Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0144
Korean J Radiol 2024;25(4):328-330

Korean Journal of Radiology
KJR

Caveats in Using Abnormality/Probability Scores from Artificial Intelligence Algorithms: Neither True Probability nor Level of Trustworthiness

Seong Ho Park[1], Eui Jin Hwang[2]

[1]Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[2]Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea

See the corresponding articles "Positive Predictive Values of Abnormality Scores from a Commercial Artificial Intelligence-Based Computer-Aided Diagnosis for Mammography" at https://doi.org/10.3348/kjr.2023.0907 and "Uncover This Tech Term: Uncertainty Quantification for Deep Learning" at https://doi.org/10.3348/kjr.2024.0108.

Keywords: Artificial intelligence; Deep learning; Abnormality score; Probability; Calibration; Uncertainty quantification; Reliability; Credibility; Trustworthiness; Confidence

One significant barrier to the adoption of artificial intelligence (AI) algorithms based on deep learning architectures in clinical practice is the inherent lack of understanding regarding why an AI algorithm produces a particular result, often referred to as its "black box" nature. When an AI algorithm generates outputs without allowing users to interrogate the decision-making process, it becomes challenging for users to adequately accept or reject the AI's results.

Providing specific probabilities and levels of credibility for AI decisions, such as "73% probability of cancer" or "85% certainty in cancer diagnosis," rather than offering a blunt AI output like "cancer" can help alleviate the black box problem. Although such information does not explain the inner workings behind an AI decision, it makes the use of AI more straightforward because users can more readily accept AI results when the results are accompanied by both high probabilities and credibility. In this context, the article by Faghani et al. [1] published in the current issue of the journal introduces an updated method known as "uncertainty quantification," which aims to demonstrate the level of uncertainty associated with an AI decision.

Many commercial AI algorithms already provide somewhat related AI output, such as abnormality scores or probability scores, alongside primary AI predictions (Fig. 1). These outputs are collectively referred to hereafter as "abnormality scores" in this article. Some unwitting users of AI may mistakenly interpret these scores as representing the true probability or level of trustworthiness of AI decisions. In reality, one should refrain from hastily interpreting such scores in this manner.

An AI algorithm initially calculates a probability-like continuous numerical output, for example 0 to 1. It subsequently converts this continuous output into categorical results such as the presence or absence of the target disease, by applying a predetermined threshold, for example 0.5 [2]. AI systems may present the raw continuous numerical output directly as an abnormality score, or adjust it in certain ways to display the score values in a more

Received: February 7, 2024 **Accepted:** February 8, 2024
Corresponding author: Seong Ho Park, MD, PhD, Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: parksh.radiology@gmail.com

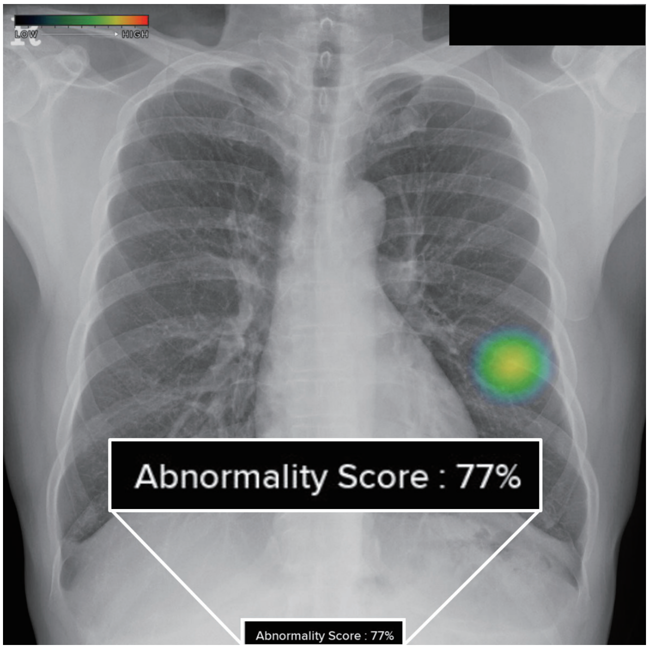
[redacted]

Fig. 1. An example of abnormality scores provided by a commercial artificial intelligence model. In this example the abnormality scores range from 0% (indicating the lowest probability) to 100% (indicating the highest probability). The presented score of 77% should not be hastily accepted as "77% true probability of the disease" or interpreted as "77% certainty in the diagnosis."

convenient range or as percentages, as depicted in Figure 1. While the continuous numerical output bears resemblance to probability, it may not necessarily represent the true probability of the disease predicted by the AI for several reasons. Firstly, in most cases the calibration performance of the continuous AI output remains unknown. Calibration performance refers to the degree of similarity of the AI-predicted probabilities to the actual probabilities [2]. It differs from the discrimination performance of an AI algorithm, often measured using the area under the receiver operating characteristic curve. It is important to note that good discrimination performance does not guarantee good calibration performance [2,3]. Despite its significance, calibration performance has received limited attention in both academic studies and regulatory evaluations of AI algorithms [3-5]. Secondly, even if an AI model demonstrates excellent calibration performance within a specific testing dataset, it may not be well calibrated for application elsewhere due to inherent heterogeneity in healthcare caused by factors such as variations in patient populations, differences in the acquisition and measurement of predictive and outcome variables, and shifts in healthcare practices over time [6,7]. Calibration performance is typically more susceptible to this heterogeneity compared to discrimination performance [6]. Looking at it from a different perspective, the optimal threshold used to convert the continuous numerical AI output into categorical results chosen in a particular testing scenario may not be applicable to individual users' practices, necessitating appropriate adjustments [8].

Given the above-described considerations, it is crucial for AI users not to interpret the abnormality score values provided by an AI algorithm as fixed probabilities. Moreover, regarding the effects of patient population characteristics, AI users should be familiar with the concepts of pretest and posttest probabilities. The true probability of a disease, known as posttest probability, is significantly influenced by the pretest probability, which represents the patient's inherent level of risk of the disease. According to Bayes' theorem, the posttest probability of disease given a particular test result (e.g., AI result) is calculated as follows [2]: pretest probability x likelihood ratio ÷ (1 – pretest probability + pretest probability x likelihood ratio). Consequently, for the same AI abnormality score, different true probabilities can be inferred depending on the patient's pretest probabilities. In simple terms, a higher posttest probability is deduced for a patient with a higher pretest probability. When human experts interpret radiology exams they typically consider various clinical findings and patient characteristics besides the radiological findings, thereby instinctively accounting for the patient's pretest probability.

Studies primarily focusing on the interpretation or utilization of AI-generated abnormality scores are relatively scarce, compared to the abundance of studies investigating the discrimination performance of AI algorithms. A study by Lee et al. [9], featured in the current issue of the journal, deals with this issue concerning the abnormality score provided by a specific commercial mammography AI. Further research in this area would be valuable and welcomed.

Taking a step further, it is important to understand that the AI-generated abnormality score does not represent the level of certainty of the AI diagnosis. As eloquently explained in articles by Faghani et al. [1,10], uncertainty is a distinct measure separate from the probability presented by an AI model. AI may present a probability with any levels of uncertainty. For instance, an AI model might indicate that a patient has a 90% probability of the target disease with a high level of uncertainty. In this case the AI prediction lacks credibility regardless of the high probability. Uncertainty quantification identifies instances where an AI model lacks sufficient information to make reliable decisions, prompting low trust in the model's predictions. In such cases re-evaluation by a human expert becomes even more critical, to ensure an accurate diagnosis. Conversely, in cases where the AI prediction is associated with low uncertainty, less stringent human supervision may be acceptable. Uncertainty quantification therefore serves as a tool to enhance transparency regarding the reliability of AI results and is likely to play a pivotal role in achieving efficient and synergistic human-AI collaboration [1,4]. Although uncertainty quantification is still an area of ongoing research and has yet to be integrated into user-level AI models, it is an important concept that AI users should be cognizant of.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Seong Ho Park. Writing—original draft: Seong Ho Park. Writing—review & editing: Eui Jin Hwang.

ORCID IDs

Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Eui Jin Hwang
https://orcid.org/0000-0002-3697-5542

Funding Statement

None

REFERENCES

1. Faghani S, Gamble C, Erickson BJ. Uncover this tech term: uncertainty quantification for deep learning. *Korean J Radiol* 2024;25:395-398
2. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800-809
3. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230
4. Zhang K, Khosravi B, Vahdati S, Erickson BJ. FDA review of radiologic AI algorithms: process and challenges. *Radiology* 2024;310:e230242
5. U.S. Food & Drug Administration. Clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions [accessed on February 7, 2024]. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology
6. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med* 2023;21:70
7. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:e489-e492
8. Hwang EJ, Goo JM, Yoon SH, Beck KS, Seo JB, Choi BW, et al. Use of artificial intelligence-based software as medical devices for chest radiography: a position paper from the Korean Society of Thoracic Radiology. *Korean J Radiol* 2021;22:1743-1748
9. Lee SE, Hong H, Kim E. Positive predictive values of abnormality scores from a commercial artificial intelligence-based computer-aided diagnosis for mammography. *Korean J Radiol* 2024;25:343-350
10. Faghani S, Moassefi M, Rouzrokh P, Khosravi B, Baffour FI, Ringler MD, et al. Quantifying uncertainty in deep learning of radiologic images. *Radiology* 2023;308:e222217

https://doi.org/10.3348/kjr.2024.0144