News and Perspectives

When the Teaching Material Is the Risk: The Governance Gap for AI-Generated Content in Medical Education

Boon-How Chew, JMIR Correspondent

Abstract

AI-generated content is rapidly being incorporated into medical education, with much of the onus for responsible use being placed on trainees and physicians. In this *News and Perspectives* op-ed, JMIR Correspondent and academic physician Boon-How Chew reports on this governance gap and outlines recommendations for greater institutional oversight into curriculum development.

Key Takeaways:

- Three types of AI-generated teaching content (representational, reasoning, and assessment) are entering medical curricula with little scrutiny because plausible content can conceal errors.
- Institutions must provide the accountability that generative tools and agents lack: learners cannot readily evaluate content they have not yet learned.
- Governance requires knowing which content is used and testing it (or the system generating it) before teaching, acting on the findings, and monitoring tool performance. Institutions differ in their capacity to do this.

Boon-How Chew, MD, MMed, PhD, is an academic physician and professor of family medicine at Universiti Putra Malaysia. His research increasingly explores the intersection of AI innovation in medicine, health care professionals education and training, and the reform of scientific writing workflows within legacy academic infrastructures.

Research in medical education is increasingly focused on whether learners can use AI safely and productively and less on whether increasingly AI-generated teaching materials—clinical images, patient vignettes, examination items—can be trusted. While AI users are admonished to critically evaluate and answer for its output, this principle has been applied more to learners and clinicians than to curriculum production.

AI agents are beginning to tutor, simulate patients, and give feedback as learners work, generating content no one can review in advance. Closing the first United Nations Global Dialogue on AI Governance in Geneva in July 2026, it was highlighted that the challenge is no longer a lack of AI principles but a lack of practical mechanisms to implement them. This is particularly relevant in medical education, where mechanisms for checking institutional teaching content barely exist.

Teaching Content and Governance

Teaching content expresses a curriculum's epistemology—its assumptions about what counts as knowledge, how that knowledge is acquired and justified, and where its limits lie. In medicine, this combines propositional biomedical science with experiential, largely tacit knowledge of disease in particular patients and circumstances. The latter develops through repeated, effortful reasoning about many and varied clinical encounters. Learners generally trust teaching content because an accountable person or institution is responsible for it.

Characteristic failures can, however, crop up in different types of teaching content:

1. *For representational content*—like images, diagrams, or slide summaries—the characteristic failure is a plausible but incorrect or skewed depiction that requires checking against a reference standard.
2. *For reasoning content*—like cases, vignettes, simulated patients, or worked differentials—the characteristic failure is a fluently expressed but incorrect conclusion. Checking it requires someone able to reach the conclusion unaided, a competence that novices have yet to acquire.
3. *For assessment content*—like test items, answer keys, rubrics, scores, or feedback—the characteristic failure is an apparently sound instrument that does not measure the intended knowledge. Checking it requires evidence of what it actually measures.

Two of these failure modes are already documented. In medical education, AI-generated images commonly carry demographic bias and clinical inaccuracies, yet rarely appear incorrect; without disclosure, they receive the trust accorded to textbook illustrations without earning it. AI-generated questions can perform like human-written ones and pass unnoticed by reviewers, but this does not establish that they test the intended knowledge, a problem known as synthetic validity. AI grading can agree closely with human markers on one question and diverge on another. Generated items and

AI grading therefore require continuing checks beyond initial approval.

Human-authored content has always contained errors, but a named author was always accountable. Generative tools cannot assume authorship accountability, and learners are poorly placed to evaluate unfamiliar content. Accountability and governance therefore rest with the institution deploying it and its capacity to (1) know what content is in use and where it came from; (2) test it, or the system generating it, before learners see it; (3) act on the findings, including withdrawal; and (4) monitor approved tools for performance drift.

Two Points of Entry, Different Consequences

Medical education uses AI in two ways: learners consult it, and institutions use it to produce teaching content and assessment materials. Governance has focused almost entirely on learner use.

Institutional use remains largely ungoverned, treated as production efficiency. Educators may check what they adopt, but these checks are often personal and unrecorded. Expert review remains essential but insufficient: an expert can confirm every item is correct yet miss bias across the set or items that measure the wrong thing. AI agents blur this distinction: learners' agents compile their own study materials, and platform agents can generate a case, run the encounter, and score it, then choose what the learner sees next.

The consequences differ. Learner misuse produces an incorrect answer that a supervisor may identify and correct. A clinical error eventually produces a signal through a patient, complaint, or audit. Misuse in curriculum production exposes every learner to flawed material presented as established knowledge and may not produce a signal for years or ever.

What Governance Actually Requires

Responsible adoption of AI-generated content in medical education requires clinical subject expertise alongside psychometrics, bias, privacy, and monitoring expertise. This also requires significant resources, which risks making governance capacity a new source of inequality—schools with the most to gain from AI-generated content are likely the least equipped to audit it.

AI-generated or substantially AI-assisted content should be logged at creation and disclosed wherever used. Before deployment, it should be checked by a subject-matter expert and audited for whether it represents the patients learners will meet, across skin tones, ages, sexes, and local disease patterns. For AI agents that generate content during use, the check should move to the system: its sources, instructions, and permitted actions, with interactions logged for sampled audit. Where AI assists content review, the institution should know which model checked what, keep the reviewing model independent of the content's supplier, and have a sample of its conclusions checked by people qualified to overturn them.

Authority to pause or withdraw content or AI agents should be assigned to named roles, such as a module coordinator or assessment committee, within existing curriculum review cycles: full review for high-stakes and widely reused materials, sampling for the rest. AI agents should work within set bounds, with grades and progression decisions finalized by a named person. Schools lacking sufficient staff could form joint review panels and share audited item banks regionally. These measures depend on critical appraisal of AI output, a proposed core AI competency that educators choosing teaching content need as much as learners.

Follow the Content to Its Source

AI-generated teaching content can fill curriculum gaps faster than human production. The concern is that errors can appear credible, content enters curricula without governance, and review is least likely where the need is greatest.

> *AI capabilities will change; the need for an accountable person and responsive governance will not.*

Governance must extend from learners' AI use to the people or tools that generate, select, or judge teaching content. The increasing adoption of AI teaching content calls for careful implementation, starting with selected cohorts, rather than further delay. AI capabilities will change; the need for an accountable person and responsive governance will not.

Keywords: artificial intelligence; generative artificial intelligence; medical education; curriculum; educational measurement; clinical decision support systems; patient safety; institutional policy

Please cite as:
Chew BH

When the Teaching Material Is the Risk: The Governance Gap for AI-Generated Content in Medical Education
J Med Internet Res 2026;28:e115991
URL: https://www.jmir.org/2026/1/e115991
doi: 10.2196/115991

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 10.Nov.2026