Editorial

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.0448
Korean J Radiol 2023;24(7):601-605

Korean Journal of Radiology
KJR

How to Determine If One Diagnostic Method, Such as an Artificial Intelligence Model, is Superior to Another: Beyond Performance Metrics

Seong Ho Park[1], Ah-Ram Sul[2], Kyunghwa Han[3], Yu Sub Sung[4]

[1]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[2]Division of Healthcare Research Outcomes Research, National Evidence-based Healthcare Collaborating Agency, Seoul, Korea
[3]Department of Radiology, Research Institute of Radiological Science and Center for Clinical Imaging Data Science, Yonsei University College of Medicine, Seoul, Korea
[4]Clinical Research Center, Asan Medical Center, Seoul, Korea

Take-home points

- The effects of a diagnostic method, such as an artificial intelligence (AI) model, on patient outcomes cannot be determined by analyzing performance metrics (such as the area under the receiver operating characteristic curve, sensitivity, specificity, or the Youden index) alone.
- Two diagnostic methods can be compared more holistically in relation to patient outcomes using an equation, Δsensitivity x prevalence + Δspecificity x (1 - prevalence) x false positive (FP)-to-true positive (TP) outcome ratio, derived using the definition of the net benefit in the decision curve analysis, where the "FP-to-TP outcome ratio" is the ratio between the absolute amounts of net loss in patient outcomes incurred by an FP decision instead of leaving the patient alone and the net outcome gain provided by a TP decision compared with neglecting the disease in the patient.
- The equation can be useful for a preliminary estimation of the effects of a diagnostic method, such as AI, on patient outcomes when direct data are not available.

Keywords: Diagnosis; Model; Artificial intelligence; Accuracy; Performance; Outcome; Conversion; Equation

Studies comparing diagnostic methods often use performance metrics, such as the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, and Youden index (sensitivity + specificity - 1), to determine the superiority of one method over another. Recently, this trend has been observed frequently, particularly in studies on artificial intelligence (AI) models. For example, studies often present higher numerical values in performance metrics to substantiate the superiority of AI-assisted practice over conventional practice without AI assistance. However, these performance metrics are limited. First, the AUROC is a measure of the average performance across all possible ranges of thresholds applied to continuous raw AI outputs. However, clinical decisions are categorical and typically binary, such as the presence or absence of a target disease. Therefore, the AUROC may not represent the accuracy of AI at a particular threshold during actual use [1-3]. Second, sensitivity and specificity do not consider disease prevalence; therefore, they do not address the actual number of patients who are positively or negatively affected. For example, a 5% increase in sensitivity and a 0.5% decrease in specificity with AI assistance translates to ten additional true positive (TP) decisions for one additional false positive (FP) decision when the prevalence is 50%. However, if the disease prevalence is 5%, the same

Received: May 14, 2023 **Revised:** June 8, 2023 **Accepted:** June 9, 2023

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

changes in sensitivity and specificity indicate additional TP decisions at the cost of approximately twice as many additional FP decisions, which is very different from the viewpoint of patient outcomes. Third, performance metrics do not consider the magnitude of patient benefits and harm caused by TP and FP decisions.

In this article, we used an equation derived from the definition of the net benefit of a prediction model adopted in the decision curve analysis [4] to explain how prevalence and benefits and harm rendered by TP and FP decisions affect patient outcomes when comparing two diagnostic methods. In addition, we created a web-based graphic tool for the visualization of the equation (https://aim-aicro.com/software/performancetooutcome). We have used a comparison between AI-assisted and conventional diagnoses as an example to familiarize AI researchers. However, these explanations are applicable for comparisons between any two diagnostic methods. The explanations in this article are applicable when the sensitivity and specificity values are known. Therefore, the selection of optimal thresholds to convert continuous raw AI output into categorical decisions was not the focus of this article, which can be found elsewhere [5,6].

Derivation of the Equation

As shown in Figure 1, we designated the ultimate patient outcomes associated with TP, FP, false negative (FN), and true negative (TN) decisions as a, b, c, and d in a simple decision tree in the same manner as the decision curve analysis [4]. Suppose a comparison exists between AI-assisted and conventional diagnoses in n patients. If we designate the changes between AI-assisted and conventional diagnoses as Δ (= AI-assisted diagnosis - conventional diagnosis), the changes in the numbers of TP, FP, FN, and TN can be written as follows:

$$\Delta \text{TP} = -\Delta \text{FN} = \Delta \text{sensitivity} \times n \times \text{prevalence} \qquad \text{Eq. (1)}$$

$$\Delta \text{TN} = -\Delta \text{FP} = \Delta \text{specificity} \times n \times (1 - \text{prevalence}) \qquad \text{Eq. (2)}$$

Using AI, ΔTP patients (alternatively, -ΔFN patients) will have outcome a instead of c, and ΔTN patients (alternatively, -ΔFP patients) will have outcome d instead of b (Fig. 2). Therefore, the expected change in the patient outcome after the use of AI compared with conventional diagnosis becomes the sum of ΔTP x (a - c) (alternatively, -ΔFN x [a - c]) and ΔTN x (d - b) (alternatively, -ΔFP x [d - b]) and can be written as "ΔTP x (a – c) - ΔFP x (d - b)" using ΔTP and ΔFP. In addition, if the use of AI causes direct

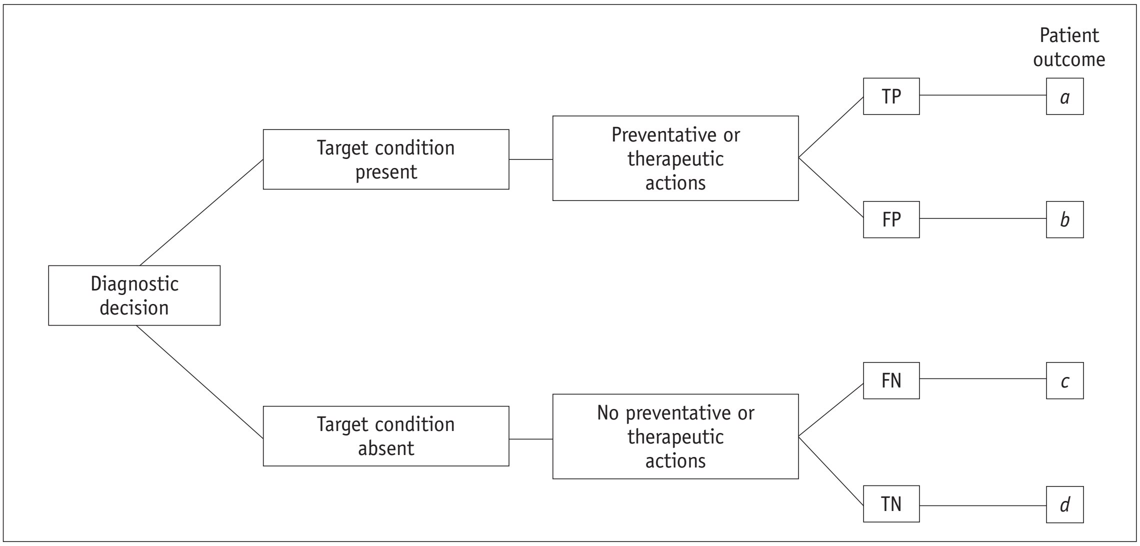

Fig. 1. Decision tree for patient management following a diagnostic decision. Preventative and therapeutic actions may result in both benefits and harm. The overall patient outcome, denoted as a, b, c, and d, is the sum of all elements of benefits and harm. For example, effective preventative or therapeutic actions that follow true positive (TP) decisions would create substantially greater benefits (i.e., intended preventative or therapeutic effects) than harmful elements (i.e., adverse effects of such actions). In contrast, preventative or therapeutic actions that follow false positive (FP) decisions would mostly create harmful elements (such as adverse effects caused by unnecessary treatments, workups, or follow-ups) and, if any, minor incidental benefit elements (such as the incidental discovery of unrelated significant diseases during the workup). FN = false negative, TN = true negative

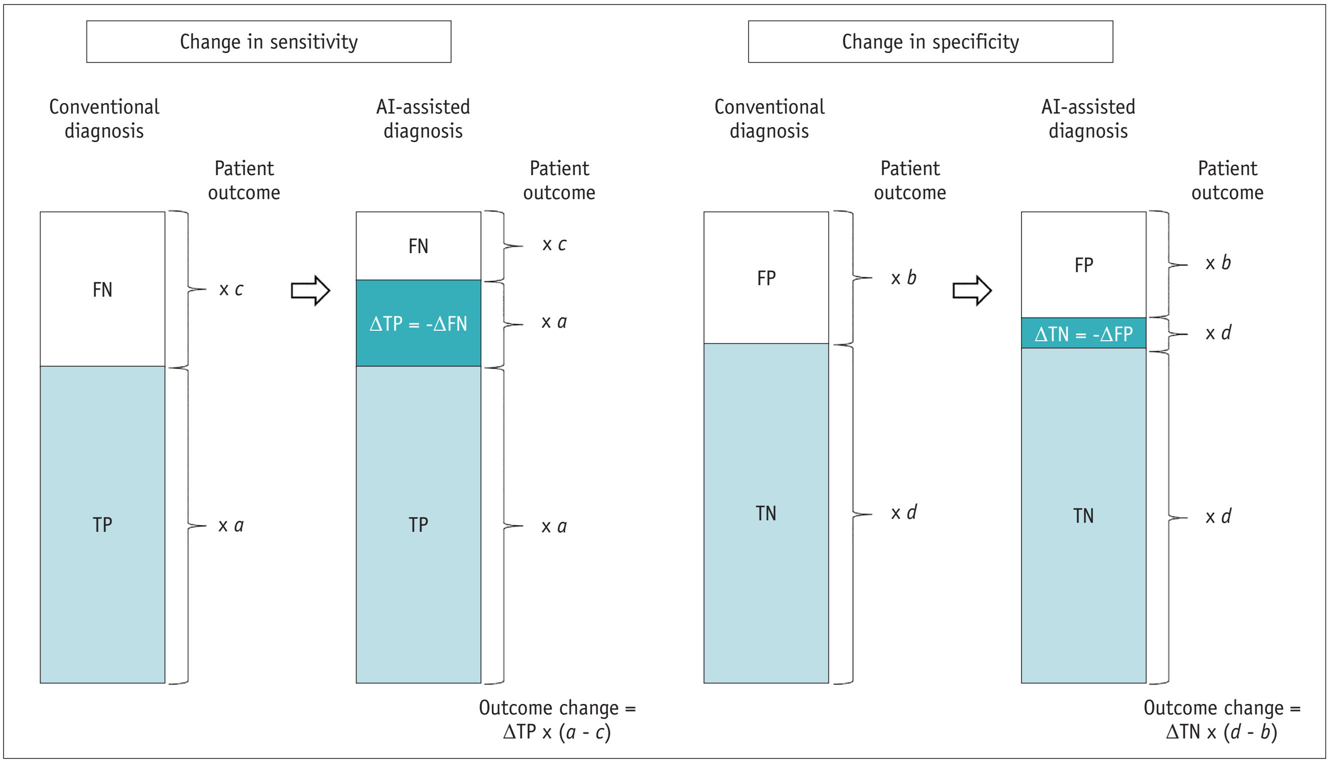

Fig. 2. Patient outcome change based on changes in sensitivity and specificity, using the comparison between artificial intelligence (AI)-assisted diagnosis and conventional diagnosis as an example. *a*, *b*, *c*, and *d* represent patient outcomes as defined in Figure 1. FN = false negative, TP = true positive, FP = false positive, TN = true negative

harm to a patient, direct AI harm must also be considered [4]. Direct AI harm differs from the indirect harm caused by an FP decision, which is already included in *b*, such as adverse effects caused by unnecessary workup, follow-up, or treatment. Because AI software tools only analyze data obtained during patient care, we can disregard direct AI harm in AI-assisted diagnoses.

We can then make the equation agnostic of the number of patients by dividing it by the total number of patients, *n*, and pull out (*a* - *c*) to the front as follows:

$$(a - c) \times \left(\frac{\Delta \mathrm{TP}}{n} - \frac{\Delta \mathrm{FP}}{n} \times \frac{(d - b)}{(a - c)}\right)$$

If we are only interested in determining whether the patient outcome change is positive or negative with the use of AI, instead of determining the quantity of outcome change, we can drop (*a* - *c*) at the beginning of the equation as follows:

$$\frac{\Delta \mathrm{TP}}{n} - \frac{\Delta \mathrm{FP}}{n} \times \frac{(d - b)}{(a - c)}$$

This equation is essentially the same as the equation for the net benefit as described in the decision curve analysis, except that the $\frac{(d - b)}{(a - c)}$ part is written using the “threshold probability” in the decision curve analysis [4,7].

We may assume that *a* (outcome for a TP) is better than *c* (outcome for an FN) and that *d* (outcome for a TN) is better than *b* (outcome for an FP); (*a* - *c*) and (*d* - *b*) have positive numerical values. Therefore, (*d* - *b*) and (a - *c*) represent the absolute amounts of net loss in the patient outcome incurred by an FP decision instead of leaving the patient alone and the net outcome gain provided by a TP decision compared to neglecting the disease in the patient, respectively, with $\frac{(d - b)}{(a - c)}$ as the ratio. Taking the surveillance of hepatocellular carcinoma as an example, an FP decision mostly causes harm compared to leaving the patient alone, including the negative consequences of unnecessary follow-up imaging tests (such as adverse effects from contrast agents, radiation exposure, and waste of time and money), harm from any invasive procedures that may follow (such as liver biopsy), harm from unnecessary treatments, and emotional distress. There could be a slight theoretical benefit from an FP decision, such as the incidental detection of unrelated significant diseases on follow-up tests. Thus, (*d* – *b*) represents the absolute amount that summarizes all the results expected in a single FP patient. Conversely, (*a* - *c*) is the summation of

https://doi.org/10.3348/kjr.2023.0448

all differences anticipated in a single TP patient compared to neglecting the diagnosis in the patient, consisting mostly of benefits from earlier diagnosis of the tumor, which may improve patient survival and make the patient eligible for less invasive treatments with less treatment-related harm compared to later-stage diagnoses and small potential harm, such as adverse effects from treatments. An accurate estimation of these specific results without direct observation is difficult. However, a rough estimation of their relative ratio, $\frac{(d - b)}{(a - c)}$, which we refer to as the "FP-to-TP outcome ratio," might be more feasible.

The previous equation can be rewritten using Eq. (1) and Eq. (2) as follows:

$$\text{net benefit} = \frac{\Delta\text{sensitivity} \times n \times \text{prevalence}}{n} + \frac{\Delta\text{specificity} \times n \times (1 - \text{prevalence})}{n} \times \text{FP-to-TP outcome ratio}$$

$$= \Delta\text{sensitivity} \times \text{prevalence} + \Delta\text{specificity} \times (1 - \text{prevalence}) \times \text{FP-to-TP outcome ratio} \qquad \text{Eq. (3)}$$

The positive and negative values calculated with the equation indicate the ultimate benefit or lack thereof, respectively, with the use of AI compared with conventional diagnosis.

Graphic Visualization

If the Δsensitivity and Δspecificity values for the use of a certain AI are available from primary research studies or their meta-analytic summaries, we can plot the net benefit against the prevalence and FP-to-TP outcome ratio using Eq. (3) (Fig. 3; a web-based program for generating the graph is available at https://aim-aicro.com/software/performancetooutcome). It may be difficult to specify the disease prevalence and FP-to-TP outcome ratio in a clinical setting where one wants to apply the AI model.

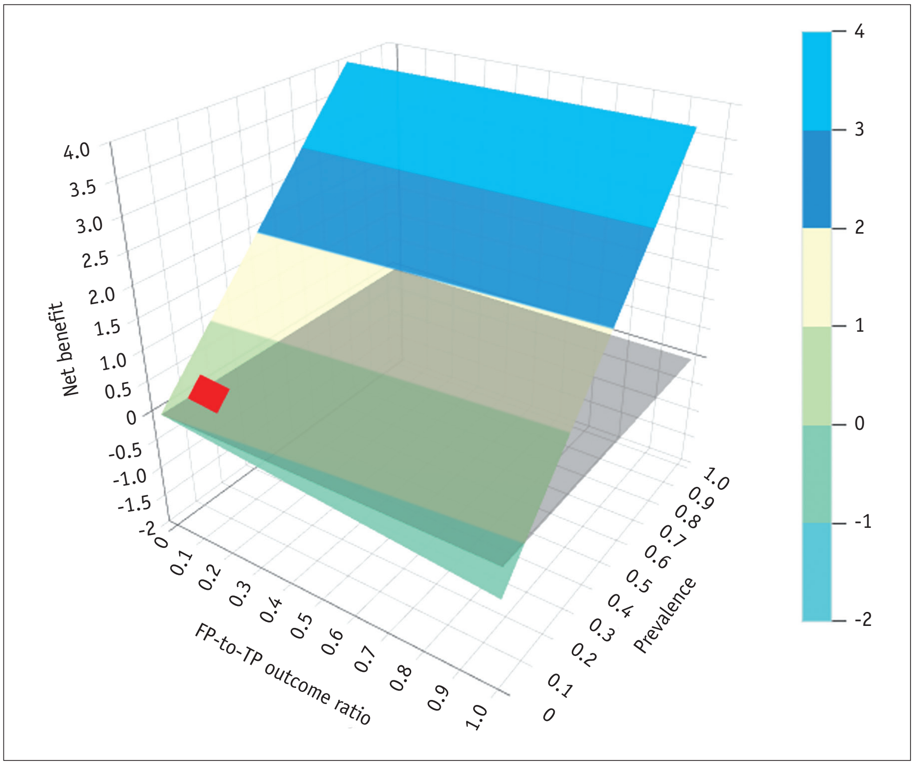

Fig. 3. Graphic visualization of net benefit. This graph is a plot of the net benefit against the prevalence and false positive (FP)-to-true positive (TP) outcome ratio, $\frac{(d - b)}{(a - c)}$, based on Eq. (3) for a hypothetical scenario in which a 5% increase in sensitivity and a 0.5% decrease in specificity were achieved using artificial intelligence (AI) assistance compared to conventional diagnosis. The red rectangular area indicates the net benefit values for using AI assistance based on the expected ranges of prevalence (5 to 10% as an example) and the FP-to-TP outcome ratio (0.05 to 0.15 as an example). When the red rectangular area is farther away from the zero plane (horizontal gray plane), compared with when close to it, greater confidence exists in the ultimate benefits of using AI assistance. The graph also presents a large tilted plane with a color gradation corresponding to the color-to-value scale provided on the right for reference. This plane illustrates the net benefit values across the entire range of both prevalence and the FP-to-TP outcome ratio. A web program for generating the graphs is available at https://aim-aicro.com/software/performancetooutcome.

However, if their ranges can be determined at least roughly, the plot can be useful because it provides direct, albeit crude, information regarding the effects of the AI model on patient outcomes, which a comparison of performance metric values cannot provide.

Figure 3 presents an example plot for a hypothetical scenario in which the use of AI results in a 5% increase in sensitivity and a 0.5% decrease in specificity compared to conventional diagnosis. If the estimated net benefit is below the zero plane, the use of AI is not advantageous. When the net benefit values are substantially above the zero plane, greater confidence exists in the ultimate benefit of using AI than when they are near the zero plane.

CONCLUSION

This article explains how to compare two diagnostic methods more holistically from the perspective of patient outcomes beyond performance metrics. This can be useful for the preliminary estimation of the effects of an AI model on patient outcomes when direct data are unavailable. Nevertheless, the direct assessment of patient outcomes in clinical trials is important because diagnostic performance improvements do not guarantee improved patient outcomes for multiple reasons [8-10].

Conflicts of Interest

The Editor-in-Chief of *Korean Journal of Radiology*, Dr. Seong Ho Park, and the Statistical Consultant for *Korean Journal of Radiology*, Prof. Kyunghwa Han, were not involved in the editorial evaluation or decision to publish this article. All remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Seong Ho Park. Funding acquisition: Seong Ho Park, Ah-Ram Sul. Methodology: all authors. Project administration: Seong Ho Park. Software: Yu Sub Sung. Supervision: Seong Ho Park. Visualization: Yu Sub Sung. Writing-original draft: Seong Ho Park. Writing-review & editing: Ah-Ram Sul, Kyunghwa Han, Yu Sub Sung.

ORCID iDs

Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Ah-Ram Sul
https://orcid.org/0000-0003-0331-5529
Kyunghwa Han
https://orcid.org/0000-0002-5687-7237
Yu Sub Sung
https://orcid.org/0000-0002-9215-735X

Funding Statement

This study was supported by the National Evidence-based Healthcare Collaborating Agency in South Korea (NECA-A-23-002).

REFERENCES

1. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2023;306:20-31
2. 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
3. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019;2:e191095
4. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565-574
5. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med (Zagreb)* 2016;26:297-307
6. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003;229:3-8
7. Vickers AJ, Cronin AM, Gönen M. A simple decision analytic solution to the comparison of two binary diagnostic tests. *Stat Med* 2013;32:1865-1876
8. Park SH, Choi JI, Fournier L, Vasey B. Randomized clinical trials of artificial intelligence in medicine: why, when, and how? *Korean J Radiol* 2022;23:1119-1125
9. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 2020;370:m3164
10. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904

https://doi.org/10.3348/kjr.2023.0448