인공지능의 임상검증, 의료기기 허가, 의료보험 적용과 관련된 핵심 원리

박 성 호¹ · 최 재 순² · 변 정 식³

¹울산대학교 의과대학 서울아산병원 영상의학과

²울산대학교 의과대학 서울아산병원 의공학과

³울산대학교 의과대학 서울아산병원 소화기내과

Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence

Seong Ho Park, MD¹ · Jaesoon Choi, PhD² · Jeong-Sik Byeon, MD³

¹Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

²Department of Biomedical Engineering, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

³Department of 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.

Key Words: Software validation; Device approval; Insurance coverage; Artificial intelligence

Received: September 22, 2020 Accepted: October 21, 2020

Corresponding author: Seong Ho Park
E-mail: seongho@amc.seoul.kr

© Korean Medical Association

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

서론

인공지능 기술이 현재 의료의 한계점들을 보완하고 의료를 발전시키는데 많은 도움을 줄 것으로 큰 기대를 받고 있으며 의료에 많은 영향을 미칠 것으로 예상된다[1,2]. 모든 의약품이나 의료기기들은 환자에게 사용하기에 앞서 안전

Figure 1. Dice similarity coefficient.

성과 유효성에 대해서 엄격한 검증 절차를 거쳐야 하며, 인공지능 의료기기도 예외가 아니다. 인공지능 의료기기는 다양한 기능과 형태로 만들어질 수 있고 기능과 형태에 따라 임상검증 방법도 부분적으로 달라지지만, 대다수의 인공지능 의료기기는 컴퓨터 보조진단(computer-aided detection [CAdE]/computer-aided diagnosis) 또는 임상 의사결정보조시스템(clinical decision support system)과 같이 의학적 진단, 판단 및 예측을 보조하는 기능을 하는 알고리즘들로 진단용 의료기기에 속한다. 따라서, 이들에 대한 임상검증 방법은 통상의 진단검사들의 임상검증 방법과 비슷한 면이 많다. 이 논문은 이러한 의료 인공지능 알고리즘의 임상검증, 의료기기 허가, 의료보험 적용과 관련된 핵심 원리를 설명하고자 한다.

인공지능 알고리즘의 성능 지표들

인공지능 알고리즘의 성능 평가에는 다양한 지표들이 이용될 수 있다. 이들 중 일부는 단지 기술적인 지표들로 의학적 연관성이 낮으며, 일부는 소수의 특정 상황에만 적용이 된다. 따라서, 모든 지표들을 이 논문에 다 열거하는 것은 효과적이지 않고, 이 논문은 의학적 연관성이 높고 많이 활용되는 지표들에 대하여 설명한다.

1. 다이스 유사성 계수

영상에서 장기나 병변을 분할하는 인공지능 알고리즘의 성능평가에 이용된다[3]. Figure 1에 정의가 설명되어 있다.

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

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

예로서, 전립선 자기공명영상에 전립선 암으로 의심되는 부분을 표시해 주는 인공지능이 있다면, 병리학적으로 확인된 실제 암 부위와 인공지능이 암으로 판단한 영역이 얼마나 겹치는지를 평가하여 인공지능의 성능을 평가할 수 있다. 다이스 유사성 계수(Dice similarity coefficient)와 비슷한 여러 가지 다른 이름의 계수들도 있다.

2. 민감도, 특이도, 수신자판단특성곡선

Figure 2처럼 인공지능 알고리즘이 결과를 이분형(예: 질환 있음 대 없음)으로 제시할 경우 일반 진단검사와 마찬가지로 실제 질환이 있을 때 인공지능 알고리즘이 질환이 있다고 판단하는 비율인 ‘민감도 = true positive / (true positive + false negative)’와 실제 질환이 없을 때 인공지능 알고리즘이 질환이 없다고 판단하는 비율인 ‘특이도 = true negative / (false positive + true negative)’로 성능을 기술할 수 있다. 인공지능 알고리즘이 결과를 이분형으로 제시한다고 하더라도 실제 내부적으로는 먼저 연속적인 수치로 결과를 출력하고(가령, 확률처럼 0에서 1까지 중 어떤 소수 값으로 출력) 이 출력 값에 임계 값을 적용하여 이분형 결과를 만들게 된다. 따라서 임계 값을 어떻게 정하는지에 따라 인공지능 알고리즘의 민감도와 특이도가 달라지게 된다. 임계 값을 높게 잡으면 민감도는 낮아지나 특이도는 증가하고 임계 값을 낮게 잡으면 민감도가 증가하나 특이도는 낮아진다. 수신자판단특성곡선은 임계 값을 변화시키며 민감도를 y축 값 1-특이도를 x축 값으로 해서 그린 그래프이다(Figure 3) [4]. 수신자판단특성곡선영역(area under the curve 또는 area under the receiver operating characteristics curve)

Figure 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 JS et al. Korean J Radiol 2019;20:749-758, with permission from the Korean Society of Radiology [4]. AUC, area under the curve.

값은 그래프 아래의 면적으로 모든 가능한 임계 값에 대한 평균 민감도 또는 평균 특이도의 의미이다. 따라서, 최대 1의 값을 갖고, 값이 클수록 이론적으로는 높은 성능을 의미한다. 하지만, 수신자판단특성곡선영역 값이 크다고 해서 인공지능 알고리즘을 실제 사용할 때 반드시 성능이 높은 것은 아니므로 주의가 필요하다. 인공지능 알고리즘을 실제 사용할 때는 임계 값을 정해야 하고 수신자판단특성곡선영역 값이 아닌 정해진 임계 값에서 나오는 민감도와 특이도가 실제 성능이 된다. 수신자판단특성곡선영역 값은 앞에서 설명한 대로 평균적 민감도 또는 특이도 값일 뿐이다. 보다 자세한 내용은 관련 문헌을 참고하기 바란다[2,5,6].

3. Free-response 수신자판단특성곡선

대장내시경 영상으로부터 대장용종을 검출하는 인공지능 알고리즘과 같이 CADe 기능의 인공지능 알고리즘에 대한 성능을 평가할 때 사용한다. 병변이 있음 대 없음의 판별뿐 아니라 병변의 위치도 맞게 찾아야 인공지능 알고리즘의 결과가 맞는 것이 된다. 대장용종을 찾는 인공지능 알고리즘이 대장용종을 가진 환자에서 용종이 있다는 결과를 제시하더라도 실제 용종이 있는 부위를 가리켜 용종이 있다고 할 때 맞는 결과를 출력한 것이지 용종은 발견하지 못하고 용

Figure 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 N et al. Proceedings of IEEE 12th International Symposium on Biomedical Imaging. New York: IEEE; 2015, with permission from IEEE [7]. CNN, convolutional neural network.

종이 없는 부위에 용종이 있다는 결과를 출력하였다면 위음성과 위양성을 동시에 만든 것이 된다. CADe 기능 알고리즘이 적용되는 상황 들에서는 보통 한 환자에 여러 개의 병변이 있을 수 있으며 위양성도 여러 개가 발생할 수 있다. 이러한 경우에는 민감도와 특이도가 아닌 민감도와 위양성의 개수로 성능을 평가하는 것이 좀 더 적절하다. 인공지능 알고리즘이 내부적으로 출력하는 연속적인 수치에 대해서 임계 값을 높게 잡으면 민감도가 낮아지고 위양성의 개수도 작아지며 임계 값을 낮게 잡으면 민감도가 증가하지만 위양성의 개수도 증가된다. Free-response 수신자판단특성곡선은 y축 값을 민감도 그리고 x축 값을 1-특이도 대신 위양성의 개수의 평균을 이용하여 그린 그래프이다(Figure 4) [7]. 평균 위양성 개수는 환자당 평균, 영상당 평균 등 여러 방식으로 계산할 수 있으며 상황에 적합한 방법을 이용한다. 이와 약간 다른 변형된 형태의 free-response 수신자판단특성곡선들도 있다. 보다 자세한 내용은 관련 문헌을 참고하기 바란다[8,9].

4. 보정정확도

앞에서 설명한 성능 지표들은 모두 판별정확도 지표들이다. 반면, 보정정확도는 인공지능 알고리즘이 확률을 제시할 때(예: 인공지능이 어떤 병변이 “암일 확률이 00%다” 식

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

CNN, convolutional neural network; AUC, area under the curve.

으로 결과를 제시하는 경우) 확률이 얼마나 정확한지를 말하는 것으로 별도의 평가방법이 필요하다[6]. 베이즈 정리(Bayes' theorem)에 따라 실제 확률은 유병률이라고도 말하는 검사 전 확률에 크게 영향을 받기 때문에, 인공지능이 검사 전 확률을 고려하지 않고 기계적으로 제시하는 확률은 많은 경우 정확하지 않을 가능성이 높다. 따라서, 사용자에게 확률을 직접 제시하는 인공지능 알고리즘에 대해서는 보정 정확도도 잘 평가하여야 하는데 보정정확도가 간과되는 경우가 많아 주의가 필요하다[6]. 보정정확도에 대한 자세한 설명은 이 논문의 범위를 넘으므로 관련 문헌을 별도로 참고하기 바란다.

의료 인공지능 알고리즘 성능의 일반화에 대한 취약성

1. 의료 인공지능 알고리즘의 과적합 현상

현재 인공지능 기술의 대표인 딥러닝(deep learning)과 같이 수학적으로 매우 복잡한 형태의 알고리즘은 데이터에 대한 의존성이 높아 학습데이터에서는 매우 높은 정확도를 보

이나 학습에 사용되지 않은 외부데이터에서는 성능이 저하되는 현상을 보인다. 이를 기술용어로 ‘과적합’ 현상이라 이야기한다[10]. 의료 인공지능 알고리즘 등에서 과적합 현상이 특히 두드러지게 나타나는 것으로 알려져 있다. 과적합을 줄이기 위해 규칙화라고 통칭되는 여러 기술적인 방법들을 이용할 수 있으나 규칙화만으로는 의료 분야의 인공지능 알고리즘들에 나타나는 과적합 문제를 해결하기 어려운 경우가 많다. 이러한 이유로, 현재 의료 분야 인공지능 알고리즘들 중 대다수는 일반화가 되지 못하는 것으로 알려져 있다[11]. 의료 인공지능 알고리즘 성능의 일반화에 대한 취약성을 보여주는 몇 가지 예를 Table 1

에 정리해 놓았다[12–16]. 이 예들에서 보듯이, 인공지능 알고리즘을 실제 진료환경에 적용하였을 때 정확도가 감소하거나, 확률이 잘 맞지 않게 되고, 인공지능의 내부 출력 값을 최종 결과로 변환하기 위한 임계 값이 맞지 않게 되는 일이 발생한다.

2. 의료 인공지능 알고리즘에서 과적합이 많이 발생하는 이유

과적합으로 인해 의료 인공지능 알고리즘의 성능이 일반화되기 어려운 현상이 발생하는 것은 인공지능 알고리즘의 학습에 사용된 데이터가 인공지능을 사용하려는 임상현장의 상황을 충분히 반영하지 못하기 때문이다[11,17–19]. 이에 여러 요인들이 관여한다. 먼저, 의료데이터는 이질성이 매우 높다. 같은 질환을 가지고 있는 환자라 하더라도 의료기관의 규모, 종별, 위치 등에 따라 환자의 연령, 성별, 질환의 중증도, 기저질환 및 동반질환 등 세부특징에 있어 차이가 있는 경우가 많다. 어떤 질환이 의심되지만 실제로는 질환을 가지고 있지 않은 환자 등에서 발견되는 유사질환이나 감별질환의 종류 및 분포 등도 의료기관마다 다른 경우가 많다. 또한, 질환의 유병률도 다양하게 나타날 수 있다. 즉, 한 병원의 상

황을 다른 병원에 그대로 적용할 수 없는 경우가 많다. 의료 기관들이 사용하는 의료장비가 다양한 것도 추가의 이질성 요소이다(가령, 전산화단층촬영이나 자기공명영상 같은 영상 장비의 경우 한 제조사의 영상에 맞춰진 인공지능 알고리즘이 다른 제조사의 영상장비로 얻은 영상에는 잘 작동하지 않을 수 있다). 아울러, 의료의 발전에 따라 치료제나 진단법이 지속적으로 변경되는 것도 의료분야 데이터의 독특한 특징이어서 시간적인 이질성을 만든다(가령, 과거에 쓰던 치료제 데이터를 포함하여 만든 인공지능 알고리즘은 치료제가 바뀐 상황에서는 올바르게 작동할 수 없을 것이며, 구형 영상장비 영상으로 학습시킨 인공지능은 신형 영상장비 영상에는 잘 작동하지 않을 수 있다). 이러한 이질성을 극복하려면 기본적으로 대량의 데이터를 가능한 많은 병원으로부터 수집하여 인공지능을 학습시켜야 하는데, 인공지능 학습을 위해 의료데이터를 정리하는 과정은 많은 경우 아무나 할 수 없고 전문 의료인의 많은 노력과 시간을 필요로 하기 때문에 데이터 수집·정리가 어렵다. 또한, 많은 물적 자원이 소요된다. 따라서, 데이터의 이질성은 매우 높는데 역설적으로 인공지능 학습에 사용할 데이터의 양은 오히려 작아서 인공지능 학습에 사용된 데이터가 인공지능을 사용하려는 임상현장을 충분히 반영하지 못하는 일이 빈번하게 발생된다. 또한, 현실의 의료데이터는 기본적으로 회색 지대와 다양한 잡음 요소가 많이 포함되어 있다. 어떤 질병 상태의 유·무에 대한 분명한 참조 표준을 구할 수 없는 경우가 있고, 질병 상태의 유·무가 숫자 0과 1처럼 구분되는 것이 아닌 연속적 변화단계의 단지 일부분일 수 있는데, 인공지능 학습에는 질병 상태의 유·무를 명확히 말할 수 있는 데이터들이 선택적으로 많이 이용된다[20]. 또한, 컴퓨터가 잘 처리할 수 있는 잡음 요소가 제거된 데이터들이 선택적으로 많이 이용된다.

3. 일반화에 대한 취약성이 갖는 의미

먼저, 인공지능 알고리즘의 성능 검증에 있어 이 논문의 뒤에서 설명하는 외부데이터를 이용한 외부검증이 매우 중요하다[6,10,11,21-29]. 인공지능 알고리즘이 실제 진료환경에서는 개발환경에서와는 다른 성능을 보일 수 있기 때문에, 실제 어떤 진료환경에서 인공지능이 얼마나 정확하게 작

동할지 알기 위해서는 직접 해당 진료환경에서 외부검증으로 확인을 해 보는 것이 가장 좋다. 그럼에도 불구하고 인공지능 알고리즘들에 대한 외부검증이 충분히 이루어지지 않아 문제가 되고 있다[30]. 또한, 인공지능의 결과를 이용할 때는 결과를 기계적으로 받아들이지 않고 임상적 상황 및 다른 임상정보를 적절히 함께 고려하여 의료진이 최종 판단을 내려야 한다. 앞에서 설명한 임계 값도 임상적 상황에 따라 적절히 조절하여야 한다. 이러한 이유로, 성능이 높은 인공지능이 예외적으로 일부 제한된 조건에서 특정 기능에 대하여 의료인을 대신할 수는 있겠지만, 기본적으로 인공지능은 의료인을 대신하는 독립적 도구가 아니라 의료인에게 정보를 제공하는 유능한 보조도구의 역할을 한다. 임상검증도 인공지능의 이러한 역할을 고려하여 계획하는 것이 바람직하다. 추가로, 인공지능 알고리즘을 개별 병원이나 특정 진료환경에서 사용하기 전에 개별 병원과 특정 진료환경의 데이터를 이용하여 추가로 학습시킴으로써 일반화에 대한 취약성을 보완할 수 있을 것이다. 반면, 오류나 편향이 있는 자료들로 추가의 학습을 시킨다면 알고리즘이 원래 가지고 있던 정확도를 오히려 훼손하게 되어 문제가 될 우려도 있다. 인공지능 알고리즘은 고정된 알고리즘과는 달리 지속적인 학습을 통해 변할 수 있기 때문에 이 논문의 뒤에서 설명하는 의료기기 허가 이후에도 지속적인 평가·관리가 필요하다. 현재의 의료기기 허가 제도와 체계는 이러한 지속적 평가·관리에 대한 구체적 고려가 없어 향후 보강이 필요하다.

인공지능 알고리즘의 성능 검증방법: 검증 데이터에 따른 분류

인공지능 알고리즘의 성능 검증방법을 검증에 사용되는 데이터의 성격에 따라 나누어 볼 수 있다. 먼저 한 가지 용어 정리가 필요하다. 검증에 해당하는 일반적인 영어단어는 ‘validation’이다. 그런데, 인공지능 분야에서는 이 단어가 인공지능 알고리즘을 만드는 과정에 있어 hyperparameter라 불리는 조건들을 미세조정하는 과정을 가리키는 전문기술용어로도 쓰인다[31]. 혼선을 피하기 위해 이 미세조정 과정을

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

‘tuning’이라 표현하기도 하지만 인공지능 관련 문헌들은 이미 전문기술용어로 정해져 있는 ‘validation’이란 표현을 더 많이 사용해 오고 있다[31]. 알고리즘 미세조정과 알고리즘 성능 검증 사이에 혼선을 피하기 위해 성능 검증을 ‘test’라는 용어로도 표현하고 아예 ‘validation test’라고 두 단어로 표현하기도 한다[32,33].

1. 내부검증

내부검증은 일반적으로 인공지능 알고리즘의 성능을 과대 평가한다. 따라서, 내부검증은 인공지능 알고리즘 성능을 검증하기 위한 목적보다는 인공지능 알고리즘을 만들면서 성능을 약식으로 확인해 보는 정도의 역할을 한다. 인공지능 알고리즘의 성능 검증에는 외부검증이 중요하며, 내부검증 결과는 외부검증의 성능 결과와 비교하는 참고자료로 활용된다. 교차검증과 분할데이터를 이용한 검증이 내부검증에 포함된다.

1) 교차검증

교차검증의 대표적인 예는 k-fold 교차검증이다[32]. 전체 데이터를 k개의 조각으로 나누고 한 조각의 데이터를 빼고 나머지의 데이터로 인공지능 알고리즘을 만들고 빼놓은 한 조각의 데이터에서 성능을 확인하는 과정을 순차적으로 한 조각씩 바꾸어 빼가면서 k번 반복한 후 k개의 성능 결과에 대해서 평균을 구하는 방식이다. 데이터가 부족할 때 인공지능 알고리즘의 성능을 조금이라도 확인해 보려는 목적으로 사용해 볼 수 있으나 사실상 모든 데이터가 인공지능 알고리즘을 만드는데 다 사용되므로 성능 검증을 위한 방법

으로는 적절하지 않다.

2) 분할데이터를 이용한 검증

분할데이터를 이용한 검증은, 전체 데이터를 세 조각으로(training set, tuning set, test set) 나누고 test set은 인공지능 알고리즘의 학습(training) 및 미세조정(tuning)에 쓰지 않고 따로 두었다가 완료된 인공지능 알고리즘의 성능을 평가하는 데만 사용하는 것을 말한다(Figure 5) [32]. 이때 데이터를 무작위로 분할할 수도 있고, 데이터수집 시기 등에 따라 분할할 수도 있다. 분할데이터를 이용하는 방법도 내부검증 방법이지만 교차검증보다는 좋은 방법이다.

2. 외부검증

외부검증은 처음 수집된 데이터를 알고리즘 성능 검증에 전혀 사용하지 않고, 별도로 독립적으로 수집된 데이터를 이용하여 인공지능 알고리즘의 성능을 평가하는 것을 말한다(Figure 5). 학습에 사용된 데이터를 제공한 의료기관이 아닌 외부 의료기관의 데이터를 이용하는 것이 대표적인 방법이다. Figure 2에 보듯이 인공지능 알고리즘의 성능을 평가하기 위해서는 인공지능 알고리즘이 진단·예측하고자 하는 상태를 갖고 있는 데이터와 갖고 있지 않은 데이터가 모두 필요하다. 이러한 검증 데이터를 모으는 방법에 따라 외부검증 연구는 크게 진단적 환자-대조군 연구와 진단적 코호트 연구로 나눌 수 있다[34].

1) 진단적 환자-대조군 연구

진단적 환자-대조군 연구는, 인공지능 알고리즘이 진단·예측하고자 하는 상태를 갖고 있는 데이터와 갖고 있지 않은 데이터를 따로 수집하는 연구설계이다. 예를 들어, 흉부 엑스레이영상을 분석해서 폐암이 있는지 없는지를 구분하는 인공지능 알고리즘의 성능을 평가할 때 폐암이 있는 영상들(case)과 없는 영상들(control)을 각각 일정 숫자만큼 수집하는 방식이다. 이렇게 데이터를 수집하면 유병률(이 예에서는 전체 데이터 중 폐암이 있는 데이터의 비율)이 실제 임상상황에서 보는 자연스러운 유병률이 아니라 임의로 지정된다.

또한, 폐암이 있는 영상들과 없는 영상들을 어떻게 수집하는가에 따라 포함되는 폐암의 크기 및 모양의 다양성 정도, 폐암이 없는 환자들에서 폐암 유사 병변의 유무나 정도, 폐암의 발견에 영향을 줄 수 있는 여러 상태나 기저질환 및 동반 질환 등의 유무나 정도 등 여러 조건들이—이러한 조건들을 통틀어 스펙트럼(spectrum)이라 한다—변할 수 있다. 일반적으로, 선택편향으로 인해 실제 임상상황에서 보는 자연스러운 스펙트럼과 차이가 있는 데이터 구성이 되는 경우가 많다[35]. 인위적 스펙트럼과 인위적 유병률은 알고리즘의 성능 평가에 영향을 미친다[2,6].

2) 진단적 코호트 연구

진단적 코호트 연구는, 인공지능 알고리즘이 진단·예측하고자 하는 상태의 유·무와 무관하게 인공지능을 적용하고자 하는 대상군의 임상적 상황을 먼저 정의하고 이 정의에 따라 데이터를 수집하는 방법이다. 앞에서 예로 사용한 흉부 엑스레이영상에 폐암이 있는지 없는지를 구분하는 인공지능 알고리즘의 경우, 가령 “OO갑년 이상의 흡연력을 가진 55세 이상의 성인” 식으로 적합기준을 만들고 이를 만족하는 환자를 연속적으로 모두 모집하거나 또는 일부를 무작위로 모집하여 흉부 엑스레이영상 검사를 하고 이를 가지고 인공지능 알고리즘의 성능을 평가하는 방식이다. 모집된 환자 중 일부는 자연스럽게 진단·예측하고자 하는 상태를 가지고 있고 일부는 가지고 있지 않을 것이다. 이러한 방식으로 데이터를 수집하면 자연스런 스펙트럼과 유병률을 갖는 데이터를 수집할 수 있고, 여기서 얻어진 인공지능 알고리즘의 성능과 임계 값은 해당 적합기준에 해당하는 임상상황으로 보다 직접적으로 적용 가능하게 된다. 진단적 환자-대조군 연구가 인위적 실험적 상황에서 성능을 평가하는 것이라면 진단적 코호트 연구는 좀 더 실제의 임상적 상황에서 성능을 평가하는 것이다. 인공지능 알고리즘을 사용하고자 하는 실제 임상적 상황을 명확하게 파악하고 이 상황을 적절하고 충분히 반영할 수 있도록 적합기준을 구체적으로 잘 만드는 것이 중요하다.

3) 진단적 환자-대조군 연구와 진단적 코호트 연구의 차이점

진단적 환자-대조군 연구와 진단적 코호트 연구는 목적에 차이가 있다. 진단적 환자-대조군 연구는 인공지능 알고리

즘의 기술적 성능을 평가하는 것이 목적으로 인공지능 알고리즘이 진단·예측하고자 하는 상태와 그렇지 않은 상태를 구별하는데 기술적으로 좋은 성능을 갖고 있는지 어느 정도 전체적으로 평가하는 것이다. 전향적 연구라면 더 좋겠지만 후향적 연구라도 인공지능 알고리즘의 사용 목적에 부합하는 다양한 난이도의 증례들을 골고루 잘 포함하는 검증 데이터를 수집한다면 기술적 성능을 적절히 평가할 수 있다. 반면, 진단적 코호트 연구는 인공지능 알고리즘의 임상적 성능을 평가하려는 목적으로 특정 임상적 상황이나 환자군에서 인공지능 알고리즘이 얼마나 정확한지 알아보기 위한 것이다. 따라서, 진단적 코호트 연구는 진단적 환자-대조군 연구와 비교해서 인공지능 알고리즘을 실제 진료에 적용할 구체적인 대상자, 즉 적응증(indication), 개념이 보다 분명하고 구체적이다. 진단적 코호트 연구를 위해서는 전향적 연구가 바람직하다. 일반적으로, 진단적 환자-대조군 연구를 통해 인공지능 알고리즘의 기술적 성능을 먼저 확인하고 이어서 다음 단계로 진단적 코호트 연구를 수행하여 임상적 성능을 평가한다. 진단적 환자-대조군 연구는 진단적 코호트 연구에 비해서 성능을 높게 평가하는 경향이 있다.

4) 표준데이터를 이용한 검증

인공지능 알고리즘의 성능 검증을 위해 표준데이터를 만들어 검증 데이터로 이용하자는 의견이 있다. 표준데이터를 모으는 과정은 일반적으로 진단적 환자-대조군 연구의 데이터 수집 과정과 성격이 비슷한 경우가 많다. 따라서, 잘 구축된 표준데이터를 인공지능 알고리즘의 기술적 성능을 평가하는 목적으로 이용해 볼 수 있을 것이다. 특히, 여러 병원으로부터 데이터를 수집하여 표준데이터를 구축한다면 좀 더 효과적일 것이다. 표준데이터의 중요한 약점의 하나는 표준데이터에만 좋은 성능을 보이는 인공지능 알고리즘이 나올 수 있다는 것이다. 표준데이터를 대학수학능력시험에 비유해 볼 수 있는데, 기출문제를 그대로 사용할 경우 제대로 된 학생 실력 평가가 될 수 없기 때문에 매년 철저한 보안을 통해 새롭게 문제를 출제하여 시험을 치른다. 마찬가지로, 표준데이터도 계속 새롭게 유지하지 않는다면 기출문제를 노출하는 상황이 될 수 있다. 아울러, 표준데이터를 성공적으로 구축하기 위해서는, 정말 대표성이 있는 표준데이터가 되

기 위해서는 어떤 조건들을 만족해야 하는지 표준데이터가 인공지능 알고리즘의 기술적 성능을 충분하고 공정하게 평가할 수 있는지 등 보다 근본적인 사항들에 대한 세밀한 사전검토가 필요하다.

인공지능 알고리즘의 성능 기준

인공지능 알고리즘의 성능을 평가할 때, 적절한 성능인지 판단하기 위한 기준을 설정해야 한다. 인공지능 알고리즘의 단독 성능을 절대적 기준(예: 90% 이상의 정확도가 요구됨)과 비교할 수도 있고, 이 경우 적절한 절대적 기준에 대한 모호함이 있을 수 있다. 대조 방법을 정하고 인공지능 알고리즘의 성능을 대조 방법의 성능과 상대적으로 비교를 하는 것이 좀 더 직관적인 방법이 될 수 있다. 1) 기존의 유사 인공지능 알고리즘이나 검사법 또는 2) 의료진을 대조 방법으로 이용하거나 3) 인공지능을 사용하지 않은 의료진을 대조 방법으로 인공지능을 사용하는 의료진과 비교를 할 수도 있다. 이 중, 앞의 두 가지는 인공지능 알고리즘 자체의 단독 성능 평가에 중점을 둔 것이고 세 번째는 인공지능은 기본적으로 의료인에게 정보를 제공하는 보조도구의 역할을 한다는 점을 반영하는 비교이다. 경험 수준이 높은 의료진과 낮은 의료진을 모두 포함하고 경험 수준에 따라 나누어 결과를 분석하면 보다 유용한 정보를 얻을 수 있을 것이다. 상대적 비교를 할 때에도, 어느 정도의 차이(예: 5% 이내의 성능 차이)를 통계학적으로 동등하다고 또는 유의한 차이가 있다고 볼 것인지 기준 설정을 하여야 한다. 절대적 성능평가를 위한 기준이나 상대적 비교를 위한 차이의 기준을 일괄적으로 정할 수는 없으며 개별 인공지능 알고리즘의 기능과 각각의 임상적 상황에 맞추어 정하여야 한다. 기준이 설정되면 잘 알려져 있는 통계학적 방법에 따라 성능 검증에 필요한 표본 수 산출이 가능하다[36-39].

인공지능 알고리즘의 임상적 유용성 검증

어떤 인공지능 알고리즘의 성능이 높다고 해서 이 인공지능

능 알고리즘을 사용함으로써 반드시 진료결과가 향상되는 것은 아니다. 인공지능 알고리즘은 기본적으로 의료진에게 정보를 제공하는 보조전산도구로 의료진과 함께 진료과정에 포함되게 되는데, 어떤 전산도구가 정말로 도움이 되려면 인공지능 자체의 성능 못지않게 인공지능이 실제 의료진의 업무 방식 및 흐름과 어떻게 조화가 되어 인공지능이 제시하는 정보가 어떠한 형태로 누구에게 어떻게 전달되는지가 중요하다. 또한, 인공지능이 제시하는 정보에 대해 의료인이 어떻게 반응하고 어떤 행위를 하는지가 진료결과에 중요한 영향을 미친다. 최종 진료결과는 진단적 판단에 이어지는 치료적·예방적 행위를 통해서 나타나는 것이므로, 적절한 치료적·예방적 행위가 따라오지 않는다면 실제 진료결과는 변하지 않을 것이며 치료적·예방적 행위는 일부 환자에서 부작용을 유발할 수도 있다. 따라서, 인공지능 알고리즘의 성능 검증을 넘어 인공지능이 실제 진료와 환자결과에 어떤 영향을 미치는지를 직접적으로 확인하는 것은 큰 의미가 있다[40]. 이를 임상적 유용성(clinical utility) 검증이라 한다. 유용성을 유효성이라 말과 혼동하지 말아야 한다. 기술적 성능, 임상적 성능, 임상적 유용성은 모두 서로 다른 수준과 성격의 유효성 지표들이다.

임상적 유용성 검증의 좋은 예로, 진통 중인 산모의 자궁 수축과 태아의 심장박동을 지속적으로 자동 감시·분석하여 태아의 문제가 의심될 경우 의사에게 주의를 실시간으로 보내주는 영국에서 개발된 인공지능 알고리즘에 대한 연구가 있다[41]. 연구자들은, 먼저 알고리즘의 정확도를 확인한 후, 고위험 산모들을 무작위 배정에 따라 두 군으로 나누어 한 군은 알고리즘의 도움을 받으며 진료를 하고 다른 군은 알고리즘 없이 기존 방식으로 진료를 한 후 두 군의 결과를 비교를 하는 임상적 유용성 검증연구를 수행하였다[41]. 알고리즘은 비정상적 심장박동 인지에 대한 높은 성능을 보였지만 아기와 산모 모두에 있어 두 군간에 진료결과에 유의한 차이는 없어서, 결과적으로 인공지능 알고리즘을 사용함으로써 환자에게 발생하는 의학적 이득 관점의 임상적 유용성을 입증할 수는 없었다.

이와 같이 인공지능 알고리즘의 임상적 유용성 검증은 기본적으로 인공지능을 사용하면서 진료할 때와 인공지능을

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 mmHg (in millimeters of mercury) x time spent below a mean arterial pressure of 65 mmHg (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

CNN, convolutional neural network; CAde, computer-aided detection.

사용하지 않고 진료할 때 환자결과에 차이가 있는지를 비교하는 것이며, 두 군 간의 비교에 있어 교란변수의 영향을 없애기 위해서 무작위 임상시험을 하는 것이 가장 이상적이다. 하지만, 무작위 임상시험이 항상 가능한 것은 아니므로 교란변수를 보정한 전향적 또는 후향적 성과분석연구를 이용할 수도 있다. 또한, 일부 인공지능 알고리즘의 경우에는 임상적 유용성을 직접 검증하지 않고 알고리즘의 성능을 적절히 검증하는 것만으로 충분한 임상검증이 이루어 질 수도 있다. 개별 인공지능 알고리즘과 개별 임상적 상황에 맞는 적절한 임상검증의 수준에 대한 의료전문가의 판단이 필요하다. 인공지능 알고리즘에 대한 무작위 임상시험의 몇 가지 사례들을 Table 2에 정리해 놓았다[41-46].

의료기기 허가 및 의료보험 관점에서 보는 인공지능 알고리즘의 임상검증

의료기기 허가-대한민국 식품의약품안전처, 미국 Food and Drug Administration, 유럽 CE Marking 등에서 내리는 허가를 의미한다-와 의료보험 적용 여부를 결정하는 데에는 학술적 원칙 외에도 사회·정치적 요인이 관여된다. 따

라서, 인공지능 알고리즘의 의료기기 허가 및 의료보험 적용은 국가, 시기, 그리고 사회적 요인에 따라 차이가 있을 수 있으며 한 가지 방법으로 설명할 수는 없을 것이다. 이 논문에서는 단지 인공지능 알고리즘의 의료기기 허가 및 의료보험 적용을 위해 일반적으로 고려되는 학술적 원칙들에 대해 간단히 설명한다. 아래 설명에 나오는 ‘기술적 성능’, ‘임상적 성능’, ‘임상적 유용성’, ‘진단적 환자-대조군 연구’, ‘진단적 코호트 연구’ 용어들의 의미에 대해서는 이 논문의 앞쪽 해당 부분에 나오는 설명을 같이 참고하기 바란다.

1. 의약품과 진단기기의 차이

인공지능 알고리즘은 대다수가 진단용 의료기기에 속하므로 진단용 의료기기의 허가과 의료보험 적용 과정을 따르게 된다. 인공지능 알고리즘의 의료기기 허가과 의료보험 적용을 이해하려면 진단용 의료기기와 의약품 간의 차이점을 먼저 이해하는 것이 도움이 된다.

허가 측면을 보면, 치료 의약품의 허가를 위해서는 일반적으로 3상 임상시험이라 불리는 임상연구를 통해 구체적으로 어떠한 환자군에서 어떤 치료 의약품을 사용하였을 때 환자의 진료결과가 향상된다는 것을 입증하는 것이 필요하다. 즉, 치료 의약품을 사용할 구체적 적응증에 해당하는

대상환자에 대해서 임상적 유용성을 검증하는 것이다. 이에 비해, 진단기기의 허가는 의약품 허가과 같은 높은 수준의 임상검증 근거를 요구하지 않으며 일반적으로 진단기기의 기술적 성능 검증에 중점을 두게 된다(물론 더 높은 수준의 임상검증 자료가 있다면 보다 충분한 평가가 이루어질 것이다).

의료보험 적용 측면을 보면, 의료보험 적용이라는 것은 환자들이 납부한 보험료를 가지고 환자들에게 도움이 되는 의료행위(즉, 의약품이나 의료기기의 사용)를 보험자가 구입해 주는 과정이다. 따라서 기본적으로 어떤 의료행위가 환자에게 도움이 된다는 임상적 유용성을 입증하는 것이 중요하다. 이때 같은 의료행위라도 어떤 환자에게는 도움이 되지만 다른 환자에게는 도움이 되지 않을 수 있기 때문에 의료행위의 대상이 되는 구체적인 임상적 적응증을 분명히 해서 의료보험을 적용하여야 한다. 올바른 적응증이 아닌, 임상적 유용성이 입증되지 않은 환자에게 어떤 의료행위를 하였을 때 의료보험을 적용한다는 것은(즉, 의료보험을 통해 환자가 비용을 지불한다는 것은) 비상식적이고 적절하지 않다. 치료 의약품의 경우, 이와 같은 의료보험 적용의 조건은 앞에서 이야기한 의약품 허가에 필요한 근거와 같은 성격이다. 따라서 치료약제는 허가를 받은 후 너무 고가가 아니라면 일반적으로 자연스럽게 의료보험 적용으로 이어진다. 하지만, 진단기기의 경우는 의료기기 허가에 사용되는 근거가 기술적 성능 검증에 그치는 경우가 많아 의료보험 적용에 필요한 임상적 유용성 근거를 만족하지 못하므로 의료기기 허가가 의료보험 적용으로 곧바로 이어지지 않는다. 즉, 의료기기 허가가 이루어지면 진단기기의 판매와 임상진료에 사용이 가능하고, 추후에 이 진단기기가 환자들에게 도움을 주는 임상적 조건과 환자 군이 추가의 임상검증을 통해 밝혀지면 이러한 적응증에 대해서 의료보험을 적용하게 되는 순서를 거친다. 예를 들면, 자기 공명영상이나 초음파검사가 기술적 성능을 인정받아 의료기기 허가를 받았다 해도 곧바로 의료보험이 적용되는 것은 아니며 이 검사들을 사용함으로써 환자에게 도움을 주는 임상적 조건과 환자 군이 밝혀지면 이들에 대해 의료보험이 적용되는 방식이다.

2. 인공지능 알고리즘의 의료기기 허가

인공지능 알고리즘의 의료기기 허가를 위해서는 적어도 기술적 성능에 대한 충분한 검증 결과가 필요하다. 인공지능 알고리즘의 기술적 성능은 앞에서 설명한 것처럼 진단적 환자-대조군 연구 형태의 외부검증을 통해 평가할 수 있다. 전향적 연구가 가능하더라도 물론 좋겠지만, 인공지능 알고리즘의 사용 목적에 부합하는 다양한 난이도의 증례들을 골고루 잘 포함하는 검증 데이터를 모을 수만 있다면 후향적 연구로도 가능하다. 기술적 성능 검증에서 인공지능 알고리즘이 적절히 작동하기 위한 데이터의 조건들(예: 사용 가능한 장비의 종류, 영상획득법 등)을 확인하고 구체적으로 명시하기도 한다. 인공지능 의료기기의 허가는, 환자를 대상으로 기기를 사용할 수 있다는 점과 이러한 목적으로 기기를 판매할 수 있다는 허가로서 보통 알고리즘의 안전성과 유효성에 대해서 기술적 성능을 통해 일정 수준의 평가를 하지만 인공지능 알고리즘의 사용이 환자에게 도움이 되고 진료현장에서 사용할 만한 가치가 있다는 것을 증명하려는 것은 아님을 분명히 이해하여야 한다[25,26,47]. 진료현장의 의료인들은 의료기기 허가를 받은 인공지능 알고리즘에 대하여 알고리즘이 정말 환자에 도움이 되는지 본격적인 임상검증과 평가를 하고 이를 통해 알고리즘이 안전하고 효과적으로 임상에 적용될 수 있도록 하여야 한다[47,48]. 더구나, 의료기기 허가 과정에서 인공지능 알고리즘의 일반화에 대한 취약성과 관련된 모든 사항을 확인하기도 어렵기 때문에 인공지능 알고리즘을 진료에 사용할 때 어떤 상황에서 결과가 정확하고 어떤 경우는 정확하지 않은 지 주의 깊게 평가를 할 필요가 있다.

식품의약품안전처가 2019년 10월 개정하여 발표한 ‘빅데이터 및 인공지능 기술이 적용된 의료기기 허가심사 가이드라인(민원인 안내서)(개정안)’과 ‘인공지능 기반 의료기기의 임상 유효성 평가 가이드라인(민원인 안내서)(개정안)’에도 이러한 학술적 원칙이 잘 반영되어 있다. 이 가이드라인은 인공지능 기반 의료기기 허가를 위한 임상시험에 사용하는 표본 데이터는 제품 개발 과정 동안 사용된 데이터와는 독립적이어야 한다고 설명한다. 즉, 외부검증이 필요하다는 것을 이야기하고 있다. 또한, 미리 확보된 신뢰성 있는 의료 데이터를 이용하여 평가하는 후향적 임상시험을 적용하는 것도 고려할 수 있으며

제품의 특성에 따라 전향적 임상시험과 후향적 임상시험 중 적합한 임상시험 방법을 설계할 수 있다고 안내한다.

3. 인공지능 의료기기에 대한 의료보험 적용

인공지능 의료기기에 대한 의료보험 적용과 관련하여, 보건복지부와 건강보험심사평가원이 2019년 12월 ‘혁신적 의료기술의 요양급여 여부평가 가이드라인’을 발표하였다. 이 가이드라인은 의료보험 적용에 대한 학술적 원칙에 약간의 유연성을 더한 것으로 보인다. 이 가이드라인에 따르면, 인공지능 알고리즘을 사용함으로써 기존 의료행위 대비 환자의 진료결과가 향상되거나 현저한 진단능력의 향상이 있는 경우 인공지능 알고리즘 의료기기를 사용한 것에 대하여 의료보험을 통한 별도의 보상이 고려될 수 있다고 설명한다(비용효과성 부분은 일반적인 사항이며 이 논문의 범위를 넘어 생략한다). 이때 적절한 근거 자료의 형태로 교란변수를 보정한 전향적 또는 후향적 성과분석연구나 무작위 임상시험이 바람직한 것으로 되어 있지만 진단적 코호트 연구를 이용한 임상적 성능의 외부검증 연구도 사례별로 인정하는 것으로 되어 있다. 즉, 의료보험 적용을 위해서는 환자의 진료결과 향상이라는 임상적 유용성이 입증되어야 한다는 점이 기본적으로 중요하나, 사례에 따라서는 특정 임상적 조건과 환자군에서 인공지능 알고리즘을 사용하였을 때 현저한 진단능력의 향상이 있다는 것을 진단적 코호트 연구를 통해 보여주는 것으로도 의료보험 적용이 가능할 수 있음을 이야기한다.

결론

의료 인공지능 알고리즘의 임상검증, 의료기기 허가, 의료보험 적용과 관련된 핵심 원리를 알아보았다. 인공지능의 판별정확도를 평가할 때는 다이슨 유사성 계수, 민감도, 특이도, 수신자판단특성곡선, free-response 수신자판단특성곡선이 널리 이용된다. 확률을 직접 제시하는 인공지능 알고리즘의 경우에는 보정정확도 또한 잘 평가하여야 한다. 현재 의학적 진단·예측을 제공하는 인공지능 알고리즘 들의 대다수는 개발 단계나 내부검증에서 보인 성능이 실제 진료 환경으

로 일반화되기 어려운 제한점을 갖고 있다. 따라서 인공지능의 임상검증을 할 때 성능의 외부검증이 매우 중요하며 인공지능은 기본적으로 독립적 도구가 아니라 의료인에게 정보를 제공하는 보조 도구임을 고려하여야 한다. 인공지능 성능의 외부검증에는 진단적 환자-대조군 연구와 진단적 코호트 연구 형태가 있다. 전자는 인공지능이 진단·예측하고자 하는 상태와 그렇지 않은 상태를 어느 정도 전체적으로 구별하는 기술적 성능을 평가하고, 후자는 실제 진료환경의 어떤 임상적 상황이나 환자 군에서의 임상적 정확도를 평가한다. 인공지능의 궁극적 임상검증은 인공지능이 환자 진료결과에 어떤 영향을 미치는지를 평가하는 것이며 이를 임상적 유용성 검증이라 한다. 이를 위해서는 무작위 임상시험이 가장 좋다. 인공지능의 의료기기 허가는 일반적으로 기술적 성능 검증에 중점을 둔다. 따라서, 의료기기 허가는 인공지능이 진료결과를 향상시키고 환자에게 도움이 된다는 것을 직접적으로 증명하는 것은 아니다. 또한, 의료기기 허가 과정에서 인공지능 알고리즘의 일반화에 대한 취약성과 관련된 모든 사항을 확인하기도 어렵다. 따라서, 인공지능 알고리즘이 의료기기 허가를 받은 후 실제 환자에 도움이 되는지 본격적인 임상평가를 하는 것은 진료현장의 의료인의 몫이다. 의료보험 적용을 위해서는 인공지능 의료기기를 사용하는 것이 환자의 진료결과를 향상시킨다는 임상적 유용성을 입증하는 것이 중요하다. 마지막으로, 의료 인공지능 알고리즘이 앞으로 임상진료에 보편적으로 활용될 것이 예상되므로, 의과대학 교육과정에서부터 이러한 내용들을 다룰 필요가 있다[49].

찾아보기말: 소프트웨어 검증; 기기허가; 보험급여; 인공지능

ORCID

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>

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

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* 2020;21:33-41.
4. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, Kwon MR, Shin JH, Hahn SY. 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. In: Proceedings of IEEE 12th International Symposium on Biomedical Imaging; 2015 Apr 16-19; New York, NY, USA. New York: 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 web site [Internet]. [place unknown]: DevChakraborty.com; 2019 [cited 2020 Sep 14]. Available from: <http://www.devchakraborty.com/>.
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 DS, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, Wong EY, Sabanayagam C, Baskaran M, Ibrahim F, Tan NC, Finkelstein EA, Lamoureux EL, Wong IY, Bressler NM, Sivaprasad S, Varma R, Jonas JB, He MG, Cheng CY, Cheung GC, Aung T, Hsu W, Lee ML, Wong TY. 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 [Internet]. Arlington: AuntMinnie.com; 2018 [cited 2020 Sep 14]. Available from: <https://www.auntminnie.com/index.aspx?sec=nws&sub=rad&pag=dis&ItemID=123871>.
15. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, Goo JM, Aum J, Yim JJ, Cohen JG, Ferretti GR, Park CM; DLAD Development and Evaluation Group. 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, Kim K, Choi C, Lee G, Yi J, Bang WC. 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, Folio LR, Summers RM, Rubin DL, Lungren MP. 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, Halpern EF, Hess CP, Schiebler ML, Weiss CR. 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, Woo H, Lee JM, Jung SE, Oh JH. 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-Chenevert A, Guest W, Chong J, Barfett J, Chepelev L, Cairns R, Mitchell JR, Cicero MD, Poudrette MG, Jaremko JL, Reinhold C, Gallix B, Gray B, Geis R; Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. 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 cha-

racteristics 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, Lim JS, Park SH. 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, Bachmann LM, Keane PA, Denniston AK. 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, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. 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, Schenk J, Terwindt LE, Hollmann MW, Vlaar AP, Veelo DP. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective non-cardiac surgery: the HYPE randomized clinical trial. *JAMA* 2020;323:1052-1060.
43. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A, Galtieri PA, Pellegatta G, Carrara S, Di Leo M, Cravioito V, Lamonaca L, Lorenzetti R, Andrealli A, Antonelli G, Wallace M, Sharma P, Rosch T, Hassan C. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* 2020;159:512-520.e7.
44. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, Lei L, Li L, Guo Z, Lei S, Xiong F, Wang H, Song Y, Pan Y, Zhou G. 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, Liu P, Li L, Song Y, Zhang D, Li Y, Xu G, Tu M, Liu X. 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, Jiang X, Huang X, Mu G, Wan X, Lv X, Gao J, Cui N, Hu S, Chen Y, Hu X, Li J, Chen D, Gong D, He X, Ding Q, Zhu X, Li S, Wei X, Li X, Wang X, Zhou J, Zhang M, Yu HG. 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 14 [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.

Peer Reviewers' Commentary

이 논문은 인공지능의 임상 검증, 의료기기 허가, 의료보험 적용과 관련된 핵심 원리에 대하여 현 단계의 범주와 이슈를 편향 없이 잘 정리하여 제시해 주고 있다. 새로운 기술인 인공지능 알고리즘의 성능 평가 지표와 기존의 통계학적 기법에 대하여 알기 쉽게 설명하고 있다. 특히 인공지능 알고리즘 검증 방법에 대하여 기존 임상 시험 방법이 어떻게 적용되는지 체계적으로 정리되어 있다. 또한, 진단 의료기기의 허가와 보험급여에 관하여 의약품과 다른 핵심 요점부터 기술적 성능 검증과 임상적 유용성 검증을 대비하여 명확히 설명하고 있다. 이 논문은 인공지능 의료기기의 개발을 준비하는 연구자들뿐만 아니라 향후 이러한 기기를 이용할 의료진들에게 많은 도움을 줄 것으로 판단된다.

[정리: 편집위원회]