Teaching Radiology Trainees to Recognize AI Failure Modes: A Six-Domain Educational Framework

Colin Wu[1], Edward Pettyjohn, MD[2], Youyou Cheng[1], Ravi Patel[1], Shreyas Meruga[3], Zhuangwei Kang, PhD[4], Sarah Pettyjohn, MD[5]

[1] University of the Incarnate Word School of Osteopathic Medicine [2] University of Illinois College of Medicine Peoria (UICOMP) [3] UT Health San Antonio, Department of Molecular Medicine [4] Independent Researcher [5] Rutgers Robert Wood Johnson Medical School

Introduction

Radiology trainees now routinely encounter artificial intelligence (AI) tools, yet residency programs provide no formal instruction on how to understand or analyze AI errors. Traditional teaching formats, such as morbidity and mortality (M&M) conferences, focus on human and procedural mistakes and offer little to no framework for algorithmic failures. As AI becomes embedded in workflows, learners need explicit training in how and why these systems fail. This project developed an M&M-inspired educational model to teach trainees a systematic method for identifying and categorizing AI errors.

Methods

We synthesized literature from AI failure science, human factors research, radiology quality-improvement pedagogy, and real-world algorithmic errors. These sources informed the design of a six-domain educational taxonomy and a case-based instructional format modeled on M&M discussions. The framework provides trainees with conceptual scaffolding to analyze AI failures, describe underlying causes, and understand interactions between data, labels, models, workflow, integration, and human interpretation.

Results

Results

The framework introduces six domains of AI error understanding for trainees:

1. Data failures such as image quality issues or drift
2. Label failures such as noisy or outdated ground truth
3. Model failures such as miscalibration or brittle thresholds
4. Integration failures involving PACS or routing problems
5. Workflow failures involving timing, preprocessing, or task design
6. Human-AI interaction failures involving automation bias or overreliance

These domains support a teachable curriculum with case prompts, guided discussion questions, and error-classification exercises modeled on M&M practices.

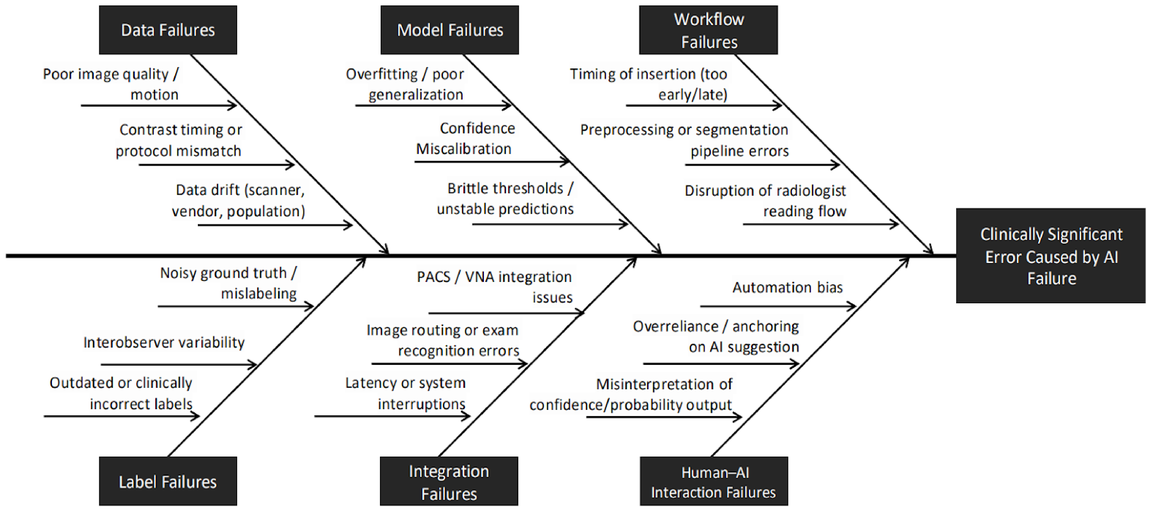

Figure 1. *Fishbone (Ishikawa) diagram illustrating a six-domain educational framework for analyzing radiology AI errors. The model introduces trainees to the major categories of algorithmic failure: data failures, label failures, model failures, integration failures, workflow failures, and human–AI interaction failures. Each domain includes representative examples that help learners recognize how different failures arise and how they influence AI performance. This diagram serves as a conceptual scaffold for teaching residents a structured approach to AI error analysis during case-based or M&M-style educational sessions.*

Conclusion

Trainees lack formal methods for analyzing AI errors, creating a critical gap in AI literacy. This M&M-style educational framework offers a structured approach for teaching learners how to recognize, classify, and interpret algorithmic failures. Integrating this model into residency teaching may improve conceptual understanding, support safer AI use, and prepare future radiologists for responsible oversight of clinical AI systems.

References

- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56.
- Oakden-Rayner L. Exploring large-scale pKelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17(1):195.
- ublic medical image datasets. *Acad Radiol.* 2020;27(1):106–112.
- Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol.* 2020;30(6):3576–3584.
- Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. *Radiology.* 2019;293(2):436–440.
- Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med.* 2020;3:41.
- Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak.* 2020;20(1):310.
- Graber ML, Kissam S, Payne VL, et al. Cognitive interventions to reduce diagnostic error: a narrative review. *BMJ Qual Saf.* 2012;21(7):535–557.