels routine? a presentation you fear getting wrong?).
- Name your tell — the observable sign that you have started outsourcing the reasoning rather than checking it.
- Write one guardrail habit, in one sentence, that you could actually keep. Keep it; the reflective-writing activity in this session builds on it.

FACILITATOR GUIDE — CASE 1

Educational purpose. This is the curriculum's primary case for independent-reasoning-first workflow and for the deskilling / mis-skilling / never-skilling vocabulary (defined in Section C, with citations). The restructure moves the learning from evaluating a fictional student's behavior to performing and appraising the workflow directly — learners generate a committed differential, then integrate, question, or override real AI output with justification.

Part 1 model answers. Problem representation: a young adult with acute exudative pharyngitis, systemic symptoms, and a relevant sexual exposure history. Most likely: streptococcal and viral pharyngitis; mononucleosis (posterior nodes, fatigue, roommate contact). Less likely / exposure-driven: gonococcal pharyngitis. Do-not-miss: acute HIV seroconversion; peritonsillar abscess and airway red flags (explicitly excluded on exam — expect groups to notice the exam was designed to let them rule these out). Plan: rapid strep antigen ± culture, heterophile testing, pharyngeal gonococcal NAAT, 4th-generation HIV antigen/antibody test; counseling around testing.

Part 3 mediation, question by question. (1) Groups typically miss gonococcal pharyngitis or acute HIV when they rushed the sexual history; the AI typically “misses” nothing listed but ranks without transparent probabilities. (2) Push for concrete verification acts: the antibiotic-timing claim against a current pharyngitis guideline; the HIV-testing window against a reference or ID guidance — not “I'd double-check it.” (3) The empiric-antibiotics-before-results recommendation is the planted discordance: guidelines generally

recommend testing-based treatment of adults. The lesson is that fluent, confident output can embed a non-concordant recommendation without flagging it. (4) The output is silent on red flags — AI omissions are as clinically important as AI statements, and silence is not reassurance. (5) The sexual-history line. Without it, gonococcal pharyngitis and acute HIV disappear — which converts the abstraction "input quality determines output reliability" into a concrete, memorable case behavior.

Health-equity teaching moment (acute HIV / gonococcal pharyngitis). The differential the AI returns does not depend on who the clinician assumes is "at risk" — it lists exposure-driven diagnoses whenever the exposure history is present. Clinicians, by contrast, demonstrably under-elicit sexual histories and under-test some populations while over-attributing risk to others. Frame the AI here as a potential equity asset: it can counter premature closure and selective risk attribution — but only for the learner who (a) gathered the nonjudgmental history that feeds it and (b) has the reasoning habits to evaluate and act on what it returns. A tool cannot correct for a history that was never taken.

Reframing never-skilling and error attribution. Introduce the three-term vocabulary (Section C) developmentally, not morally. Never-skilling is not a character flaw in lazy students; it is the predictable result of placing a fluent tool upstream of an unbuilt skill, and its defining danger is invisibility — the affected learner has no felt sense of the missing capability. Likewise, when AI-assisted care goes wrong, professional and legal attribution lands on the clinician who acted, not the tool (see Section C, automation bias; Section D on signed notes) — so the guardrails discussion in Part 4 is about habit design and system design, not willpower. Anticipate and validate the honest answer "I will absolutely be tempted"; the goal is a tell and a guardrail, not a promise of abstinence. The curriculum's through-line applies: stewardship, not abstinence — the AI is a tool, not a team member.

Common pitfalls. (1) Groups turning Part 3 into a medicine seminar on pharyngitis — redirect to the appraisal; the medicine is deliberately simple. (2) Groups concluding "never use the AI" — reject this explicitly; the demonstrated workflow (commit, consult, compare, verify, decide) is the intended takeaway. (3) In Part 4, students writing rules for other people. Redirect to first person.

Debrief prompt. "What did comparing your differential against the AI's teach you about your own reasoning that neither alone would have?"

Case 2 — "The Note You Sign"

Objectives 1, 3, and 4 • Groups of 5–7 • 30–35 minutes • Urgent care

The setting

Friday evening in a hospital-affiliated urgent care clinic. Dr. Lin is the supervising physician; you are the student working with her. The clinic recently deployed a sanctioned ambient AI scribe: with the patient's verbal agreement, a room microphone captures the visit and an AI drafts the note in near-real time for the clinician to review and sign. Tonight you will do what you will be asked to do on clerkships within two years: check the AI's draft against what was actually said, correct it, and decide what you would be willing to sign.

The patient

Mrs. Alvarez is 62. She speaks Spanish as her first language and conversational English, and declined an interpreter tonight ("my English is okay"). She has come in because she has felt "run down" for about a month and her daughter insisted. She mentions a recent normal mammogram. Dr. Lin obtained verbal agreement for the scribe at the start of the visit.

Exhibit 1 — Verbatim transcript (excerpt, history)

Dr. Lin: So tell me about this tiredness. When did it start?

Mrs. Alvarez: Maybe one month? I am tired all the day. My husband says I look tired all the time. I say it's the age, but…

Dr. Lin: Is it worse at any particular time of day?

Mrs. Alvarez: The nights are not good. Some nights I am sleeping in my chair, in the living room. Because in the bed, when I lie down flat, I get — how do you say — like I cannot get the air. So in the chair, sitting, is better for me.

Dr. Lin: Mm-hm. And your appetite?

Mrs. Alvarez: Not so much hungry. My shoes also are tight in the evening, my ankles. From standing in the kitchen, I think.

Dr. Lin: Okay. Any fevers, chills, night sweats?

Mrs. Alvarez: No, no fever. I take my pill for the pressure — the little white one — every morning, that one I never forget.

Dr. Lin: Good. Let's take a look at you.

Exhibit 2 — The AI scribe's draft note (same excerpt)

HPI: 62-year-old female presents with approximately one month of generalized fatigue. Patient reports feeling tired throughout the day, confirmed by family members. Symptoms attributed by patient to aging. Sleep is fair; patient sometimes sleeps in a recliner by preference. Appetite is mildly decreased. Patient denies fever, chills, or night sweats. She denies chest pain or palpitations.

ROS: Constitutional: fatigue, decreased appetite. Cardiovascular: negative. Respiratory: negative for shortness of breath or cough. All other systems reviewed and negative.

Medications: Hydrochlorothiazide 25 mg daily for hypertension. Adherent.

Allergies: No known drug allergies.

Task 1 — Line-by-line audit (10 minutes)

In pairs, compare Exhibit 2 against Exhibit 1. Sort every claim in the draft note into four categories, citing the transcript line that supports your classification:

- Accurate — supported by the transcript.
- Omitted — said by the patient, absent from the note. Star any omission that is clinically significant.
- Fabricated — asserted in the note but never said by anyone. (AI summarization systems generate fluent text; when the conversation has gaps, they can fill them with plausible content. This is the documentation form of hallucination — see Section C.)
- Transformed — based on something said, but altered in meaning, certainty, or specificity.

Exhibit 3 — Verbatim transcript (excerpt, close of visit)

Dr. Lin: Your exam is mostly reassuring. I want to get some blood tests and I'd like you to come back next week.

Mrs. Alvarez: So it is nothing serious? Because my daughter, she worries.

Dr. Lin: The tests will tell us more. The front desk will set up your follow-up.

Mrs. Alvarez: Okay. And the tiredness — I can take vitamins?

Dr. Lin: Vitamins won't hurt. Let's see the labs first.

Exhibit 4 — The AI scribe's visit summary (patient instructions)

Assessment: Fatigue, likely multifactorial; age-related deconditioning and mild anemia suspected. Plan: CBC, CMP, TSH ordered. Patient reassured that no serious cause is suspected. Patient advised to begin a daily multivitamin. Return to clinic in one week; return precautions reviewed.

Task 2 — The do-not-miss hunt and the corrected note (10 minutes)

- As a group: what is the single most clinically significant piece of information in the two transcripts, and where did it go in the AI's documentation? Why might a summarization system rank it as low-salience when a clinician would not?
- Audit Exhibit 4 the same way you audited Exhibit 2. Which sentences were never said? Which could change what the patient does tonight — and which could mislead the covering clinician who reads this note this weekend?
- Draft the corrections: (a) the HPI and ROS sentences you would rewrite, (b) the two questions you would go back into the room and ask, and (c) the one sentence you would say to Mrs. Alvarez before she leaves. You have five minutes — the length of time you will actually have in clinic.

Task 3 — Your workflow (5 minutes)

Individually, write the note-validation workflow you will actually use when a scribe drafts under your name — at least three steps, fast enough to survive a busy clinic. Compare with a partner; steal the best step of theirs.

FACILITATOR GUIDE — CASE 2

Educational purpose. Anchors the objectives on scribe function and limitations, note review/correction/validation with responsibility for errors, omissions, and tone, and the influence of AI documentation on attention and patient-centered communication. The transcript-versus-summary format exists so learners practice the actual clerkship skill: auditing an AI draft against ground truth, then owning the corrected product.

Task 1 model audit (Exhibit 2 vs. Exhibit 1). Accurate: one-month fatigue; family corroboration; decreased appetite; no fever/chills/night sweats; antihypertensive adherence. Omitted (starred): orthopnea — “when I lie down flat, I get like I cannot get the air” became “sleeps in a recliner by preference,” and evening ankle swelling vanished entirely; together these are the case's do-not-miss thread (possible heart failure) and the reason the visit's disposition could be wrong. Fabricated: “denies chest pain or palpitations” (never asked), “all other systems reviewed and negative” (no ROS was taken), “Hydrochlorothiazide 25 mg” (the patient said “the little white pill for the pressure” — the scribe resolved ambiguity by inventing specificity), “No known drug allergies” (never discussed — the classic downstream hazard for a covering prescriber). Transformed: “sleeps in a recliner by preference” (meaning inverted: symptom → preference); “Cardiovascular: negative” and “Respiratory: negative for shortness of breath” (her positional dyspnea, described in conversational language, was normalized to a negative).

Task 2 model answers. (1) The orthopnea-plus-edema thread. Summarization models rank salience by pattern-typicality of language, not clinical meaning; symptoms described idiomatically (“I sleep in my chair,” “my shoes are tight”), in second-language phrasing, or never verbally flagged by the physician are systematically at risk of being dropped or normalized. Note the physician-behavior half of this: Dr. Lin never said “that sounds like orthopnea — tell me more,” so the system had no salience signal (see Section D). (2) Exhibit 4 fabrications: “mild anemia suspected” (never said), “reassured that no serious cause is suspected” (she was told the opposite: “the tests will tell us more” — tone transformed into false reassurance), “advised to begin a daily multivitamin” (“vitamins won't hurt” transformed into a recommendation), “return precautions reviewed” (they were not — the most dangerous line for tonight, since she actually needs precautions about worsening breathlessness). (3) Strong corrections re-open the

ROS, ask directly about positional dyspnea and swelling, and say one honest sentence to the patient about why the labs matter and when to seek care tonight.

Health-equity teaching moment (Mrs. Alvarez's ESL status). Speech-recognition and summarization accuracy are not uniform across accents, dialects, and second-language phrasing; error rates are documented to be higher for some speaker groups (see references — Koenecke et al.). Mrs. Alvarez's symptoms arrived in exactly the linguistic register the system handles worst, and she had declined an interpreter — a common real-world configuration. Balance the framing: documentation AI can also be an equity asset — it can free clinician attention for patients who need more communication work, and some tools draft multilingual patient instructions — but the benefit accrues only when the clinician has the verification habits to catch what the tool does to non-standard speech. The corrected note, not the draft, is where equity is protected.

Reframing never-skilling and error attribution. If learners lean on "the scribe made the error," redirect: the signed note is the clinician's professional statement, whoever drafted it — responsibility does not transfer to the tool (Section C; Section D). If learners lean the other way ("never use scribes"), redirect equally: the documented benefits (documentation burden, attention returned to the patient) are real; the skill being trained is validation, not avoidance. Frame skill erosion developmentally: a clinician who stops composing summaries loses the summarization-as-understanding step; a learner who never composes them never builds it. The mitigation is a workflow, not vigilance-by-willpower.

Common pitfalls. (1) Groups diving into the heart-failure workup — the medicine stays simple by design; redirect to the documentation workflow. (2) Groups treating fabrication as malfunction — it is expected behavior of fluent summarization systems, which is precisely why review-before-signature is the standard. (3) Groups missing the tone transformations (false reassurance) — prompt with: "Read Exhibit 4 as the daughter. What does she now believe?"

Debrief prompt. "What is one sentence you will say out loud in every scribe-recorded visit, knowing that what you flag verbally shapes what the note remembers?"

Section C — Concepts used in these cases

Plain-language definitions, written for learners with no prior AI coursework. Citations support the definitions; all require verification before publication per consortium policy.

CLINICAL DECISION SUPPORT (CDS)

Software that takes clinical information as input and returns diagnostic or management suggestions — differentials, risk scores, alerts, recommended workups. Some CDS is rule-based; newer tools use machine learning or large language models. CDS output is an input to clinical judgment, never a verdict: its quality is bounded by the quality of the data entered and of the data it was trained on (Sutton et al., 2020; Cabitza et al., 2017).

AMBIENT LISTENING / AI SCRIBE

A system that records the audio of a clinical encounter, transcribes it, and uses a language model to draft a clinical note or visit summary for clinician review. Documented benefits include reduced documentation burden and more face-to-face attention (Tierney et al., 2024); documented limitations include omission of conversationally phrased findings, fabrication, and uneven accuracy across accents and speech patterns (Koenecke et al., 2020). Section D details the mechanics.

AI-POWERED DOCUMENTATION AIDS

The broader family of drafting tools inside the EHR: autocomplete and smart phrases, AI-drafted replies to patient messages, and chart summarizers that condense prior records. Distinct from scribes in that they

work from text already in the record — which means they can also propagate and amplify existing chart errors (copy-forward risk).

AUTOMATED SKILL-TRACKING PLATFORMS

Systems that log and analyze learner performance data — procedural counts, checklist scores, sometimes video — and flag patterns for review. Learners should always ask: what is captured, who sees it, and what decisions does it feed?

DESKILLING, MIS-SKILLING, NEVER-SKILLING

Deskilling: a clinician who had a skill loses it through disuse after a tool absorbs the task. An observational colonoscopy study found lower unassisted detection after AI exposure (Budzyń et al., 2025); causation and generalization to other skills remain uncertain (concept: Cabitza et al., 2017).

Mis-skilling: a skill is learned, but learned wrong, because the tool shaped how it was practiced.

Never-skilling: a learner never develops a foundational skill because the tool performed the cognitive work before the learner ever did. It is the variant specific to trainees, and the hardest to detect — the affected clinician has no memory of the skill they lack. These curricula therefore sequence independent reasoning before AI consultation, so the tool verifies and extends reasoning rather than replacing its development.

HALLUCINATION / FABRICATION

Confidently stated false content produced by a generative model — invented findings, citations, doses, or statements never made. It is a predictable property of systems that compose plausible text rather than retrieve verified facts, which is why generative output requires verification against primary sources, and AI-drafted notes require verification against the actual encounter. Distinguish this generative-AI failure mode from machine-learning/CDS failure modes (algorithmic bias, drift, validation gaps) — different mechanisms, different safeguards.

AUTOMATION BIAS

The human tendency to over-trust automated output: accepting a machine suggestion in place of one's own assessment (commission), or missing what the machine missed because silence read as reassurance (omission) (Goddard et al., 2012). The standing defense in both cases: form your own impression first, then compare.

Section D — Faculty one-pager: how ambient AI scribes actually work

A preparation sheet for facilitators who have not personally used an ambient scribe. Learners will ask operational questions; this page contains the answers that matter.

The pipeline

Microphone → automatic speech recognition (audio to text) → a language model condenses the transcript into a note in the clinic's template → clinician reviews, edits, and signs in the EHR. Each stage can introduce error: mishearing at transcription; omission, transformation, and fabrication at summarization. The draft usually arrives within minutes of the visit ending.

The spectrum of tools — and why it matters

- Institutionally sanctioned, EHR-integrated systems. Covered by institutional agreements and HIPAA business-associate terms; audio handling, retention, and data use are contractually specified. This is the deployment your learners will most often see on clerkships.

- Third-party applications used by individual physicians. Privacy and data-ownership terms are whatever the app's consumer agreement says; institutional oversight may be absent. Learners should know these exist and that institutional policy governs whether they may be used.
- **Patient-facing consumer apps.** Increasingly, patients arrive recording or summarizing the visit with their own AI tools. Consent, privacy, and data-ownership implications differ at every point of this spectrum, and faculty should expect questions about each — including “can the patient record me?”, which is governed by state recording law and institutional policy, and which raises the point below.

Facts to have ready

- **Accuracy is not uniform across patients.** Speech recognition performs measurably worse for some accents, dialects, and second-language speech patterns (Koenecke et al., 2020), and summarization handles conversational, idiomatic symptom descriptions worse than clinical language. Performance claims averaged over a population say little about the encounter in front of you.
- Accuracy also depends on the physician's verbal behavior. Summarization systems take salience cues from what is said aloud. A clinically significant finding that the physician never verbally flags or follows up — as in Dr. Lin's encounter with Mrs. Alvarez — may not register as high-salience to the scribe, regardless of tool quality. Teaching implication: thinking aloud (“that sounds like orthopnea — say more”) is now both good communication and good documentation practice.
- **Responsibility does not move.** Physicians retain full legal and professional responsibility for the accuracy of any note they sign, regardless of how it was generated. “The scribe wrote it” has no standing — which is why review-before-signature is the skill this session trains.
- **Consent practices vary.** There is currently no universal standard for whether and how patients are asked before ambient recording; institutions range from signage to scripted verbal agreement to written consent. Teach learners to know and follow their site's practice, and to answer patient questions honestly.
- Patient privacy and physician privacy are distinct. Students often conflate them. The patient's concern is their health information: where audio goes, how long it is kept, who can access it, whether it trains future models. The physician's (and student's) concern is their own recorded speech: an ambient system also records the clinician, and that record may be discoverable or reviewed. Be prepared to address the two separately.

References

The following readings support the session. Distinguish empirical findings from conceptual frameworks, and check current local policy when applying guidance.

- **Cabitza F, Rasoini R, Gensini GF.** Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517–518. (Deskilling and over-reliance failure modes; conceptual anchor for the three-term vocabulary.)
- **Sutton RT, Pincock D, Baumgart DC, et al.** An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digital Medicine. 2020;3:17. (CDS definition, benefits, and failure modes.)
- **Tierney AA, Gayre G, Hoberman B, et al.** Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst. 2024. (Implementation evidence for scribe benefits; grounds the balanced framing in Case 2.)
- **Koenecke A, Nam A, Lake E, et al.** Racial disparities in automated speech recognition. PNAS. 2020;117(14):7684–7689. (Documented accuracy variation across speaker groups; grounds the equity moment in Case 2.)
- **Goh E, Gallo R, Hom J, et al.** Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Network Open. 2024;7(10):e2440969. (Independent-reasoning-first workflow; clinician–AI integration is a skill, not a default.)

- **Goddard K, Roudsari A, Wyatt JC.** Automation bias: a systematic review of frequency, effect mediators, and mitigators. JAMIA. 2012;19(1):121–127. (Automation bias definition and mitigation.)
- **Budzyń K, Romańczyk M, Kitala D, et al.** Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. Lancet Gastroenterol Hepatol. 2025;10:896–903. doi:10.1016/S2468-1253(25)00133-5. — Lower unassisted adenoma detection was observed after AI exposure; the observational design does not establish that AI caused the decline or that this generalizes to diagnostic reasoning.
- **Obermeyer Z, Powers B, Vogeli C, Mullainathan S.** Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453. (Algorithmic bias mechanism, for the CDS pitfalls discussion.)
- **Biro J, Handley JL, Cobb NK, et al.** Accuracy and safety of AI-enabled scribe technology: instrument validation study. J Med Internet Res. 2025;27:e64993. doi:10.2196/64993. — Simulated encounter study identifying omissions, additions, and wrong outputs. For deployed outpatient experience, see Memon S, Brand A, Taylor B, et al. Performance, acceptability, and impact of ambient listening scribe technology in an outpatient context: a mixed methods trial evaluation. BMC Health Serv Res. 2026;26:182. doi:10.1186/s12913-025-13954-5. Staff-reported errors are distinct from independently measured note-error rates.