Review Article | Technology, Experiment, and Physics

eISSN 2005-8330
https://doi.org/10.3348/kjr.2021.0048
Korean J Radiol 2021;22(3):442-453

Korean Journal of Radiology
KJR

Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence

Seong Ho Park[1], Jaesoon Choi[2], Jeong-Sik Byeon[3]

[1]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea; Departments of [2]Biomedical Engineering and [3]Gastroenterology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Artificial intelligence (AI) will likely affect various fields of medicine. This article aims to explain the fundamental principles of clinical validation, device approval, and insurance coverage decisions of AI algorithms for medical diagnosis and prediction. Discrimination accuracy of AI algorithms is often evaluated with the Dice similarity coefficient, sensitivity, specificity, and traditional or free-response receiver operating characteristic curves. Calibration accuracy should also be assessed, especially for algorithms that provide probabilities to users. As current AI algorithms have limited generalizability to real-world practice, clinical validation of AI should put it to proper external testing and assisting roles. External testing could adopt diagnostic case-control or diagnostic cohort designs. A diagnostic case-control study evaluates the technical validity/accuracy of AI while the latter tests the clinical validity/accuracy of AI in samples representing target patients in real-world clinical scenarios. Ultimate clinical validation of AI requires evaluations of its impact on patient outcomes, referred to as clinical utility, and for which randomized clinical trials are ideal. Device approval of AI is typically granted with proof of technical validity/accuracy and thus does not intend to directly indicate if AI is beneficial for patient care or if it improves patient outcomes. Neither can it categorically address the issue of limited generalizability of AI. After achieving device approval, it is up to medical professionals to determine if the approved AI algorithms are beneficial for real-world patient care. Insurance coverage decisions generally require a demonstration of clinical utility that the use of AI has improved patient outcomes.

Keywords: *Software validation; Device approval; Insurance coverage; Artificial intelligence*

INTRODUCTION

Artificial intelligence (AI) technology is expected to be of substantial help in medicine by overcoming current limitations and developing innovative solutions and will likely have a great impact on healthcare in the future [1,2]. All pharmaceuticals and medical devices, including AI devices, must be subjected to a rigorous clinical validation process to ensure safety and efficacy prior to use on patients. There is a wide range of AI devices for use in healthcare, and the methods used for clinical validation vary according to their form and function. Most are classified as diagnostic devices, as they are algorithms used to assist with diagnosis, decision-making, and prediction, such as computer-aided detection (CADe), computer-aided diagnosis, or clinical decision support systems. As such, methods for their clinical validation resemble those for common diagnostic tests. In this article, we aim to explain the key principles of clinical validation, device approval, and insurance coverage decisions for AI algorithms in healthcare.

Received: January 15, 2021 **Revised:** January 15, 2021
Accepted: January 15, 2021
This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI17C2410).
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, Korea.
• E-mail: seongho@amc.seoul.kr

Korean Journal of Radiology
KJR

Performance Indicators of AI Algorithms

There are a variety of indicators that may be used to evaluate the performance of AI algorithms. Some are technical indicators with little medical relevance, and others apply only to specific situations. Therefore, instead of presenting a comprehensive list of all indicators, we have focused on frequently used indicators with high medical relevance.

Dice Similarity Coefficient

The Dice similarity coefficient is used to evaluate AI algorithms that perform segmentation of organs or lesions on medical images [3]. Its definition is illustrated in Figure 1. For example, if there is an AI algorithm that can display the area suspected of prostate cancer on prostate magnetic resonance imaging (MRI), its performance can be evaluated by measuring the degree of overlap between the pathologically confirmed cancerous region and the area identified as cancer by the algorithm. There are several other coefficients similar to the Dice similarity coefficient.

Sensitivity, Specificity, Receiver Operating Characteristic Curve

As shown in Figure 2, if an AI algorithm presents a binary result (e.g., presence vs. absence of a disease), its performance can be described, as in general diagnostic tests, in terms of sensitivity = true positive/(true positive + false negative), i.e., the proportion of subjects identified as positive by the AI out of all disease-positive subjects, and specificity = true negative/(false positive + true negative), i.e., the proportion of subjects identified as negative by the AI out of all disease-negative subjects. Even though the result an AI algorithm gives is presented as a binary classification, it is preceded by the process of outputting the result as a continuous number (for example, a decimal range between 0 and 1 as in probability). A threshold is then applied to convert it into a binary result. The sensitivity and specificity of the AI algorithm vary depending on how the threshold is set. If the threshold is set high, sensitivity decreases and specificity increases. If the threshold is set low, the sensitivity increases and the specificity decreases. A receiver operating characteristic (ROC) curve is a graph drawn by plotting the sensitivity on the y-axis and 1 - specificity on the x-axis, while varying the threshold value (Fig. 3) [4]. The value of the area under the curve (AUC) or area under the ROC (AUROC) curve is the mean sensitivity or specificity for all possible threshold values. Its maximum value is 1. In theory, the higher the value, the higher the diagnostic accuracy. Interpretations should be made carefully, however, because a higher AUROC value of an AI algorithm is not necessarily equivalent to higher performance of the AI in practice. Given that a

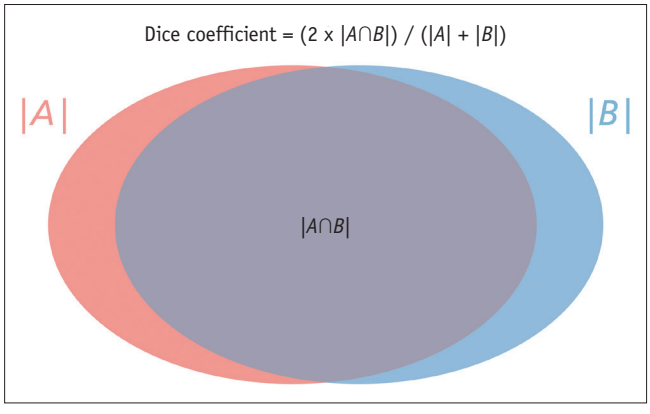

Fig. 1. Dice similarity coefficient.

		Disease state to diagnose/predict according to reference standard	
		Present	Absent
AI result	Present	TP	FP
	Absent	FN	TN

Fig. 2. Diagnostic cross-table (also referred to as confusion matrix). AI = artificial intelligence, FN = false negative, FP = false positive, TN = true negative, TP = true positive

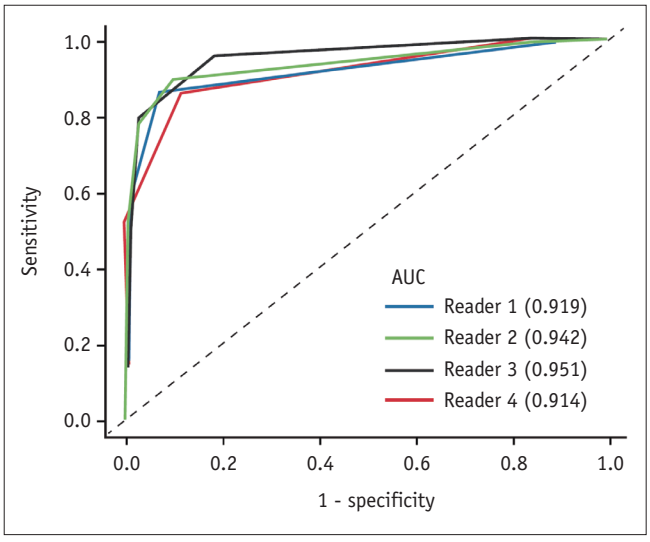

Fig. 3. Exemplary receiver operating characteristic curves that show the performance of four readers in interpreting breast ultrasonography assisted by a deep-learning algorithm. Adapted from Choi et al. Korean J Radiol 2019;20:749-758, with permission from the Korean Society of Radiology [4]. AUC = area under the curve

https://doi.org/10.3348/kjr.2021.0048

particular threshold value is required when using an AI algorithm in practice, the sensitivity and specificity values for the given threshold, not the mean AUROC value, are the algorithm's actual measures of performance. The AUROC value is merely the mean sensitivity or specificity value. For further details, see the relevant literature [2,5,6].

Free-Response ROC Curve

The free-response ROC (FROC) curve is used to evaluate the performance of AI algorithms with a CADe function, such as those for detecting colonic polyps on colonoscopy images. The AI algorithm output is correct when both the presence of a lesion and the localization of the lesion site are proven correct. When the AI algorithm for detecting colonic polyps indicates there is a polyp in a patient with colonic polyps, its diagnosis is correct only when it also detects the correct lesion site. If it fails to detect a polyp in an area where there is one and indicates there is a polyp in an area where there is none, it has produced both false-negative and false-positive results. In diagnostic tasks where a CADe-enabled algorithm is applied, there may be multiple lesions in a patient, and a CADe-enabled algorithm may present multiple false positives. In this case, it would be more appropriate to evaluate the algorithm's diagnostic accuracy using sensitivity and the number of false positives instead of sensitivity and specificity. If the threshold value is set too high for the algorithm's internal continuous output values, the sensitivity decreases, as does the number of false positives; if the threshold value is set too low, the sensitivity increases, but the number of false positives increases as well. The FROC curve is a graph drawn by plotting the sensitivity on the y-axis and the mean number of false positives instead of 1 - specificity on the x-axis (Fig. 4) [7]. The mean number of false positives can be calculated in several ways, depending on the situation. For example, they can be calculated using the mean number per patient or per image. There are also slightly modified forms of the FROC method. For further details, see the relevant literature [8,9].

Calibration Accuracy

The performance indicators described above are all indicators of discrimination accuracy. Calibration accuracy, on the other hand, which describes how similar the predicted probability values presented by an AI algorithm (for example, "The probability of this lesion to be cancerous is X%") are to the actual probabilities, should be evaluated separately [6]. According to the Bayes' theorem of probability, the actual probability is greatly influenced by the pretest probability, also referred to as disease prevalence. It follows that a probability presented by an AI algorithm that does not take into consideration the pretest probability is likely inaccurate. Therefore, a rigorous evaluation of calibration accuracy is required for all AI algorithms that present probabilities directly to users. Particular care should be taken, as calibration accuracy is often overlooked when evaluating the performance of an AI algorithm [6]. It goes beyond the scope of this paper to go into details about calibration accuracy. For further details, see the relevant literature.

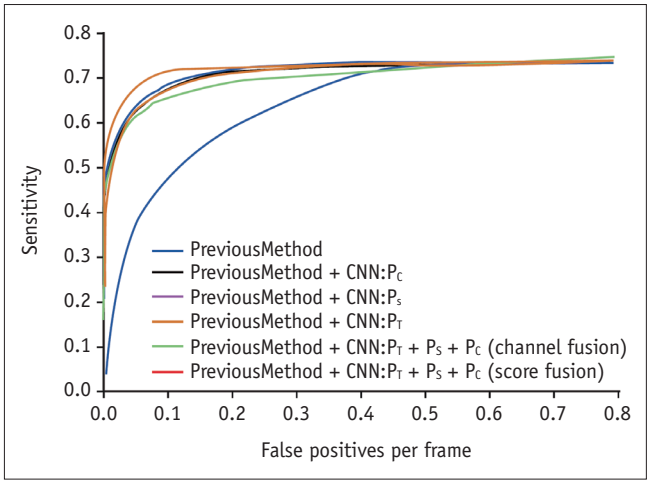

Fig. 4. Exemplary free-response receiver operating characteristic curves that show the performance of six methods of detecting polyps in colonoscopy videos. The x-axis is the mean number of false positives per image frame. A curve closer to the left upper corner indicates a higher performance, for example, a higher performance of the red curve than the blue curve. Adapted from Tajbakhsh et al. Proceedings of IEEE 12th International Symposium on Biomedical Imaging. New York: IEEE; 2015, with permission from IEEE [7]. CNN = convolutional neural network

Limited Generalizability of AI Algorithm Performance in Healthcare

Overfitting in AI Algorithms for Medical Diagnosis/Prediction

Machine learning algorithms characterized by high dimensionality and mathematical complexity, such as deep learning which represents the current AI technology, have strong data dependence. Therefore, they tend to have excellent accuracy in training data, but their performance deteriorates in external data not used for training. This phenomenon is called 'overfitting' [10]. It

https://doi.org/10.3348/kjr.2021.0048

Korean Journal of Radiology
KJR

is well known that AI algorithms for medical diagnosis/prediction are particularly prone to overfitting. There are several techniques to reduce overfitting, collectively termed regularization, but regularization alone is often insufficient to address overfitting in AI algorithms for medical diagnosis/prediction. For this reason, most current AI algorithms in medicine may fail to generalize [11]. Table 1 shows some examples of the limited generalizability of AI algorithms for medical diagnosis/prediction [12-16]. As shown in these examples, in real-world clinical settings, the diagnostic accuracy of these AI algorithms decreases or the presented probability becomes incorrect, and the threshold value set for converting the internal output value into the final result does not fit.

Reasons for High Overfitting in AI Algorithms for Medical Diagnosis/Prediction

The fundamental reason behind the high overfitting and limited generalizability of AI algorithms for medical diagnosis/prediction is their failure to sufficiently reflect the real-world situations in the data sets used to train the AI algorithms [11,17-19]. Several factors are involved in this phenomenon. First, medical data are highly heterogeneous. Even if patients have the same disease, their other characteristics such as age, sex, disease severity, underlying conditions, or comorbidities often differ across the capacity, type, and location of the hospitals. The variety and distribution of similar diseases or differential diagnoses found in patients suspected of a particular disease but who do not have the disease also often differ across hospitals. Disease prevalence may also vary from one hospital to another. Simply put, the situation of one hospital often cannot be applied to another. The fact that hospitals utilize different devices also contributes to the data heterogeneity. For example, in the case of imaging devices, such as computed tomography and MRI, an AI algorithm tuned to the image properties of one vendor may not work well on the images obtained with the scanners from another manufacturer. Furthermore, advances in healthcare equipment and technologies lead to constant changes in therapeutic agents and diagnostic tools, which creates temporal heterogeneity. For example, AI algorithms trained with data including treatment agents used in the past cannot function correctly in situations where different therapeutic agents are used; likewise, AI that has been trained on images from old imaging devices may not work properly on images from new devices. To overcome this data heterogeneity, it is necessary to train AI with a huge amount data systematically collected from as many hospitals as possible, which is a labor- and time-intensive procedure requiring committed medical professionals and a large amount of material resources. The paradoxical combination of high data heterogeneity and insufficient data for training AI algorithms often results in a situation where the training

Table 1. Examples of Limited Generalizability of the Performance of Artificial Intelligence Algorithms for Medical Diagnosis/Prediction

Author	Algorithm	Result
Zech et al. [12]	CNN algorithm to detect pneumonia on chest radiographs	AUC of 0.931 in internal testing compared with 0.815 in external testing
Ting et al. [13]	CNN algorithm to detect referable diabetic retinopathy on retinal photographs	AUC ranging from 0.889 to 0.983 when tested externally at 10 different hospitals
Ridley [14]	CNN algorithm to detect intracranial hemorrhage on noncontrast head computed tomography scans	Sensitivity, specificity, and AUC of 98%, 95%, and 0.993, respectively, when tested internally compared with 87.1%, 58.3%, and 0.834, respectively, when tested on a real-world data set
Hwang et al. [15]	CNN algorithm to distinguish normal chest radiographs from abnormal chest radiographs that contain any of the four types of pathologies including malignancy, tuberculosis, pneumonia, and pneumothorax	When externally tested at five different hospitals with a single fixed threshold applied to the raw algorithm output, the specificity indicated a wide range from 56.6% to 100%, while the sensitivity was less variable ranging from 91.3% to 100%
Lee et al. [16]	CNN algorithm to categorize hepatic fibrosis (F0, F1, F2–3, and F4 according to METAVIR scoring) on B-mode ultrasonography images	Accuracy of 83.5% in internal testing compared with 76.4% in external testing

AUC = area under the curve, CNN = convolutional neural network

https://doi.org/10.3348/kjr.2021.0048

data sets do not sufficiently reflect the clinical settings in which an algorithm is intended to be used. In addition, real-world medical data contain diverse gray areas and noise elements. There are cases in which no clear reference standards are available to determine the presence or absence of a certain disease. The presence vs. absence of a disease, as described in a binary variable (0 or 1), may only represent a few specific points along a continuous phase of change in the development and progress of an entire disease process. Nevertheless, data allowing a clear binary disease classification (present/absent) are often selectively used for AI training purposes [20]. Also, data from which noise elements are removed for more efficient processing by computer programs are used preferentially.

Implications of the Limited Generalizability

First, when evaluating AI algorithm performance, it is important to perform external validation using external data, as discussed further later [6,10,11,21-29]. Given that an AI algorithm's performance in a clinical envirionment may differ from when it was developed, it is best to conduct external validation directly in the target clinical environment. Nevertheless, insufficient external validation of AI algorithms frequently poses problem [30]. Second, instead of blindly accepting the result presented by an AI algorithm, medical professionals should make final decisions after due consideration to the clinical situation and other relevant information. The threshold value described above should also be properly tuned to the clinical situation. For these reasons, while high-performance AI might replace medical professionals for specific functions under limited conditions, AI is not an autonomous tool to replace a medical professional; its role is limited to providing competent assistance and information to the medical professional. Third, as a method to improve the generalizability of an AI algorithm, an additional training round may be administered using data from target hospitals and specific clinical settings prior to its use in the practice. There are concerns, however, that an AI algorithm's initial accuracy may be impaired if it is trained with additional data that include errors and biases. Unlike locked software algorithms, AI algorithms can change through continuous learning, and continuous evaluation and management are required, even after device approval. As the current device-approval system has no concrete provisions for continuous evaluation and management, this aspect should be addressed in the future.

Evaluating AI Algorithm Performance: Classification according to Data Used

The methods of evaluating AI algorithms' performance can be classified according to the characteristics of the data used for the evaluation. Before explaining these classifications, it is necessary to understand the term 'validation' clearly. In addition to its ordinary meaning (i.e., verification or confirmation), as used in this article, validation is also a technical term in the machine learning field, referring to the process of adjusting hyperparameters when making AI algorithms [31]. The process of adjusting hyperparameters is also called tuning to avoid confusion; however, validation is more widely used to indicate the procedure in AI-related literature [31]. On the other hand, validation test or test is used instead of validation to indicate verification of algorithm performance and distinguish it from the process of adjusting hyperparameters [32,33].

Internal Validation

Internal validation tends to overestimate the performance of AI algorithms. Therefore, internal validation has a role in checking the algorithm performance while developing it rather than confirming the performance of a finished model. External validation is required to determine the performance of AI algorithms. Results from internal validation can be used to compare the results of external validation. Cross-validation and split-sample validation are categorized as types of internal validation.

Cross-Validation

A well-known example of cross-validation is k-fold cross-validation [32]. The original data are split into k number of groups; one group is retained as the testing data, and the remaining groups form the training data. At each iteration, one group after another is used as the testing data until every group has been used once. Finally, the mean of all results is obtained. This method can be used for a preliminary evaluation of an algorithm's performance when the original data size is small. However, it is considered inadequate for algorithm performance validation.

Split-Sample Validation

In split-sample validation, the original data are split into three sets (training set, tuning set, and test set). The test set is not used for training and tuning of the AI algorithm;

https://doi.org/10.3348/kjr.2021.0048

Korean Journal of Radiology
KJR

it is used to test the performance of the trained and tuned AI algorithm (Fig. 5) [32]. The data can be split randomly or stratified according to the data-collection period. Split-sample validation is better suited for internal validation than cross-validation.

External Validation

External validation refers to evaluation of an AI algorithm's performance using data collected independently instead of the original data (Fig. 5). Typically, a data set provided by external hospitals is used instead of the one that provided the training data. As shown in Figure 2, to evaluate the performance of an AI algorithm, two categories of data are required: one that contains the condition targeted by the AI algorithm for diagnosis or prediction and one that does not. Depending on the method of collecting these validation data, external validation studies can be largely divided into diagnostic case-control and diagnostic cohort studies [34].

Diagnostic Case-Control Study

In diagnostic case-control studies, samples with and without the target condition to be diagnosed or predicted by an AI algorithm are collected separately. For example, when evaluating the performance of an AI algorithm that discriminates the presence or absence of lung cancer by analyzing chest X-ray images, a certain number of chest X-ray images with lung cancer (case) and without lung cancer (control) are collected. When data is collected in this way, prevalence (in the example case, the proportion of chest X-rays with lung cancer among the total number of chest X-ray images) is artificially designated, unlike the natural prevalence observed in real-world settings. Moreover, the method by which the images with and without lung cancer are collected may affect the degree of variation in the size and shape of the included lung cancers, the presence or variety of lung lesions that may mimic lung cancer in patients without lung cancer, and the presence or degree of various conditions or underlying diseases and comorbidities that can affect the discovery of lung cancer, all of which are collectively referred to as 'spectrum.' Data collected in this case-control manner often differ from the natural spectrum in clinical settings due to selection bias [35]. This artificial spectrum and prevalence affect the evaluation of algorithm performance [2,6].

Diagnostic Cohort Study

In a diagnostic cohort study, the clinical setting in which an AI algorithm will be applied is predefined; data are collected based on this definition, regardless of the presence/absence of the disease to be diagnosed or predicted by the AI algorithm. In the case of an AI algorithm discriminating the presence or absence of lung cancer on chest X-rays, the eligibility criteria (for example, "adults 55 years of age and older with X pack-year smoking history") are defined. The AI algorithm performance is assessed on X-ray images taken from patients who are continuously recruited or randomly selected from those satisfying the eligibility criteria. Some of the recruited patients may have lung cancer, and others may not. In this

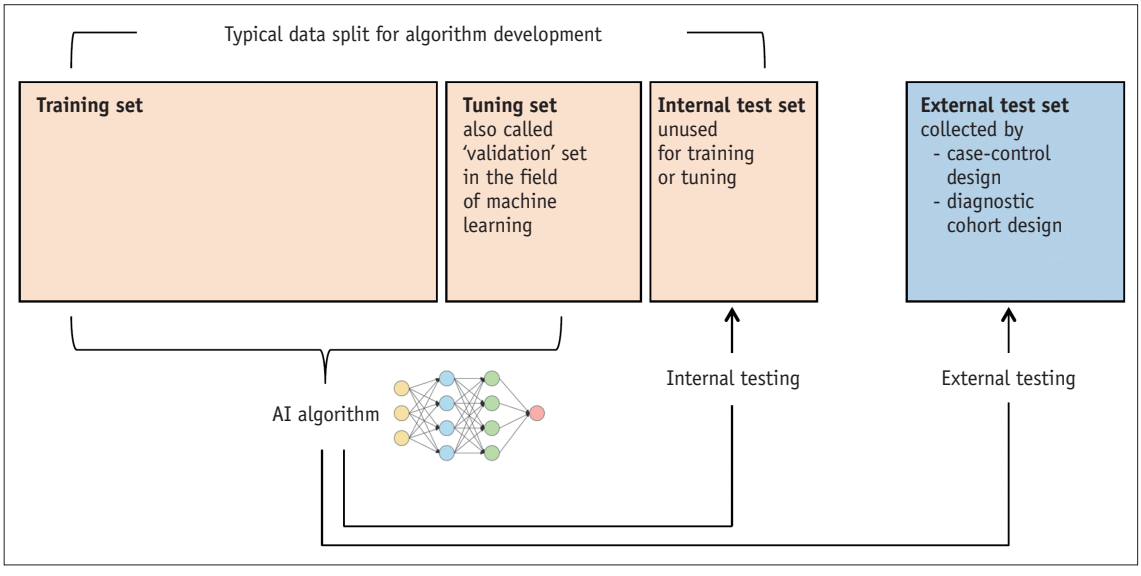

Fig. 5. Typical data sets used for development and testing of an AI algorithm. AI = artificial intelligence

https://doi.org/10.3348/kjr.2021.0048

way, data with the natural spectrum and prevalence can be collected, and the performance and the threshold value determined in the validation study can be more directly applied to the clinical setting defined by the eligibility criteria. Whereas a diagnostic case-control study evaluates performance in a somewhat artificial experimental setting, a diagnostic cohort study evaluates performance in a more realistic clinical environment. It is essential to clearly understand the actual clinical setting for which the AI algorithm is intended when determining the concrete eligibility criteria to reflect the clinical setting adequately.

Differences between Diagnostic Case-Control Study and Diagnostic Cohort Study

Diagnostic case-control and diagnostic cohort studies are designed for different purposes. The former aims to evaluate the overall 'technical performance' of an AI algorithm for the intended diagnosis/prediction. Although prospective studies offer certain advantages, retrospective studies can also be used to validate the technical performance of an AI algorithm, subject to the availability of a validation data set containing well-distributed examples of various difficulty levels matching the purpose of the AI algorithm. Diagnostic cohort studies aims to evaluate the 'clinical performance' of an AI algorithm in specific clinical settings or patient groups. Therefore, compared to a diagnostic case-control study, a diagnostic cohort study has a clearer and more concrete notion of the clinical target population, i.e., clinical indication. For diagnostic cohort studies, a prospective study design is recommended. In general, the technical performance of an AI algorithm is first tested via a diagnostic case-control study, then the clinical performance is tested via a diagnostic cohort study. There is a tendency for performance to be rated higher in diagnostic case-control studies than in diagnostic cohort studies.

Validation Using Standard Data Sets

Some are of the opinion that standard data sets should be used for performance validation of AI algorithms. The process of collecting standard data is often similar to that of collecting data for diagnostic case-control studies. Therefore, well-established standard data sets may prove suitable for evaluating the technical performance of AI algorithms. Such standard sets would be rendered more efficient when collected from many different hospitals. One of the major drawbacks of standard data is that AI algorithms that perform well only on standard data can emerge. A standard data set can be compared to the College Entrance Exam (Korean equivalent of SAT). Every year, new questions are formulated under thorough security measures, because reusing questions does not allow for the proper evaluation of a student's performance. Likewise, a standard data set that is not continuously updated may be subject to the "leaking test questions" effect. Moreover, to successfully generate a genuinely representative standard data set, meticulous prior research is needed on fundamental issues such as the conditions required for a standard data set to test the performance of AI algorithms sufficiently and objectively.

Passing Criteria for AI Algorithm Performance Evaluation

When evaluating AI algorithm performance, a criterion for determining whether the performance is adequate should also be prepared. This can be done by comparing the stand-alone performance of an AI algorithm with an absolute criterion (e.g., 90% or higher accuracy), which may involve ambiguity. It may be more intuitive to prepare a control method against which you can check the performance of an AI algorithm. These may include comparing AI with existing similar AI algorithms, other tests, or medical professionals as the control, as well as comparing medical professionals using AI with those not using AI as the control. While the focus of former is on the performance of AI algorithms themselves, the latter reflects the role of AI as an auxiliary tool providing information to medical professionals. The medical professionals may include both experienced and unexperienced doctors, and more useful information will likely be derived when the comparative results are analyzed according to their level of experience. In a relative comparison with a control, it is necessary to set the performance-difference criteria (e.g., less than 5%) at which the two compared results are considered statistically equivalent or significantly different. The passing criterion for absolute performance evaluation or for the relative comparison with a control cannot be set uniformly; it must be set according to the function of each AI algorithm and each clinical setting. Once the criterion is established, the sample size needed for performance evaluation can be calculated using well-known statistical methods [36-39].

https://doi.org/10.3348/kjr.2021.0048

Korean Journal of Radiology
KJR

Evaluation of the Clinical Utility of AI Algorithms

High accuracy does not necessarily mean an AI algorithm can improve clinical outcomes. AI algorithms are computerized aids that provide information to medical professionals to assist them in the clinical decision process. For any computerized tool to be useful, how the tool is integrated into the workflow is critical besides its performance. An AI algorithm must deliver information to the right person in the right way. Likewise, the coordinating doctor's response to the information from AI and the actions taken greatly affect the outcomes of patient care. Since the final clinical outcomes are achieved through the therapeutic or prophylactic actions taken based on diagnostic decisions, if no therapeutic or prophylactic actions are taken, no effects are made to the clinical outcomes. On the other hand, therapeutic or prophylactic actions can also bring about adverse reactions in some patients. Therefore, it is crucial to directly assess the effect of AI on clinical outcomes apart from its performance [40]. Such an evaluation is called validation of 'clinical utility.' Utility and efficacy are not interchangeable terms. Technical performance, clinical performance, and clinical utility are all indicators of efficacy of different levels and characters.

An example of validating clinical utility is provided in a study on an AI algorithm developed in the United Kingdom that monitors and analyzes the uterine contractions of a woman in labor and the heartbeat of the fetus and sends a real-time alert to the doctor when a fetal problem is suspected [41]. After verifying the accuracy of the algorithm, the investigators randomly assigned high-risk women in labor into AI-aided and AI-unaided groups to compare the outcomes of care [41]. Although the AI algorithm showed high accuracy in recognizing abnormal heartbeats, no significant difference in clinical outcome was observed between the experimental and control groups for both fetuses and mothers. As a result, the study could not demonstrate the clinical utility of the AI algorithm in terms of medical benefits to patients.

As shown in this example, validating the clinical utility of AI algorithms involves determining any differences in patient outcomes between AI-aided and AI-unaided patient care. Ideally, a randomized clinical trial should be conducted to prevent the effect of confounding variables in the intergroup comparison. However, since randomized clinical trials are not always possible, the results of prospective or retrospective observational research adjusted for confounding variables may be used. There are also AI algorithms that may be clinically validated sufficiently via performance validation alone without clinical utility validation. The appropriate level of clinical validation tailored to individual AI algorithms and clinical settings should be determined by medical experts. Table 2 provides a few examples of randomized controlled trials examining AI algorithms [41-46].

Clinical Validation of AI Algorithms from the Viewpoint of Device Approval and Insurance Coverage

Device approval, issued by entities such as the Korean Ministry of Food and Drug Safety (MFDS), the US Food and Drug Administration, and the European Commission (CE Marking), and decisions surrounding insurance coverage involve not only scientific principles but also sociopolitical factors. Therefore, AI device approval and insurance coverage may vary according to country, period, and social factors, and they cannot be explained from one perspective. This paper provides scientific principles alone related to AI device approval and insurance coverage. For the definitions of technical performance, clinical performance, clinical utility, diagnostic case-control study, and diagnostic cohort study, see explanations in the corresponding parts of this article.

Differences between Pharmaceuticals and Diagnostic Devices

Most medical AI algorithms are diagnostic devices and thus subject to the process of diagnostic device approval and insurance coverage. For a proper understanding of AI algorithms in this context, it is useful to clarify the difference between diagnostic devices and pharmaceutical agents.

For the approval of a pharmaceutical agent, it is generally necessary to prove, in a phase III clinical trial, that a given pharmaceutical agent improves patient treatment outcomes when used within a specific patient population. In other words, the clinical utility should be demonstrated for a particular indication. In contrast, approval of diagnostic devices does not require high-level clinical evidence applied to pharmaceutical agents, and generally focuses on technical performance validation. Of course, higher-level clinical validation data, if available, may enable a more thorough evaluation.

https://doi.org/10.3348/kjr.2021.0048

Table 2. Examples of Randomized Controlled Trials that Compared Practice with and without Artificial Intelligence Algorithms

Author	Algorithm	Patient	Primary Outcome
Wijnberge et al. [42]	Non-deep learning, machine learning algorithm that continuously analyzes arterial pressure waveform during surgery and warns if hypotensive event is expected within the next 15 minutes	Adult patients (≥ 18 years old) scheduled to undergo an elective noncardiac surgery under general anesthesia with need for continuous invasive blood pressure monitoring per arterial line	Time-weighted average of hypotension during surgery defined as hypotension below a mean arterial pressure of 65 mm Hg (in millimeters of mercury) x time spent below a mean arterial pressure of 65 mm Hg (in minutes) divided by total duration of operation (in minutes)
INFANT Collaborative Group [41]	Non-deep learning, machine learning algorithm that continuously analyzes cardiotocographic data and delivers color-coded alerts to physicians when abnormalities are noted	Women in labor who require continuous electronic fetal heart rate monitoring	Rate of poor neonatal outcome (intrapartum stillbirth or early neonatal death excluding lethal congenital anomalies, or neonatal encephalopathy, admission to the neonatal unit within 24 h for ≥ 48 h with evidence of feeding difficulties, respiratory illness, or encephalopathy with evidence of compromise at birth), and developmental assessment at age 2 years in a subset of surviving children
Repici et al. [43], Wang et al. [44], Wang et al. [45]	CNN-based CADe algorithm that detects polyps on colonoscopy images	Patients undergoing screening, surveillance, or diagnostic colonoscopy	Adenoma detection rate (percentage of patients with at least one histologically proven adenoma or carcinoma)
Wu et al. [46]	CNN-based algorithm that monitors occurrence of blind spots during esophagogastroduodenoscopy examination	Patients undergoing esophagogastroduodenoscopy	Rate of blind spots (number of unobserved sites/views from a total of 26 different sites/views in a patient as defined by the investigators) during endoscopic examination

CADe = computer-aided detection, CNN = convolutional neural network

Insurance coverage is an act that an insurer pays for medical services (i.e., the use of pharmaceutical agents or medical devices) delivered to policyholders (i.e., patients) who pay premiums. Therefore, it is important to demonstrate the clinical utility (i.e., patient benefit) of the medical services provided. Given that a medical service can be useful to one patient and useless to another, its indications must be specified when applying insurance coverage. Coverage of a medical service provided to patients (i.e., payment for medical fees indirectly by patients through insurance premium) in whom clinical utility has not been proven is unusual and unreasonable. In the case of therapeutic agents, the conditions for insurance coverage are essentially the same as those for their approval. Therefore, after approval, reasonably priced drugs are generally automatically covered by insurance. For diagnostic devices, however, approval is usually issued after technical performance validation, falling short of the requirements for clinical utility validation needed for insurance coverage. For this reason, device approval is not automatically associated with insurance coverage. A diagnostic device approved by regulatory agencies can be marketed and used for clinical practice. Later, if further clinical testing reveals clinical conditions in which the device is beneficial for patients, insurance coverage may include these indications. For example, even if MRI and ultrasonography have been approved by their proven technical performances, they are not covered by insurance until their use has proven beneficial for patient care in more specified clinical conditions and patient populations.

Device Approval for AI Algorithms

For device approval of an AI algorithm, its technical performance validation must be sufficiently documented at least. An AI algorithm's technical performance validation can be performed through external validation such as diagnostic case-control study. A prospective study is advantageous, if possible, but a retrospective study can also provide technical performance validation of an AI algorithm, subject to the availability of a validation

https://doi.org/10.3348/kjr.2021.0048

Korean Journal of Radiology
KJR

data set containing well-distributed examples of various difficulty levels matching its purpose. The conditions under which the AI algorithm operates well should be clarified and documented during the technical performance validation (e.g., devices and image acquisition methods, etc. that work well with the AI). It should be noted that AI device approval is merely permission to use the device on patients and to bring it to the market for that purpose. In other words, although a certain level of safety and efficacy of the AI should be demonstrated (typically through technical performance validation), AI device approval does not indicate whether the AI device is beneficial or valuable for patient care [25,26,47]. Medical professionals involved in patient care should conduct further clinical validation and evaluation of the approved AI device to verify its clinical utility and ensure its safe and efficacious clinical application [47,48]. Moreover, as it is difficult to investigate all matters related to generalizability of the AI algorithm during the device-approval process, it is important to further clarify the circumstances under which the AI output is accurate/inaccurate.

These scientific principles are adopted in the Guidelines for Big Data- and AI-based Medical Device Approval revised and released by the MFDS for the general public. These guidelines state that the sample data used in clinical investigations for an AI-based device seeking device approval should comprise independent data sets other than those used during the product development process. In other words, external validation is required. Use of reliable retrospective data sets are allowed for the validation for device approval by MFDS, if appropriate. Applicants may decide whether a prospective or retrospective clinical study design is suitable for the product.

Insurance Coverage for AI-Based Medical Device

Regarding insurance coverage for AI-based devices, the Health Insurance Review and Assessment Service under the Korean Ministry of Health and Welfare released the Guidelines for the Evaluation for Medical Insurance Coverage for Innovative Medical Technology in December 2019. These guidelines added some flexibility to the scientific principles of medical insurance coverage. They state that, when improved patient outcomes or significant improvement in diagnostic accuracy with the use of an AI-based device compared to conventional care is verified, extra compensation through insurance coverage may be considered (the demonstration of cost-effectiveness is also included in the guidelines; we omit it because it is beyond the scope of this paper). While results from a prospective or retrospective research on patient outcomes adjusting for confounding variables or a randomized clinical trial is recommended for the evaluation, a diagnostic cohort study may also be accepted on a case-by-case basis for external validation of the clinical performance of AI. That is, for insurance coverage, clinical utility should be demonstrated in the form of improved patient outcomes; however, under certain circumstances, demonstration in a diagnostic cohort study of a significant improvement in diagnostic accuracy with the use of an AI-based device for a specific clinical condition/patient population may also satisfy the conditions for insurance coverage.

CONCLUSIONS

We examined the key principles of clinical validation, device approval, and insurance coverage of AI algorithms for medical diagnosis/prediction. When evaluating the discrimination performance of AI, the Dice similarity coefficient, sensitivity, specificity, ROC curve, and FROC curve are widely used. In the case of an AI algorithm presenting probability directly, calibration performance should be evaluated as well. Most currently available AI algorithms for medical diagnosis and prediction have limited generalizability to real-world healthcare settings of their performance shown in their development stage or through internal validation. This highlights the importance of external validation of performance in the clinical validation of AI algorithms. It should also be considered that AI is generally not meant to serve as a stand-alone tool. It is an auxiliary tool providing information to the medical professional. For external validation of AI performance, diagnostic case-control and diagnostic cohort studies may be conducted. The former evaluates the technical performance of an AI algorithm, while the latter evaluates the clinical performance in samples representing the target patients in real-world clinical scenarios. The ultimate clinical validation of an AI algorithm lies in the evaluation of its effect on patient outcomes. A randomized clinical trial is ideal for this validation of clinical utility. AI-device approval generally focuses on technical performance validation. Therefore, it is not used to determine whether the AI is beneficial for patient care and improves patient outcomes. Also, it is difficult to investigate all matters related to generalizability of the AI algorithm during the

 https://doi.org/10.3348/kjr.2021.0048

device-approval process. After achieving device approval, it is up to medical professionals to determine whether the approved AI algorithms are beneficial for real-world patient care. To obtain insurance coverage, it is essential to demonstrate clinical utility in the form of improved patient outcomes. As the use of AI algorithms for medical diagnosis/prediction is likely to increase in the future, the topics discussed herein should be introduced into medical-school curriculums [49].

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Acknowledgments

This article is a republication of the original paper published in Korean in the Journal of the Korean Medical Association (J Korean Med Assoc 2020;63:696-708), translated into English with the original publisher's consent.

Author Contributions

Conceptualization: Seong Ho Park. Funding acquisition: Jaesoon Choi. Writing—original draft: Seong Ho Park. Writing—review & editing: Jaesoon Choi, Jeong-Sik Byeon.

ORCID iDs

Seong Ho Park
https://orcid.org/0000-0002-1257-8315
Jaesoon Choi
https://orcid.org/0000-0002-6817-618X
Jeong-Sik Byeon
https://orcid.org/0000-0002-9793-6379

REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56
2. Park SH, Lim TH. Artificial intelligence: guide for healthcare personnel. Seoul: Koonja, 2020
3. Do S, Song KD, Chung JW. Basics of deep learning: a radiologist's guide to understanding published radiology articles on deep learning. *Korean J Radiol Korean J Radiol* 2020;21:33-41
4. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. *Korean J Radiol* 2019;20:749-758
5. Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol* 2004;5:11-18
6. 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
7. Tajbakhsh N, Gurudu SR, Liang J. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. Proceedings of 2015 IEEE 12th International Symposium on Biomedical Imaging; 2015 Apr 16-19; New York, USA: IEEE; 2015; p. 79-83
8. Moskowitz CS. Using free-response receiver operating characteristic curves to assess the accuracy of machine diagnosis of cancer. *JAMA* 2017;318:2250-2251
9. Chakraborty DP. Welcome to Prof. Dev Chakraborty's FROC methodology. Devchakraborty.com Web site. http://www.devchakraborty.com/. Published 2019. Accessed September 14, 2020
10. Mutasa S, Sun S, Ha R. Understanding artificial intelligence based radiology studies: what is overfitting? *Clin Imaging* 2020;65:96-99
11. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195
12. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018;15:e1002683
13. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017;318:2211-2223
14. Ridley EL. Deep-learning algorithms need real-world testing. Auntminnie.com Web site. https://www.auntminnie.com/index.aspx?sec=nws&sub=rad&pag=dis&ItemID=123871. Published 2018. Accessed September 14, 2020
15. 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
16. Lee JH, Joo I, Kang TW, Paik YH, Sinn DH, Ha SY, et al. Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. *Eur Radiol* 2020;30:1264-1273
17. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. *Lancet Digit Health* 2019;1:e157-e159
18. Park SH, Kim YH, Lee JY, Yoo S, Kim CJ. Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review. *Sci Ed* 2019;6:91-98
19. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D,

Harvey H, et al. Preparing medical imaging data for machine learning. *Radiology* 2020;295:4-15
20. Adamson AS, Welch HG. Machine learning and the cancer-diagnosis problem-no gold standard. *N Engl J Med* 2019;381:2285-2287
21. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. *Radiology* 2020;294:487-489
22. Mehta MC, Katz IT, Jha AK. Transforming global health with AI. *N Engl J Med* 2020;382:791-793
23. Nevin L; PLOS medicine editors. Advancing the beneficial use of machine learning in health care and medicine: toward a community understanding. *PLoS Med* 2018;15:e1002708
24. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. *JAMA Netw Open* 2018;1:e182658
25. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science* 2019;363:810-812
26. Park SH, Do KH, Choi JI, Sim JS, Yang DM, Eo H, et al. Principles for evaluating the clinical implementation of novel digital healthcare devices. *J Korean Med Assoc* 2018;61:765-775
27. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019;26:1651-1654
28. Yu KH, Kohane IS. Framing the challenges of artificial intelligence in medicine. *BMJ Qual Saf* 2019;28:238-241
29. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Can Assoc Radiol J* 2018;69:120-135
30. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean J Radiol* 2019;20:405-410
31. Kim DW, Jang HY, Ko Y, Son JH, Kim PH, Kim SO, et al. Inconsistency in the use of the term "validation" in studies reporting the performance of deep learning algorithms in providing diagnosis from medical imaging. *PLoS One* 2020;15:e0238908
32. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A clinician's guide to artificial intelligence: how to critically appraise machine learning studies. *Transl Vis Sci Technol* 2020;9:7
33. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271-e297
34. Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. *Radiology* 2019;290:272-273
35. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem* 2005;51:1335-1341
36. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA* 2014;312:1342-1343
37. Eng J. Sample size estimation: how many individuals should be studied? *Radiology* 2003;227:309-313
38. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998;7:371-392
39. Ahn S, Park SH, Lee KH. How to demonstrate similarity by using noninferiority and equivalence statistical testing in radiology research. *Radiology* 2013;267:328-338
40. Liu X, Cruz Rivera S, 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. *Nat Med* 2020;26:1364-1374
41. INFANT Collaborative Group. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. *Lancet* 2017;389:1719-1729
42. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020;323:1052-1060
43. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* 2020;159:512-520
44. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020;5:343-351
45. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68:1813-1819
46. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* 2019;68:2161-2169
47. Park SH. Regulatory approval versus clinical validation of artificial intelligence diagnostic tools. *Radiology* 2018;288:910-911
48. Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. *JAMA* 2020 Sep [Epub]. https://doi.org/10.1001/jama.2020.9371
49. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? *J Educ Eval Health Prof* 2019;16:18