HealthManagement
Promoting Management and Leadership

Making AI Severity Scores More Reliable in Radiology

AI severity scores are increasingly integrated into radiological workflows, offering a quantifiable measure of the likelihood of pathology. While these tools have demonstrated potential in enhancing diagnostic performance, their current implementation often lacks the context necessary for accurate interpretation. The lack of clarity surrounding how these scores are produced and what they truly signify can hinder their effectiveness. Six human factors—ranging from inconsistent systems to perceptual mismatches—highlight the limitations of AI scores used in isolation. Understanding these constraints and identifying ways to address them is essential to ensure that AI supports rather than complicates radiologists' decision-making.

Interpretative Variability and Cognitive Challenges

Radiologists are confronted with a growing number of AI tools, each presenting its own severity score system. These scoring systems differ not only between different AI vendors but also between updates of the same system. An AI score of 3 may have different implications depending on the algorithm used or whether the system has recently been updated. Such inconsistencies demand a significant cognitive effort from radiologists who may be required to interpret several score types in the same clinical session. This variability can generate confusion, increase mental load and contribute to alarm fatigue.

Even if a system maintains consistency over time, radiologists themselves bring their own variability to the interpretation process. Two clinicians may interpret the same score differently, leading to diverging decisions regarding further testing or treatment. Additionally, a single radiologist's interpretation may shift over time as they gain more experience with the system or adapt to its updates. This interaction complicates the ability to rely on scores without further guidance. These inconsistencies reduce the potential for AI scores to provide reliable, standardised support across radiological practice.

Lack of Context and Distribution Awareness

Another major limitation lies in the opacity of score distributions. Without knowing where a particular score fits within a broader dataset, it becomes nearly impossible to assess the true significance of the number. A score that appears moderate may correspond to a high or low level of actual risk, depending on the underlying distribution, which may differ between training and local clinical populations. This lack of shared reference points leaves room for misinterpretation and hinders confident decision-making.

Must Read:** **How AI Strategy Shapes Diagnosis in Radiology

Moreover, AI systems often employ visual aids, such as coloured heatmaps, to illustrate the severity of findings. These visual cues are not universally understood and may be culturally or perceptually misleading. For example, red might imply danger in some settings but not in others, and red-green colourblindness affects a significant proportion of the population. Display hardware variability and degradation further compound the risk of misinterpretation. All of these factors point to the conclusion that AI scores, when presented without context or standardisation, lack the clarity necessary for effective clinical integration.

Towards a Solution: Embedding Diagnostic Value

To overcome these challenges, a structured approach is proposed: supplementing AI severity scores with the false discovery rate (FDR) and false omission rate (FOR) for each score threshold. The FDR represents the probability that a positive AI finding is incorrect, while the FOR indicates the chance that a negative result misses a true pathology. Together, these metrics provide clinicians with contextual probabilities tied to their specific patient populations. These values can be derived from retrospective local data, ensuring their relevance and improving

interpretative reliability.

Presenting FDR and FOR values alongside AI scores offers radiologists a clearer picture of the real diagnostic implications of a given result. This approach acknowledges that raw scores alone are insufficient and that interpretability hinges on transparency and localisation. If radiologists can consistently assess the risk of false positives and negatives at each threshold, decision-making will become more evidence-based and standardised. This could also improve both inter- and intra-rater reliability, addressing the variability that currently undermines confidence in AI-assisted diagnostics.

A robust experimental design has been proposed to test this hypothesis. Radiologists would evaluate the same set of images under two conditions—first, with AI scores alone, and second, with scores accompanied by corresponding FDR and FOR values. Their accuracy, consistency and diagnostic decisions would then be compared across these two conditions. The outcomes of such a study would provide empirical support for implementing these additional metrics into clinical AI systems, aligning with regulatory efforts to increase transparency and safety.

AI severity scores have the potential to support radiological diagnosis, but their effectiveness is limited when used in isolation. Variability in scoring systems, radiologist interpretation and score meaning all contribute to uncertainty. Without context, these scores cannot be relied upon to guide clinical decisions. A promising solution lies in augmenting scores with the false discovery and omission rates tied to local data, offering radiologists a clearer, standardised understanding of what each score truly means. This approach promises to improve diagnostic performance, reduce cognitive burden and facilitate more consistent clinical decisions across the board.

Source: **European Radiology Experimental**

Image Credit: **iStock**

Published on : Tue, 22 Jul 2025