



Current Landscape and Commercialization of AI Models in Musculoskeletal Imaging

근골격계 영상 AI 모델의 현황과 상용화

Chang Ho Kang, MD*

Department of Radiology, Korea University Anam Hospital, Korea University College of Medicine, Seoul, Korea

Received June 29, 2025
Revised July 29, 2025
Accepted September 21, 2025

*Corresponding author

Chang Ho Kang, MD
Department of Radiology,
Korea University Anam Hospital,
Korea University College of Medicine,
73 Goryeodae-ro, Seongbuk-gu,
Seoul 02841, Korea.

Tel 82-2-920-6540
Fax 82-2-929-3796
E-mail mallecot@gmail.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

AI-based software as a medical device (SaMD) using deep learning applications for musculoskeletal diseases is being clinically implemented in South Korea, although it is still in its early stages in the musculoskeletal field compared to other fields of radiology, such as neuroradiology, chest, and breast imaging. AI models for detecting various fractures, estimating pediatric bone age, calculating geometric skeleton measurements, grading arthritis, and osteoporosis screening have shown high diagnostic performance, and many of these applications are now commercially available for use in clinical practice. Many studies have documented the feasibility of using an AI model for detecting joint pathology on MRI and interpreting spine MRIs. This review provides information on the domestic and international commercialization status of AI-based SaMD for musculoskeletal imaging and beneficial considerations for its application in clinical practice, helping readers who are interested in the field application of musculoskeletal imaging AI models in their decision-making.

Index terms Artificial Intelligence; Deep Learning; Musculoskeletal Imaging; Software as a Medical Device; Commercialization

서론

최근 몇 년간 인공지능(이하 AI) 기술은 의료영상전송시스템(picture archiving and communication systems; 이하 PACS), 전자의무기록(electronic medical records) 등을 통한 대규모 의료데이터를 원료로 의료영상 분야에서 혁신적인 변화를 가져왔으며 새로운 의료서비스 가치 창출을 모색하고 있다. 전 세계적으로 의료 인공지능을 개발하는 다양한 기업들이 등장하여 공학자, 의료진, 의료기관과의 협력 하에 임상현장의 미충족 수

요를 반영할 수 있는 독립형 인공지능 기반 솔루션을 제공하거나 의료기기에 인공지능을 탑재하는 방식으로 새로운 기술의 상용화가 진행 중이다.

2024년 개최된 Radiological Society of North America (이하 RSNA)를 예로 들면 영상의학 및 영상진료에서의 AI 적용과 관련된 최첨단 연구가 발표되었고 거의 모든 영상의학 영역에서 상용화 모델의 전시 및 임상적용 그리고 미래 모델이 ‘Building Expertise with the Latest in Medical Imaging AI’의 주제 아래 소개되었다. 전시회에 참가한 업체 650개 중 208개 업체가 AI 관련 기술을 선보였다. RSNA는 2017년부터 AI 관련 챌린지를 개최하며(1) 첨단 의료 기술 개발을 촉진하고 있다. 2022년에는 CT 영상의 경추 골절 진단 챌린지를 통하여 기존 모델의 성능을 능가하는 알고리즘이 소개되었고(2), 2024년에는 ‘미국신경영상의학회(American Society of Neuroradiology)’와 협업체 요추 퇴행성 질병을 테마로 제시했다. ‘요추 퇴행성 분류 AI 챌린지’에서는 요추 MRI 영상을 기반으로 퇴행성 척추 상태를 평가하는 AI 모델 개발을 목표로 관련 데이터 세트를 제공해 연구를 지원하였다.

의료 영상에서 입증된 딥러닝 기법의 우수성은 영상의학에 혁신을 일으키고 현재 영상진료의 모습을 바꾸고 있다. 딥러닝 방법은 자기공명영상(MRI) 재구성, 인체 조직 분할, 병소 검출, 영상 진단/분류 등 다양한 근골격계 영상 분야에 적용되고 있는데, 현재 기술 발전 상황을 고려하였을 때 임상 현장의 안착이 가장 유망한 근골격계 적용 분야는 골절 검출, 골연령 추정, 하지 및 척추 정렬 측정, 관절염 등급 분류, 골다공증 기회검진, 척추 및 대관절 질환 MRI 판독 보조 등으로 생각된다.

본 종설에서는 근골격계 영상 인공지능 소프트웨어 의료기기(software as a medical device; SaMD)의 국내외 상용화 현황과 향후 임상 진료에 적용이 기대되는 딥러닝 모델에 대한 정보를 제공함으로써, 근골격계 영상 AI 모델의 현장 적용에 관심이 있는 독자들의 의사 결정에 도움이 되고자 한다.

제품허가와 의료현장 도입

2024년 12월까지 미국 식품의약국(U.S. Food and Drug Administration; 이하 FDA)에서 승인 받은 AI 기반 의료기기(AI/machine learning-enabled medical devices)는 1,016개이며 이 중 811개의 제품이 radiology 카테고리에서 확인되어 약 80%의 비중을 보였다. 근골격계 분야의 영상 AI 모델은 2018년 5월 손목골절 의료영상 진단 보조 소프트웨어인 Imagen사의 OsteoDetect 제품이 FDA 시판 승인(de novo pathway)을 받은 것을 시작으로 다양한 신체 부위의 골절, 무릎 골관절염, 골격계 계측, 척추영상 판독 보조, 골다공증 등의 진료 영역으로 확장되고 있다. 정형외과/척추 수술의 계획 수립 및 치료와 연관된 제품을 제외하고 저자가 확인한 주요 근골격계 영상 분야 AI 모델은 다음과 같다(Table 1).

한편 국내의 상황을 살펴보면 2018년 골연령 판독을 보조하는 제품이 국내 최초로 인공지능 의료기기로서 식품의약품안전처의 허가 승인을 받은 것을 시작으로 2022년 47건, 2023년 64건, 2024년 108건을 포함 매년 인공지능 의료기기 허가가 급속히 증가하고 있다. 2023년 9월 30일 기

Table 1. U.S. Food and Drug Administration-Cleared Musculoskeletal AI/Machine Learning-Enabled Medical Devices as of December 2024

Company	Product	Modality	Headquarter	Released	URL	Key-Features
Avicenna.AI	CINA-CSpine	CT	France	2024	https://avicenna.ai/solutions/spine/	Detection and triage of cervical spine fractures
RemedyLogic	RAI	MRI	USA	2024	https://www.remedyllogic.com/	Spine MRI interpretation
Ever Fortune.AI	EFai BAPXR	Radiography	Republic of China	2024	https://www.everfortuneai.com.tw/en/	Bone age assessment
Rho	16 Bit	Radiography	Canada	2024	https://www.16bit.ai/rho	Opportunistic screening of osteoporosis
ImageBiopsy Lab	LAMA	Radiography	Austria	2023	https://www.imagebiopsy.com/product/ib-lab-lama-fda?utm_lang=en	Length and angle measurements of the lower limb
Medical Metrics	Spine CAMP	Radiography	USA	2023	https://medicalmetrics.com/spinecamp/	Quantitative spine measurements & motion analysis
SmartSoft Healthcare	Columbo	MR	Bulgaria	2022	https://columbo.me/	Spine MRI interpretation
Nanox.AI	HealthOST	CT	Israel	2022	https://www.nanox.vision/ai-bone-solution/	Detects vertebral compression fractures & measures bone mineral density to identify patients at risk
MRIguidance	BoneMRI	MR	Netherlands	2022	https://mriguidance.com/	3D CT-like image generation from MRI
Aidoc Medical	BriefCase for Rib Fracture Triage	CT	Israel	2021	https://www.aidoc.com/solutions/radiology/	Rib fracture triage, prioritization, notification
Radiobotics	RBknee	Radiography	Denmark	2021	https://radiobotics.com/	Diagnosis of OA: osteophytes, subchondral sclerosis, and joint space narrowing
Shanghai United Imaging Intelligence	uAI EasyTriage-Rib	CT	China	2021	https://eu.united-imaging.com/en/product-service/software-and-solution/software	Rib fractures (detects multiple and acute fractures)
Gleamer	BoneView	Radiography	France	2020	https://www.gleamer.ai/solutions/boneview/	Bone trauma X-ray interpretation in various body regions
Imagen Technologies	FractureDetect	Radiography	USA	2020	https://imagen.ai/ai-software/	Detecting hard-to-spot fractures in upper and lower extremities
Zebra Medical Vision	HealthVCF	CT	Israel	2020	http://www.zebra-med.com/	Vertebral compression fracture detection and prioritization
Annalise.AI	Enterprise CXR Triage Trauma	Radiography	Australia	2020	https://annalise.ai	5 acute findings: pneumothorax, tension pneumothorax, pleural effusion, pneumoperitoneum, and vertebral compression fracture
Aidoc Medical	BriefCase for CSF Triage	CT	Israel	2019	https://www.aidoc.com/solutions/radiology/	Cervical spine fracture triage, prioritization, notification
AZmed	Rayvolve	Radiography	France	2019	https://azmed.co/	Detecting fractures, dislocations, and joint effusions on X-rays
Medimaps	TBS IN (Osteo)	Radiography	Switzerland	2019	https://www.medimapsgroup.com/tbs-osteo/	Bone microarchitecture (trabecular bone score) computing
ImageBiopsy Lab	KOALA	Radiography	Austria	2019	https://www.imagebiopsy.com/	Knee OA labeling assistant
Imagen Technologies	Osteodetect	Radiography	USA	2018	https://imagen.ai/ai-software/	Detect and localize distal radius fractures

OA = osteoarthritis

준으로 한국 식품의약품안전처에 허가, 인증 및 신고된 AI 기반 의료기기는 총 186개로, 이 중 138개가 영상의학과와 관련된 제품이었다. 근골격계 영상분야에서도 매년 꾸준히 사용 허가 승인 제품이 축적되고 있는데, 흉부, 신경, 유방 영상의학 분야와 비교 시 의료현장의 도입 상황은 아직 초기 단계인 것으로 보인다. 의료기기 안심책방(<https://emedi.mfds.go.kr/>)을 참조하여 근골격계 영상분야 인공지능 의료기기 소프트웨어를 조사한 자료를 제공하니(Table 2) 기술 개발의 아이디어 발아, 의료기기의 임상검증, 혹은 인허가 제품의 병원도입 시 참고하기 바란다. 본 자료에서 치료 계획 수립 용 소프트웨어는 제외하였으며 모델의 업그레이드에 의한 중복은 최신 인허가 제품으로 통일하였고 사용 목적이 의료영상 분석처럼 구체적인 임상적용 시나리오가 명기되지 않은 경우도 역시 제외하였다.

미국과 유럽의 AI 기반 근골격 영상 의료기기 인허가 상황과 비교하면 우리나라의 근골격계 영상 분야에서 AI 모델의 개발 속도와 제품군은 골절 영역을 제외하고 큰 차이가 없는 것으로 보인다. 하지만 병원 현장 도입의 관점에서 우리나라의 인허가 제품이 시장에 안착한 경우는 골연령 판독 보조 솔루션을 제외하고 거의 확인이 되고 있지 않다. 아직까지 식품의약품안전처 허가를 받은 근골격계 분야 인공지능 소프트웨어 솔루션들이 국내에서 어느 정도 사용되고 있는지 정확한 데이터는 없다.

새로운 의료기술이 의료현장으로 널리 도입되는 데 중요한 요소가 의료보험 적용 여부인데(3), 특히 건강보험 당연지정제를 채택하고 있는 우리나라의 경우에는 국민건강보험 적용 여부가 매우 중요하다. 하지만 영상의학 분야 인공지능 의료기기 소프트웨어의 대부분은 기존에 사용하던 영상검사들을 분석, 판독, 이용하는 방식에 변화를 주는 진단보조 기능들이어서 새로운 검사나 새로운 행위에 해당하지 않아 기존의 틀안에서는 새로운 의료기술이 보상을 받을 수 없었다. AI 기반 영상의료기술이 현저한 진단능력 향상과 환자의 궁극적 진료 결과를 향상시킬 수 있다는 임상적 가치를 입증한 경우라면 급여 건강보험 적용이 가능하나, 현재까지 개발된 제품을 적용하여 새로운 가치를 창출하거나 궁극적인 치료효과성을 입증하는 것은 일반적으로 쉽지 않다.

이에 정부는 AI 의료기기의 시장 진출 지원을 위하여 혁신의료기술평가, 혁신의료기기 통합심사평가 제도, 의료현장 선진입 후 평가와 같은 효율성을 높이는 정책을 도입하였고 2023년 인공지능 기반 혁신의료기술 건강보험 임시등재 운영 규정을 발표하였다. 식품의약품안전처, 건강보험심사평가원, 한국보건 의료연구원의 혁신의료기기 통합심사를 통과한 인공지능 의료기기는 한시적으로 건강보험에 등재되어 수가 보상을 받을 수 있게 되었다(4). 제도 시행 직후 2024년 지정된 29개 혁신의료기기 제품 유형을 살펴보면 인공지능 기술을 활용한 '영상검출 진단보조 소프트웨어' 등 '의료기기 소프트웨어'가 23개(79%)로 가장 높은 비중을 차지했다. 혁신의료기술 평가트랙이 적용되는 경우 영상 판독 수가 범위 내에서 보상을 받고, 정식 보험등재 시에는 환자에게 이익이 되거나 비용 절감 효과 등이 입증되는 경우 추가 가치를 인정한다. 2023년 10월에 AI 뇌경색 유형 분류 솔루션이 국내 첫 건강보험 수가 적용 사례로 선정되었다.

급여권에 진입하는 또 다른 트랙인 신의료기술 평가 유예 제도는 조기 도입이 필요한 새로운 의료 기술에 대해 평가를 유예함으로써 조기에 의료현장에서 비급여로 사용할 수 있도록 허용하는

Table 2. Korea Ministry of Food and Drug Safety-Approved Musculoskeletal AI-Based Software as Medical Devices as of March 2025

Company (Korean)	Product	Released	Modality	Intended Use
PROMEDIUS (프로메디우스 주식회사)	PROS CXR: OSTEO	2025	Radiography	Chest X-ray based osteoporosis screening and diagnosis
Sillasystem (주식회사 신라스티스)	Pine-DXS	2024	Radiography	Cobb angle measurements
PROMEDIUS (프로메디우스 주식회사)	PROS CXR: COBB	2024	Radiography	Cobb angle measurements
CorelineSoft (주식회사 코어라인소프트)	AVIEW SpineBH	2024	CT	Auto-spine analysis (assigning the vertebral numbers and measuring the height and bone density of the vertebral bodies)
DEEPNOID (주식회사 딥노이드)	DEEP-KNEE	2024	MRI	Knee MRI analysis
CRESKOM (주식회사 크레스콤)	MediAI-OA	2024	Radiography	Knee osteoarthritis Kellgren-Lawrence grading
Basgenbio (주식회사 바스젠바이오)	Dr.L.Spine-D-P	2024	Radiography	Lumbar compression fracture detection
Basgenbio (주식회사 바스젠바이오)	BAS-OS-10	2024	Radiography	Ankle and foot fracture detection
JLK (주식회사 제이엘케이)	MediHub-Shoulder	2024	Not reported	Shoulder image analysis
Connective (코넥티브 주식회사)	CONNEVO KOA EXP	2024	Radiography	Knee abnormality detection
VUNO (주식회사 뷰노)	VUNO Med-Spine X-ray	2023	Radiography	Lumbar compression fracture detection
Basgenbio (주식회사 바스젠바이오)	Dr.Knee-A-P	2023	Not reported	Knee image analysis
ClariPi (주식회사 클라리파이)	ClariOsteo	2023	CT	Lumbar spine analysis (measuring the height and bone density of the vertebral bodies)
CorelineSoft (주식회사 코어라인소프트)	AVIEWPsoas	2023	CT	Psoas muscle segmentation on abdomen-pelvis CT
DEEPNOID (주식회사 딥노이드)	DEEP:SPINE-AS	2022	Radiography	Cobb angle measurements, Vertebral compression fracture detection and grading
Basgenbio (주식회사 바스젠바이오)	Dr.Wrist	2022	Not reported	Wrist image analysis
HealthHub (주식회사 헬스허브)	LE-AI	2022	Radiography	Lower extremity scanogram analysis
VUNO (주식회사 뷰노)	VUNO Med-Spine-CT	2022	CT	Spine CT auto-analysis
VUNO (주식회사 뷰노)	VUNO Med-Knee MR	2022	MRI	Knee MRI analysis
CRESKOM (주식회사 크레스콤)	MediAI-BA	2022	Radiography	Wrist fracture detection
ClariPi (주식회사 클라리파이)	ClariSarco	2022	CT	Skeletal muscle and psoas muscle volume analysis
Ravence (주식회사 레이언스)	Xmaru Pro CXR	2020	Radiography	Bone age assessment
CRESKOM (주식회사 크레스콤)	MediAI-BA	2020	Radiography	Bone age assessment
HealthHub (주식회사 헬스허브)	Boneage.io	2020	Radiography	Bone age assessment
DEEPNOID (주식회사 딥노이드)	DEEP:SPINE	2019	Radiography	Lumbar compression fracture detection
VUNO (주식회사 뷰노)	VUNO Med-BoneAge	2018	Radiography	Bone age assessment

제도다. 현저한 진단 능력 향상 및 오류 감소 효과 등의 가능성을 인정받아 신의료기술로 분류돼 평가 유예를 받은 첫 번째 사례는 유방촬영술에서 유방암 의심부위를 검출하고 유방 치밀도를 분류하는 솔루션 ‘루닛 인사이트 MMG’이다. 평가 유예 신의료기술은 비급여 상한액을 제한받지 않아 국내 시장에서의 빠른 확장을 기대할 수 있으나 환자 혜택보다는 기업과 의료기관의 이익 추구를 용이하게 만들 가능성이 크다는 비판도 제기되고 있다.

근골격계 영상 AI 모델의 상용화 분야

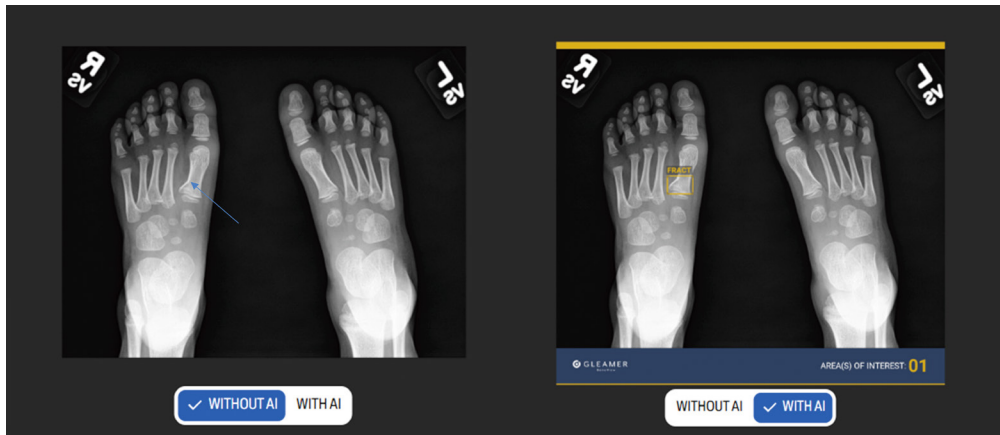
골절

근골격계 분야에서 영상 AI 모델의 초기 적용 분야가 골절이다. 근골격계 영상을 주로 판독하는 영상의학과 의사나 정형외과 의사의 경우 판독의 정확도가 높지만, 비전공 의사들의 경우 판독에 어려움을 겪는 경우가 많다. 일례로 영상판독 전문의료진이 부족한 군의료 현장에 AI 기반 의료영상 판독 솔루션의 활용을 목표로 과기정통부와 국방부가 ‘디지털 뉴딜’ 사업인 ‘AI 융합 프로젝트 (AI+X)’의 일환으로, 군에서 발생위험이 높은 척추 압박골절과 손목 골절의 의료영상 판독 솔루션 개발을 촉진한 바 있다.

영상의학 영역에서 컴퓨터 보조 영상진단을 위한 최초의 FDA 승인 인공지능 소프트웨어가 일반촬영에서 원위 요골골절을 검출하고 국부화하는 솔루션(OsteoDetect, Imagen Technologies)인데 기업의 홍보자료에 의하면 소프트웨어의 도움을 받은 경우 임상치의 판독 오류가 47% 감소했다고 보고하였다(5). 골절 진단에서 일반촬영의 역할은 여전히 주요하며, 모델개발의 관점에서 공개 데이터셋의 활용 및 학습 데이터 구축이 용이하여 다양한 신체 부위의 골절 검출 연구 자료가 축적되었다(6).

최근 연구에 따르면 제품화된 골절 검출 소프트웨어가 영상의학과와 타과 의료진 모두에서, 그리고 성인 및 소아 환자에서 골절 검출 능력을 향상시키는 것으로 나타났으며(Fig. 1), 높은 민감도와 특이도를 보이고 시간을 절약하는 효과도 있었다(7-10). 딥러닝을 이용한 골절 검출 연구에 대한 체계적 문헌고찰 및 메타분석 결과, 딥러닝 모델은 숙련된 판독자와 유사하거나 능가하는 성능을 보이는 것으로 나타났다(11). 이 메타분석에서는 39개의 연구가 선택되었는데, 이 중 29개 연구는 일반촬영 영상, 6개 연구는 CT 영상, 4개는 골다공증 관련 척추골절 검출을 위하여 골밀도 영상을 대상으로 하였다. 딥러닝 모델들의 종합(pooled) 민감도, 특이도 및 area under curve (이하 AUC)는 각각 90%, 92%, 97%로 숙련된 판독자와 유사한 수준의 성능을 보였는데, 비전문 판독자가 참여한 연구에서는 AI 모델이 민감도(95% 대 88%)와 AUC(98% 대 96%)에서 더 좋은 결과를 보였다. 그리고 다기관 설계를 사용한 모델이 단일기관 설계를 사용한 모델에 비해 민감도(92% 대 88%)와 특이도(94% 대 91%)가 더 높게 나타났다. Kuo 등(12)의 또 다른 메타분석에서도 유사한 결론이 도출되었는데, AI와 판독자가 골절 검출에서 대등한 결과를 보였으며, AI 참조 판독 시 판독자의 정확도가 향상됨을 보고하였다. 그러나 분석에 포함된 42개 연구 중 13개 연구만이 외부 검증 결과를 제시했고, 단 1개의 연구만이 전향적 환경에서 모델 성능을 평가했기 때문에 이러한

Fig. 1. Using an AI solution (BoneView, Gleamer), the buckle fracture (arrow) of the first metatarsal bone could be easily missed or detected.



관찰 결과가 실제 임상 환경과 관련성이 부족할 수 있다는 점을 지적하고 있다. Krogue 등(13)은 고관절 골절 검출 모델을 참조하였을 때 전공의와 전문의 모두에서 정확도가 향상됨을 입증하였으며, 특히 영상의학과 전공의는 AI 참조 판독 시 펠로우십 교육을 받은 전문가 수준에 도달함을 보여주었다. 이는 국내 인공지능 기업의 손목골절 검출 솔루션의 식품의약품안전처 인허가 임상 시험에서도 확인이 되었는데, 진단이 까다로운 주상골 골절 진단에서 AI 판독 결과를 참고하였을 때 영상의학과 전공의와 전문의 모두에서 민감도와 AUC가 매우 큰 폭으로 상승하였다(14).

다수의 인공지능 기업이 척추, 골반, 상하지, 늑골 등의 부위에서 골절을 검출하는 다양한 적응증의 상용화 제품을 유럽과 미국에서 출시하고 있는데, 2023년 발표된 한 시장조사 자료에 의하면 (15) 총 15개의 업체에서 개발한 21개의 골절 관련 상용화 AI 제품이 확인되며, 3개 제품은 시제품/시장 출시 전 단계에 있었다. 상용화 제품의 대부분(14/21)은 일반촬영에서 골절 진단을 위해 개발되었으며, 나머지는 CT 기반 솔루션이었다. 성인과 소아 모두에게 사용하도록 명시된 제품은 단 3개에 불과하며, 나머지는 모두 성인을 적응증으로 하고 있다. 모든 제품은 의료진의 판독을 보조하거나 분류하는 것을 목적으로 하고 있다. 미국과 유럽에서 허가받은 주요 일반촬영 골절 검출 제품군을 Table 3에 정리하였다.

상용화 제품의 개발 정보, 성능 및 검증에 대한 증거, 그리고 목표 대상 집단에 대한 정보는 항상 명확하지는 않아서 특정 임상 환경에서 얼마나 잘 작동할지, 그리고 구매할 “가치”가 있는지 알기가 어려운데, 골절 검출 AI 제품의 현장 도입 시 도움이 될 만한 정보와 상용화 제품의 진단 성능에 대한 연구를 소개하니 참고하기 바란다.

미국과 유럽에서 제품 허가를 받은(FDA-cleared, Conformité Européenne [CE]-marked) 일반촬영 골절 검출 상용화 제품 7종을 적용한 17개 연구의 메타분석에 의하면(16) 대부분의 AI 솔루션이 늑골과 척추를 제외한 해부학적 부위에서 우수한 진단 정확도를 보였으며, AI 참조 판독 시 가장 높은 성능을 보고하였다. 본 메타분석에서 BoneView 제품이 가장 많이 검증되었는데, 종합 민감도 0.90 (95% confidence interval [이하 CI]: 0.85, 0.94) 및 특이도 0.89 (95% CI: 0.87, 0.92)를 나타냈다. 그리고 AI 단독 판독(stand-alone AI) 연구 5개 중 4개는 우수한 민감도(>90%)와 중

Table 3. Conventional Radiography Based Commercial Products for Detection of Fracture

Company	Product	Released	CE-Certified	FDA-Cleared
Radiobotics	RBfracture	2022	O	X
Annalise-AI	Enterprise CXR Triage Trauma	2020	O	O
Milvue	Suite - SmartUrgences	2020	O	X
Gleamer	BoneView Trauma	2020	O	O
Qure.AI	qMSK	2020	O	X
AZmed	Rayvolve	2019	O	O
Imagen Technologies	FractureDetect	2020	X	O
	OsteoDetect	2018	X	O

CE = Conformité Européenne, FDA = Food and Drug Administration

간 정도의 특이도(80%–90%)를 보였고, 1개는 매우 낮은 민감도(<60%)와 우수한 특이도(>95%)를 보였다. AI 단독 판독과 AI 참조 판독 간에는 민감도에서 유의미한 차이가 없었지만 특이도는 AI 참조 판독에서 유의미하게 더 높았다. AI 비참조 판독은 AI 참조 및 AI 단독 판독에 비해 민감도가 낮았으며, 특이도는 AI 단독 판독에 비해 높았고, AI 참조 판독과는 유의미한 차이가 없었다. 그리고 앞서 기술한 것처럼 전반적으로 대부분의 해부학적 부위에서 높은 민감도와 중간에서 높은 수준의 특이도가 나타났지만, 늑골골절과 척추골절의 경우 진단 정확도(늑골골절-낮은 종합 민감도 0.66, 척추골절-낮은 종합 특이도 0.63)가 낮은 것으로 나타났다. 따라서 늑골과 척추의 골절 진단 모델 개발의 경우 더 광범위한 훈련 세트와 검증 테스트가 필요한 것으로 보인다. 이외에도 기업 펀딩 연구($n=4$)는 지원을 받지 않은 연구($n=11$)보다 민감도가 다소 높고(5% 상승) 특이도가 낮음(3% 감소) 흥미로운 결과도 보여주고 있다. 이는 기업 펀딩의 경우 대상자 선정과 평가지표를 포함한 연구 설계를 솔루션의 장점을 강조하는 방향으로 최적화한 것이 요인으로 추측된다.

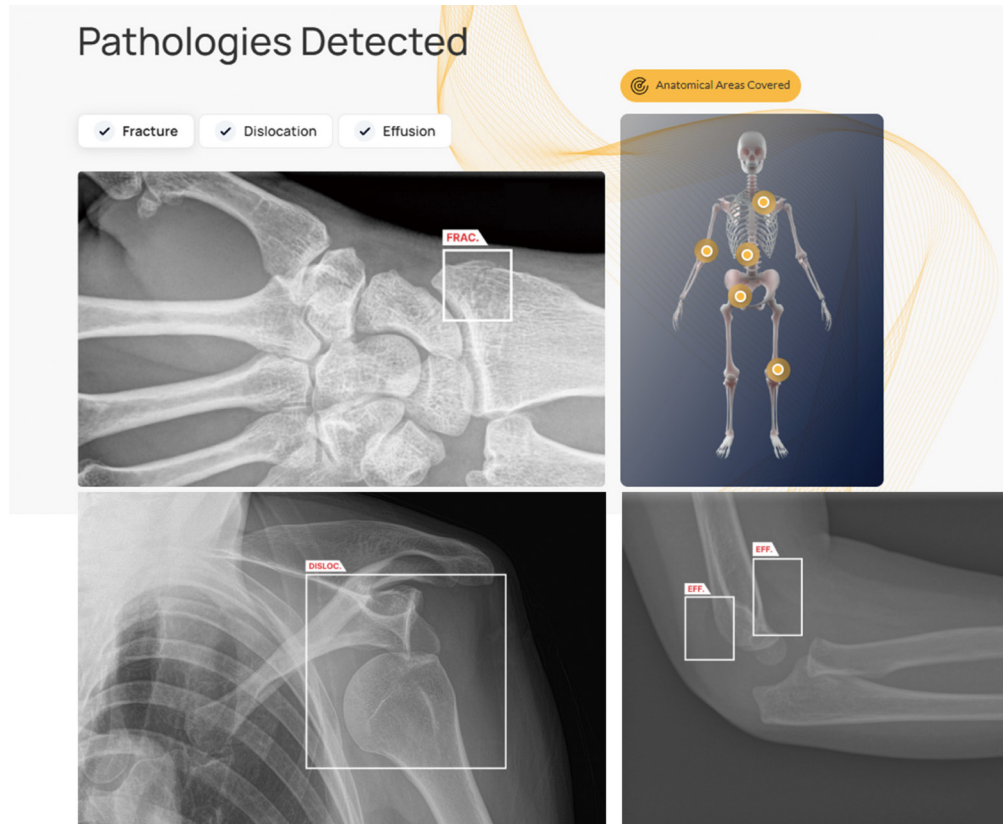
현재 상용화된 AI 알고리즘은 우수한 성능을 보여주고 있지만 상당한 한계를 가지고 있다. 예를 들어, 골절과 탈구가 복합된 경우 AI 모델에서 오류가 발생할 수 있다. 즉 AI는 단일 병소의 평가에는 탁월하지만 골절과 연관된 주요 이상 소견을 배제하거나 포함하는 인간의 추론 및 빠른 인지 능력은 아직 온전히 재현할 수 없다. 하지만 골절과 수반된 몇몇 소견들도 검출할 수 있는 모델이 제품화되기 시작하였고(Fig. 2), AI 모델 검출 결과와 영상 의학과 전문의의 판독 간의 불일치를 비교하여 판독에서 누락된 외상성 이상(골절, 탈구, 국소 골병변, 관절삼출)의 상당 부분이 AI 모델을 통해 진단될 수 있음이 보고된 바 있다(17).

골절 검출 AI 알고리즘을 개발하는 기업과 현장 도입을 고려하는 의료진은 누적된 연구 결과와 현 상용화 제품의 한계 및 성능 범위를 잘 파악하여 자동화의 효과가 잘 발휘될 수 있도록 임상업무 흐름 상의 미충족 수요를 찾는 것과 동시에 최근 관심을 받는 환자중심결과(patient-centered outcome; PCO)를 고려한 솔루션의 실사용증거(real world evidence; RWE) 구축이 중요할 것으로 보인다.

골연령

국내 최초 AI 의료기기 인허가 제품이 골연령 판독 보조 솔루션(VUNOMed-BoneAge, 주식회

Fig. 2. Product leaflet for AI software (Rayvolve, AZmed) detecting all types of fractures, dislocations, and joint effusions for all age groups (including pediatric) on the plain radiography of the following areas: hand, wrist, forearm, elbow, arm, shoulder, foot, ankle, leg, knee, femur, clavicle, hip, humerus, pelvis, tibia, ribs and spine. The software seamlessly integrates into a local PACS and sorts exams by priority order.



사 뷰노)이다. 골연령 판독은 좌측 손목과 손의 단일 이미지를 기반으로 비교적 표준화된 소견이 존재하기 때문에 딥러닝 알고리즘을 학습시키는데 이상적인 분야로(18) 국내외에서 상당수의 딥러닝 골연령 판독 제품이 이미 상용화되었다.

전통적으로 골연령 평가는 Greulich-Pyle (이하 GP) 및 Tanner-Whitehouse (이하 TW) 방법과 같은 좌측 손과 손목의 목측 관찰에 의존해 왔다. 이러한 방법들은 시간이 많이 걸리고 관찰자 간 및 관찰자 내 변이가 발생하게 된다. 특히 GP 방법의 경우 특정 골연령을 대표하는 표준 영상에서 어떤 부위와 소견에 중점을 두느냐에 따라 판독 결과의 차이가 크게 나타날 수 있다. 이러한 단점을 보완하고자 소개된 TW 방법은 손과 손목의 여러 부위의 성숙도를 여러 단계로 구분하여 스코어를 제시하는데 GP 방법과 비교 시 복잡하고 판독 시간이 더 길어서 의료현장에서의 선호도는 높지 않은 편이다. 그리고 부위별 성숙도의 단계 또한 목측 평가이므로 판독 재현성의 문제는 여전히 내포되어 있다. 따라서 골연령 평가의 이러한 한계 때문에 자동 판독에 대한 수요는 이전부터 존재해 왔다.

1950년 영국의 수학자 앨런 튜링이 AI 개념을 최초로 제시 후 비선형 문제를 해결할 수 있는 알고리즘의 고안과 함께 1980년대 2차 AI 붐이 일면서 골연령 판독의 초기 자동화가 시도되어 컴퓨

터 보조 수치골 및 수골 분석 시스템이 개발되었다(19-21). 우리나라에서는 2009년 표준 영상으로부터 추출된 정규화된 형태 모델에 의존하는 자동화 기술이 발표되었는데(22) 이러한 시스템들은 반자동화되었거나 여전히 시간이 많이 소요되는 등 임상적 유용성이 떨어져 시장 진입에는 한계가 있었다. 올해 노벨상 수상자인 제프리 힌튼의 2006년 심층 신뢰 신경망(deep belief network; DBN) 발표, 인터넷 보급에 기반한 풍부한 데이터 그리고 향상된 컴퓨터 성능을 기반으로 한 딥러닝의 등장과 발전은 자동화된 골연령 평가의 정확성과 효율성을 획기적으로 향상시켰다. 특히 합성곱 신경망(convolutional neural network; 이하 CNN)과 순환 신경망(recurrent neural network; RNN)은 영상 분류, 영상 분할, 패턴 인식에 효과적임이 입증되었고 3차 AI 붐과 함께 다양한 AI 기반 골연령 판독 알고리즘이 등장하였다(18, 23-25).

다양한 전통적 머신러닝 및 딥러닝 기법(회귀 기반 방법, 서포트 벡터 머신, 베이지안 네트워크, 결정 트리(decision tree), K-최근접 이웃(K-nearest neighbors), 인공신경망(artificial neural network; ANN), 합성곱 신경망(CNN)의 골연령 솔루션을 리뷰한 메타분석 연구에 의하면 골연령 자동 판독의 가중평균 성능은 평균절대오차(mean absolute error; 이하 MAE) 9.96개월의 참조표준과의 차이를 보였는데(26), 이는 현재 상용화된 제품들과 비교 시 성능이 다소 떨어진다.

2009년 소개되어 유럽을 중심으로 현재 약 200여 개 병원에서 널리 사용되고 있는 AI 골연령 상용화 솔루션의 선구자인 BoneXpert (Visiana, Holte, Denmark)는 13개의 뼈(요골, 척골, 세 손가락의 11개 단골)를 기반으로 왼손 일반촬영 사진을 분석하여 골연령을 판정한다(27). 영상이 BoneXpert AI 소프트웨어 서버에 전송되면, 소프트웨어는 분석된 각 뼈의 규칙적인 형태와 밀도, texture 분포를 학습한 능동적 형태 모델을 적용하여 13개의 뼈를 분석하며 최종 골연령은 GP 방법이나 TW 방법을 통해 결정된다. Booz 등(28)에 따르면, BoneXpert (version 2.1)가 판정한 골연령과 참조표준 간의 상관관계($r = 0.99$)는 의료진이 판정한 골연령과 참조표준 간의 상관관계($r = 0.90$)보다 유의미하게 높았으며, 임상현장에서 판독 시간을 대폭 감소시키는 것으로 보고되었다(28).

유럽에서 개발된 세 종류의 AI 기반 골연령 판독 프로그램(BoneXpert [v3.2.2], PANDA [v1.13.21], BoneView [v2.3.1.1])의 성능을 비교 분석한 연구를 살펴보면 세 가지 솔루션 모두 참조표준과 높은 상관관계($R^2 \geq 0.98$)를 보여 유럽 코카서스계 소아의 주요 연령대에서 매우 신뢰할 수 있는 GP 방법의 골연령 추정치를 제공한다(29). 전체 그룹에서 BoneXpert는 BoneView (Gleamer, Paris, France)와 PANDA (ImageBiopsy Lab, Vienna, Austria)보다 낮은 평균제곱근 오차(root mean square error; RMSE)를 나타냈는데(각각 0.62년, 0.65년, 0.75년), 임상현장에서 가장 흔한 골연령 검사 연령대(여성 4.8–15.5세, 남성 4.9–17.0세)에서는 이러한 차이가 줄었다. 솔루션 간의 차이는 저 연령대와 고 연령대에서 두드러졌는데 가장 흔한 연령대에서는 BoneXpert와 BoneView의 경우 오차가 연령대에 따라 약간의 차이를 보였으나 PANDA는 차이가 없었다. 이외에도 흥미로운 결과는 AI 솔루션 간의 표준 편차는 의료진 간의 표준 편차보다 낮았는데 이는 AI 골연령 판독의 높은 신뢰성을 시사한다.

국내 제품이 포함된 또 다른 상용화 제품 2개(BoneXpert v3.1.4., VUNO Med-BoneAge v1.1.)의 비교 연구(판독자 간의 일관된 평가를 보장하기 위해 표준화된 방법론을 적용함)에서도 두 알

고리즘 모두 전문가에 의해 구축된 참조표준과 강한 상관관계($r = 0.99$)를 보였으며, 두 알고리즘 간의 RMSE 차이는 유의미하지 않았고, 의료진의 RMSE와도 차이가 없었다. 알고리즘은 아동의 연령대와 성별에 따라 골연령 예측에 약간의 차이를 보였지만, 이는 임상적으로 무시할 만한 수준이라고 보고하였다(30).

다음으로 한국 식품의약품안전처의 승인을 받은 상용화 모델들에 대하여 간략히 살펴보겠다. VUNO Med-BoneAge는 GP 방법으로 분석된 국내 의료기관의 18,940장 왼손 일반촬영 영상을 이용하여 학습되었다. 분석할 영상이 입력되면 시스템은 유사한 영상과 함께 확률에 기반하여 GP 방식에서 정의한 연령구간(남자 31개, 여자 27개) 중 가장 유사한 골연령을 최상위 3순위까지 제안한다. 솔루션 개발에 기반이 된 연구에 의하면 1순위 골연령의 정확도는 69.5%였으며, 1순위, 2순위, 3순위 골연령을 모두 고려했을 때 정확도는 93%였다(25). 그리고 AI 참조 판독 시 시간이 29% 단축되는 것으로 보고하였다. 엄밀히 말하면, VUNO Med-BoneAge는 딥러닝 모델이 제시한 세 가지 골연령 결과 및 영상 중 하나를 의료진이 선택해야 하므로 완전 자동화 시스템이라기보다는 반자동 시스템이었는데, 현재는 모델을 업데이트하여 Top 3 골연령의 가중평균치를 계산한 골연령 결과값(VUNO score)을 제공한다(Fig. 3).

국내 최초 TW3 방식의 골연령 분석 소프트웨어인 Boneage.io (헬스허브)는 만 7세 이상-15세 미만 연령대의 소아 및 청소년의 골연령을 측정하는 AI 기반 솔루션으로 요골, 척골 및 1, 3, 5 손가락의 11개 단뼈를 포함한 13개 뼈의 성장판 영역을 자동으로 찾고 해당 부위의 뼈 성숙도에 따른 등급 점수로부터 골 성숙점수(RUS score; analysis of the radius, ulna, phalanges, and metacarpal bones)를 산출 후 변환표를 이용하여 골연령을 예측한다. 골연령표가 0.1년 단위로 제시되기에 GP 방법과 비교하여 보다 정밀한 결과 값을 제시한다. 솔루션 개발에 기반이 된 연구에 의하면 참조표준 기준으로 알고리즘의 평균절대오차(MAE)와 평균 제곱근오차(RMSE)는 각각 0.46년과 0.62년이었으며 1년 이내의 정확성은 97.6%를 보고하였다(23).

2020년 제품 허가를 받은 MediAI-BA (크레스콤)는 GP와 TW3 방법의 한계를 극복하기 위해 TW3 방법의 13개 영역 대신, 7개 영역(요골, 척골, 세 번째 손가락의 원위지골, 중간지골, 근위지골, 중수골, 첫 번째 손가락의 중수골)을 대상으로 하는 변형된 TW 방법과 GP 방법에 기반하여 하이브리드 방식으로 알고리즘을 개발하였다(Fig. 3). GP 방법의 전체 손 이미지 분석으로 TW3 방법에 포함되지 않은 모든 영역을 커버할 수 있으며, 변형된 TW 방법의 개별 관심영역(region of interest; 이하 ROI)의 성숙도를 기존의 9개 구간이 아닌 데이터 기반으로 보다 세밀하게 분류하여 평가함으로써 성능의 향상을 도모하였는데, 식품의약품안전처 인허가 임상시험 결과 참조표준과의 차이는 평균절대편차(mean absolute difference; MAD) 0.39세였다(31).

하지 및 척추 정렬 측정

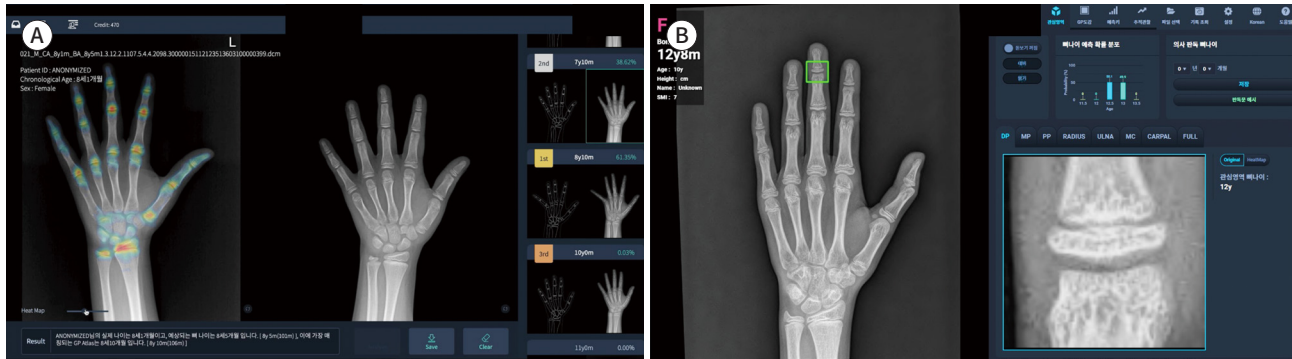
정확하고 표준화된 하지/척추 기하학 및 정렬 측정은 수술 전후에 요구되는 중요한 작업이다. 그러나 이러한 영상기반 측정은 다양한 지표표를 포함하고 있어 단순하지 않고 시간 소모적이고 반복적인 작업으로 임상의료진 뿐만 아니라 영상의학과 의사에게도 부담이 되고 있어 자동화의 요

Fig. 3. AI-aided bone age estimation using VUNO Med-BoneAge (A) and mediAI-BA (B).

A. VUNO Med-BoneAge provides three most likely GP bone age standards with the highest probabilities. The latest version also added a bone age (VUNO score), which is a weighted average of the top three bone ages. Image courtesy of VUNO Inc., VUNO Med®-BoneAge™

B. MediAI-BA provides integrated bone age (12 y 8 m in the left upper corner) based on the detailed ROIs and a holistic image to complement the limitations of GP and TW methods. Distal interphalangeal joint of the 3rd finger among the detailed ROIs was selected and shown to be 12y of bone age as local maturity.

GP = Greulich-Pyle, ROI = region of interest, TW = Tanner-Whitehouse



구가 꾸준히 있어 왔다.

하지길이 불일치(leg length discrepancy; 이하 LLD)는 다리 길이가 서로 다른 상태로, LLD의 유병률은 약 90%로 추정되며 대부분은 그 정도가 크지 않아 임상적으로 유의미하지 않지만, 차이가 클 경우 관절염, 엉덩이 통증, 무릎 통증, 허리 통증, 척추 측만증 및 보행 패턴 변화와 같은 고통스럽고 장애를 초래할 수 있는 합병증으로 이어질 수 있다(32-34). 다리와 무릎의 각도 부정렬 또한 골관절염의 발생 및 진행과 관련된 또 다른 요인이다. 하지길이 불일치(LLD) 및 부정렬은 하지의 전후면 일반촬영 사진을 통해 일련의 길이와 각도 측정으로 평가할 수 있는데(Fig. 4A, B), 포괄적인 영상학적 평가를 수행하기 위하여 시간이 많이 소요된다는 점 그리고 문헌에서 검증된 수많은 측정에 필요한 정확한 랜드마크를 기억해야 한다는 점이 난관이다.

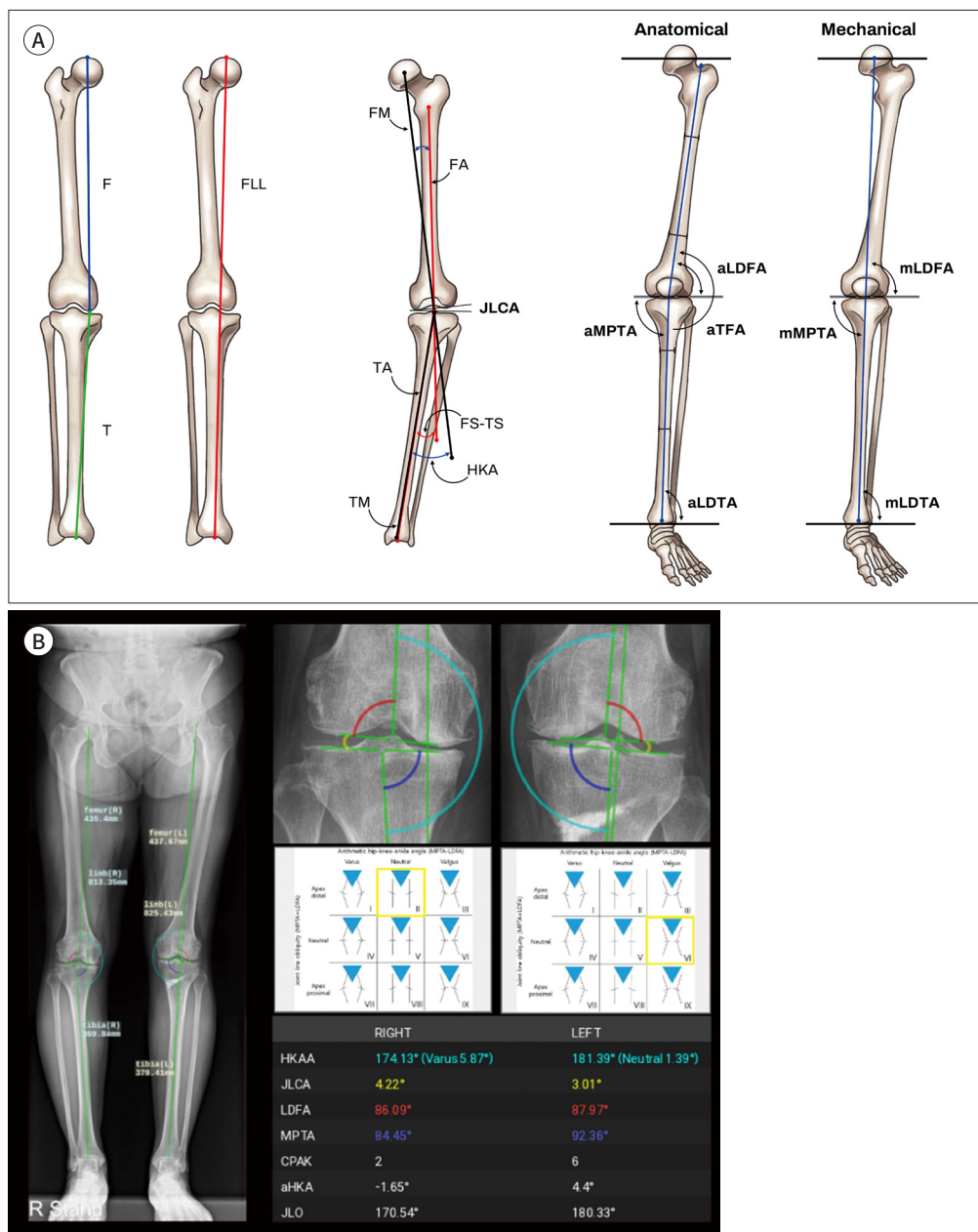
하지정렬 측정 자동화 연구들은 성능의 향상을 위하여 다양한 알고리즘을 시도하고 있다. 소아 환자의 하지길이 불일치를 평가하기 위해 Zheng 등(35)은 U-Net을 적용하여 일반촬영 영상에서 대퇴골과 경골을 분할한 다음 다리 길이를 계산하였고, Schock 등(36)은 체중부하 양측 하지영상을 사용하여 해부학적-기계적 각도와 고관절-무릎-발목 각도를 계산하여 하지정렬을 자동으로 분석하는 딥러닝 방법을 제안했다. 효율적인 영상 분할을 위해 개발된 합성곱 신경망인 U-Net을 사용하여 각도를 정량적으로 측정하는 데 필요한 이진(binary) 마스크 이미지를 생성하였으며, 모델의 과적합을 방지하기 위해 다양한 데이터 증강 기술을 채택하였다. Tack 등(37)은 하지정렬 평가를 위한 랜드마크의 위치를 정하기 위해 다단계 접근 방식을 제안하였다. 먼저 객체 인식(object detection) 알고리즘인 YOLO를 적용하여 전체 하지영상에서 관심영역(ROI)을 검출하고, 개별 ROI 내의 랜드마크는 Residual Network (ResNet)를 사용하여 찾았으며 평균방사형오차(mean radial error)를 손실 함수로 사용하여 회귀 오차를 최소화했다. 그러나 대조도가 저하된 일반촬영 사진에서 고관절-무릎-발목 각도 측정에 상당한 오차가 발생한다는 제한점도 밝히고 있다. Lee 등(38)은 정형외과용 하드웨어가 이식된 환자를 포함하여 하지길이를 자동으로 측정하는 모델을 제

Fig. 4. Lower extremity measurement.

A. Schematic overview of measurements performed on long leg radiographs. These are crucial for diagnosing conditions like LLD and genu varus/valgus, especially in the context of total knee replacement surgery and osteoarthritis research.

B. Automatic lower extremity measurements. The DL software provides representative joint alignment index along with CPAK type suggestion.

aHKA = hip-knee-ankle angle (anatomical angle), aLDFA = anatomical lateral distal femoral angle, aLDTA = anatomical lateral distal tibial angle, aMPTA = anatomical medial proximal tibial angle, aTFA = anatomic tibiofemoral angle, CPAK = coronal plane alignment of the knee, DL = deep learning, F = femur length, FA = femoral anatomical axis, FLL = full leg length, FM = femoral mechanical axis, FS-TS = femoral shaft-tibial shaft angle (anatomic angle), HKA = hip-knee-ankle angle (mechanical angle), HKAA = hip-knee-ankle angle (mechanical angle), JLCA = joint line convergence angle, JLO = joint line obliquity, LDFA = lateral distal femoral angle, LLD = leg length discrepancy, mLDFA = mechanical lateral distal femoral angle, mLDTA = mechanical lateral distal tibial angle, mMPTA = mechanical medial proximal tibial angle, MPTA = medial proximal tibial angle, T = tibia length, TA = tibial anatomical axis, TM = tibial mechanical axis



안했는데, 알고리즘은 다단계 접근 방식(ROI 검출, 뼈 분할, 랜드마크 검출 및 다리 길이 계산)로 구성되었으며 하드웨어가 이식된 환자에서도 영상의학과 전문의와 유사한 성능을 보고하였다. 이 연구는 국내 IT기업과 영상의학과 교수 그리고 공학자가 협업하여 모델을 개발한 케이스이다.

이외에도 최근 국내 영상의학과와 정형외과 연구진이 무릎 정렬 평가의 주요 각도 지표를 자동으로 측정하는 딥러닝 기반 알고리즘을 개발하여 우수 학술지에 게재하였는데, 내부 기관과 외부 기관 검증에서 판독자와 알고리즘 간의 하지 정렬 측정값은 유의미한 차이를 보이지 않았고, AI 자동 측정에 걸린 시간은 데이터 로딩 시간을 포함하여 환자 당 평균 50 초로 수동 측정보다 3.44 배 빠른 속도였다고 보고하였다(39). 각도뿐만 아니라 길이 측정을 포함한 13개 지표 모두에서 뛰어난 성능을 보인 하지 정렬 자동 측정 소프트웨어도 국내 연구진에 의하여 소개된 바 있다(40).

FDA 510(k) 승인을 받은 AI 소프트웨어(LAMA, ImageBiosy Lab)를 활용한 AI 보조 하지 측정이 영상의학과 의사가 측정한 결과만큼 정확하면서도 소요 시간이 획기적으로 감소한다는 최신 연구 결과가 발표되었는데(41), AI가 시간을 절약할 수 있는 것은 잘 알려져 있지만, 하지 영상 평가에서 절약할 수 있는 시간의 규모를 구체적으로 제시하고 있는 것이 흥미롭다. 이 연구에서 AI는 13개 측정 지표 중 9-12개에서 우수한 결과를 보여주었으며, 다리 길이 측정의 정확도가 각도 측정보다 높았는데, 세 명의 의료진과 비교했을 때 전체적인 성능은 비슷하다고 보고하였다. 그러나 시간 면에서는 큰 차이가 있었는데, 세 명의 의료진은 각각 250초, 282초, 236초의 측정 시간을 기록한 반면, AI는 평균 38초가 소요되었다.

척추측만증은 척추가 옆으로 휘어지고 회전하는 변형이 나타나는 상태로 체중부하 일반촬영이 척추측만증을 평가하는 일차적인 방법이다. 일반촬영에서 척추측만증을 포함한 다양한 척추변형을 평가하는 여러 가지 지표 중 정면상에서 측정하는 Cobb 각도가 가장 많이 사용되는 영상학적 마커로 본 종설에서는 이 부분에서 초점을 맞춘다.

Cobb 각도가 10° 이상인 경우 측만증에 해당하며(42), 일반적으로 10°-20°는 경도 측만증, 20°-40°는 중등도 측만증, 40° 이상은 중증 측만증으로 진단한다. Cobb 각도의 측정은 약 4°에서 8° 정도의 관찰자 간 및 관찰자 내 변동을 가지며, 이는 척추측만증을 판단하는 10° 기준선을 고려할 때 중요한 차이이다(43). 이러한 부정확성 때문에 Cobb 각도는 AI 적용의 주요한 대상이 되고 있으며, 다양한 연구에서 Cobb 각도 자동화를 시도한 모델을 생성하려는 노력이 진행 중이다.

딥러닝의 등장 전 컴퓨터 보조 측정도 우수한 성능을 보이긴 하였으나 판독자가 적절한 끝척추와 끝종판을 수동으로 선택해야 하는 번거로움이 있었다. 합성곱 신경망(CNN)을 사용한 Cobb 각도 측정 알고리즘의 등장과 함께 작업의 완전 자동화가 가능해지긴 하였으나 아직도 임상 현장에서의 실제 수요를 온전히 만족시키고 있지는 못하는 것으로 보인다. 즉 많은 연구가 청소년기 특발성 척추측만증에만 초점을 맞추었으며, 성인척추변형(adult spinal deformity)를 대상으로 한 연구 결과는 제한적이다. 이뿐만 아니라 소곡선(minor curve) 측정과 다양한 자세에서의 정확성을 검증한 연구 결과 데이터도 많이 부족한 상태이다.

최근 Cobb 각도 측정 알고리즘 개발에 있어서 여러 코호트의 학습데이터를 활용하는 것이 특정 코호트만을 대상으로 한 경우보다 더 효율적인 AI 학습 방법이 될 수 있음을 입증한 연구가 발표

되었다(44). Kato 등(44)은 소아 및 성인척추변형의 Cobb 각도 측정의 정확성을 향상시키기 위해 1,612개(소아와 성인 약 2대 1 비율)의 전체 척추 일반촬영 영상을 학습 데이터로 사용하여 세 가지 독립적인 AI 알고리즘(소아+성인 학습, 소아 학습, 성인 학습 AI)을 개발하고, 알고리즘의 정확성을 평가하기 위해 내부 검증 데이터셋을 대상으로 각 AI 알고리즘과 네 명의 척추 전문가의 수동 측정 평균 간의 평균절대오차(MAE)와 군내상관계수(intraclass correlation coefficient; 이하 ICC)를 계산했다. 세 가지 AI 알고리즘 중에서 소아+성인 학습 모델이 가장 높은 정확성(MAE 2.8°, ICC 0.97)을 보였다.

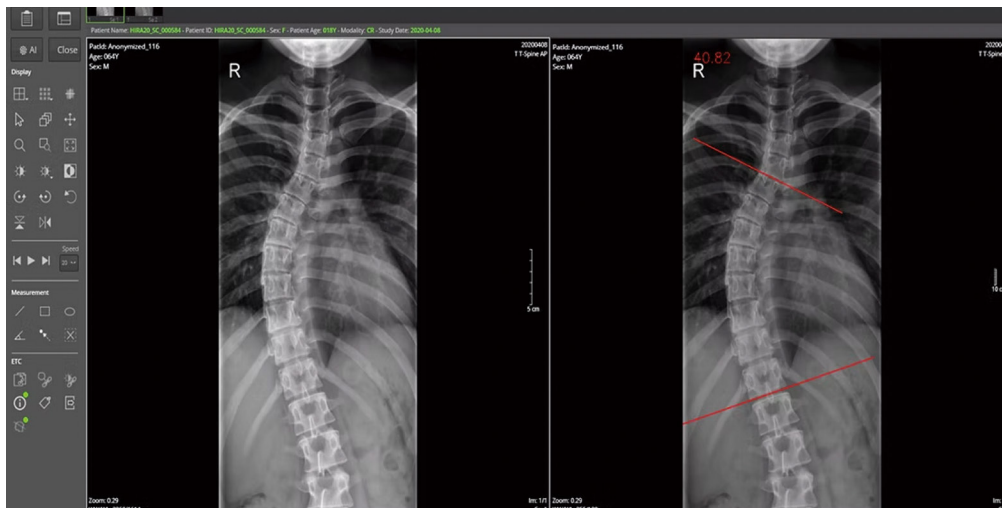
우리나라의 척추 정렬 영상 측정용 AI 모델의 인허가 현황을 살펴보면 현재까지 2개의 제품이 확인이 되는데(Fig. 5) 학술지에 성능을 보고한 연구 결과가 검색되지 않아서, 유럽 규제기관의 인증을 받은 상용화 AI 모델 두 가지(BoneMetrics, Gleamer; IB Lab SQUIRREL™, ImageBiopsy Lab)를 소개한다. BoneMetrics 솔루션은 소아와 성인의 비율이 대등한 집단을 대상으로 두 명의 영상의학과 의사와 한 명의 정형외과 의사에 의하여 구축된 참조표준과 비교 시 주곡선의 MAE가 2.6° (소아 1.9° 대 성인 3.3°)의 성능을 입증하였고(45), IB Lab SQUIRREL™ 솔루션은(Fig. 6) 척추 측만증 환자 182명의 250개 커브를 대상으로 한 검증에서 194개의 커브에서 솔루션 분석이 가능하였으며 4명의 영상의학과에 의해 구축된 참조표준과 비교 시 평균 0.16°의 Cobb 각도 차이가 있음을 보고하였다(46).

관절염 등급 분류

근골격계를 전공하는 의료진 사이에서도 판독편차가 크게 발생하는 대표적인 등급판정 시스템이 무릎관절염 심각도를 평가하는 Kellgren-Lawrence (이하 KL) grading이다. 1957년 논문 발표 이후 1961년 World Health Organization (WHO)이 역학연구 목적의 영상학적 도구로 승인한

Fig. 5. Automatic Cobb angle measurements by the deep learning software (DEEP:SPINE, Deepnoid Inc.) on spine radiographs.

The software detects the superior and inferior endplates of the upper and lower end vertebrae in frontal images of the spine and measures the Cobb angle.



바 있다. 등급판정 시 주요 소견에 대한 정의가 다양한 용어로 기술되어 있으며 양성 기준 또한 모호하여 판독편차의 취약성에 대한 비판이 있음에도, 일반촬영을 통한 무릎관절염 등급판정 시스템은 현재까지도 임상현장에서 그 역할이 감소하지 않고 널리 사용되고 있다.

이러한 배경과 함께 양질의 학습 데이터와 참조표준을 수집할 수 있는 공개 데이터세트(Osteoarthritis Initiative [이하 OAI], Multicenter Osteoarthritis Study [MOST])가 잘 구축되어 있어서, 다수의 AI 모델이 개발되었고 내부 및 외부 검증이 이루어졌다(47-54). CNN 알고리즘을 기반으로 개발된 KL grading 모델의 성능지표를 보고한 주요 연구를 살펴보면 다음과 같다(Table 4). 알고리즘, 학습 및 테스트 데이터세트, 그리고 성능 평가 방법 등의 차이로 인해 무릎관절염 등급 분류를 위해 고안된 다양한 딥러닝 모델을 직접 비교하는 것은 주의를 필요로 하는데, KL 등급별 성능에 있어서 무릎관절염이 진행된 KL 3등급과 4등급은 우수한 결과를 KL 2등급 이하에서는 다소 만족스럽지 못한 결과가 나타나는 경향성은 확인이 된다.

최근 KL 등급별 AI 판독의 메타분석 연구가 발표되었는데(55), 총 19개의 연구를 분석한 결과 KL 0, 1, 2등급의 민감도는 각각 86.7%, 64.0%, 75.0%이었고 KL 3, 4등급은 84.8%, 90.3%으로, 특히 KL 4등급에 대한 민감도는 매우 높지만 KL 1과 2등급에 대한 민감도는 낮아 개선의 필요성을 강조하였다. 골극과 관절간격을 기반으로 하는 KL 등급 시스템은 종종 KL 1과 KL 2등급 간에 영상학적 해석의 중복을 초래하며, 이는 판독자를 혼란스럽게 할 뿐만 아니라 딥러닝 모델에서 정

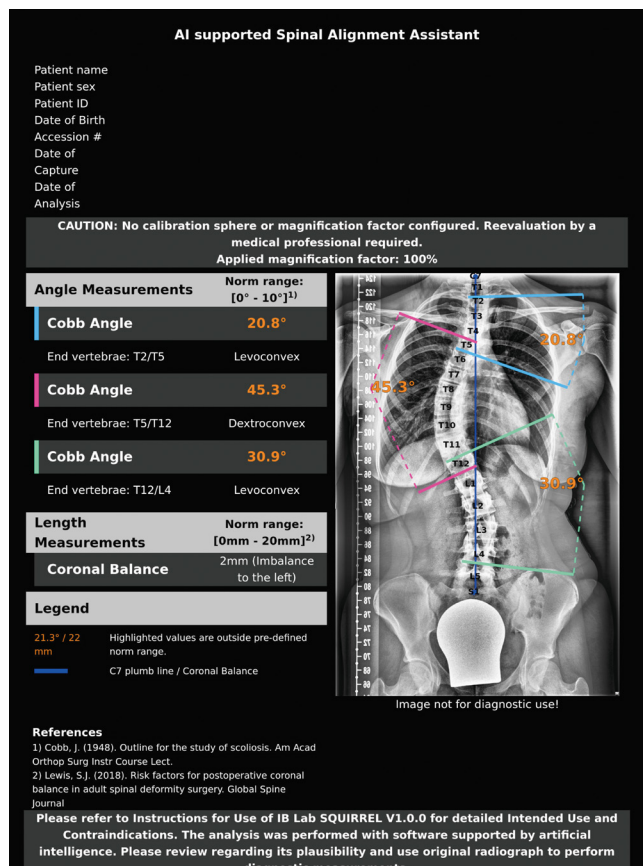


Fig. 6. Exemplary Conformité Européenne-marked IB Lab SQUIRREL™ (ImageBiopsy Lab, Austria) printed report.

The solution is indicated for use in patients aged 10 and over undergoing full spine radiographs.

Table 4. Comparison of Performance Metrics of KL Grading of AI Models

	AUC	Kappa	Accuracy	Sensitivity	Specificity	Test
Tiulpin et al. (47)	0.93					External (OAI)
Liu et al. (48)			0.83	0.78	0.95	Internal
Norman et al. (49)				KL0, 1 = 71% KL2 = 69% KL3 = 86% KL4 = 85%	KL0, 1 = 86% KL2 = 84% KL3 = 98% KL4 = 99%	Internal
Nguyen et al. (50)		0.79		53%		External (MOST)
Tiulpin & Saarakkala (51)		0.82	0.67			External (MOST)
Kim et al. (52)	KL0, 1 = 0.80 KL2 = 0.69 KL3 = 0.89 KL4 = 0.95					Internal
Thomas et al. (53)		0.86	0.66			Internal (OAI)
Brejneboel et al. (54)	0.84	0.88	KL0 = 79% KL1 = 73% KL2 = 83% KL3 = 92% KL4 = 96%	KL0 = 100% KL1 = 27% KL2 = 75% KL3 = 94% KL4 = 86%	KL0 = 76% KL1 = 98% KL2 = 85% KL3 = 91% KL4 = 100%	External

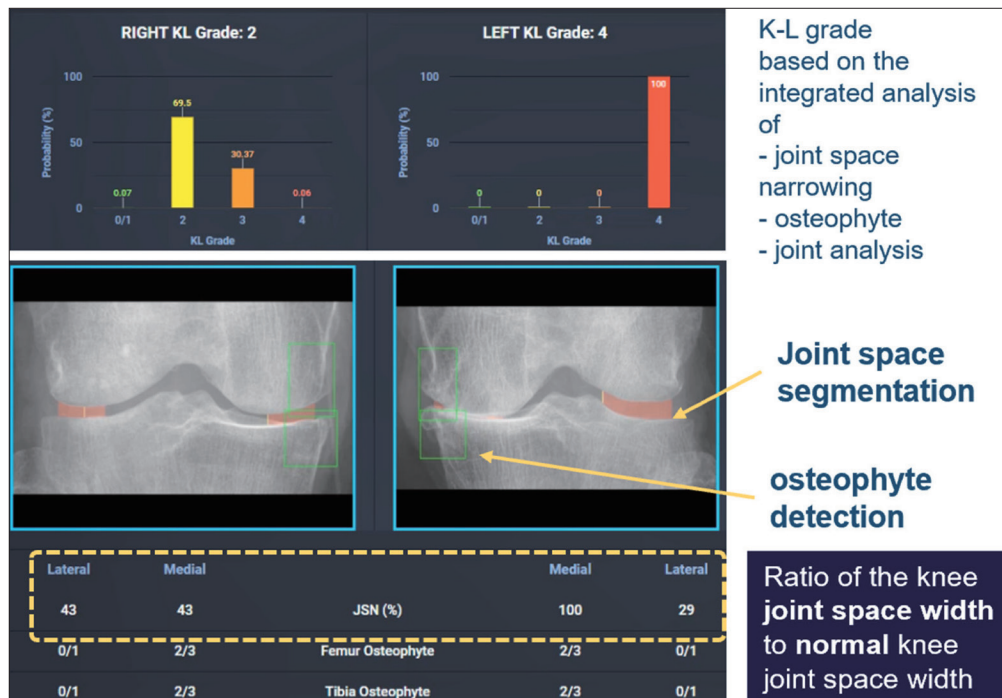
AUC = area under curve, KL = Kellgren-Lawrence, MOST = Multicenter Osteoarthritis Study, OAI = Osteoarthritis Initiative

확한 매개변수 설정을 방해한다. KL 1등급과 2등급 사이의 영상소견의 유사성, 즉 관절간격 감소량, 골극 모양/크기, 연골하 경화증 및 뼈 변형 유무는 결과 해석에 영향을 미치는 교란 요인으로, 현재의 CNN 모델은 이 부분에서 어려움을 겪고 있으며, 종종 작은 데이터셋에서 과적합되어 새로운 데이터에서 일반화에 실패한다. 또한 KL 등급 시스템의 주관성으로 인해 연구 간 해석에 불일치가 발생하는데(56), Kohn 등(57)은 KL 2등급에 대한 정의가 “경미한 관절간격 협착이 있는 명확한 골극”에서 “관절간격 감소가 없는 명확한 골극”으로 진화하고 있음을 강조하기도 하는 등, 논란의 여지가 있는 등급 정의는 연구 결과 및 모델 성능에 편향을 초래한다. 하지만 AI 솔루션의 등급 판정 성능 이슈와 별개로 AI 판독과 숙련된 근골격계 영상의학과 전문의 간의 일치도가 근골격계 영상의학과 전문의 간의 일치도와 유사한 것으로 보고되고 있어 향후 전문성과 경험에 의한 판독 편차의 취약성을 보완해 주는 AI 역할이 기대가 된다.

우리나라에서 KL 등급은 의료보험 적용의 중요한 기준이지만 등급 판정과 관련된 논란과 한계가 있다; 건강보험심사평가원은 KL 등급을 기반으로 무릎관절염의 질환 진행 정도를 평가하고, 관절치환술의 보험수가 적용 여부를 판단하고 있다. 몇몇 연구에서 KL 등급의 관찰자 간 및 관찰자 내 신뢰도는 훈련을 통해 더욱 향상될 수 있음이 밝혀지기도 했지만, 관찰자 간 및 관찰자 내 불일치는 실제 임상현장에서 여전히 존재한다(56, 58, 59). 따라서 KL 등급은 판독자의 경험과 전문성에 따라 달라질 수 있으며, 이는 환자 치료 방향의 차이로 이어질 수 있다.

이에 건강보험심사평가원은 2021년부터 과학기술정보통신부 산하 정보통신산업진흥원(National IT Industry Promotion Agency; NIPA)이 주관하는 ‘AI 융합 의료영상 진료·판독시스템 구축’에 같이 참여하여 3년 과제로 슬관절염 솔루션의 개발을 공모한 바 있다. 여기서 개발된 ‘Me-

Fig. 7. Exemplary Korea Ministry of Food and Drug Safety-approved MediAI-OA (CRESCOM) report.



diAI-OA' 제품이 현재 무릎관절염 보험 급여 심사 업무에 활용되고 있다(Fig. 7). 인허가 임상시험은 두 기관의 영상의학과와 정형외과에서 수행되었으며 KL 등급별 및 부위별 골극 진단에 대한 솔루션의 성능을 검증하여, 2024년 식품의약품안전처로부터 2등급 의료기기 제조 허가를 받았다.

미국 FDA와 유럽 CE 승인 제품인 Knees Osteoarthritis Labeling System (KOALA, ImageBiopsy Lab)은 내측 및 외측 관절 최소 폭, KL 등급, 관절 간격 협착, 골극 및 경화증에 대한 OARSI 등급을 판정하여 리포트를 생성할 수 있는데(Fig. 8) (60), 이 소프트웨어를 사용하면 영상의학과 의사와 정형외과 의사가 무릎 골관절염 등급을 판정하는 데 있어 일치율과 정확도가 높아지는 것으로 나타났다(61, 62).

류마티스 내과 영역에서도 일반촬영 판독이 진료에 핵심이 되는 임상상황이 다수 존재한다. 특히 강직척추염의 주요 소견 중 하나인 천장관절염은 영상 판독 시, 양성 소견이 미묘하게 나타나 진단 자체가 모호할 수 있고, 등급판정의 객관성과 재현성이 떨어지는 한계가 있어, 강직척추염 확진 및 보험 적용에 어려움을 겪는 경우가 드물지 않다. 따라서 천장관절염의 등급판정은 KL 등급판정의 예처럼 인공지능 기술을 이용한 판독보조 솔루션의 임상현장 도입이 유망한 주제이다. 아직까지 FDA와 국내 식품의약품안전처 허가 제품은 확인되고 있지 않아서 유럽과 국내 연구진이 발표한 연구개발 단계의 모델을 간략히 소개하겠다.

다수의 유럽 국가와 의료기관에서 수집된 코호트(Patients With Axial Spondyloarthritis: Multicountry Registry of Clinical Characteristics [PROOF], German Spondyloarthritis Inception Cohort [GESPIC])를 바탕으로 개발된 천장관절염(sacroiliitis) 일반촬영 진단 모델의 검증 및 테스트 세트의 AUC는 각각 0.97 및 0.94, 민감도와 특이도는 검증의 경우 88%와 95%, 테스트

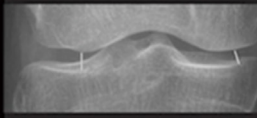

Kellgren & Lawrence (KL) Grade		
KL-Grade (0-4)	1	
OARSIS Grade		
Joint Space Narrowing (0-3)	1	
Sclerosis (0-3)	0	
Osteophytosis (0-3)	1	
Joint Space Measurements		
Laterality	Lateral	Medial
Joint Space Width [mm]	4.6	4.4
		
Image not for diagnostic use!		
Standardized Joint Space Measurements		
Standardized JSW [mm]	4.7	4.5
Compartment Imbalance	51%	49%
Overview		
		
The analysis was performed with software supported by artificial intelligence. Please review regarding its plausibility and use original x-ray to perform diagnostic measurements.		

Fig. 8. Exemplary AI-KOALA printed report. AI-aided radiographic diagnosis of knee osteoarthritis leads to a higher association between clinical findings and diagnostic ratings. Adapted from Neubauer et al. *J Clin Med* 2023;12:744 (60).

트 세트의 경우 92%와 81%로 우수한 성능을 보였고, 모델과 참조표준 사이의 일치도(Cohen's kappa)는 검증과 테스트 세트에서 각각 0.79와 0.72였다(63, 64) 본 연구의 강점은 다양한 의료기관의 영상을 사용하여 이질적인 훈련 데이터세트를 구성하여 모델의 일반화 가능성을 높였는데 테스트 데이터에서의 성능이 크게 감소하지 않았다는 점이다. 그리고 테스트 데이터는 환자뿐만 아니라 판독자의 측면에서도 검증 데이터와 독립적이었다.

국내 기업과 내과 교수진의 협업으로 개발된 천장관절염 영상 판독 AI 모델의 유효성을 밝힌 논문이 국제 학술지에 게재된 바 있는데(65), 201개의 레이어로 구성된 DenseNet 알고리즘을 사용하였고, 모든 작업은 해당 기업의 노코드/로우코드 플랫폼 ‘딥파이(DEEP:PHI, ver.2.7.6; Deepnoid Inc.)’를 활용하였다. 해당 알고리즘은 Modified New York criteria 등급별 정확도를 제시하였고 모든 등급에 걸쳐 95% 이상의 강력한 성능을 보여주었다. 하지만 학습 데이터세트 그리고 내부 검증과 테스트 데이터가 모두 단일기관의 자료로만 구성되어 있어서, 외부 의료기관의 ‘처음 보는 데이터’에서의 성능은 향후 검증이 필요할 것으로 보인다. 환자군과 영상의 특성에 성능이

좌우되는 인공지능의 속성을 고려하면, 특히 강직척추염의 경우, 일반촬영 영상 등급판정의 불완전성과 다양한 촬영방법에 의한 골반영상의 이질성이 상용화 모델의 개발에 있어서 극복해야 할 요소로 생각된다.

골다공증 기회검진

골다공증은 뼈의 밀도가 감소하고 구조가 약해져 골절 위험이 증가하는 질병으로, 특히 노인 인구에서 흔하게 발생한다. 골다공증 진단의 최적표준(gold standard)인 골밀도 검사는 비용이 많이 들고 접근성이 제한적이다. 우리나라의 경우 국가건강검진 시 골밀도 검사가 포함되어 있지 않은 만 53세 이하 여성과 건강보험이 적용되지 않는 만 69세 이하 남성의 경우 골다공증이 있어도 특별한 증상이 없다 보니 검사를 받는 경우가 많지 않아 조기 진단이 어려웠다. 이에 AI를 활용한 영상 기반의 기회검진(opportunistic screening)이 주목받고 있다.

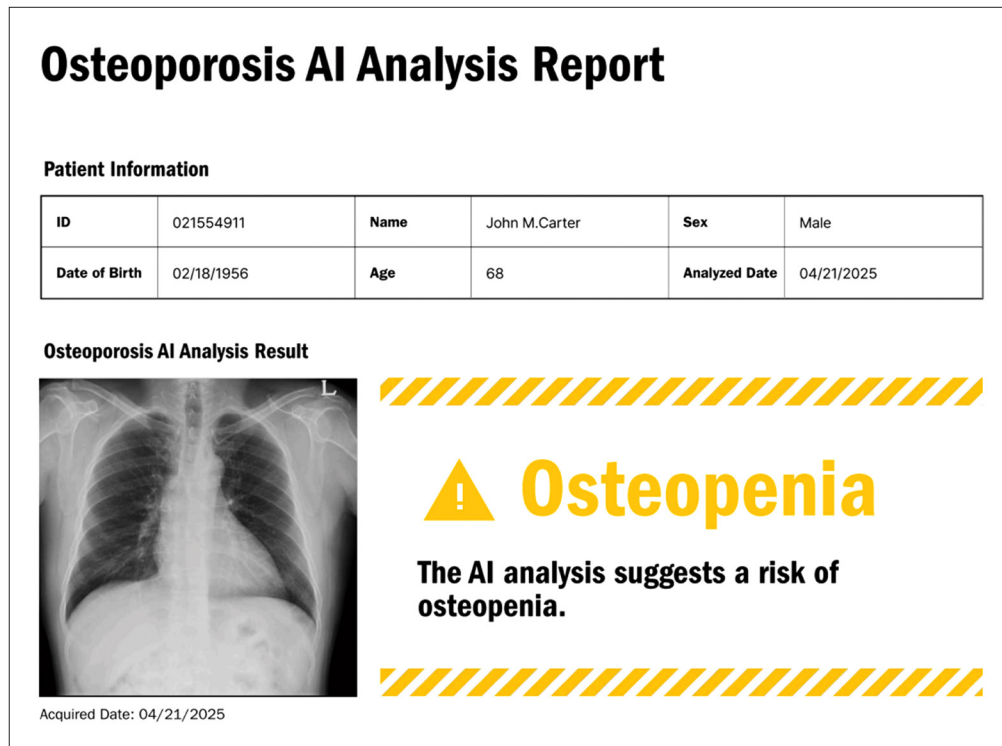
기회검진은 환자가 별도로 골다공증 검사를 받지 않더라도, 기존의 영상 데이터를 활용하여 진단할 수 있어서 의학적 효용성뿐만 아니라 비용 효율성 측면에서도 큰 장점을 가진다. 예를 들면 흉부 일반촬영 사진이나 흉복부 CT 같은 다빈도 영상검사에서 골다공증 발생 가능성이 높은 환자들을 예측하고 무증상의 혹은 간과되긴 쉬운 골다공증 골절을 검출한다면, 골다공증을 적시에 진단하고 치료받을 수 있도록 할 수 있을 것으로 기대된다.

흉부 일반촬영 사진을 이용한 국내 연구를 살펴보면(66) 연구진은 건강검진센터에서 7년간 수집된 약 13,000여 개의 흉부 일반촬영과 골밀도 데이터를 이용하여 모델을 개발하였으며, 골다공증 진단의 성능은 AUC 0.91 (내부검증), 0.88 (외부검증)을 보고하였다. 경사 가중치 클래스 활성화 맵(gradient-weighted class activation map; Grad-CAM)에 기반한 모델 해석 결과는 흉부 일반촬영 사진의 특정 영역(견갑골, 상완골두, 하부 흉추 및 요추, 뼈 주변 연조직, 근위부 늑골, 폐 등)에 초점을 맞추는 것으로 나타나 의학적 관련성이 다소 불분명하였고, 연구 데이터가 단일 기관에서 수집된 점 그리고 연구에 사용된 흉부 일반촬영 사진에 비정상 소견이 없는 경우에만 한정된 점 등이 연구의 제한점이다. 향후 다양한 기관의 데이터를 포함한 대규모 연구와 모델 해석에 대한 연구를 통한 임상적 유효성의 추가 입증이 기대된다.

최근 국내에서 흉부 일반촬영 사진을 활용한 AI 기술로 골다공증 여부를 확인할 수 있는 선별 및 진단보조 소프트웨어가 출시되어 본격적인 의료현장 도입이 가시화될 것으로 보인다. 국내 의료 AI 기업 프로메디우스의 골다공증 스크리닝 AI 소프트웨어 'PROS® CXR: OSTEO'가 2024년 11월 골다공증 분야에서 국내 첫 혁신의료기기로 지정되었고 2025년 3월 식품의약품안전처의 허가를 획득하였다(Fig. 9).

골다공증은 고관절 및 척추 골절의 위험을 증가시키는데, 상당수의 골다공증성 골절(fragility fracture)은 골다공증을 진단하는 데 사용되는 역치보다 높은 골밀도 값을 가진 환자에게서 발생한다(67-69). 게다가 골다공증성 척추 골절은 골격의 파손성(fragility)을 파악하는 데 매우 중요한 정보임에도 불구하고 임상적으로 인지되지 않거나 간과되는 경우가 흔하다. 세계 골다공증 학회에서는 골다공증 환자의 척추 압박골절 중 약 66%가 발견되지 않거나 보고되지 않는 것으로 추정

Fig. 9. Chest radiography-based osteoporosis diagnosis support AI software (PROS® CXR: OSTEO, PROMEDIUS). The domestic version has been approved as high and low risk at osteoporosis, while the overseas version is currently in the approval process as a 3-Class model (normal, osteopenia, osteoporosis).



하고 있다(70). 따라서 다른 임상적 이유로 시행된 흉부 및 복부/골반 CT에서 척추 압박골절을 검출하는 AI 모델은 골다공증의 조기 진단 및 개입을 촉진하여 골절 위험과 의료비용을 줄일 수 있는 잠재력을 가지고 있음을 시사하며, 이는 AI가 기존 의료 행위를 뛰어넘는 새로운 가치를 창출할 수 있다는 의미이기도 하다.

헬스케어 전문 컴퓨터 비전 스타트업 Zebra Medical Vision은 의료진이 흔히 간과하는 척추 압박골절 병변을 검출하는 것을 돕기 위해 HealthVCF 솔루션을 개발하여 2020년 미국 식품의약국(FDA)으로부터 정식 허가를 받은 바 있는데, 미국 의사협회(American Medical Association; AMA)가 척추 골절이 아닌 다른 질환을 의심하여 시행된 흉복부 CT에서 척추 압박골절을 찾아내는 솔루션에 대하여 Category III CPT® 코드를 부여하였다(의료영상 인공지능의 첫 current procedural terminology [이하 CPT] 코드). CPT 코드는 미국에서 각각의 진료 행위에 부여하는 코드로 이것이 중요한 이유는 CPT 코드를 바탕으로 보험 수가를 정하기 때문이다. Class I은 임상적인 의미가 충분하다고 인정되는 정식 코드이며 Class III는 임시 코드로 안전성과 유효성을 아직 인정받지 못한 신의료기술에 부여된다. 나스닥 상장사인 의료영상 기술 기업 나녹스(NNOX)의 자회사 Nanox.AI는 HealthVCF를 인수하여 낮은 골밀도를 계산하고 검출된 척추 압박골절의 심각도를 측정할 수 있는 HealthOST라는 이름의 업데이트된 버전을 출시하여 2022년 FDA 510(k) 승인을 받았으며 유럽 CE 인증도 획득하였다(Fig. 10).

HealthVCF 솔루션의 진단 성능을 평가한 연구에 의하면(71), 척추 압박골절 진단의 민감도는

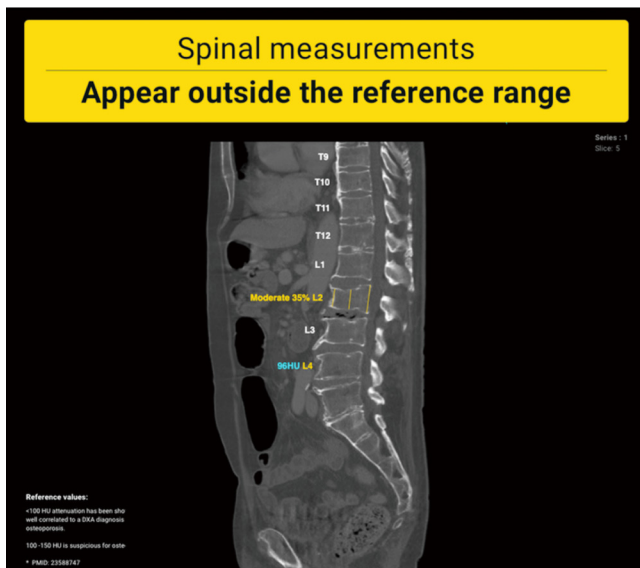


Fig. 10. AI bone solution (HealthOST, Nanox.AI) detects vertebral compression fractures and measures bone mineral density to identify patients at risk.

HealthOST is indicated for use in patients aged 50 and over undergoing CT scan for any clinical indication that includes at least four vertebrae in the T1-L4 portion of the spine (for vertebral height loss) and T11-L4 (for bone attenuation) portions of the spine.

0.66 (중등도/중증-Genant 2/3의 경우 0.78), 특이도는 0.90 (중등도/중증-Genant 2/3의 경우 0.87)이었다. 하지만, 알고리즘이 상당수의 척추 압박골절을 검출하지 못했고(낮은 민감도), 일부 이미지에 대한 평가가 불가능했던 점, 그리고 단일 기관에서 수집된 데이터를 사용하였고, 후향적 연구 설계라는 점이 제한점이다. 그리고 임상현장에서 중요한 판단 요소인 척추 압박골절의 급성과 만성 여부는 구분하지 못한다는 것이 아쉬운 점이다.

관절 질환(어깨와 무릎) MRI

어깨 질환의 흔한 원인인 회전근개 파열은 AI 보조 판독의 유망한 적용 분야 중 하나이다. 회전근개 파열 평가는 다양한 힘줄 침범의 조합, 힘줄 파열의 깊이와 크기, 다양한 파열 형태와 건인 패턴을 포함하는 복잡하고 광범위한 영상 진단을 요한다. 따라서 딥러닝 모델은 회전근개 파열의 MRI 진단에서 영상의학 전문의의 정확도와 일관성을 향상시킬 수 있는 잠재력을 가지고 있다.

회전근개 파열 MRI 진단에 3D CNN 모델을 적용한 연구에 의하면(72), AI의 파열 이분법 진단 정확도가 92.5%로 어깨 전공 전문의(76.4%)와 일반 정형외과 전문의(68.3%)보다 월등히 높았으며, 이외에도 파열의 위치와 크기 그리고 모양에 대한 3D 정보도 제공하여 판독 보조 및 임상진료의 관점에서 유용성이 기대가 되는 결과이다. 하지만 알고리즘이 극상근 건만을 대상으로 하고 있고 단일 기관의 단일 기종 MRI 검사 프로토콜에 기반한 학습 및 검증 데이터라는 점은 약 2,000명이 넘는 MRI 데이터라는 강점에도 불구하고 모델의 상용화를 위해서는 다른 기관의 다양한 기종 및 검사 프로토콜에서 어느 정도 성능을 보일지 즉 일반화에 대한 추가 검증이 필요하다.

2023년 Lin 등(73)은 회전근개 파열이 발생한 힘줄의 종류와 파열 등급 그리고 다양한 MRI 장비를 고려한 디자인의 연구를 발표하였다. 극상근, 극하근, 견갑하근 건 파열에 대한 전체 AUC는 각각 0.93, 0.89, 0.90이었고, 극상근, 극하근, 견갑하근 전층 파열에 대해 각각 0.98, 0.99, 0.95의 AUC로 부분층 파열보다 우수한 성능을 보였다. 그리고 다중 시퀀스 데이터 입력(4-view input)은

단일 시퀀스 입력보다 극하근과 견갑하근 건 파열에 대해 높은 AUC를 보인 반면, 관상사면 지방 억제 영상과 다중 시퀀스 데이터 입력의 알고리즘 비교 시 극상근 건 파열의 AUC는 차이가 없다고 보고하였다. 이 연구는 두 기관의 5개 MRI 장비(Vendor; Siemens, Philips, GE, Toshiba, Hitachi Systems, field strength; 0.3T-3T)로부터 11,925개 어깨 MRI 검사 데이터를 확보하여 수행된 점이 강점이다.

MRI에서 관절내장증(internal derangement of joints)의 병변을 탐지하고 특성화하기 위한 다수의 딥러닝 알고리즘이 무릎관절을 대상으로 개발되었으며, 이는 다른 관절에 비하여 무릎 MRI 검사의 수가 많고, 고품질 영상, 표준화된 검사 자세, 잘 정의된 병변의 종류와 특성 등의 요인과 연관이 있다. 2024년 11월 '의사 vs 인공지능... '관절 MRI' 진단 정확도, 어느 쪽이 높았을까?'라는 제목의 기사가 헬스조선에 게재되어 저자의 눈길을 끌었다. 중국 연구진이 12가지 무릎관절 질환 MRI 데이터셋을 이용하여 AI 모델을 개발하고 영상의학과 의료진을 대상으로 검증하였다는 것이 기사의 내용인데, AI 기반 무릎관절 MRI 모델이 몇몇 특정 병소의 진단을 넘어서(예를 들면 관절연골 병변의 검출) 다양한 해부학적 부위의 병변을 포괄적으로 분석하여 진단 결과를 제시하기 시작했다는 점이 흥미롭다.

2025년 발표된 무릎 MRI에 적용된 딥러닝 모델 54개의 메타분석에 의하면(74), 대상 질환의 빈도는 전방십자인대 손상(35%), 반월판 병소(24%), 일반적인 이상 소견(20%), 골관절염/연골병변(17%), 관절삼출(2%), 기타(2%) 순이었다. 한편 모델 학습에 이용된 데이터 출처는 54개 연구의 50% 이상이 로컬, 28% ($n = 15$)는 MRNet 혹은 OAI와 같은 오픈 데이터베이스였다. MRI 기종과 검사 프로토콜은 다양한데 3T 장비와 2D 영상 그리고 시상면 영상이 높은 비율을 차지하고 있었고 기종과 MRI 데이터의 형태를 구체적으로 밝히지 않은 연구도 적지 않았다. 이 연구에서 보고한 딥러닝 모델의 성능지표는 민감도 88.7%, 특이도 90.1%, AUC 92.1%, 분류 정확도 88.3%로 이는 다양한 무릎 병변에 대해 평균화한 값이다. 그리고 특정 병변을 대상으로 학습된 모델이 모든 지표에서 더 개선된 결과 값을 보였다. 즉 특정 병변에 초점을 맞춘 딥러닝 모델의 미세 조정이 일반적인 스크리닝 모델과 비교하여 모델 성능을 향상시키는데 유리하다는 점을 시사한다.

2024년 12월 현재 관절 질환 MRI 영상분석 인공지능 소프트웨어의 규제기관 허가 상황을 검색해 보면 유럽 CE 인증을 받은 소프트웨어가 2건, 미국 FDA 승인 제품은 확인되지 않았고, 국내 식품의약품안전처 허가 제품은 3건이 등록되어 있는데 해당 국내 기업의 홈페이지에서는 소개되고 있지 않아 상용화 단계에 진입한 것은 아닌 것으로 보인다.

프랑스 AI 스타트업 INCEPTO사의 무릎 MRI 진단 솔루션 'KEROS'는 유럽에서 Class IIa 제품 허가를 받은 상용화 모델로 전방십자인대, 내측측부인대, 반월판과 관절연골의 병변을 검출하고 특성화하며, 관절삼출, 오금낭종, 뼈부종도 탐지한다. 또한, 슬개골 불안정성의 주요 영상지표(In-sall-Salvati 비율, trochlea dysplasia 측정 값, TT-TG)도 제공한다. 그리고 자체 개발한 PACS 통합 시스템을 사용하여 AI 결과값을 구조화된(structured) 혹은 서술형(narrative) 판독지에 pre-filling할 수 있게 하여 임상현장의 사용성을 높인 점은 주목할 만하다. 본 솔루션 개발에 기반이 된 외부기관 검증 연구 결과에 의하면 반월판 파열의 AUC는 0.83, 전방십자인대 파열의 AUC는 0.94

이다(75, 76). 이는 지난 몇 년간 학술지에 발표된 결과들과 비교하면(77-80) 다소 낮은 수치인데, 하지만 잘 분리된 외부데이터셋에서 수행된 일반화(generalizability) 성능에 초점을 맞춘 연구인 점을 감안하여 판단을 할 필요가 있다.

척추 퇴행성 질환 MRI

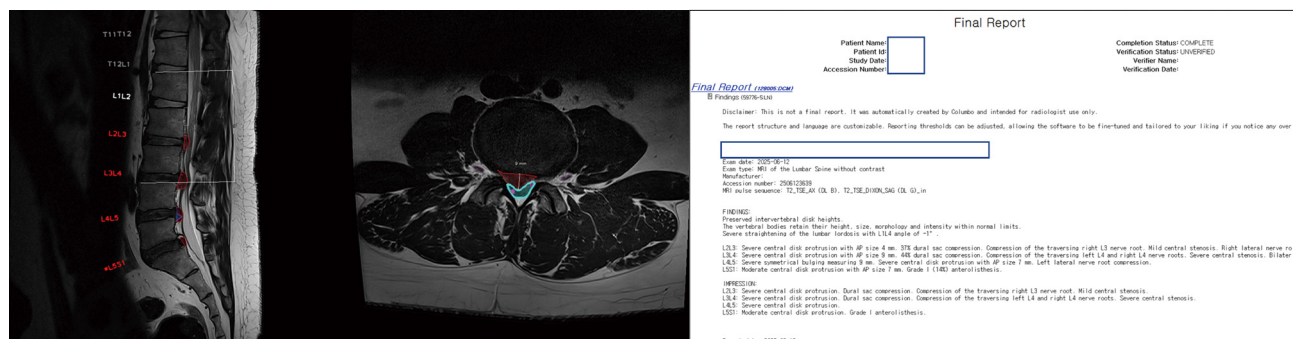
AI는 영상분석 자동화와 데이터 기반 의사결정을 통해 척추 영상 및 척추 질환 환자 치료에 변화를 일으키기 시작하였다(81). 척추 영상의 딥러닝 적용은 초기에는 척추 분할, 척추 라벨링, 추간판 병변 분류, 척추체 골절 검출 및 분류를 자동화하는 데 중점을 두었는데, 모델 개발이 꾸준히 임상진료 적용으로 확대되면서 척추 퇴행성 질환 대상의 상용화된 MRI 모델(RAI-RemedyLogic, New York, USA; CoLumbo-SmartSoft Helathcare, Varna, Bulgaria)이 출시되었다(Fig. 11).

MRI에서 주요 퇴행성 변화(Pfirrmann 등급, 추간판 높이, Modic 변화, 종판결손, 요추 중심관 협착증, 척추전방전위증)를 분석한 초기의 인공지능 연구는 영국 옥스퍼드 그룹이 개발한 SpineNet이 대표적이다(82). CNN 기반으로 유럽 4개국의 1,806명 환자 데이터를 학습에 이용하였고, 203명 환자를 대상으로 내부 검증이 수행되었다. 추간판 검출 및 라벨링에서 95.6%의 정확도를 달성하였고 중심관협착증과 척추전방전위증의 유무 판단은 95% 내외의 뛰어난 결과를 보였으나 다단계의 Pfirrmann 등급 판정의 정확도는 70%로 저조하였다. 추가 연구에서 SpineNet은 4점 등급 척도를 사용하여 축상면 영상에서 중심관 협착증 등급을 판정하도록 학습되었는데(83), 다수 등급(정상, 경증, 중등도, 중증) 분류에 대해 65.7%의 일치도를 보였고, 정상, 경증 또는 중등도 대 중증 협착증의 이분형 등급에 대해 94%를 달성했다.

AI 모델은 요추 척추협착증을 진단하고 중증도를 판정하는데 상당한 진전을 이루어, 임상현장의 실제 판독을 반영한 요추의 중심관, 측와동 및 신경공의 축상면/시상면 평가로 확장되었다(84-86). 2018년 Lu 등(84)은 중심관 및 신경공의 요추 척추협착증을 등급화하기 위한 딥러닝 알고리즘(Deep Spine)을 개발했으며, 이 알고리즘은 미리 정의된 등급 체계 없이 기존 판독지의 자연어 처리 기법을 이용하였다. 2021년 Hallinan 등(85)은 중심관, 측와동, 신경공 세 부위를 검출하고 각 부위의 협착증 등급을 판정하는 모델을 개발하여 내부와 외부 검증을 하였다. 네 명의 영상의

Fig. 11. Example of automatic spine MRI interpretation by the AI software (CoLumbo, SmartSoft Helathcare) on sagittal and axial T2-weighted images.

Disc hernias from L2-3 to L5-S1 were detected and measured with spinal stenosis grading at L2-3 and L3-4.



학과 의사가 미리 정의된 등급(정상, 경미, 중등도, 심각)에 기반하여 정교하게 학습데이터를 구축하였다. 내부 검증에서 이분법적 분류와 4등급 분류 결과, 모델(9개 모델의 평균)과 참조표준과의 κ 값은 중심관에서 0.96 vs. 0.82, 측와동에서 0.92 vs. 0.72, 신경공에서 0.89 vs. 0.75, 외부 검증에서는 중심관 0.95 vs. 0.66, 측와동 0.96 vs. 0.77, 신경공 0.96 vs. 0.83를 보고하여 모든 관심 영역(중심관, 측와동, 신경공)에 대해 딥러닝 모델이 이분법적 분류에서는 거의 완벽한 일치를 보였다. 이 연구에 활용된 딥러닝 알고리즘의 코드(Spine AI, <https://github.com/NUHS-NUS-SpineAI/SpineAI-Detect-Classify-LumbarMRI-Stenosis>)가 공개되어 있으며 저자들은 이 모델을 이용하여 후향적으로 판독업무의 효율성과 생산성에 관한 연구를 수행한 바 있다. Lim 등(86)에 의하면 AI 참조는 요추 척추협착증의 판독시간을 크게 단축시키고(근골격영상 전문의; 62% 감소, 일반영상 전문의; 69% 감소, 영상의학 전공의; 74% 감소) 판독자 간 일치도를 향상시킴으로써(AI 참조 시 중심관협착 4등급 분류의 경우 κ 값이 최대 0.39에서 0.71로 상승), 판독업무에 상당한 이점을 제공할 수 있음을 입증하였다.

앞서 기술한 2022년 FDA 승인을 받은 솔루션 CoLumbo는 요추의 다양한 퇴행성 병변의 진단 보조 목적으로 개발되었는데, 두 개의 독립된 연구에서 중심관 협착증 및 디스크 탈출증/팽윤, 신경근 압박, 전방전위증을 진단하는 데 있어 우수한 성능을 보고하였다(87, 88). 2024년 FDA 승인을 받은 솔루션 RAI는 미국 영상의학 전문의와 척추 전공 임상과의 주도로 개발되었으며 척추 라벨링, 분할, 주요 지표 측정, 퇴행성 변화 분류, 판독지 생성 등 CoLumbo와 대동소이한 기능과 서비스를 제공한다. 최근 발표된 본 솔루션 검증 연구에서 인상적인 결과를 입증하였는데(89), 중심관 협착증의 경우, 이진 분류에서 모델의 민감도, 특이도, AUC가 각각 0.97, 0.86, 0.96 (영상의학과 의사 평균 0.79, 0.89, 0.84), 측와동 협착증에서 모델의 민감도, 특이도, AUC가 각각 0.85, 0.79, 0.91 (영상의학과 의사 평균 0.71, 0.89, 0.81), 신경공 협착증에서는 모델의 민감도, 특이도, AUC가 각각 0.94, 0.84, 0.95 (영상의학과 의사 평균 0.88, 0.88, 0.88)로 AI가 영상의학과 의료진의 판독보조 역할을 훌륭히 수행할 수 있음을 시사한다.

국의 AI 스타트업이 실시한 설문조사에 따르면, 영상의학 의사들은 합병증이 없는 경우에 요추 MRI 판독에 평균 약 17.2분이 필요하다고 생각하며, 복잡한 경우에는 약 30분이 걸린다고 응답했다. 그러나 실제 임상 상황에서는 평균 13.6분에 MRI 판독을 수행하고 있으며, 이로 인하여 진단의 오류와 누락이 발생할 수 있음을 우려하고 있는데, 한국의 영상의학과 의사들이 처한 현실과 크게 다르지 않은 것으로 생각된다. 척추 퇴행성 질환의 판독은 반복적이고 시간이 많이 걸리는 과정이어서 AI를 활용한 업무 생산성 향상과 판독의 정확도 및 일관성 확보는 향후 그 가치가 높을 것으로 기대된다.

결론

최근 몇 년간 근골격계 영상 분야에서 딥러닝 기술이 발전하면서 다양한 인공지능 소프트웨어 의료기기가 개발되어 임상 현장에 적용되고 상용화되기 시작했다. 딥러닝 모델은 검사 프로토콜

설정, 영상 획득, 영상 해석 및 판독지 작성 등 영상 진료의 전반적인 단계에 활용될 수 있다. 본 종설에서는 근골격계 영상 분석 및 판독 AI 모델의 연구 개발 현황과 상용화 솔루션에 초점을 맞추어 골절 검출, 골연령 판독, 하지 및 척추 정렬 측정, 관절염 등급 분류, 골다공증 기회검진, 관절 및 척추 MRI 진단 분야를 중심으로 살펴보았다.

다양한 연구와 검증을 통해 근골격계 영상 딥러닝 모델의 높은 성능이 확인되었지만, 연구 방법론적인 한계, 특히 외부 검증 부족과 의료진 평가의 불확실성으로 인해 실제 임상 적용에는 신중한 접근이 필요하다. 특히 인공지능 성능이 환자군과 영상 특성에 따라 달라질 수 있다는 점을 고려하면 객관적인 일반화 성능 평가가 필수적인데 현재까지는 상용화 모델의 실사용 데이터 검증 및 인공지능 의료가 간 성능 비교에 대한 정보는 제한적인 상황이다. 또한, 상용화 과정에서 AI 모델의 임상 환경 내 통합, 도입 후 성능 및 유효성 모니터링, 법적 책임 문제, 의료진의 수용성 확보 등 해결해야 할 과제들이 남아있다.

결론적으로, 근골격계 영상 AI 모델의 상용화는 임상 현장의 미충족 수요를 반영한 영역에서 이미 시작되었다. 여러 도전 과제가 존재하지만 기술 발전과 더불어 근골격계 영상 진단 분야에서 AI 모델 활용은 더욱 확대될 것으로 예상된다. 판독 보조 도구를 넘어 인공지능의 가치를 극대화할 수 있는 근골격계 적응증 및 활용 방안에 대한 지속적인 논의와 연구가 필요하며, 인공지능 소양을 갖춘 근골격계 영상의학 전문의의 적극적인 역할이 중요하다.

Conflicts of Interest

Chang Ho Kang is a chief medical officer of Crecsom Inc., a startup company, the eventual products and services of which will be related to the subject matter of the article.

ORCID iD

Chang Ho Kang  <https://orcid.org/0000-0003-2385-7245>

Funding

None

REFERENCES

1. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mammonov AB, Bilbily A, Cicero M, et al. The RSNA pediatric bone age machine learning challenge. *Radiology* 2019;290:498-503
2. Lee GR, Flanders AE, Richards T, Kitamura F, Colak E, Lin HM, et al. Performance of the winning algorithms of the RSNA 2022 cervical spine fracture detection challenge. *Radiol Artif Intell* 2024;6:e230256
3. Chen MM, Golding LP, Nicola GN. Who will pay for AI? *Radiol Artif Intell* 2021;3:e210030
4. Health Insurance Review and Assessment Service. Release of guideline for health insurance registration of digital treatment devices and artificial intelligence (AI). Available at: <https://www.hira.or.kr/bbsDummy.do?pgmid=HIRAA020041000100&brdScnBltno=4&brdBltno=10957&pageIndex=1&pageIndex2=1#none>. Published 2023. Accessed April 12, 2024
5. Anderson PG, Baum GL, Keathley N, Sicular S, Venkatesh S, Sharma A, et al. Deep learning assistance closes the accuracy gap in fracture detection across clinician types. *Clin Orthop Relat Res* 2023;481:580-588
6. Ruitenbeek HC, Oei EH, Visser JJ, Kijowski R. Artificial intelligence in musculoskeletal imaging: realistic clinical applications in the next decade. *Skeletal Radiol* 2024;53:1849-1868
7. Guermazi A, Tannoury C, Koppel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology* 2022;302:627-636

8. Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, et al. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic Study. *Radiology* 2021;300:120-129
9. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard NE, et al. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. *Skeletal Radiol* 2022;51:2129-2139
10. Nguyen T, Maarek R, Hermann AL, Kammoun A, Marchi A, Khelifi-Touhami MR, et al. Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. *Pediatr Radiol* 2022;52:2215-2226
11. Zhang X, Yang Y, Shen YW, Zhang KR, Jiang ZK, Ma LT, et al. Diagnostic accuracy and potential covariates of artificial intelligence for diagnosing orthopedic fractures: a systematic literature review and meta-analysis. *Eur Radiol* 2022;32:7196-7216
12. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. *Radiology* 2022;304:50-62
13. Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiol Artif Intell* 2020;2:e190023
14. Lee KC, Choi IC, Kang CH, Ahn KS, Yoon H, Lee JJ, et al. Clinical validation of an artificial intelligence model for detecting distal radius, ulnar styloid, and scaphoid fractures on conventional wrist radiographs. *Diagnostics (Basel)* 2023;13:1657
15. Pauling C, Kanber B, Arthurs OJ, Shelmerdine SC. Commercially available artificial intelligence tools for fracture detection: the evidence. *BJR Open* 2023;6:tzad005
16. Husarek J, Hess S, Razaiean S, Ruder TD, Sehmisch S, Müller M, et al. Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy. *Sci Rep* 2024;14:23053
17. Regnard NE, Lanseur B, Ventre J, Ducarouge A, Clovis L, Lassalle L, et al. Assessment of performances of a deep learning algorithm for the detection of limbs and pelvic fractures, dislocations, focal bone lesions, and elbow effusions on trauma X-rays. *Eur J Radiol* 2022;154:110447
18. Tajmir SH, Lee H, Shailam R, Gale HI, Nguyen JC, Westra SJ, et al. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. *Skeletal Radiol* 2019;48:275-283
19. Michael DJ, Nelson AC. HANDX: a model-based system for automatic segmentation of bones from digital hand radiographs. *IEEE Trans Med Imaging* 1989;8:64-69
20. Pietka E, McNitt-Gray MF, Kuo ML, Huang HK. Computer-assisted phalangeal analysis in skeletal age assessment. *IEEE Trans Med Imaging* 1991;10:616-620
21. Tanner JM, Oshman D, Lindgren G, Grunbaum JA, Elsouki R, Labarthe D. Reliability and validity of computer-assisted estimates of Tanner-Whitehouse skeletal maturity (CASAS): comparison with the manual method. *Horm Res* 1994;42:288-294
22. Yoo JW, Lee JM, Kim WY. [A bone age assessment method based on normalized shape model]. *J Korea Multimed Soc* 2009;12:383-396. Korean
23. Son SJ, Song Y, Kim N, Do Y, Kwak N, Lee MS, et al. TW3-based fully automated bone age assessment system using deep neural networks. *IEEE Access* 2019;7:33346-33358
24. Lee H, Tajmir S, Lee J, Zissen M, Yesiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. *J Digit Imaging* 2017;30:427-441
25. Kim JR, Shim WH, Yoon HM, Hong SH, Lee JS, Cho YA, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *AJR Am J Roentgenol* 2017;209:1374-1380
26. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: a systematic literature review and meta-analysis. *PLoS One* 2019;14:e0220242
27. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. *IEEE Trans Med Imaging* 2009;28:52-66
28. Booz C, Yel I, Wichmann JL, Boettger S, Al Kamali A, Albrecht MH, et al. Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. *Eur Radiol Exp* 2020;4:6
29. Pape J, Rosolowski M, Pfäffle R, Beeskow AB, Gräfe D. A critical comparative study of the performance of

- three AI-assisted programs for bone age determination. *Eur Radiol* 2025;35:1190-1196
30. van Leeuwen KG, Schalekamp S, Rutten MJCM, Huisman M, Schaefer-Prokop CM, de Rooij M, et al. Comparison of commercial AI software performance for radiograph lung nodule detection and bone age prediction. *Radiology* 2024;310:e230981
 31. Lee KC, Lee KH, Kang CH, Ahn KS, Chung LY, Lee JJ, et al. Clinical validation of a deep learning-based hybrid (Greulich-Pyle and modified Tanner-Whitehouse) method for bone age assessment. *Korean J Radiol* 2021; 22:2017-2025
 32. Gurney B. Leg length discrepancy. *Gait Posture* 2002;15:195-206
 33. Raczkowski JW, Daniszewska B, Zolynski K. Functional scoliosis caused by leg length discrepancy. *Arch Med Sci* 2010;6:393-398
 34. Gofton JP, Trueman GE. Studies in osteoarthritis of the hip. II. Osteoarthritis of the hip and leg-length disparity. *Can Med Assoc J* 1971;104:791-799
 35. Zheng Q, Shellikeri S, Huang H, Hwang M, Sze RW. Deep learning measurement of leg length discrepancy in children based on radiographs. *Radiology* 2020;296:152-158
 36. Schock J, Truhn D, Abrar DB, Merhof D, Conrad S, Post M, et al. Automated analysis of alignment in long-leg radiographs by using a fully automated support system based on artificial intelligence. *Radiol Artif Intell* 2020;3:e200198
 37. Tack A, Preim B, Zachow S. Fully automated assessment of knee alignment from full-leg X-rays employing a "YOLOv4 and resnet landmark regression algorithm" (YARLA): data from the osteoarthritis initiative. *Comput Methods Programs Biomed* 2021;205:106080
 38. Lee CS, Lee MS, Byon SS, Kim SH, Lee BI, Lee BD. Computer-aided automatic measurement of leg length on full leg radiographs. *Skeletal Radiol* 2022;51:1007-1016
 39. Lee HS, Hwang S, Kim SH, Joon NB, Kim H, Hong YS, et al. Automated analysis of knee joint alignment using detailed angular values in long leg radiographs based on deep learning. *Sci Rep* 2024;14:7226
 40. Moon KR, Lee BD, Lee MS. A deep learning approach for fully automated measurements of lower extremity alignment in radiographic images. *Sci Rep* 2023;13:14692
 41. Archer H, Reine S, Xia S, Vazquez LC, Ashikyan O, Pezeshk P, et al. Reliability assessment of leg length and angular alignment on manual reads versus artificial intelligence-generated lower extremity radiographic measurements. *Clin Imaging* 2024;113:110233
 42. Kim H, Kim HS, Moon ES, Yoon CS, Chung TS, Song HT, et al. Scoliosis imaging: what radiologists should know. *Radiographics* 2010;30:1823-1842
 43. Gstoettner M, Sekyra K, Walochnik N, Winter P, Wachter R, Bach CM. Inter- and intraobserver reliability assessment of the Cobb angle: manual versus digital measurement tools. *Eur Spine J* 2007;16:1587-1592
 44. Kato S, Maeda Y, Nagura T, Nakamura M, Watanabe K. Comparison of three artificial intelligence algorithms for automatic Cobb angle measurement using teaching data specific to three disease groups. *Sci Rep* 2024; 14:17989
 45. Hayashi D, Regnard NE, Ventre J, Marty V, Clovis L, Lim L, et al. Deep learning algorithm enables automated Cobb angle measurements with high accuracy. *Skeletal Radiol* 2025;54:1469-1478
 46. Chen K, Stotter C, Klestil T, Mitterer JA, Lepenik C, Nehrer S. Fully automated measurement of Cobb angles in coronal plane spine radiographs. *J Clin Med* 2024;13:4122
 47. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci Rep* 2018;8:1727
 48. Liu B, Luo J, Huang H. Toward automatic quantification of knee osteoarthritis severity using improved faster R-CNN. *Int J Comput Assist Radiol Surg* 2020;15:457-466
 49. Norman B, Padoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J Digit Imaging* 2019;32:471-477
 50. Nguyen HH, Saarakkala S, Tiulpin A. Automatic knee osteoarthritis severity assessment from plain radiographs with limited data. *Osteoarthritis Cartilage* 2020;28(Supplement 1):S314-S315
 51. Tiulpin A, Saarakkala S. Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. *Diagnostics (Basel)* 2020;10:932
 52. Kim DH, Lee KJ, Choi D, Lee JI, Choi HG, Lee YS. Can additional patient information improve the diagnostic performance of deep learning for the interpretation of knee osteoarthritis severity. *J Clin Med* 2020;9:3341

53. Thomas KA, Kidziński Ł, Halilaj E, Fleming SL, Venkataraman GR, Oei EHG, et al. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiol Artif Intell* 2020;2:e190065
54. Brejnebo MW, Hansen P, Nybing JU, Bachmann R, Ratjen U, Hansen IV, et al. External validation of an artificial intelligence tool for radiographic knee osteoarthritis severity classification. *Eur J Radiol* 2022;150:110249
55. Zhao H, Ou L, Zhang Z, Zhang L, Liu K, Kuang J. The value of deep learning-based X-ray techniques in detecting and classifying K-L grades of knee osteoarthritis: a systematic review and meta-analysis. *Eur Radiol* 2025;35:327-340
56. Wright RW; MARS Group. Osteoarthritis classification scales: interobserver reliability and arthroscopic correlation. *J Bone Joint Surg Am* 2014;96:1145-1151
57. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clin Orthop Relat Res* 2016;474:1886-1893
58. Damen J, Schiphof D, Wolde ST, Cats HA, Bierma-Zeinstra SM, Oei EH. Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (cohort hip and cohort knee) study. *Osteoarthritis Cartilage* 2014;22:969-974
59. Günther KP, Sun Y. Reliability of radiographic assessment in hip and knee osteoarthritis. *Osteoarthritis Cartilage* 1999;7:239-246
60. Neubauer M, Moser L, Neugebauer J, Raudner M, Wondrasch B, Führer M, et al. Artificial-intelligence-aided radiographic diagnostic of knee osteoarthritis leads to a higher association of clinical findings with diagnostic ratings. *J Clin Med* 2023;12:744
61. Nehrer S, Ljuhar R, Steindl P, Simon R, Maurer D, Ljuhar D, et al. Automated knee osteoarthritis assessment increases physicians' agreement rate and accuracy: data from the osteoarthritis initiative. *Cartilage* 2021; 13(1_suppl):957S-965S
62. Smolle MA, Goetz C, Maurer D, Vielgut I, Novak M, Zier G, et al. Artificial intelligence-based computer-aided system for knee osteoarthritis assessment increases experienced orthopaedic surgeons' agreement rate and accuracy. *Knee Surg Sports Traumatol Arthrosc* 2023;31:1053-1062
63. Bressemer KK, Vahldiek JL, Adams L, Niehues SM, Haibel H, Rodriguez VR, et al. Deep learning for detection of radiographic sacroiliitis: achieving expert-level performance. *Arthritis Res Ther* 2021;23:106
64. Rudwaleit M, Haibel H, Baraliakos X, Listing J, Märker-Hermann E, Zeidler H, et al. The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. *Arthritis Rheum* 2009; 60:717-727
65. Lee KH, Lee RW, Lee KH, Park W, Kwon SR, Lim MJ. The development and validation of an AI diagnostic model for sacroiliitis: a deep-learning approach. *Diagnostics (Basel)* 2023;13:3643
66. Jang M, Kim M, Bae SJ, Lee SH, Koh JM, Kim N. Opportunistic osteoporosis screening using chest radiographs with deep learning: development and external validation with a cohort dataset. *J Bone Miner Res* 2022; 37:369-377
67. Compston JE, McClung MR, Leslie WD. Osteoporosis. *Lancet* 2019;393:364-376
68. Schuit SC, van der Klift M, Weel AE, de Laet CE, Burger H, Seeman E, et al. Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam study. *Bone* 2004;34:195-202
69. Siris ES, Chen YT, Abbott TA, Barrett-Connor E, Miller PD, Wehren LE, et al. Bone mineral density thresholds for pharmacological intervention to prevent fractures. *Arch Intern Med* 2004;164:1108-1112
70. Kutsal FY, Ergin Ergani GO. Vertebral compression fractures: still an unpredictable aspect of osteoporosis. *Turk J Med Sci* 2021;51:393-399
71. Page JH, Moser FG, Maya MM, Prasad R, Pressman BD. Opportunistic CT screening-machine learning algorithm identifies majority of vertebral compression fractures: a cohort study. *JBMR Plus* 2023;7:e10778
72. Shim E, Kim JY, Yoon JP, Ki SY, Lho T, Kim Y, et al. Automated rotator cuff tear classification using 3D convolutional neural network. *Sci Rep* 2020;10:15632
73. Lin DJ, Schwier M, Geiger B, Raithel E, von Busch H, Fritz J, et al. Deep learning diagnosis and classification of rotator cuff tears on shoulder MRI. *Invest Radiol* 2023;58:405-412
74. Mead K, Cross T, Roger G, Sabharwal R, Singh S, Giannotti N. MRI deep learning models for assisted diagnosis of knee pathologies: a systematic review. *Eur Radiol* 2025;35:2457-2469
75. Tran A, Lassalle L, Zille P, Guillin R, Pluot E, Adam C, et al. Deep learning to detect anterior cruciate ligament tear on knee MRI: multi-continental external validation. *Eur Radiol* 2022;32:8394-8403

76. Rizk B, Brat H, Zille P, Guillin R, Pouchy C, Adam C, et al. Meniscal lesion detection and characterization in adult knee MRI: a deep learning model approach with external validation. *Phys Med* 2021;83:64-71
77. Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with mask-RCNN. *Diagn Interv Imaging* 2019;100:235-242
78. Roblot V, Giret Y, Bou Antoun M, Morillot C, Chassin X, Cotten A, et al. Artificial intelligence to diagnose meniscus tears on MRI. *Diagn Interv Imaging* 2019;100:243-249
79. Chang PD, Wong TT, Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. *J Digit Imaging* 2019;32:980-986
80. Santomartino SM, Kung J, Yi PH. Systematic review of artificial intelligence development and evaluation for MRI diagnosis of knee ligament or meniscus tears. *Skeletal Radiol* 2024;53:445-454
81. Lee S, Jung JY, Mahatthanatrakul A, Kim JS. Artificial intelligence in spinal imaging and patient care: a review of recent advances. *Neurospine* 2024;21:474-486
82. Jamaludin A, Lootus M, Kadir T, Zisserman A, Urban J, Battié MC, et al. ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J* 2017;26:1374-1383
83. Ishimoto Y, Jamaludin A, Cooper C, Walker-Bone K, Yamada H, Hashizume H, et al. Could automated machine-learned MRI grading aid epidemiological studies of lumbar spinal stenosis? Validation within the Wakayama spine study. *BMC Musculoskelet Disord* 2020;21:158
84. Lu JT, Pedemonte S, Bizzo B, Doyle S, Andriole KP, Michalski MH, et al. Deep spine: automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. Available at: <https://proceedings.mlr.press/v85/lu18a.html>. Published 2018. Accessed April 21, 2025
85. Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, Thian YL, et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. *Radiology* 2021;300:130-138
86. Lim DSW, Makmur A, Zhu L, Zhang W, Cheng AJL, Sia DSY, et al. Improved productivity using deep learning-assisted reporting for lumbar spine MRI. *Radiology* 2022;305:160-166
87. Georgiev R, Novakova M, Bliznakova K. Clinical assessment of CoLumbo deep learning system for central canal stenosis diagnostics. *Eurasian J Med Oncol* 2023;7:42-48
88. Lehnen NC, Haase R, Faber J, Rüber T, Vatter H, Radbruch A, et al. Detection of degenerative changes on MR images of the lumbar spine with a convolutional neural network: a feasibility study. *Diagnostics (Basel)* 2021;11:902
89. Tumko V, Kim J, Uspenskaia N, Honig S, Abel F, Lebl DR, et al. A neural network model for detection and classification of lumbar spinal stenosis on MRI. *Eur Spine J* 2024;33:941-948

근골격계 영상 AI 모델의 현황과 상용화

강창호*

딥러닝 기반의 인공지능 소프트웨어 의료기기(software as a medical device; 이하 SaMD)는 영상의학에 혁신을 일으키며 신경, 흉부, 유방 분야에서 영상 진료의 모습을 바꾸고 있다. 근골격계 영역에서도 상용화 모델이 국내 임상 현장에 진입하기 시작하였는데 다른 영상의학 분야에 비해 아직 초기 단계에 머물러 있다. 골절 검출, 골연령 추정, 골격계 측정, 관절염 등급 분류, 골다공증 스크리닝, 척추 및 대관절 질환 MRI 판독 보조에서 인공지능 모델의 우수한 진단 성능이 입증되었으며, 이들 중 상당수는 규제기관의 제품 허가를 거쳐 상용화되기 시작하였다. 본 종설에서는 근골격계 영상 인공지능 소프트웨어 의료기기(SaMD)의 개발과 국내외 상용화 현황을 알아보고, 임상 적용을 위한 유의한 고려 사항에 대한 정보를 제공하여 근골격계 영상 AI 모델의 현장 적용에 관심 있는 독자들이 합리적인 의사결정을 하는 데 도움을 주고자 한다.

고려대학교 의과대학 고려대학교 안암병원 영상의학과