From ROC to Report: A Framework for Safely Communicating AI-Derived Probability and Uncertainty in Radiology

Colin Wu[1], Ravi Patel[1], Youyou Cheng[1], Sina Ashraf[1], Justin Kuyn[2], Edward Pettyjohn, MD[3], Shreyas Meruga[4], Zhuangwei Kang, PhD[5], Sarah Pettyjohn, MD[6]

[1] University of the Incarnate Word School of Osteopathic Medicine; [2] University of Colorado Boulder; [3] University of Illinois College of Medicine Peoria (UICOMP); [4] UT Health San Antonio, Department of Molecular Medicine; [5] Independent Researcher; [6] Rutgers Robert Wood Johnson Medical School

Introduction

AI systems increasingly provide numerical probabilities (for example, “87% malignant”), yet there is no standardized guidance on how radiologists should communicate these values in reports. Exact percentages may imply unwarranted certainty, misrepresent model calibration, and increase medicolegal risk. This project aimed to develop a practical framework for translating AI-derived risk, confidence, and uncertainty into radiology report language.

Methods

We conducted a practice-focused analysis integrating concepts from AI calibration science, uncertainty quantification, radiology reporting conventions, numerical cognition, and medicolegal guidance. Known limitations of probabilistic AI output were synthesized to develop a structured lexicon and communication strategy suitable for clinical imaging practice.

Results

Five findings informed the framework.

1. Calibration varies across anatomy, patient factors, and imaging conditions; even well-calibrated models may produce unreliable case-level probabilities.
2. Exact percentages are frequently over-interpreted by referring clinicians due to numerical anchoring.
3. Model reliability changes across scanners, protocols, and postoperative or treated anatomy, making single numeric predictions misleading.
4. Precise values may increase medicolegal exposure by implying validated accuracy and radiologist endorsement.
5. Exact numbers may overshadow radiologist expertise and contextual interpretation.

Using these insights, we developed a communication framework that includes:

- Categorical probability terms (low, intermediate, moderate-to-high) instead of exact percentages.
- Statements describing model limitations, including uncalibrated or out-of-distribution output.
- Language directing clinicians to integrate AI information with radiologist assessment rather than treating it as a standalone probability.

AI Model Probability Output	Recommended Categorical Interpretation	Example Radiology Report Language	Implementation Considerations
<10%	Low likelihood	“AI analysis suggests a low likelihood of malignancy. Findings should be interpreted in conjunction with imaging features and clinical context.”	Avoid reporting exact percentages to prevent numerical anchoring. Low probabilities may still occur in poorly calibrated models.
10–40%	Indeterminate likelihood	“AI analysis indicates indeterminate likelihood of malignancy. Radiologist interpretation and clinical correlation remain essential.”	Intermediate probabilities are most susceptible to misinterpretation by referring clinicians.
40–70%	Moderate likelihood	“AI analysis suggests a moderate likelihood of malignancy. Final interpretation should integrate radiologist assessment and clinical information.”	Moderate probabilities frequently reflect model uncertainty or feature ambiguity.
>70%	High likelihood	“AI analysis suggests a high likelihood of malignancy; however, AI predictions should be interpreted within the full imaging and clinical context.”	High numeric outputs may reflect model overconfidence and should not imply diagnostic certainty.
Any probability with poor reliability indicators (e.g., low confidence score, OOD detection, artifact)	Uncertain or unreliable output	“AI output may be unreliable due to imaging conditions outside the model’s validated domain (e.g., artifact, postoperative anatomy, or protocol variation). Radiologist interpretation should guide clinical decision-making.”	Addresses dataset shift, scanner variability, treated anatomy, and distribution drift.

Conclusion

AI-generated percentages are often misinterpreted and poorly calibrated at the case level. A structured lexicon using categorical probability ranges, explicit uncertainty statements, and radiologist-centered integration provides a safer and more accurate way to communicate AI-derived information in radiology reports.

References

1. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
2. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur Radiol Exp*. 2018;2(1):35. doi:10.1186/s41747-018-0061-6
3. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. JMLR.org; 2017:1321-1330.
4. Price WN 2nd, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. *JAMA*. 2019;322(18):1765-1766. doi:10.1001/jama.2019.15064
5. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9
6. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science*. 1974;185(4157):1124-1131. doi:10.1126/science.185.4157.1124
7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7