EISSN 1305-3612

DIR

Diagnostic and Interventional Radiology

TSR 1924
TURKISH SOCIETY
OF RADIOLOGY

dirjournal.org

VOLUME **31**
ISSUE **5**
September 2025

Diagnostic and Interventional Radiology

Official Journal of the Turkish Society of Radiology

E-ISSN: 1305-3612

www.dirjournal.org

Editor in Chief

Mehmet Ruhi Onur, MD
Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye
ORCID ID: 0000-0003-1732-7862

Section Editors and Scientific Editorial Board

Abdominal Imaging

İlkay S. İdilman, MD
Department of Radiology, Hacettepe University Faculty of Medicine, Ankara, Türkiye
ORCID ID: 0000-0002-1913-2404

Sonay Aydın, MD
Department of Radiology, Erzincan Binali Yıldırım University Faculty of Medicine, Erzincan, Türkiye
ORCID ID: 0000-0002-3812-6333

Artificial Intelligence and Informatics

Burak Koçak, MD
Department of Radiology, University of Health Sciences, Başakşehir Çam and Sakura City Hospital, İstanbul, Türkiye
ORCID ID: 0000-0002-7307-396X

Tuğba Akıncı D'Antonoli, MD
Institute of Radiology and Nuclear Medicine, Cantonal Hospital Baselland, Liestal, Switzerland
ORCID ID: 0000-0002-7237-711X

Breast Imaging

Serap Gültekin, MD
Department of Radiology, Gazi University Faculty of Medicine, Ankara, Türkiye
ORCID ID: 0000-0001-6349-3998

Chest and Cardiovascular Imaging

Furkan Ufuk, MD
Department of Radiology, The University of Chicago, Chicago, USA
ORCID ID: 0000-0002-8614-5387

Hybrid Imaging and Nuclear Medicine

Evrim Bengi Türkbey, MD
Radiology and Imaging Sciences, Clinical Center, National Institutes of Health Bethesda, Maryland, United States
ORCID ID: 0000-0002-5216-3528

Interventional Radiology

Barbaros Çil, MD, FCIRSE
Department of Radiology, Koç University School of Medicine, İstanbul, Türkiye
ORCID ID: 0000-0003-1079-0088

Bahri Üstünsöz, MD
Department of Radiology, LSUHSC (Louisiana State University Health Science Center) School of Medicine, New Orleans, United States
ORCID ID: 0000-0003-4308-6708

James Milburn, MD
Department of Radiology, Ochsner Medical System, New Orleans, Louisiana, USA
ORCID ID: 0000-0003-3403-2628

Musculoskeletal Imaging

Zeynep Maraş Özdemir, MD
Department of Radiology, İnönü University Faculty of Medicine, Malatya, Türkiye
ORCID ID: 0000-0003-1085-8978

Neuroradiology

Gülgün Yılmaz Ovalı, MD
Department of Radiology, Celal Bayar University Faculty of Medicine, Manisa, Türkiye
ORCID ID: 0000-0001-8433-5622

Erkan Gökçe, MD
Department of Radiology, Tokat Gaziosmanpaşa University Faculty of Medicine, Tokat, Türkiye
ORCID ID: 0000-0003-3947-2972

Pediatric Radiology

Meltem Ceyhan Bilgici, MD
Department of Radiology, 19 Mayıs University Faculty of Medicine, Samsun, Türkiye
ORCID ID: 0000-0002-0133-0234

Evrim Özmen, MD
Department of Radiology, Koç University Hospital, İstanbul, Türkiye
ORCID ID: 0000-0003-3100-4197

Publication Coordinator

Nermin Tunçbilek, MD
Department of Radiology, Trakya University Faculty of Medicine, Edirne, Türkiye
ORCID ID: 0000-0002-8734-1849

Biostatistical Consultant

İlker Ercan, PhD
Department of Biostatistics, Uludağ University School of Medicine, Bursa, Türkiye
ORCID ID: 0000-0002-2382-290X

Publication Services

Galenos Publishing, İstanbul, Türkiye

Past Editors

Editors in Chief

Mustafa Seçil, MD (2016-2023)
Nevzat Karabulut, MD (2011-2016)
Üstün Aydıngöz, MD (2010-2011)
Okan Akhan, MD (2001-2010)
Ferhun Balkancı, MD (1999-2001)
Aytekin Besim, MD (1994-1999)*

** Dr. Aytekin Besim actually served as the General Coordinator. His work in this capacity, however, was in effect that of an Editor in Chief.*

Editors

Ayşenur Cila, MD (2001-2002)
Suat Kemal Aytaç, MD (1997-2001)
Erhan Ilgıt, MD (1994-2001)
Okan Akhan, MD (1994-2001)
Ferhun Balkancı, MD (1994-2000)
Serdar Akyar, MD (1994-1997)

Section Editors

Section Editorship was established in 2002 at the tenure of Dr Okan Akhan, Editor in Chief.

Abdominal Imaging

Bengi Gürses, MD (2020-2023)
Mehmet Ruhi Onur, MD (2016-2023)
Barış Türkbey, MD (2014-2020)
Mustafa N. Özmen, MD (2012-2018)
Murat Acar, MD (2015-2016)
Mustafa Seçil, MD (2011-2016)
Ahmet Tuncay Turgut, MD (2011)
Deniz Akata, MD (2007-2011)
Ayşe Erden, MD (2002-2011)
Okan Akhan, MD (2002-2010)
Hakan Özdemir, MD (2002-2010)

Diagnostic and Interventional
Radiology

Official Journal of the
Turkish Society of Radiology

E-ISSN: 1305-3612
www.dirjournal.org

Artificial Intelligence and Informatics
Barış Türkbey, MD (2020-2023)

Breast Imaging
Mustafa Erkin Arıbal, MD (2016-2023)
Sibel Kul (2015-2018)
Ayşenur Oktay, MD (2009-2014)
Ayşegül Özdemir, MD (2004-2009)

Cardiovascular Imaging
Uğur Bozlar, MD (2016-2023)
Muşturay Karçaaltıncaba, MD (2007-2010)
Mecit Kantarcı (2010-2016)

Chest Imaging
Nevzat Karabulut, MD (2010-2014)
Çetin Atasoy, MD (2007-2010)
Macit Arıyürek, MD (2002-2007)
Figen Demirkazık, MD, (2014-2018)

General Radiology
Ersin Öztürk, MD (2014-2017)
Utku Şenol, MD (2010-2013)
Oğuz Dicle, MD (2007-2010)

Interventional Radiology
Cüneyt Aytekin, MD (2016-2023)
Bora Peynircioğlu, MD (2012-2015)
Levent Oğuzkurt, MD (2011-2014)
Fatih Boyvat, MD (2007-2010)
İsmail Oran, MD (2015-2019)

Musculoskeletal Imaging
Hatice Tuba Sanal, MD (2016-2023)
Fatih Kantarcı, MD (2014-2016)
Ayşenur Oktay, MD (2011-2013)
Üstün Aydıngöz, MD (2002-2011)
Berna Dirim Mete (2016-2017)

Neuroradiology and Head & Neck Imaging
Kubilay Aydın, MD (2016-2023)
Nafi Aygün, MD (2016-2023)
Kader Karlı Oğuz, MD (2011-2015)
Süleyman Men, MD (2007-2013)
Muhteşem Ağıldere, MD (2002-2011)

Nuclear Medicine
A. Cahid Civelek, MD (2016-2023)
Oktay Sarı, MD (2015)
Akın Yıldız, MD (2011-2014)

Pediatric Radiology
Korgün Koral, MD (2016-2023)
Murat Kocaoğlu, MD (2016-2023)
Ensar Yekeler, MD (2014-2016)
Suat Fitöz, MD (2007-2013)

Diagnostic and Interventional Radiology (Diagn Interv Radiol) is a bimonthly periodical of the Turkish Society of Radiology and the content of the journal is available at https://www.dirjournal.org/. It is peer-reviewed and adheres to the highest ethical and editorial standards. The editors of the journal endorse the Editorial Policy Statements Approved by the Council of Science Editors Board of Directors (https://cse.memberclicks.net/). The journal is in compliance with the Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals published by the International Committee of Medical Journal Editors (updated May 2022, www.icmje.org).

First ten volumes of Diagnostic and Interventional Radiology have been published in Turkish under the name of Tanısal ve Girişimsel Radyoloji (Index Medicus® abbreviation: Tani Girisim Radyol), the current title's exact Turkish translation.

Diagnostic and Interventional Radiology is an open access publication, and the journal's publication model is based on Budapest Open Access Initiative (BOAI) declaration. All published content is available online, free of charge at https://www.dirjournal.org/. Authors retain the copyright of their published work in Diagnostic and Interventional Radiology. The journal's content is licensed under a Creative Commons Attribution-NonCommercial (CC BY-NC) 4.0 International License which permits third parties to share and adapt the content for non-commercial purposes by giving the appropriate credit to the original work.

Please refer to the journal's webpage **(https://dirjournal.org/)** for "Aims and Scope", "Instructions to Authors" and "Instructions to Reviewers".

The editorial and publication processes of the journal are shaped in accordance with the guidelines of the ICMJE, WAME, CSE, COPE, EASE, and NISO.
Diagnostic and Interventional Radiology is indexed in **SCI-Expanded, Pubmed/Medline, Pubmed Central, TUBITAK ULAKBIM TR Index, DOAJ, HINARI, EMBASE, CINAHL, Scopus, Gale and CNKI.**

The journal is published online.

Owner: Can Çevikol on behalf of Turkish Society of Radiology

Responsible Manager: Mehmet Ruhi Onur

Contact Information
Diagnostic and Interventional Radiology Turkish Society of Radiology
Hoşdere Cad., Güzelkent Sok., Çankaya Evleri, F/2, 06540
Ankara, Türkiye
E-mail: info@dirjournal.org
Phone: +90 (312) 442 36 53 **Fax:** +90 (312) 442 36 54

Publisher Contact
Address: Molla Gürani Mah. Kaçamak Sk.
No: 21/1 34093 İstanbul, Türkiye
Phone: +90 (530) 177 30 97
E-mail: info@galenos.com.tr/yayin@galenos.com.tr
Web: www.galenos.com.tr **Publisher Certificate Number:** 14521

Online Publication Date: September 2025

EISSN 1305-3612

International scientific journal published bimonthly.

Turkish Society of Radiology is one of the foremost medical specialty organizations in Türkiye. It was formed by the merger of the two main radiology societies of Türkiye, one of which was founded in 1924. The Society is based in Ankara, Türkiye.

TSR
1924
TURKISH SOCIETY OF RADIOLOGY

DIR
Diagnostic and Interventional Radiology

Volume 31 • Issue 5

September 2025

www.dirjournal.org

Contents

ABDOMINAL IMAGING

405 Original Article **Evaluating artificial intelligence for a focal nodular hyperplasia diagnosis using magnetic resonance imaging: preliminary findings** *Mecit Kantarcı, Volkan Kızılgöz, Ramazan Terzi, Ahmet Enes Kılıç, Halime Kabalcı, Önder Durmaz, Nil Tokgöz, Mustafa Harman, Ayşegül Sağır Kahraman, Ali Avanaz, Sonay Aydın, Gülsüm Özlem Elpek, Merve Yazol, Bülent Aydınlı*

416 Original Article **Diffusion kurtosis versus diffusion-weighted magnetic resonance imaging in differentiating clear cell renal cell carcinoma and renal angiomyolipoma with minimal fat: a comparative study** *Yarong Lin, Wenrong Zhu, Qingqiang Zhu*

423 Original Article **Diagnostic performance of magnetic resonance imaging for lateral pelvic lymph node metastasis in patients with rectal carcinoma: a meta-analysis and systematic review** *Xiaolong Liu, Keping Liao, Peng Wang, Yongqiang Gao, Yongxin Du*

ARTIFICIAL INTELLIGENCE AND INFORMATICS

430 Original Article **Deep learning for named entity recognition in Turkish radiology reports** *Abubakar Ahmad Abdullahi, Murat Can Ganiz, Ural Koç, Muhammet Batuhan Gökhan, Ceren Aydın, Ali Bahadır Özdemir*

440 Original Article **Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis** *Burak Koçak, Fadime Köse, Ali Keleş, Abdurrezzak Şendur, İsmail Meşe, Mehmet Karagülle*

456 Original Article **Automatic bone age assessment: a Turkish population study** *Samet Öztürk, Murat Yüce, Gül Gizem Pamuk, Candan Varlık, Ahmet Tan Cimilli, Musa Atay*

465 Original Article **Gastrointestinal bleeding detection on digital subtraction angiography using convolutional neural networks with and without temporal information** *Derek Smetanick, Sailendra Naidu, Alex Wallace, M-Grace Knuttinen, Indravadan Patel, Sadeer Alzubaidi*

BREAST IMAGING

474 Original Article **The role of the Kaiser score system in uncertain malignant potential (B3) breast lesions: a pilot study** *Fatma Çelik Yabul, Hafize Otçu Temur, Bahar Atasoy, Serdar Balsak, Alpay Alkan, Şeyma Yıldız*

HEAD AND NECK IMAGING

480 Letter to the Editor **Letter to editor: dual-energy computed tomography-based volumetric thyroid iodine quantification: correlation with thyroid hormonal status, pathologic diagnosis, and phantom validation** *Ahmet Gürkan Erdemir*

INTERVENTIONAL RADIOLOGY

482 Original Article **Factors effecting the success of retrograde tibiopedal access and recanalization in infrapopliteal artery occlusions** *Cemal Aydın Gündoğmuş, Hande Özen Atalay, Vugar Samadli, Levent Oğuzkurt*

489 Original Article **Flow-diverting stents in the management of extracranial carotid artery aneurysms** *Celal Cinar, Erol Akgul, Alperen Elek, Mahmut Kusbeci, Egemen Ozturk, Hasan Bilen Onan, Irem Islek, Mohammad Naim Forogh, Mohammad Nawas Nasiri, Ismail Oran*

496 Technical Note **Utilization of a steerable microcatheter and adjunctive techniques for prostatic artery embolization in anatomically challenging vesicoprostatic trunks** *Hippocrates Moschouris, Çağın Şentürk, Konstantinos Stamatiou*

502 Letter to the Editor **Proposal for training: the educational value of a musculoskeletal embolization patellar tendinopathy model** *Emeric Gremen, Julien Ghelfi, Marylène Bacle, Julien Frandon*

MODALITY-BASED (UC, CT, MRI, PET-CT) IMAGING

505 Invited Review **New imaging techniques and trends in radiology** *Mecit Kantarcı, Sonay Aydın, Hayri Oğul, Volkan Kızılgöz*

NEURORADIOLOGY

518 Original Article **Impact of a computed tomography-based artificial intelligence software on radiologists' workflow for detecting acute intracranial hemorrhage** *Jimin Kim, Jinhee Jang, Se Won Oh, Ha Young Lee, Eun Jeong Min, Jin Wook Choi, Kook-Jin Ahn*

PEDIATRIC RADIOLOGY

526 Original Article **Feasibility study of computed high b-value diffusion-weighted magnetic resonance imaging for pediatric posterior fossa tumors** *Semra Delibalta, Barış Genç, Meltem Ceyhan Bilgici, Kerim Aslan*

Full text of these articles can be accessed online at www.dirjournal.org or through PubMed (https://www.ncbi.nlm.nih.gov/pmc/journals/2754/).

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243095

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Evaluating artificial intelligence for a focal nodular hyperplasia diagnosis using magnetic resonance imaging: preliminary findings

Mecit Kantarcı[1,2]
Volkan Kızılgöz[1]
Ramazan Terzi[3]
Ahmet Enes Kılıç[3]
Halime Kabalcı[4]
Önder Durmaz[1]
Nil Tokgöz[5]
Mustafa Harman[6]
Ayşegül Sağır Kahraman[7]
Ali Avanaz[8]
Sonay Aydın[1]
Gülsüm Özlem Elpek[9]
Merve Yazol[5]
Bülent Aydınlı[8]

[1]Erzincan Binali Yıldırım University Faculty of Medicine, Department of Radiology, Erzincan, Türkiye

[2]Atatürk University Faculty of Medicine, Department of Radiology, Erzurum, Türkiye

[3]Digital Transformation Office, Presidency of the Republic of Türkiye, Ankara, Türkiye

[4]Erzincan Binali Yıldırım University Faculty of Medicine, Department of Anatomy, Erzincan, Türkiye

[5]Gazi University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

[6]Ege University Faculty of Medicine, Department of Radiology, İzmir, Türkiye

[7]İnönü University Faculty of Medicine, Department of Radiology, Malatya, Türkiye

[8]Akdeniz University Faculty of Medicine, Department of General Surgery, Antalya, Türkiye

[9]Akdeniz University Faculty of Medicine, Department of Pathology, Antalya, Türkiye

Corresponding author: Volkan Kızılgöz

E-mail: volkankizilgoz@gmail.com

Received 25 October 2024; revision requested 04 December 2024; last revision received 09 February 2025; accepted 2 March 2025.

Epub: 26.03.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.243095

PURPOSE

This study aimed to evaluate the effectiveness of artificial intelligence (AI) in diagnosing focal nodular hyperplasia (FNH) of the liver using magnetic resonance imaging (MRI) and compare its performance with that of radiologists.

METHODS

In the first phase of the study, the MRIs of 60 patients (30 patients with FNH and 30 patients with no lesions or lesions other than FNH) were processed using a segmentation program and introduced to an AI model. After the learning process, the MRIs of 42 different patients that the AI model had no experience with were introduced to the system. In addition, a radiology resident and a radiology specialist evaluated patients with the same MR sequences. The sensitivity and specificity values were obtained from all three reviews.

RESULTS

The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the AI model were found to be 0.769, 0.966, 0.909, and 0.903, respectively. The sensitivity and specificity values were higher than those of the radiology resident and lower than those of the radiology specialist. The results of the specialist versus the AI model revealed a good agreement level, with a kappa (κ) value of 0.777.

CONCLUSION

For the diagnosis of FNH, the sensitivity, specificity, PPV, and NPV of the AI device were higher than those of the radiology resident and lower than those of the radiology specialist. With additional studies focused on different specific lesions of the liver, AI models are expected to be able to diagnose each liver lesion with high accuracy in the future.

CLINICAL SIGNIFICANCE

AI is studied to provide assisted or automated interpretation of radiological images with an accurate and reproducible imaging diagnosis.

KEYWORDS

Artificial intelligence, deep learning, liver lesion, focal nodular hyperplasia, magnetic resonance imaging

Focal nodular hyperplasia (FNH) is the second most common benign tumor of the liver after hemangioma. The prevalence of FNH was found to be 0.4% to 3% in autopsy series.[1] FNH is believed to result from arterial malformations, and 60%–80% of cases are asymptomatic and are discovered incidentally.[2,3] The imaging characteristics of FNH correspond well with histological properties and are observed as a solitary well-circumscribed lobulated mass in a cross-sectional imaging study (Figure 1).[4] Magnetic resonance imaging (MRI) has a higher sensitivity than ultrasound and computed tomography (CT) imaging and a specificity of almost 100%.[5] In MRI, a typical FNH is a solitary, well-defined, unencapsulated lesion with central scar formation.[6] Approximately 35%–70% of FNH lesions do not have

You may cite this article as: Kantarcı M, Kızılgöz V, Terzi R, et al. Evaluating artificial intelligence for a focal nodular hyperplasia diagnosis using magnetic resonance imaging: preliminary findings. *Diagn Interv Radiol.* 2025;31(5):405-415.

these imaging features; they might have a pseudo capsule mimicking a true capsule, show washout-like hepatocellular carcinoma (HCC), or have no scar formation.[7,8] The hepatobiliary phase (HBP) of MRI provides important data for the diagnosis of FNH, and 73%–90% of these lesions are observed with iso-intensity or hyperintensity in the HBP.[9] Even though HCC and hepatic adenoma are usually hypointense in the HBP, these lesions may have upregulated hepatocyte-specific membrane transport proteins and, thus, may be observed as an iso- or hyperintense lesion in HBP images.[4]

Artificial intelligence (AI) is becoming a widespread method to interpret radiological images for research purposes, even in daily practice. It is expected to provide assisted or automated interpretation of radiological images with an accurate and reproducible imaging diagnosis. After obtaining images of the patients, AI may quickly interpret them and make critical diagnostic decisions for numerous patients. This may provide a quick and accurate diagnosis of many lesions located in different organs or systems in the future. Thus, all scientific studies targeted at developing AI for use as a diagnostic assistant can be considered a contribution to this topic. As a diagnostic tool, AI has been used in the detection and characterization of diffuse diseases or focal lesions of the liver and pancreas in recent studies. It has been applied to different imaging techniques, including ultrasound, CT, and MRI.[10]

The aim of the present study was to determine the effectiveness of AI in detecting the presence of FNH lesions of the liver and compare this diagnostic capacity of AI with that of radiologists. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated, considering the radiological and pathological results of the patients as the gold standard.

Main points

- The targeted long-term result is automated interpretation with an accurate diagnosis using artificial intelligence (AI) models for liver lesions; this study is part of the AI education program focusing on a specific liver lesion.
- A new scoring system is established to train the AI model to distinguish focal nodular hyperplasia (FNH) from other liver lesions.
- The AI model used in this research achieved sensitivity and specificity values higher than those of a radiology resident and lower than those of a radiology specialist for the diagnosis of FNH.

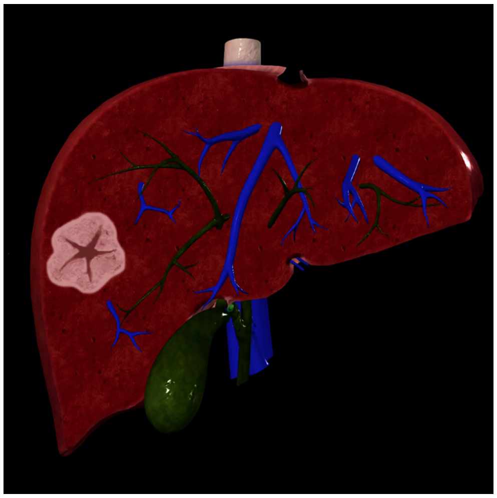


Figure 1. Schematic of classic focal nodular hyperplasia, a solitary well-circumscribed lobulated mass with central scar tissue. This illustration has been created by the Adobe Photoshop program (Adobe Inc., 2021. Adobe Photoshop, https://www.adobe.com/products/photoshop.html) based on the figures provided by the Complete Anatomy program (3D4 Medical, 2021; Complete Anatomy; retrieved from https://3d4medical.com)

Methods

Patients and the study workflow

This study was approved by the Ethics Committee of Erzincan Binali Yıldırım University (clinical trial number: 2023-13/6, date: 22.06.2023) and the requirement for informed consent was waived by the ethics committee due to the retrospective nature of the study. The study population constituted patients who had undergone MRI, and abdominal MRIs of 30 patients were used in the initial phase. In the first phase, the MRIs of patients (n = 30) who had been histologically diagnosed with FNH were introduced to the AI system using a segmentation program. In addition, the abdominal MRIs of 30 patients with no liver lesions were segmented using the same program. A scoring system was used to diagnose FNH. Then, 42 patients with various lesions, including FNH (n = 13), HCC (n = 5), low-grade dysplastic nodules (n = 1), hepatic adenoma (n = 3), biliary hamartoma (n = 1), primary hepatic neuroendocrine tumor (n = 1), colon cancer metastasis (n = 2), breast cancer metastasis (n = 2), stomach cancer metastasis (n = 1), pancreatic cancer metastasis (n = 1), hydatid cyst (n = 1), complex cyst (n = 1), biliary cystadenoma (n = 2), hemangioma (n = 4), simple cyst (n = 3), and a normal liver were reviewed by AI and two radiologists (a specialist with 18 years of experience and a radiology resident with 2.5 years of experience) independently in randomized order (Figure 2). Following the AI interpretations, sensitivity, specificity, the PPV, and the NPV were calculated. Then, the accuracy of the results from the AI model and the two radiologists were compared. The radiological diagnosis (stable lesions with typical imaging features in follow-up examinations or typical imaging findings with primary tumor) and the histological results (obtained by biopsy procedures) were taken as the gold standard to reveal the sensitivity and specificity values.

Diagnosing focal nodular hyperplasia

A standardized method was used to simplify the interpretation, and for reproducibility and repeatability regarding the FNH diagnosis, only the axial plane images, including T1-weighted, T2-weighted, and T1-weighted enhanced (arterial, portal, and venous phase) images, and HBP images were evaluated. Typically, FNH is hypointense or isointense on T1-weighted images and hyperintense or isointense on T2-weighted images, showing intense contrast medium enhancement in the arterial phase and retaining contrast in the portal and venous phases.[11] The central FNH scar is best seen on MRI. The scar is monitored as hypointense on a pre-contrast T1-weighted sequence, substantially hyperintense on T2-weighted images, and becomes hyperintense on HBP images because of the accumulation of the contrast medium in the fibrous tissue. Most FNH lesions are iso- or hyperintense on HBP images (Figure 3).[7]

Magnetic resonance imaging protocol and selected sequences

All the MRIs were acquired using a 1.5T MRI scanner (Magnetom Era, Siemens, Erlangen, Germany) with a standard abdominal coil. The axial sequences, including the T1-weighted and T2-weighted images, as well as the contrast-enhanced phases, were evaluated. The MRIs were segmented using dedicated software, ensuring the precise identification of focal lesions. All the contrast-enhanced T1-weighted images were obtained using gadoxetate disodium (Primovist®) through intravenous injection at a dosage of 0.1 mmol/kg (maximum dose, 20 mL) and a rate of 2 mL/s, followed by saline flush (50 mL at the rate of 2 mL/s). Postcontrast images were analyzed, including the late arterial phase (15–20s postinjection), portal venous phase (60–70s postinjection), delayed phase (3–5 min postinjection), and HBP (20 min postinjection). The axial plane T1-and T2-weighted, arterial, portal, venous, and HBP enhanced T1-weighted MRIs were introduced to the AI model and interpreted by the two radiologists through the liver lesion diagnostic process in relation to FNH. The following technical parameters were applied to both the enhanced and non-enhanced series: T2 weighted: time of repetition (TR): 1,200 ms, time of echo (TE): 95 ms, number of excitations (NEX): 1, slice thickness: 6 mm; T1 weighted: TR: 6.94 ms, TE: 2.39 ms, NEX: 1, slice thickness: 3.3 mm.

Segmentation

The segmentation process was performed by an anatomist and a radiologist (with 3 and 27 years of experience, respectively) using the same monitor as that used for segmentation of the abdominal MRIs and at the same time. The radiologist decided on the presence and locations of the liver lesions for each patient. The anatomist had learned about image maps, anatomical details, and liver lesions from an experienced radiologist. The anatomist consulted with the radiologist at every step. Each patient MR examination was also checked at the end of the segmentation session by the experienced radiologist for every segmented anatomical part or liver lesion. Segmentation of the axial images was performed with 3D Slicer software (v5.3.0, http://www.slicer.org) manually. The liver borders, FNH lesions (if any), lesions other than FNH, the main branches of the portal veins, and the hepatic veins were segmented on six sequences in the axial plane, as described before in the diagnosing FNH section. Only the focal lesions were tagged, and fibrosis or other diffuse parenchymal signal alterations were not segmented. The main portal veins and main hepatic veins were segmented in each patient. All FNH lesions in the liver were segmented if the patient had more than one lesion. The FNH lesions were segmented based on the lesion borders, and scar formation was also segmented in typical FNH lesions. The FNH lesion, scar formation of the FNH lesion, liver, main portal vein, main hepatic veins, and lesions other than FNH were tagged with different colors before being introduced to the AI model. Based on the axial slices, three-dimensional (3D) reconstruction images were also obtained by the segmentation program (Figure 4). After the AI training session, a radiology resident (with 2.5 years of experience), a radiology specialist (with 18 years of experience), and the AI model evaluated the random dataset that included the FNH and other lesions.

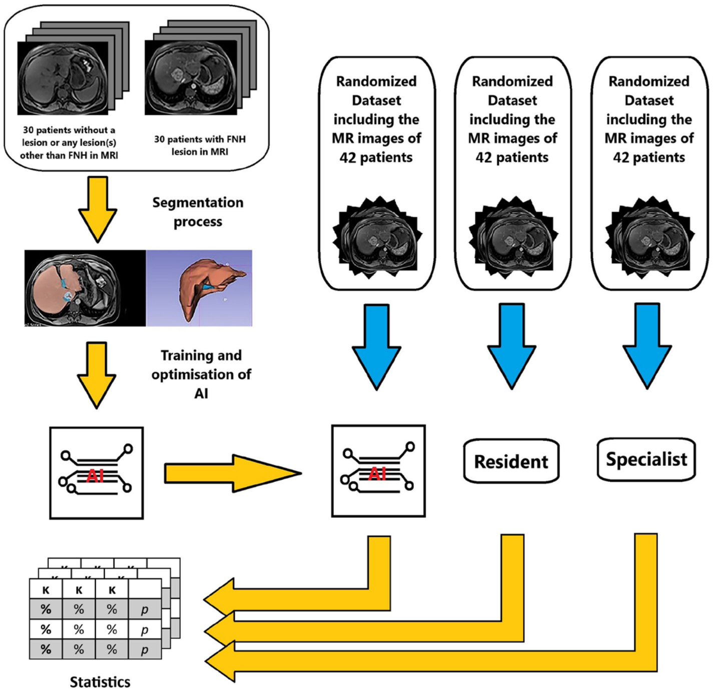

30 patients without a lesion or any lesion(s) other than FNH in MRI
30 patients with FNH lesion in MRI
Segmentation process
Training and optimisation of AI
AI
Randomized Dataset including the MR images of 42 patients
Randomized Dataset including the MR images of 42 patients
Randomized Dataset including the MR images of 42 patients
AI
Resident
Specialist
K
K
K
%
%
%
p
%
%
%
p
%
%
%
p
Statistics

Figure 2. Study workflow. After the segmentation process and training the AI model, a randomized dataset was evaluated by AI, a radiology resident, and a radiology specialist independently. AI, artificial intelligence; FNH, focal nodular hyperplasia; MRI, magnetic resonance imaging.

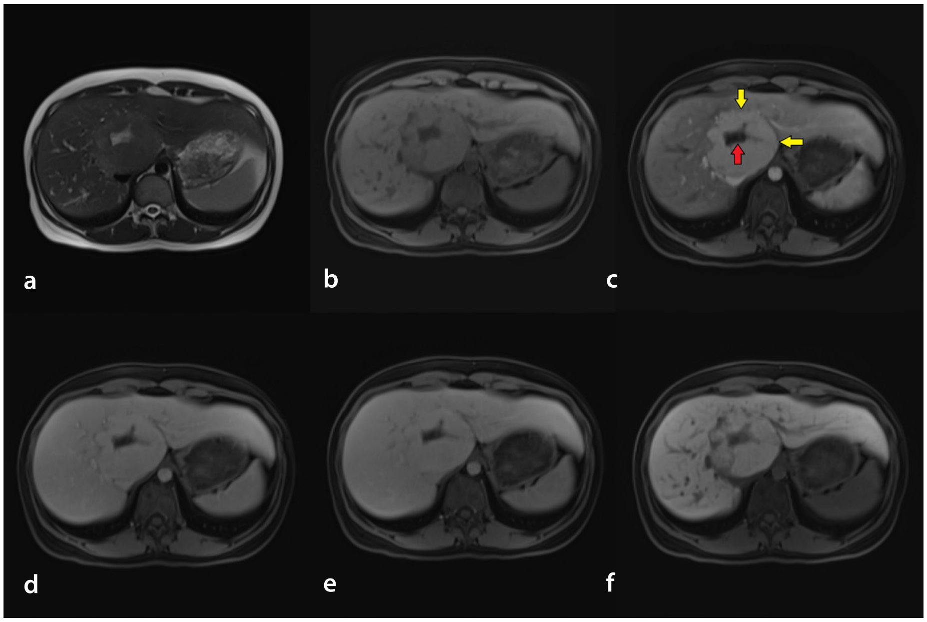


Figure 3. Classic focal nodular hyperplasia with radial scar tissue. Axial plane magnetic resonance images with T2-weighted **(a)**, T1-weighted pre-contrast **(b)**, arterial phase contrast-enhanced T1-weighted **(c)**, portal phase contrast-enhanced T1-weighted **(d)**, venous phase contrast-enhanced T1-weighted **(e)**, and hepatobiliary phase **(f)** images (yellow arrows indicate the lesion in an enhanced T1-weighted axial plane image, and the red arrow shows the typical central scar of the lesion).

Artificial intelligence protocol and data preprocessing

The workflow developed for FNH detection with AI from MRIs is presented in Figure 5. The workflow consists of two stages: segmentation and the FNH detection process. The process from dataset preparation to FNH detection with AI is explained in detail in this section.

The MRI data provided were converted from nrrd format to .nii.gz format, and a data standard was created. For the 3D modeling, the data were converted to Medical Segmentation Decathlon format.[12] To produce a more generalizable result, the five-fold cross-validation method was applied instead of random split for the algorithms.

Deep learning architectures

Organs such as the liver, veins, and gallbladder can be detected in MRIs thanks to deep learning architecture such as object recognition, semantic segmentation, and instance segmentation. In this study, a decision-making process was used to focus on the intensity of FNH so that it could be detected by AI. Thus, the use of segmentation algorithms was deemed more appropriate. Moreover, it was decided to use 3D segmentation algorithms instead of two-dimensional (2D) segmentation algorithms to access the temporal information between MRI slices. In this study, the nnU-Net deep learning algorithm, a deep learning-based semantic segmentation model developed with both 2D and 3D U-Net configurations, was used.[13] We chose to use this algorithm for this study because it can automatically configure appropriate preprocessing, network architecture, training parameters, and post-processing processes according to the data in the medical imaging.

Artificial intelligence-training and testing

The 3D nnU-Net model training was performed in three categories—the liver, vein, and FNH—using model configurations prepared based on the data of 60 patients, 30 with FNH and 30 without. Model training was conducted with the five-fold cross-validation method. Thirty nnU-Net models were trained in five-fold form over six phases: T1 weighted, T2 weighted, arterial, portal, and HBP. The hyperparameters used for model training are shown in Table 1. Optimal model selection was made according to the highest average validation Dice score. The most successful 3D nnU-Net model selected was tested on 30 test patients.

Artificial intelligence-evaluation metrics

The metrics used to evaluate the segmentation model performance provide a key tool for measuring the sensitivity, accuracy, and overall effectiveness of the developed model. In this study, the Dice score (Sørensen–Dice coefficient) metric was used. The Dice score is a metric that measures how well the region predicted by the model overlaps with the actual labeled region. This metric, used to evaluate the similarity between two clusters, is calculated with the following formula:[14]

$$Dice\ Score = 2x\frac{|Prediction \cap Ground\ truth|}{|Prediction| + |Ground\ truth|}$$

In this formula, prediction represents the segmentation region predicted by the model, and ground truth represents the ground truth region. The Dice score has a value between 0 and 1, with a value closer to 1 indicating greater overlap. A high Dice score indicates that the model performs segmentation correctly, whereas a low value indicates that the model's predictions are incompatible with the actual data.

Artificial intelligence-registration

Six phases were used to decide whether a patient had an FNH liver lesion. When six deep learning models are developed for six phases and used separately, a situation occurs if the lesion can be found in one phase

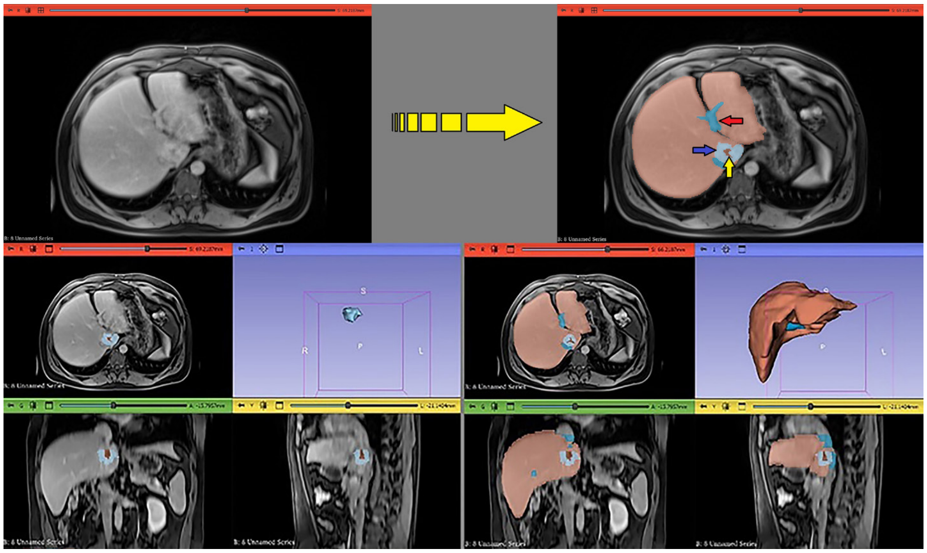
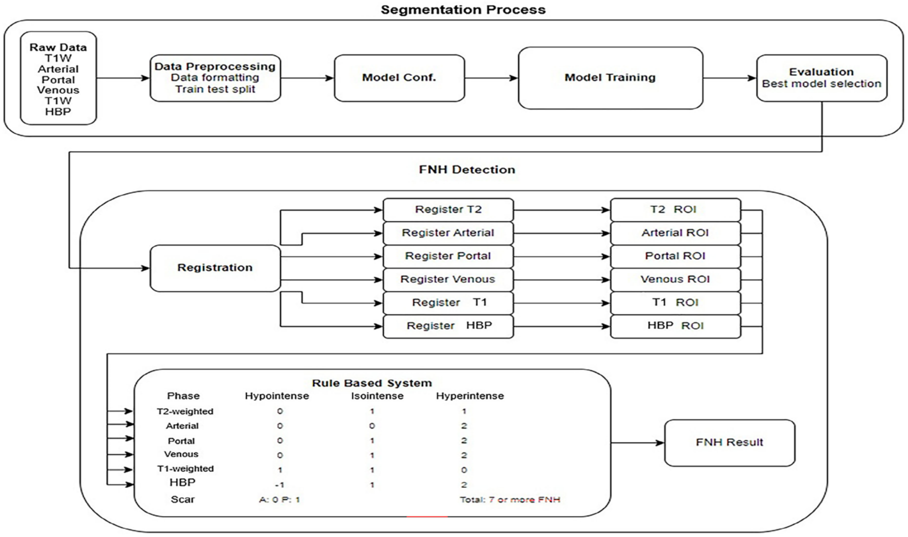


Figure 4. Magnetic resonance images after segmentation of the anatomical structures and focal nodular hyperplasia (FNH). After segmentation of all of the axial slices, either FNH or anatomical formation of the liver has been coded and tagged as a space-occupying structure by the segmentation program. In the right upper corner, the left main portal vein (red arrow), FNH (blue arrow), and central FNH scar formation (yellow arrow) are tagged with different colors after segmentation.



Rule Based System

Phase	Hypointense	Isointense	Hyperintense
T2-weighted	0	1	1
Arterial	0	0	2
Portal	0	1	2
Venous	0	1	2
T1-weighted	1	1	0
HBP	-1	1	2
Scar	A: 0 P: 1		Total: 7 or more FNH

Figure 5. Artificial intelligence workflow. T1W, T1-weighted; HBP, hepatobiliary phase; ROI, region-of-interest; FNH, focal nodular hyperplasia.

and not in another. In this case, deficiencies in the evaluation exist in terms of AI. Therefore, a 3D registration process was used in this study. The 3D registration process is used to align the position and orientation of images in the 3D space. This process is generally performed to obtain geometric harmony between a reference (base) and a moving image. In this study, a reference phase and the other five phases were registered separately. Since the most successful deep learning model was developed on the arterial phase, the reference phase was determined as the arterial phase. The registration process shown in Figure 6a has been produced automatically in 3D Slicer (Figure 6). The elastix registration method was used in this process.[15]

Region-of-interest extraction

When performing phase checks for FNH, specialist physicians make decisions by focusing on the surroundings of the FNH region. However, deep learning models segment all the relevant locations for the liver, vein, and FNH. To solve this, the region-of-interest (ROI) extraction post-processing method was used. For ROI extraction, as shown in Figure 6b, the region segmented by the deep learning model as FNH was increased by 30%, and only the liver and vein segmentations around the FNH label were obtained. Since the arterial phase is the reference phase, the regions predicted by the deep learning model in the arterial phase were mapped onto the other five registration phases, and ROI extraction was completed for the six phases.

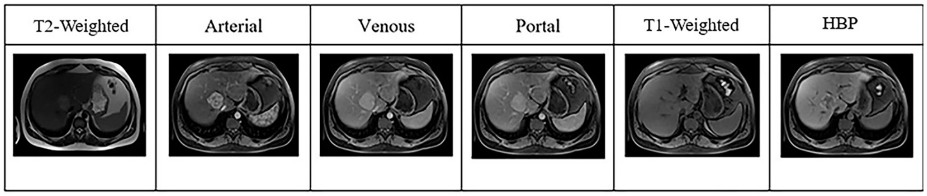
Rule-based system

The average pixel intensity was measured using the signal intensity of the liver, vein, and FNH segmentations within the ROI regions, six phases apart, and extracted. To make an intensity decision, the liver average pixel intensity of each phase was compared with the FNH average pixel intensity. To determine the lesion as hypo-, iso-, or hyperintense relative to the liver tissue, the surrounding liver parenchyma (the adjacent 30% of the area of the lesion) was considered (Figure 7). A comparison table for the intensity decision and the scoring system for each phase is shown in Table 1, and the decision regarding the presence of FNH is made according to the MR intensity obtained. To enable AI to determine the presence of FNH, a strict pattern must be followed. A lack of information in the literature and the absence of any widely used or accepted rule to enable AI to decide accurately, compelled the researchers to find a new pathway. Therefore, a new scoring system was established based on the MR signal features of the lesion. The images of patients in the training session (the images of 30 patients with at least one FNH lesion) were used for the preliminary testing to optimize the scoring system. According to this scoring system, for the unenhanced series, 1 point was allocated to iso- or hyperintensity in T2-weighted images and 1 point to hypo- or iso-intensity in T1-weighted images. For the dynamic contrast-enhanced series, the signal intensity was identified relative to the surrounding liver parenchyma. According to this rule, hyperintensity in the arterial phase and iso- or hyperintensity in the portal, venous, and HBPs were all allocated 1 point. A lesion with scar tissue was considered as 2 points. In total, 7 or more points were considered to be FNH according to the morpholog-

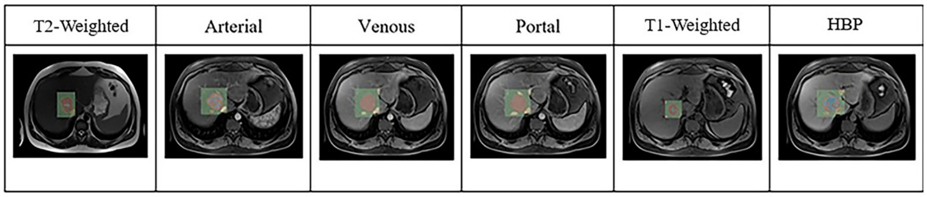
Table 1. Model hyperparameters, intensity decision, and the rule-based system of this study

Model hyperparameters			
Hyperparameters		**Values**	
Epoch number		1,000	
Batch size		2	
Learning rate		Poly learning rate scheduler (initial learning rate: 0.01)	
Optimizer		Stochastic gradient descent	
Momentum		0.99	
Weight decay		3e-05	
Loss function		Robust cross entropy loss Memory efficient soft Dice loss	
Intensity decision (SI unit)			
Value (liver SI–FNH SI)		**Decision of intensity**	
Value < −10		Hypointense	
−10< value <10		Isointense	
Value >10		Hyperintense	
Rule-based system			
T2 weighted	0	1	1
T1 weighted	1	1	0
Arterial phase	0	0	2
Portal phase	0	1	2
Venous phase	0	1	2
Hepatobiliary phase	−1	1	2
Scar	Absence of the scar scored as 0 points, and presence of the scar scored as 1 point		
Result: 7 or more points were considered focal nodular hyperplasia			

SI, signal intensity; FNH, focal nodular hyperplasia.



a-Registration example of the MRI phases



b-ROI extraction of the MRI phases

Figure 6. Registration process of each magnetic resonance sequence and the region-of-interest extraction. HBP, hepatobiliary phase; ROI, region-of-interest; MRI, magnetic resonance image.

ical appearance in the MRIs. If one or more lesions in the liver were consistent with FNH on the MRIs, the patient was accepted as FNH positive by each reviewer.

Statistical analysis

All statistical analyses were calculated using IBM SPSS statistics v22.0 for Windows. Sensitivity, specificity, the PPV, the NPV, and accuracy were calculated using the chi-square test, with the radiological and histological results considered as the gold standard. The area under curve (AUC) values were calculated and presented as 95% confidence intervals (CIs). Cohen's kappa analysis was performed to reveal the agreement levels between reviewers, and Koo et al.'s[16] classification method was used to represent the agreement levels. According to these agreement levels, values less than 0.50 indicated poor agreement, values between 0.50 and 0.75 showed moderate agreement, values between 0.75 and 0.90 revealed good agreement, and values greater than 0.90 demonstrated excellent agreement.[16] A *P* value < 0.05 was considered to represent a statistically significant difference.

Results

The training of the 30 nnU-Net models was conducted in the form of five-fold cross validation for six phases: T2 weighted, pre-contrast (T1 weighted), arterial, portal, venous, and HBP. The nnU-Net deep learning algorithm automatically adjusts model configurations according to the data. Figure 2 shows the Dice score and mean validation Dice score for the liver, vein, and FNH classes over five-fold means ± standard separately for each phase.

Among the 30 nnU-Net models trained, the results were given singularly for each MR sequence. The arterial phase images had the highest performance in terms of the average validation Dice score (0.7998), and this sequence was chosen as the best model in the category of both FNH class success and high average validation Dice score (Table 2).

In this study, 5 of the 13 FNH lesions were typical FNH lesions with scar formation. The list of patients with the histological and interpretation results of each reviewer are presented in Table 3. Two patients had more than one FNH lesion in the liver, as seen in the dataset presented in Table 3. The dimensions of the FNH lesions measured on the axial plane are presented in Table 4. The mean was 2.78 ± 1.84 for the FNH dimensions.

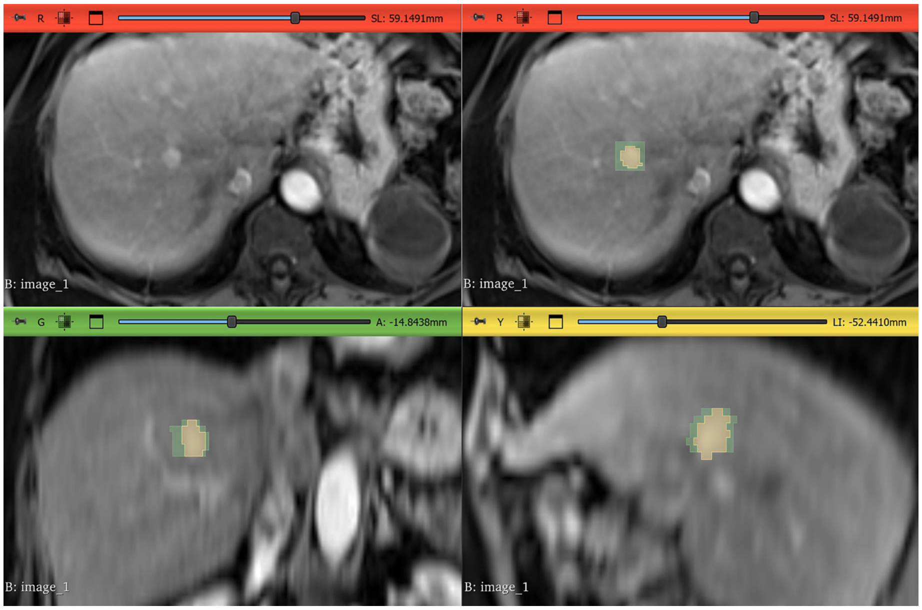


Figure 7. Surrounding liver parenchyma (the adjacent 30% of the area of the lesion) was considered to determine the lesion as hypo-, iso-, or hyperintense relative to the liver tissue.

Table 2. Fold mean results

Sequence	Liver	Vein	FNH	EMA Dice	Mean val Dice
T2 weighted	0.949 ± 0.002	0.586 ± 0.05	0.254 ± 0.253	0.571 ± 0.08	0.512 ± 0.03
T1 weighted	0.954 ± 0.01	0.734 ± 0.02	0.136 ± 0.160	0.589 ± 0.04	0.560 ± 0.01
Arterial phase	0.955 ± 0.01	0.680 ± 0.02	0.733 ± 0.246	0.780 ± 0.09	0.712 ± 0.05
Venous phase	0.961 ± 0.01	0.746 ± 0.04	0.645 ± 0.221	0.767 ± 0.08	0.671 ± 0.07
Portal phase	0.946 ± 0.04	0.752 ± 0.03	0.608 ± 0.246	0.759 ± 0.09	0.665 ± 0.07
HBP	0.948 ± 0.02	0.629 ± 0.10	0.436 ± 0.262	0.651 ± 0.09	0.633 ± 0.06

FNH, focal nodular hyperplasia; EMA, exponential moving average; Val, validated; HBP, hepatobiliary phase.

Table 3. Reviewer results for each patient ("+" indicates the presence of FNH, and "–" represents the absence of FNH in the liver)

	Reference diagnosis	RD	H	Resident	AI	Specialist
1	Normal liver	+		–	–	–
2	Hemangioma	+		–	–	–
3	Hydatid cyst	+		–	–	–
4	PHNT		+	–	–	–
5	Simple cyst	+		–	–	–
6	Hemangioma	+		+	–	–
7	Breast cancer metastasis		+	–	–	–
8	FNH	+		+	+	+
9	FNH	+		+	–	–
10	Stomach cancer metastasis		+	–	–	–
11	Hemangioma	+		+	–	–
12	Hepatocellular carcinoma		+	–	+	–
13	Biliary hamartoma		+	–	–	–
14	Complex cyst		+	–	–	–
15	Low-grade dysplastic nodule		+	–	–	–
16	FNH	+		+	+	+
17	Breast cancer metastasis	+		–	–	–
18	Colon cancer metastasis		+	–	–	–

The liver interpretations on the MRIs according to each reviewer were compared with the histopathological results regarding the presence of FNH. The sensitivity, specificity, PPV, and NPV obtained from the radiology resident, the radiology specialist, and the AI model are presented in Table 5. According to these results, the diagnostic parameters of the AI model were better than those of the resident and lower than those of the specialist.

The results of the radiology resident and the AI model showed poor agreement (κ = 0.374), and the results of the radiology resident and the radiology specialist indicated good agreement (κ = 0.602). The results of the radiology specialist and the AI model revealed good agreement (κ = 0.777) (Table 6). The AUC values with 95% CI were 0.794 (0.630–0.959) for the radiology resident, 0.833 (0.682–0.983) for the AI model, and 0.944 (0.851–1.000) for the radiology specialist (Table 7). The accuracy values were 0.833, 0.905, and 0.952 for the radiology resident, AI model, and radiology specialist, respectively.

Discussion

The AI model used in this study had 76.9% sensitivity, 96.6% specificity, a 90.9% PPV, and a 90.3% NPV for the diagnosis of FNH of the liver. The AI results were better than those of the radiology resident and lower than those of the radiology specialist. Additionally, the AI results indicated a good level of agreement with the specialist.

FNH is a conservatively managed lesion for most patients, and surgery is not required in the management of this condition. Only patients with pedunculated, exophytic, or expanding lesions are considered for surgery.[17] Hepatic adenoma, however, is treated by surgery because of its well-known complications, including spontaneous hemorrhage and malignant transformation.[18] HCC is another lesion that occurs in the differential diagnosis of FNH, and HCC may also occur in a non-cirrhotic liver.[19] The spectrum of patients that AI will evaluate should comprise all these lesions as well as cirrhotic livers that may have diagnostic challenges. Another important discussion point is distinguishing hepatic adenomas from FNH lesions. This may not be easy to accomplish using MRIs. Most adenomas (reported to be between 75% and 90%) are hypointense in the HBP, whereas FNH is iso- or hyperintense compared with the surrounding liver parenchyma, and these different lesion properties make the diagnosis easier in daily practice.

Table 3. Continued

	Reference diagnosis	RD	H	Resident	AI	Specialist
19	Colon cancer metastasis	+		–	–	–
20	Hepatic adenoma		+	+	+	+
21	Hemangioma	+		–	+	–
22	Simple cyst	+		–	–	–
23	Pancreas cancer metastasis	+		–	–	–
24	Hepatocellular carcinoma		+	–	–	–
25	FNH	+		–	+	+
26	FNH	+		+	–	+
27	Simple cyst	+		–	–	–
28	FNH	+		+	+	+
29	Biliary cystadenoma		+	–	–	–
30	Biliary cystadenoma		+	–	–	–
31	FNH	+		+	+	+
32	FNH	+		–	+	+
33	FNH		+	–	+	+
34	FNH	+		+	–	+
35	FNH		+	–	+	+
36	FNH	+		+	+	+
37	FNH	+		+	+	+
38	Hepatocellular carcinoma		+	–	–	–
39	Hepatic adenoma	+		–	–	–
40	Hepatocellular carcinoma		+	–	–	–
41	Hepatocellular carcinoma		+	–	–	–
42	Hepatic adenoma		+	–	–	–

AI, artificial intelligence; RD, radiological diagnosis (stable in follow-up examinations or typical imaging findings with primary tumor), H, histologically confirmed lesions; PHNT, primary hepatic neuroendocrine tumor; FNH, focal nodular hyperplasia.

Table 4. Dimensions of the FNH lesions in axial plane

	Patient number*	Lesion 1 (TR × AP)	Lesion 2 (TR × AP)	Lesion 3 (TR × AP)
1	Patient number 8	3.62 × 2.97 cm		
2	Patient number 9	1.56 × 1.39 cm		
3	Patient number 16	1.47 × 1.42 cm		
4	Patient number 25	3.39 × 2.79 cm		
5	Patient number 26	0.80 × 0.89 cm		
6	Patient number 28	1.98 × 1.59 cm	2.96 × 2.54 cm	6.65 × 7.50 cm
7	Patient number 31	7.17 × 5.49 cm		
8	Patient number 32	5.41 × 5.15 cm		
9	Patient number 33	3.51 × 3.30 cm		
10	Patient number 34	1.28 × 1.49 cm	1.94 × 2.01 cm	
11	Patient number 35	2.24 × 1.66 cm		
12	Patient number 36	0.91 × 0.86 cm		
13	Patient number 37	1.75 × 2.33 cm		

*Represents the patient numbers in the study, also shown in Table 3; TR, maximum transvers diameter of the lesion; AP, maximum anteroposterior diameter of the lesion. TR, time of repetition; AP, anterioposterior.

The uptake and excretion of hepatocyte-specific contrast agents into the biliary system is facilitated by hepatocyte-specific membrane transport proteins, which are not present in other cells. HCC and hepatic adenoma are usually hypointense in the HBP; however, these lesions may have upregulated hepatocyte-specific membrane transport proteins, which make them appear as iso- or hyperintense lesions in HBP images. Approximately

25% of inflammatory hepatic adenomas and 40–80% of beta-catenin-activated hepatic adenomas are reported to appear as iso- or hyperintense on HBP images, and this overlap makes diagnosis challenging. Moreover, beta-catenin-activated hepatic adenomas have the highest risk for malignant transformation (40%).[4]

This study is a step forward in using AI to diagnose one of the most common hepatic nodular lesions, FNH. Not only typical but also atypical nodular hyperplasia lesions, which have been histologically confirmed, were evaluated through AI as a reviewer. The AI model provided a relatively high sensitivity value along with 96.6% specificity in diagnosing FNH in the liver with one or more nodular lesions, including non-FNH lesions.

In the literature, researchers have included many parameters in their studies. These include the contrast curve, gray-level histogram, and gray-level co-occurrence matrix texture properties, as well as risk factors, such as the presence of steatosis, known primary tumors, or cirrhosis, and MR sequences such as dynamic contrast-enhanced T1-weighted with T2-weighted images, for the classification of focal liver lesions.[20] In the present study, a simplified approach, using only certain MR sequences that were unaware of other risk factors or medical conditions, was used to understand the diagnostic success of AI. The nnU-Net deep learning algorithm was chosen for this study, which is considered to be highly impactful for object identification with successful segmentation capabilities. This algorithm was designed to optimize 2D or 3D image segmentation tasks and is usable for any given input geometrical type. This deep learning modality optimally segments organs using CT images based on the use of differences of densities.[21] In this research, signal intensity was used as the indicator of FNH lesions using the same algorithm and both 2D and 3D U-Net configurations.

In this study, arterial phase images had the highest performance for the average validation Dice score. Dice scores were important for determining the anomaly and starting to implement further calculations to reveal the lesion characterization regarding the presence of FNH. The ground truth-based border drawn by the radiologists was analyzed along with a prediction based on the border of the model. The Dice score represents the overall segmentation performance and indicates the success of the segmentation through the prediction ability of the model. The Dice score ranges from 0 (no overlap compared with the segmented borders of the radiologist) to 1 (perfect overlap compared with the segmented borders of the radiologist). This method was used in the literature for similar purposes, such as the segmentation of HCC in the liver.[22]

Table 5. Sensitivity, specificity, and positive and negative predictive values of the reviewers with 95% confidence intervals

Results for the resident				
Radiology resident		Radiology resident		Total
		Positive	Negative	
Pathology result	Positive	9	4	13
	Negative	3	26	29
Total		12	30	42
Sensitivity: 0.692 (0.388 ± 0.896) Specificity: 0.897 (0.715 ± 0.972) Positive predictive value: 0.750 (0.428 ± 0.933) Negative predictive value: 0.867 (0.683 ± 0.956)			$P < 0.001$	
Results for artificial intelligence				
AI		AI		Total
		Positive	Negative	
Pathology result	Positive	10	3	13
	Negative	1	28	29
Total		11	31	42
Sensitivity: 0.769 (0.459 ± 0.938) Specificity: 0.966 (0.803 ± 0.998) Positive predictive value: 0.909 (0.571 ± 0.995) Negative predictive value: 0.903 (0.730 ± 0.974)			$P < 0.001$	
Results for the radiology specialist				
Radiology specialist		Radiology specialist		Total
		Positive	Negative	
Pathology result	Positive	12	1	13
	Negative	1	28	29
Total		13	29	42
Sensitivity: 0.923 (0.620 ± 0.995) Specificity: 0.966 (0.803 ± 0.998) Positive predictive value: 0.923 (0.620 ± 0.995) Negative predictive value: 0.966 (0.803 ± 0.998)			$P < 0.001$	

AI, artificial intelligence.

Table 6. Kappa values for the comparison of each reviewer

	Value	Asymptotic standard error[a]	Approximate T[b]	Approximate significance
Resident vs. AI	0.374	0.155	2.428	0.015
Specialist vs. AI	0.777	0.106	5.037	<0.001
Resident vs. specialist	0.602	0.135	3.905	<0.001

[a], not assuming a null hypothesis; [b], using the asymptotic standard error assuming a null hypothesis. AI, artificial intelligence.

Table 7. Area under the curve for each reviewer's results

Test result variable (s)	Area	Standard error	*P* values	Asymptotic 95% confidence interval	
				Lower bound	Upper bound
Resident	0.794	0.084	0.003	0.630	0.959
AI	0.833	0.077	0.001	0.682	0.983
Specialist	0.944	0.048	0.000	0.851	1.000

AI, artificial intelligence.

According to the results presented in Table 2, each imaging phase demonstrates distinct characteristics in segmenting different parts of the liver. For instance, the HBP exhibited the highest overall liver segmentation performance, with a Dice score of 0.948 ± 0.02, whereas the portal phase achieved the best vein segmentation performance, with a Dice score of 0.752 ± 0.03. Similarly, the most effective segmentation for FNH was observed in the arterial phase, yielding a Dice score of 0.733 ± 0.246. In summary, based on the mean Dice score across phases, the arterial phase proved most effective in segmenting the three liver components, achieving a Dice score of 0.712 ± 0.05. This might be a result of the increased signal difference between the lesion and the surrounding parenchyma. The ability of the model to distinguish the lesion borders was considered superior for the arterial phase images. Consequently, the arterial phase was chosen as the foundational model. Specifically, the first-fold model of the arterial phase, which achieved a mean Dice score of 0.7998, was selected as the base model for FNH detection and segmentation. Subsequently, the scoring system was considered for analyzing the lesion, particularly for the diagnosis of FNH.

There are attempts in the literature to use autotomized AI models to diagnose focal liver lesions. In Goehler et al.'s[23] study, the researchers tried to detect liver metastases and evaluate changes in tumor size on consecutive MR examinations. A convolutional neural network (CNN) and Kuhn–Munkres algorithm were used for 64 patients with neuroendocrine tumors with two consecutive liver MR examinations using gadoxetic acid. The results of this study indicated that this evaluation system was 91% concordant with the radiologists' decision, and the sensitivity and specificity were 0.85 and 0.92, respectively. In addition, the model was capable of assessing the interval change in tumor burden between two MRI examinations.[23] A computer-assisted diagnosis system, the liver artificial neural network (ANN), was analyzed by Zhang et al.[24] regarding its feasibility for identifying focal liver lesions. Using an ANN technique, this system classified the liver lesions into five categories. Their investigation used 320 MRIs (from 80 patients); however, the system was human assisted, and a radiologist had to delineate an ROI for the lesion. The five hepatic categories for the lesions in their study were cavernous hemangioma, HCC, hepatic cyst, dysplasia in cirrhosis, and metastasis. This liver ANN system was developed to assist the radiologists, giving a second opinion with a training accuracy of 100% and a testing accuracy of 93%.[24] For the diagnosis of focal liver lesions, Hamm et al.[25] performed a study using multi-phasic MRIs, and 92% sensitivity, 98% specificity, and 92% accuracy were achieved with their CNN. In the same study, the model displayed a sensitivity of 90% for the diagnosis of HCC, whereas the radiologist achieved 70%.[25] Jansen et al.[20] utilized a system of automatic classification to classify focal liver lesions using MRIs and the risk factors for a more accurate diagnosis. They achieved an overall accuracy for focal liver lesions of 0.77. The sensitivity and specificity values for hepatic hemangioma were 84% and 82%, respectively, for hepatic cyst, 93% and 93%, for hepatic adenoma, 80% and 78%, for HCC, 73% and 56%, and for metastasis, 62% and 77%.[20] Zhen et al.[26] analyzed the efficiency of a deep learning-based tool based on the fact that dynamic contrast-enhanced MRI provides the most precise diagnosis of hepatic tumors. In their analysis, enhanced and unenhanced MRIs, along with relevant patient clinical information, were used. The results indicated that the deep learning-based system differentiated malignant from benign focal liver lesions well using only unenhanced images (AUC: 0.946; 95% CI: 0.914–0.979 vs. AUC: 0.951; 95% CI: 0.919–0.982, $P = 0.664$). Moreover, the performance of the deep learning-based system was improved when combining unenhanced images with clinical data to classify malignancies as metastatic tumors (AUC = 0.998; 95% CI: 0.989–1.000), HCC (AUC: 0.998; 95% CI: 0.989–1.000), HCC (AUC: 0.985; 95% CI: 0.960–1.000), and other primary malignancies (AUC: 0.963; 95% CI: 0.896–1.000). Compared with the pathological examination, the agreement was 91.9%, and the sensitivity and specificity values for almost every liver lesion category achieved the same accuracy as those of experienced radiologists.[26] A study by Stollmayer et al.[27] used deep learning with 2D and 3D networks to diagnose FNH, HCC, and liver metastases on hepatocyte-specific contrast-enhanced MRIs. In total, 216 MRIs from 69 patients were analyzed. Overall, the 2D model performed better, with AUCs of 0.990, 0.966, and 0.960, respectively, for the investigated liver lesions.[27] Wang et al.'s[28] CNN-based model differentiated various focal liver lesions as either benign or malignant. Then, detailed classification was performed depending on tumor types. A total of 557 images were separated into a training and a testing set, and the AUCs for the classifications were 0.969 and 0.919, respectively. Seven focal liver lesions—liver cyst, cavernous hemangioma, hepatic abscess, FNH, HCC, intrahepatic cholangiocarcinoma, and hepatic metastasis—were investigated in their research using seven MR sequences (T2 weighted, diffusion weighted, apparent diffusion coefficient, T1 weighted, late arterial phase, portal venous phase, and delayed phase), and the accuracy for performing the seven-way classification was 79.6%.[28] The present study focused on a specific lesion, and it is difficult to compare the results with those of other studies in which some of the lesions were grouped and some were focused on distinguishing benign and malignant lesions.

To the best of our knowledge, this is the first study focused solely on the presence of FNH of the liver, and the results indicate promising results for the future. The sensitivity and specificity of the AI model were 76.9% and 96.6%, respectively, which were lower than those of the radiology specialist. The AUC value was 0.833 (95% CI: 0.682–0.983) and the accuracy was 0.905 for the AI model for indicating the presence of FNH using six MR sequences. These AUC values seem to be lower than some values from previous studies, and a higher accuracy value was obtained than from some other investigations. However, this study cannot be compared exactly with the other studies mentioned above. Datasets from previous studies including various lesions cannot be compared with the dataset from this study since this research was solely focused on FNH lesions. The AI results were better than those of the radiology resident; however, they were lower than those of the radiology specialist. Nonetheless, it was remarkable that a good agreement level was indicated between the radiology specialist and the AI model according to the results of this study. This might be caused by the lack of diversity among the FNH lesions in the MRIs experienced by the AI model in the training session. The radiology specialist's years of experience cannot be compared with the AI's training image dataset, which only included 30 patients with FNH lesions. This gap between AI and the radiology specialist might be compensated for by introducing a larger number and greater variety of FNH lesions to the AI model.

Important factors such as feasibility, ethical concerns, precision, safety, and overall acceptability influence the application speed of auto-diagnosis systems in medicine. Collaboration between healthcare professionals and AI-based diagnostic systems remains a mandatory objective for succeeding in this difficult task, and AI can still not replace skilled diagnosticians.[29]

There are limitations to this study, which must be considered when interpreting its outcomes. First, the presence of FNH was determined with only six MR sequences on axial planes to standardize, simplify, and easily compare the interpretation results. It also helped the standardization of the segmentation process, which should have been performed meticulously as part of a long-lasting process. However, a standard interpretation of the liver MRIs needed all the sequences obtained during the imaging procedure. If the patient had one or more lesions consistent with FNH, they were accepted as FNH positive by each reviewer, and chi-square tests were performed using these results. The AI model used in the study indicated the results regarding the presence or absence of FNH as an outcome. To compare the results and calculate interobserver reliability accurately, the study was planned in this way. This methodological approach might be criticized in terms of its appropriateness for indicating the sensitivity and specificity values. A lesion-based model rather than patient-based evaluation results would provide more accurate outcomes. The detection of the lesion was based on signal properties and dynamic enhancement patterns, but the borders of the lesion were underestimated. A morphological approach using the border attributes would be a more realistic approach, similar to the routine radiological liver interpretations on MRIs. Having more patients with hepatic adenomas and HCC to evaluate the ability of AI in distinguishing FNH from other lesions would be better. Some of the patients in this study were used in the AI training process, and some were not suitable for the investigation because of motion artifacts or image distortions. Moreover, we could only share the results of patients confidently diagnosed either radiologically or histologically. Although a variety of lesions with different histological and imaging features were evaluated in this study, additional studies with larger sample sizes are needed to confirm the results of this investigation. Due to the extremely detailed and very long-lasting process of segmentation, the proximal branches of the hepatic and portal veins were mapped to introduce them to the AI model. It was expected that the more distant segments would be perceived by the AI model, as it was part of the program. To minimize the AI model's possible segmentation and interpretation errors, the more distant segments of the vessels might also be drawn manually.

In conclusion, the AI model provided remarkable sensitivity, specificity, PPV, and NPV results regarding the detection of FNH in this study. The potential of AI should not be underestimated since this current investigation indicated that AI achieved better results than a radiology resident. Through multidisciplinary studies based on the increasing interest of physicians and engineers, AI might become a crucial element in diagnostics and play a major role in the detection and characterization of liver lesions. Targeted studies focused on specific lesions may be combined in the same diagnostic tool, using the experience of all focal lesions of the liver to widen the spectrum of lesions recognized by AI.

Footnotes

Conflict of interest disclosure

Sonay Aydın, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

1. Pompili M, Ardito F, Brunetti E, et al. Benign liver lesions 2022: Guideline for clinical practice of Associazione Italiana Studio del Fegato (AISF), Società Italiana di Radiologia Medica e Interventistica (SIRM), Società Italiana di Chirurgia (SIC), Società Italiana di Ultrasonologia in Medicina e Biologia (SIUMB), Associazione Italiana di Chirurgia Epatobilio-Pancreatica (AICEP), Società Italiana Trapianti d'Organo (SITO), Società Italiana di Anatomia Patologica e Citologia Diagnostica (SIAPEC-IAP) - Part II - Solid lesions. *Dig Liver Dis.* 2022;54(12):1614-1622. [Crossref]
2. Basturk O, Farris AB, Adsay NV. Chapter 15 - Immunohistology of the pancreas, biliary tract, and liver, editor(s): David J. Dabbs. *Diagnostic Immunohistochemistry (Third Edition)*. 2011; p.541-592 [Crossref]
3. Ding Z, Lin K, Fu J, et al. An MR-based radiomics model for differentiation between hepatocellular carcinoma and focal nodular hyperplasia in non-cirrhotic liver. *World J Surg Oncol.* 2021;19(1):181. [Crossref]
4. LeGout JD, Bolan CW, Bowman AW, et al. Focal nodular hyperplasia and focal nodular hyperplasia-like lesions. *Radiographics*. 2022;42(4):1043-1061. [Crossref]
5. Kamel IR, Liapi E, Fishman EK. Focal nodular hyperplasia: lesion evaluation using 16-MDCT and 3D CT angiography. *AJR Am J Roentgenol*. 2006;186(6):1587-1596. [Crossref]
6. Giambelluca D, Taibbi A, Midiri M, Bartolotta TV. The "spoke wheel" sign in hepatic focal nodular hyperplasia. *Abdom Radiol (NY)*. 2019;44(3):1183-1184. [Crossref]
7. European Association for the Study of the Liver (EASL). EASL clinical practice guidelines on the management of benign liver tumours. *J Hepatol.* 2016;65(2):386-398. [Crossref]
8. Murakami T, Tsurusaki M. Hypervascular benign and malignant liver tumors that require differentiation from hepatocellular carcinoma: key points of imaging diagnosis. Liver Cancer. 2014;3(2):85-96. [Crossref]
9. Suh CH, Kim KW, Kim GY, Shin YM, Kim PN, Park SH. The diagnostic value of Gd-EOB-DTPA-MRI for the diagnosis of focal nodular hyperplasia: a systematic review and meta-analysis. *Eur Radiol*. 2015;25(4):950-960. [Crossref]
10. Berbís MA, Paulano Godino F, Royuela Del Val J, Alcalá Mata L, Luna A. Clinical impact of artificial intelligence-based solutions on imaging of the pancreas and liver. *World J Gastroenterol*. 2023;29(9):1427-1445. [Crossref]
11. Hussain SM, Terkivatan T, Zondervan PE, Lanjouw E, de Rave S, Ijzermans JN, de Man RA. Focal nodular hyperplasia: findings at state-of-the-art MR imaging, US, CT, and pathologic analysis. *Radiographics.* 2004;24(1):3-17;discussion 18-9. [Crossref]
12. Antonelli M, Reinke A, Bakas S, et al. The medical segmentation decathlon. *Nat Commun.* 2022;13(1):4128. [Crossref]
13. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2021;18(2):203-211. [Crossref]
14. Bertels J, Eelbode T, Berman M, et al. Optimizing the Dice Score and Jaccard index for medical image segmentation: theory and practice. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II. 2019. p. 92-100. [Crossref]
15. https://github.com/lassoan/SlicerElastix#slicerelastix [Crossref]
16. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155-163. Erratum in: *J Chiropr Med.* 2017;16(4):346. [Crossref]
17. Hanna EJ, Ismail N, Arsalane A, Quenum C, Moumen A, Sacrieru D, Kabbej M, Khadra J. A case of liver rupture in a patient with focal nodular hyperplasia at 33 weeks of gestation: a multidisciplinary management. *Gynecology and Obstetrics Clinical Medicine.* 2022; 2: 46-48. [Crossref]
18. Tsilimigras DI, Rahnemai-Azar AA, Ntanasis-Stathopoulos I, et al. Current approaches in the management of hepatic adenomas. *J Gastrointest Surg*. 2019;23(1):199-209. Erratum in: *J Gastrointest Surg*. 2020;24(1):232. [Crossref]

19. Schütte K, Schulz C, Poranzke J, et al. Characterization and prognosis of patients with hepatocellular carcinoma (HCC) in the non-cirrhotic liver. *BMC Gastroenterol*. 2014;14:117. [Crossref]

20. Jansen MJA, Kuijf HJ, Veldhuis WB, Wessels FJ, Viergever MA, Pluim JPW. Automatic classification of focal liver lesions based on MRI and risk factors. *PLoS One*. 2019;14(5):e0217053. [Crossref]

21. Pettit RW, Marlatt BB, Corr SJ, Havelka J, Rana A. nnU-Net deep learning method for segmenting parenchyma and determining liver volume from computed tomography images. *Ann Surg Open*. 2022;3(2):e155. [Crossref]

22. Duc VT, Chien PC, Huyen LDM, et al. Deep learning model with convolutional neural network for detecting and segmenting hepatocellular carcinoma in CT: a preliminary study. *Cureus*. 2022;14(1):e21347. [Crossref]

23. Goehler A, Harry Hsu TM, Lacson R, et al. Three-dimensional neural network to automatically assess liver tumor burden change on consecutive liver MRIs. *J Am Coll Radiol*. 2020;17(11):1475-1484. [Crossref]

24. Zhang X, Kanematsu M, Fujita H, et al. Application of an artificial neural network to the computer-aided differentiation of focal liver disease in MR imaging. *Radiol Phys Technol*. 2009;2(2):175-182. [Crossref]

25. Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. *Eur Radiol*. 2019;29(7):3338-3347. [Crossref]

26. Zhen SH, Cheng M, Tao YB, et al. Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data. *Front Oncol*. 2020;10:680. [Crossref]

27. Stollmayer R, Budai BK, Tóth A, et al. Diagnosis of focal liver lesions with deep learning-based multi-channel analysis of hepatocyte-specific contrast-enhanced magnetic resonance imaging. *World J Gastroenterol*. 2021;27(35):5978-5988. [Crossref]

28. Wang SH, Han XJ, Du J, et al. Saliency-based 3D convolutional neural network for categorising common focal liver lesions on multisequence MRI. *Insights Imaging*. 2021;12(1):173. [Crossref]

29. Popa SL, Grad S, Chiarioni G, et al. Applications of artificial intelligence in the automatic diagnosis of focal liver lesions: a systematic review. *J Gastrointestin Liver Dis*. 2023;32(1):77-85. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.242880

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Diffusion kurtosis versus diffusion-weighted magnetic resonance imaging in differentiating clear cell renal cell carcinoma and renal angiomyolipoma with minimal fat: a comparative study

Yarong Lin
Wenrong Zhu
Qingqiang Zhu

Department of Medical Imaging, Northern Jiangsu People's Hospital Affiliated to Yangzhou University, Yangzhou, China

PURPOSE
To quantitatively compare the diagnostic values of conventional diffusion-weighted imaging and diffusion kurtosis imaging (DKI) in differentiating clear cell renal cell carcinoma (ccRCC) and renal angiomyolipoma with minimal fat (RAMF).

METHODS
Sixty-eight patients with ccRCC and 18 patients with RAMF were retrospectively studied. For DKI and apparent diffusion coefficient (ADC), respiratory-triggered echo-planar imaging sequences were acquired in the axial plane (three *b*-values: 0, 1000, 2000 s/mm^2; one *b*-value: 2000 s/mm^2). Mean diffusivity (MD), fractional anisotropy (FA), mean kurtosis (MK), kurtosis anisotropy (KA), radial kurtosis (RK), and ADC were evaluated. The diagnostic efficacy of various diffusion parameters in predicting ccRCC and RAMF was compared.

RESULTS
The ADC and MD values of ccRCCs were higher than those of RAMFs ($P < 0.05$), whereas comparable FA, MK, and KA values were observed between ccRCCs and RAMFs ($P > 0.05$). Moreover, the RK values of RAMFs were higher than those of ccRCCs ($P < 0.05$). Receiver operating characteristic (ROC) curve analyses showed that MD values had the highest diagnostic efficacy in differentiating ccRCCs from RAMFs. In pairwise comparisons of ROC curves and diagnostic efficacy, DKI parameters demonstrated better diagnostic accuracy than ADC in differentiating between ccRCCs and RAMFs ($P < 0.05$).

CONCLUSION
DKI analysis demonstrates superior performance than ADC analysis in differentiating ccRCC and RAMF.

CLINICAL SIGNIFICANCE
DKI technology may serve as an additional non-invasive biomarker for the differential diagnosis of renal tumor types.

KEYWORDS
Kidney, renal cell carcinoma, diffusion kurtosis imaging, angiomyolipoma, differential diagnosis.

Corresponding author: Qingqiang Zhu

E-mail: zhuqingqiang1983@163.com

Received 04 June 2024; revision requested 21 July 2024; accepted 06 April 2025.

Epub: 21.07.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.242880

Clear cell renal cell carcinoma (ccRCC) is the predominant subtype of RCC, comprising approximately 70% of all RCC cases.[1] Angiomyolipomas that are predominantly composed of smooth muscle cells, those with a mixture of all three components (smooth muscle, fat, and blood vessels), or those exhibiting prominent cystic changes may be challenging to differentiate from epithelial neoplasms preoperatively.[2]

Renal angiomyolipoma with minimal fat (RAMF) is generally considered a benign lesion. In contrast, ccRCC is a malignant tumor with the potential for metastasis and life-threatening

You may cite this article as: Lin Y, Zhu W, Zhu Q. Diffusion kurtosis versus diffusion-weighted magnetic resonance imaging in differentiating clear cell renal cell carcinoma and renal angiomyolipoma with minimal fat: a comparative study. *Diagn Interv Radiol.* 2025;31(5):416-422.

consequences. The management strategies for RAMF and ccRCC may also differ substantially. For instance, RAMF, being benign, often allows for a biopsy followed by regular surveillance. However, ccRCC, given its malignant nature, typically necessitates surgical resection.

Advancements in imaging technology have substantially transformed the management of renal masses by enabling the detection and characterization of even very small lesions. However, conventional computed tomography (CT) and magnetic resonance imaging (MRI) still face limitations in distinguishing atypical malignant from benign lesions. Therefore, identifying a simple yet accurate method to differentiate renal carcinomas from benign lesions remains the critical objective of this study.

Apparent diffusion coefficient (ADC) assessment has also shown benefits in distinguishing renal tumor types. One meta-analysis of 17 studies demonstrated that ADC values can help differentiate benign from malignant RCC tumors.[3] However, there is ongoing concern that ADCs obtained from conventional monoexponential diffusion-weighted imaging (DWI) may not accurately reflect true diffusivity because of the influence of microcirculation.[4,5]

The diffusion kurtosis imaging (DKI) model, first described in 2005, is believed to provide a more complete mathematical representation of tissue microstructural complexity than the standard monoexponential model.[6-8] It attempts to account for diffusion variation and capture non-Gaussian diffusion behavior as a reflective marker of tissue heterogeneity.[9] The aim of the current study was to produce a quantitative comparison of the potential of various diffusion parameters obtained from DWI and DKI in differentiating ccRCC and RAMF.

Main points

- Diffusion kurtosis imaging parameters demonstrated better diagnostic accuracy than apparent diffusion coefficient (ADC) in differentiating between clear cell renal cell carcinomas (ccRCCs) and renal angiomyolipomas with minimal fat (RAMFs).
- The ADC and mean diffusivity values of ccRCCs were higher than those of RAMFs.
- The radial kurtosis values of RAMFs were higher than those of ccRCCs.

Methods

Participants

This retrospective study was approved by the institutional review committee of Northern Jiangsu People's Hospital Affiliated with Yangzhou University (protocol number: 20130701, date: 7/1/2013 to 9/1/2022), and the requirement for written informed consent was waived. The study covered the period from July 1, 2013, to September 1, 2022. A total of 117 adult patients who underwent routine MRI examinations and DKI assessment followed by partial or radical nephrectomy between July 2013 and September 2022 were retrospectively enrolled (Figure 1).

The exclusion criteria were as follows: (a) lesions without histopathological confirmation of ccRCC or RAMF (n = 13); (b) lesions requiring antiangiogenic therapy (n = 6); (c) tumor recurrence (n = 7); (d) a low signal-to-noise ratio (SNR) (n = 5; SNR <7.2 for b = 2000 s/mm^2). This retrospective study was approved by our institutional review board, with a waiver of the requirement for written informed consent.

Magnetic resonance imaging technique

MRI examinations were performed using a 3.0-T MR scanner (GE Signa EXCITE HD, Milwaukee, WI, USA) with a 6-channel array body coil and a 24-channel phased-array spine coil integrated into the scanner table. For DKI, a single-shot echo-planar imaging (EPI) sequence was applied in the axial plane using respiratory triggering via a respiratory belt, with three *b*-values (0, 1000, 2000 s/mm^2) and 30 diffusion directions. For ADC, respiratory-triggered EPI sequences were acquired in the axial plane (one *b*-value: 2000 s/mm^2). Other imaging parameters were as follows: 24 axial slices covering both kidneys; echo time: 73.9 ms; repetition time: 5000 ms; number of excitations: 4; matrix: 192 × 192; field of view: 400 mm. Array spatial sensitivity encoding technique, a parallel imaging method, was applied with an acceleration factor of 4.

Imaging analysis and statistics

The acquired images were transferred to an offline workstation for processing using Automated Image Registration software, version 4.6.4. (GE Signa EXCİTE HD, GE Healthcare, Milwaukee, WI, USA). Prior to the quantification of DKI and ADC, non-rigid co-registration and smoothing were performed using a 3 × 3 kernel matrix. All DWIs were first co-registered to the b0 image using the affine model. Then, registered DWIs with *b*-values of 1000 and 2000 s/mm^2 and ADCs with a *b*-value of 2000 s/mm^2 were averaged over 30 diffusion-encoding directions.

Afterward, the two averaged DWIs were co-registered to the b0 image using the affine model, and the registered averaged DWIs were set as a reference volume for further registrations. Finally, the initial DWIs with a *b*-value of 2000 s/mm^2 were co-registered to the corresponding reference volume using a non-rigid model. The registered DWIs were then spatially smoothed using a Gaussian filter with a full width at half-maximum of 2 mm. With our DKI and ADC protocol, we obtained parametric maps related to diffusional kurtosis: mean diffusivity (MD), fractional anisotropy (FA), mean kurtosis (MK), kurtosis anisotropy (KA), radial kurtosis (RK), and ADC. The assessment of renal tumors and region-of-interest (ROI) positioning was

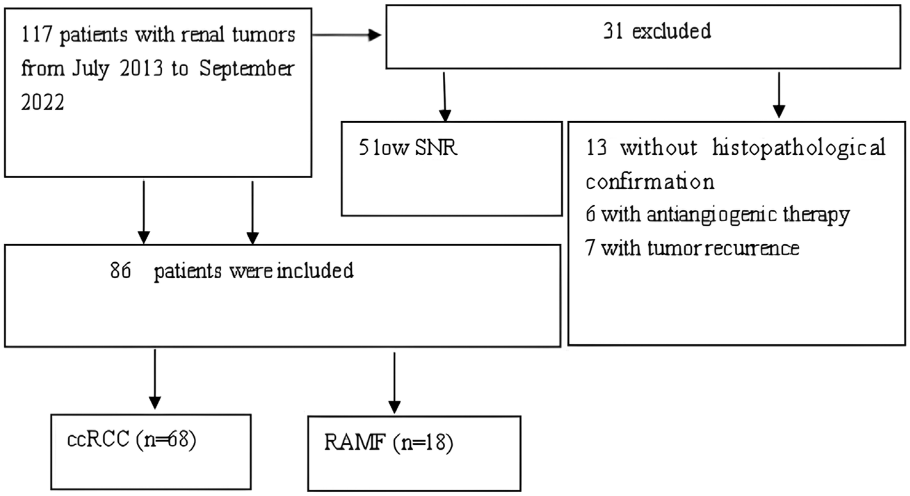

117 patients with renal tumors from July 2013 to September 2022
31 excluded
5 low SNR
13 without histopathological confirmation
6 with antiangiogenic therapy
7 with tumor recurrence
86 patients were included
ccRCC (n=68)
RAMF (n=18)

Figure 1. Patient inclusion and exclusion flowchart. SNR, signal-to-noise ratio; ccRCC, clear cell renal cell carcinoma; RAMF, renal angiomyolipoma with minimal fat.

conducted by two radiologists with 5 and 10 years of clinical experience in interpreting MRI, respectively. Both observers were blinded to the patients' clinical information and tumor histology. Lesion location, the number of layers on which the tumor appeared most prominent across different sequences, imaging characteristics of the renal tumors, and the ROI plotting method were considered.

The two observers, each with 5 and 10 years of diagnostic experience, analyzed all the parameter maps in conjunction with the DKI and ADC images. They were blinded to the pathologic diagnosis and reached a consensus on their analysis.

Free-hand ROIs were delineated around the most solid portion of each tumor (covering approximately two-thirds of the solid area) on the DKI and ADC maps. This was performed on three to five representative slices by the same two radiologists using ImageJ software (National Institutes of Health, Bethesda, MD, USA). The region with lower T2 signal intensity was identified as the most solid part in heterogeneous tumors. Strong hyperintensity on T2WI or T1WI indicated tissue necrosis or hemorrhage, and such regions were excluded. Mean values for ADC, MD, FA, MK, KA, and RK for each ROI were calculated using ImageJ software. The readers independently assessed images derived from the DKI and ADC examinations during two separate sessions, with an interval of more than four weeks between sessions to mitigate potential recall bias.

Statistical analysis

Statistical analysis was conducted using SPSS version 23.0 statistical software (SPSS, Chicago, IL, USA). Numeric data were expressed as means and standard deviations (±), and categorical data were expressed as percentages. Evaluated DKI and ADC features were compared between ccRCC and RAMF using the independent-sample t-test. A P value <0.05 was considered statistically significant.

To assess the diagnostic performance of DKI and ADC parameters in differentiating ccRCC from RAMF, we calculated the diagnostic accuracy for both tumor types. The highest Youden index value was used to determine the optimal diagnostic point, and the DeLong method[10] was applied to compare area under the curves. Intraclass correlation coefficients (ICCs) were used to assess interobserver agreement for ADC and DKI parameter measurements, with 95% confidence intervals (CIs). ICCs were interpreted as follows: ≤ 0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, perfect agreement.

The comparison of ICCs between observers with 5 and 10 years of experience was performed using a self-lifting resampling technique with 200 repetitions. This method was employed to estimate the mean ICC and 95% CI for each observer group. Retest reliability was calculated for individual observers as well as for the entire group, and comparisons were made using the Z-test for ICC. A P value <0.05 was considered statistically significant.

Results

Population demographics

A total of 86 patients with pathologically confirmed ccRCC and RAMF were included, comprising 68 patients (38 men and 30 women) with ccRCC and 18 patients (12 men and 6 women) with RAMF. The mean age at diagnosis was slightly lower in patients with RAMF (49.8 years; range 39 to 62 years) than in those with ccRCC (52.1 years; range 36 to 76 years). There was no difference in clinical manifestations between ccRCC and RAMF, such as mean age, sex, flank pain, palpable mass, and fever (all $P > 0.05$), except for hematuria (73 vs. 2, $P < 0.01$).

Apparent diffusion coefficient and diffusion kurtosis imaging parameters of the renal tumors

The ADC (Figure 2, Table 1) and MD (Figure 3, Table 1) values of ccRCCs were higher than those of RAMFs ($P < 0.05$). The RK (Figure 4) values of RAMFs were higher than those of ccRCCs (Figure 5, $P < 0.05$), whereas comparable FA, MK, and KA values were found between ccRCCs and RAMFs (Figure 6, Table 1; $P > 0.05$).

Diagnostic performance of multiple parameters

Receiver operating characteristic (ROC) curve analyses showed that MD (Figure 7, Table 2) and RK (Figure 8, Table 2) values had higher diagnostic efficacy than ADC values in differentiating ccRCCs from RAMFs. MD values demonstrated the highest diagnostic efficacy. For pairwise comparisons of ROC curves and diagnostic performance, ADC was inferior to MD and KA ($P < 0.05$).

The agreement of diffusion parameters in the 86 cases, both for individual observers and overall, was perfect for all parameters (ADC, MD, FA, MK, KA, and RK). Retest reliability, assessed by an independent repeat evaluation by two observers with 5 and 10 years of experience, was shown to be excellent (Table 3). In addition, there was no statistically significant difference in retest reliability between the two observers (Table 4).

Discussion

The ADC and MD values of ccRCCs were higher than those of RAMFs ($P < 0.05$), whereas comparable FA, MK, and KA values were found between ccRCCs and RAMFs ($P > 0.05$). Moreover, the RK values of RAMFs were

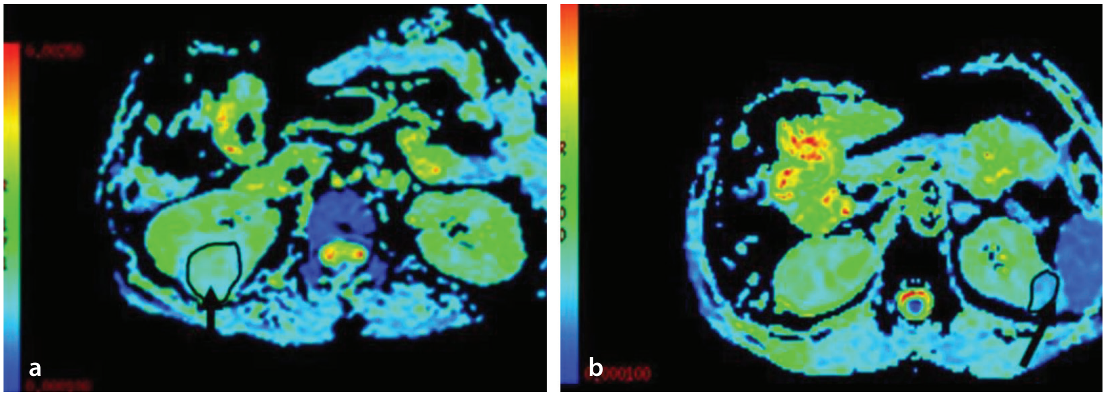


Figure 2. Apparent diffusion coefficient (ADC) features of clear cell renal cell carcinoma (ccRCC) **(a)** and renal angiomyolipoma with minimal fat (RAMF) **(b)**; ADC values were higher for ccRCC (0.89) and lower for RAMF (0.53).

Table 1. Diffusion kurtosis imaging and apparent diffusion coefficient parameters in clear cell renal cell carcinoma and renal angiomyolipoma with minimal fat

Parameters	ccRCC	RAMF	*P* values
ADC	0.81 ± 0.11	0.55 ± 0.18	<0.05
MD	2.13 ± 0.42	1.21 ± 0.26	<0.01
FA	0.17 ± 0.05	0.18 ± 0.04	>0.05
MK	0.92 ± 0.21	0.87 ± 0.16	>0.05
KA	0.99 ± 0.23	0.88 ± 0.19	>0.05
RK	0.66 ± 0.08	0.91 ± 0.24	<0.05

ccRCC, clear cell renal cell carcinoma; RAMF, renal angiomyolipoma with minimal fat; ADC, apparent diffusion coefficient; MD, mean diffusivity; FA, fractional anisotropy; MK, mean kurtosis; KA, kurtosis anisotropy; RK, radial kurtosis.

Table 2. Diagnostic test characteristics of diffusion parameters in differentiating clear cell renal cell carcinoma from renal angiomyolipoma with minimal fat

Parameters	AUC (95% CI)	Cut-off value	Sensitivity	Specificity	Accuracy
ccRCC (n = 68) vs. RAMF (n = 18)					
ADC ($\times 10^{-3}$ mm^2/s)	0.810 (0.821–0.933)	≥0.72	79.4% (54/68)	66.7% (12/18)	79.1% (68/86)
MD	0.943 (0.889–0.991)	≥1.83	94.1% (64/68)	83.3% (15/18)	94.2% (81/86)
RK	0.863 (0.808–0.921)	≤0.68	86.7% (59/68)	77.8% (14/18)	84.9% (73/86)

AUC, area under the curve; ccRCC, clear cell renal cell carcinoma; RAMF, renal angiomyolipoma with minimal fat; ADC, apparent diffusion coefficient; MD, mean diffusivity; RK, radial kurtosis.

Table 3. Intraclass correlation coefficients for measurements of apparent diffusion coefficient, mean diffusivity, fractional anisotropy, mean kurtosis, kurtosis anisotropy, and radial kurtosis by total observers

Parameters	Observer ICC
ADC	0.931 (0.911–0.952)
MD	0.951 (0.929–0.991)
FA	0.893 (0.871–0.911)
MK	0.916 (0.911–0.949)
KA	0.926 (0.901–0.963)
RK	0.911 (0.901–0.936)

ADC, apparent diffusion coefficient; MD, mean diffusivity; FA, fractional anisotropy; MK, mean kurtosis; KA, kurtosis anisotropy; RK, radial kurtosis; ICC, Intraclass correlation coefficient.

Table 4. Intraclass correlation coefficients for measurements of apparent diffusion coefficient, mean diffusivity, fractional anisotropy, mean kurtosis, kurtosis anisotropy, and radial kurtosis by individual observers

Parameters	Individual observer (first vs. second)	*P* values
ADC	0.903 (0.881–0.923) vs. 0.933 (0.907–0.968)	>0.05
MD	0.933 (0.907–0.963) vs. 0.947 (0.913–0.963)	>0.05
FA	0.886 (0.863–0.902) vs. 0.906 (0.882–0.922)	>0.05
MK	0.893 (0.886–0.921) vs. 0.921 (0.912–0.952)	>0.05
KA	0.907 (0.886–0.938) vs. 0.931 (0.912–0.966)	>0.05
RK	0.901 (0.883–0.926) vs. 0.923 (0.911–0.946)	>0.05

ADC, apparent diffusion coefficient; MD, mean diffusivity; FA, fractional anisotropy; MK, mean kurtosis; KA, kurtosis anisotropy; RK, radial kurtosis.

higher than those of ccRCCs ($P < 0.05$). ROC curve analyses showed that MD values had the highest diagnostic efficacy in differentiating ccRCCs from RAMFs. For pairwise comparisons of ROC curves and diagnostic efficacy, ADC was inferior to DKI analysis ($P < 0.05$).

DKI is a dimensionless measure that quantifies the deviation of the water diffusion displacement profile from the Gaussian distribution of unrestricted diffusion, providing a measure of the degree of diffusion hindrance or restriction.[11] It has been shown to offer superior sensitivity over conventional DTI.[12] An appealing aspect of incorporating DKI into routine clinical practice is that it can be performed in a straightforward manner, as the sequence is performed in essentially the same manner as a standard DWI sequence,[13] aside from the generally higher *b*-values required.

In a recent study, Lanzman et al.[14] highlighted the potential of DTI for non-invasive functional assessment of transplanted kidneys. They also demonstrated significant differences in FA values of the medulla between

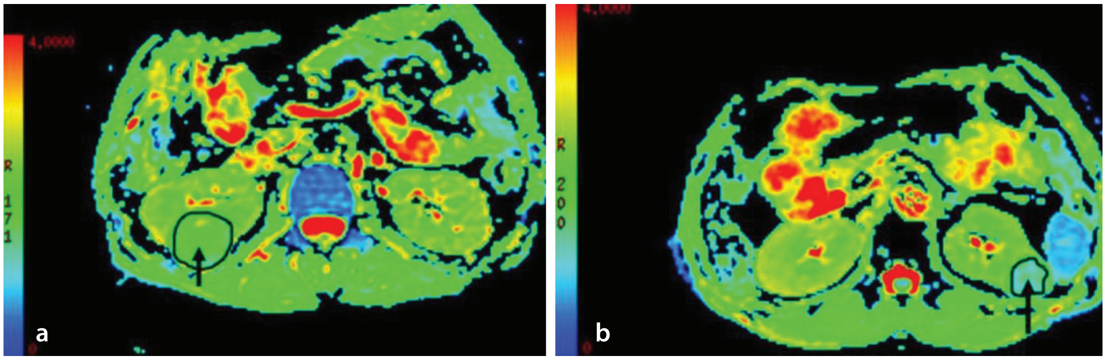


Figure 3. Mean diffusivity (MD) features of clear cell renal cell carcinoma (ccRCC) **(a)** and renal angiomyolipoma with minimal fat (RAMF) **(b)**; MD values were higher for ccRCC (2.17) and lower for RAMF (1.42).

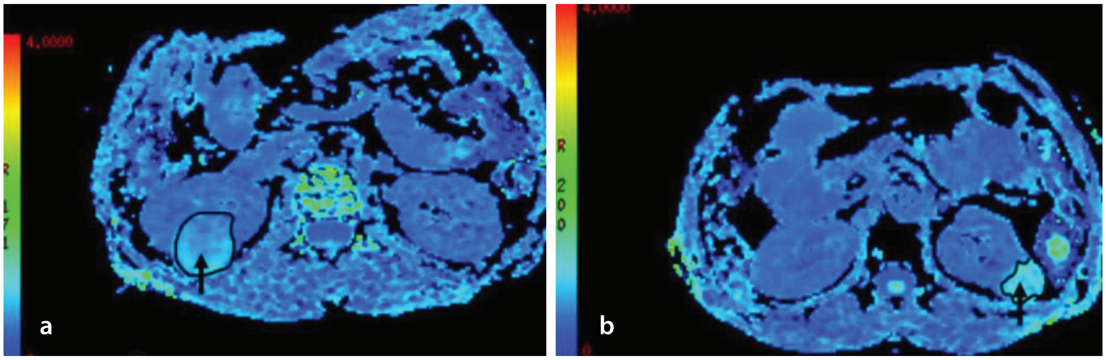


Figure 4. Radial kurtosis (RK) features of clear cell renal cell carcinoma (ccRCC) **(a)** and renal angiomyolipoma with minimal fat (RAMF) **(b)**; RK values were lower for ccRCC (0.71) and higher for RAMF (0.97).

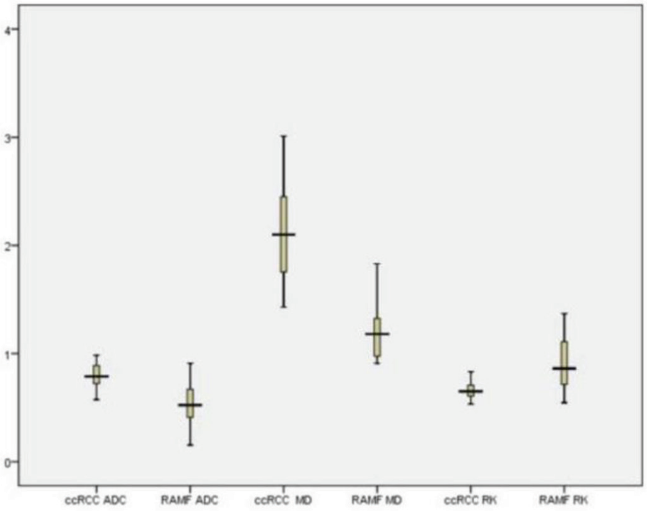


Figure 5. Box-and-whisker plots showing the distribution of apparent diffusion coefficient, mean diffusivity, and radial kurtosis parameters, with significant differences between clear cell renal cell carcinoma and renal angiomyolipoma with minimal fat. ccRCC, clear cell renal cell carcinoma; RAMF, renal angiomyolipoma with minimal fat; ADC, apparent diffusion coefficient; MD, mean diffusivity; RK, radial kurtosis.

allograft recipients with severely impaired renal function and those with moderate or mild impairment. Comparing MK values of normal kidneys with those of patients with various renal diseases may help evaluate the clinical significance of renal kurtosis values and the role of renal DKI.[15] For instance, in RCC, DKI may provide additional diagnostic information. Since DKI has been proven to be more sensitive to tissue microstructure than FA measures, DKI of the kidney might be useful in evaluating conditions involving renal tumors.[16] Notohamiprodjo et al.[17] reported that higher *b*-values and a greater number of directions improve the accuracy of diffusion measurements. In our study, we demonstrated that *b*-values in the range of 0 to 2000 s/mm^2 with 30 diffusion-encoding directions are sufficient in abdominal DKI to observe the departure of the diffusion signal from monoexponential behavior.

In our study, statistically significant differences were observed in the MD and ADC values between ccRCC and RAMF. Many authors attribute higher MD and ADC to higher cellularity. Tissue-free water content and structural differences can influence MD and ADC. Increases in MD and ADC due to micronecrosis or altered viscosity of the medium may counterbalance decreased MD and ADC values in ccRCC.[18] ccRCC is rich in lipid content; cholesterol, neutral lipids, and phospholipids are abundant in pathology.[19]

An increase in the number of cells or a decrease in cell volume leads to an increase in the diffusion limitation of water molecules, which results in an increase in RK.[20] Necrotic areas within the tumor and surrounding edema change the diffusion characteristics, usually with lower RK values in the necrotic areas and higher RK values in the edematous areas. As illustrated in our study, RAMF showed greater RK values than ccRCC, with a significant difference consistent with the understanding that RAMF has greater viscosity and restricted water diffusion due to the presence of hemorrhagic walls or hemosiderin deposition.

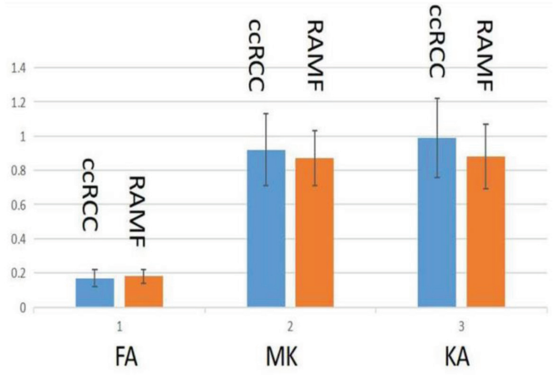


Figure 6. Bar graph showing the distribution of fractional anisotropy, mean kurtosis, and kurtosis anisotropy parameters, without significant differences between clear cell renal cell carcinoma and renal angiomyolipoma with minimal fat. ccRCC, clear cell renal cell carcinoma; RAMF, renal angiomyolipoma with minimal fat; FA, fractional anisotropy; MK, mean kurtosis; KA, kurtosis anisotropy.

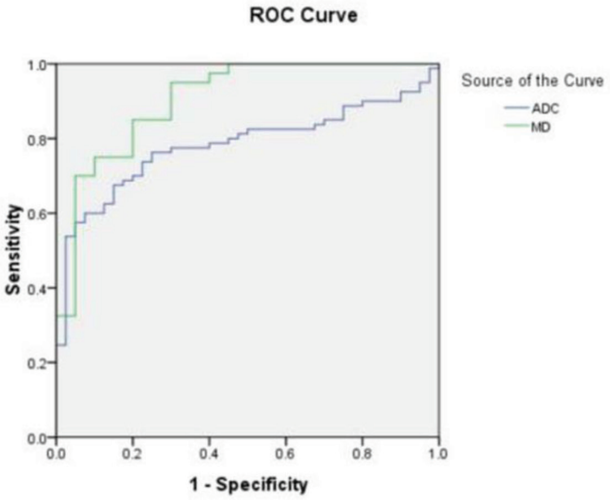


Figure 7. Receiver operating characteristic curves showing the diagnostic performance of apparent diffusion coefficient and mean diffusivity parameters in differentiating clear cell renal cell carcinoma from renal angiomyolipoma with minimal fat. ROC, receiver operating characteristic; ADC, apparent diffusion coefficient; MD, mean diffusivity.

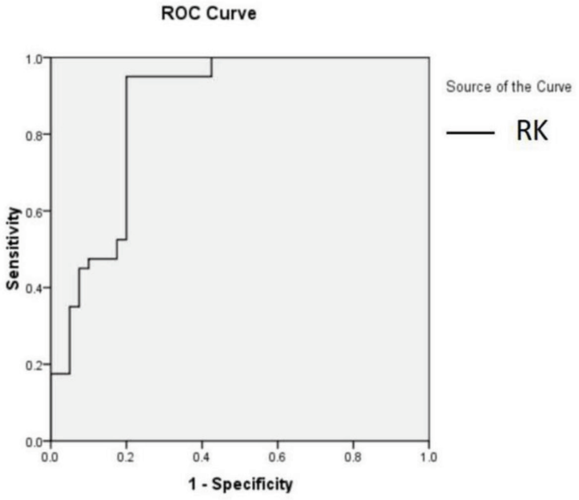


Figure 8. Receiver operating characteristic curves showing the diagnostic performance of radial kurtosis in differentiating clear cell renal cell carcinoma from renal angiomyolipoma with minimal fat. ROC, receiver operating characteristic; RK, radial kurtosis.

The MD and RK parameters showed greater discrimination of renal tumor types than the ADC parameters, perhaps because the latter includes both microcirculation and tissue cellularity information.[21] These two sources of information may affect the ADC measurement in opposing ways, decreasing sensitivity and specificity.[22] Moreover, the additional MD and RK parameters provide specific information on non-Gaussian diffusion behavior, offering a more accurate measurement of tissue diffusion.[23]

Retest reliability was evaluated through an independent repeat assessment conducted by two observers with 5 and 10 years of experience, respectively. The results demonstrated excellent retest reliability. Furthermore, no statistically significant difference in retest reliability was observed between the two observers. This finding suggests that the stability of DKI in evaluating microstructural differences in ccRCC and RAMF is not influenced by the observers' level of experience. Such consistency is highly conducive to the clinical adoption and broader application of DKI technology.

The main limitation of our study is the small number of patients in each renal tumor type, especially RAMF. Further studies with larger populations are recommended to validate our findings. We acknowledge additional limitations in the current study. As a single-center, retrospective analysis, the findings may be influenced by the specific characteristics of the sample population and the inherent biases associated with retrospective data collection. Therefore, the reliability of our results should be confirmed through well-designed prospective studies and multicenter investigations.

Notably, our study did not include comparisons with other subtypes of RCC or with renal oncocytomas. As a result, it would be overly speculative to extrapolate our findings to differentiate renal oncocytomas from other types of renal tumors. However, papillary and chromophobe RCCs, as well as renal oncocytomas, are generally less likely to be confused with ccRCC or RAMF on CT and/or MRI. ccRCC and RAMF typically exhibit hypervascularity and heterogeneous enhancement, whereas papillary and chromophobe RCCs are characterized by hypovascularity. In contrast, renal oncocytomas are often identified by a central stellate scar, homogeneous enhancement, and a spoke-wheel pattern of enhancement, which are considered characteristic features.

In conclusion, DKI parameters demonstrated better performance than ADC in differentiating ccRCC and RAMF. This new technique can potentially be used as another non-invasive biomarker for the differential diagnosis of renal tumor types.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Jomoto W, Takaki H, Yamamoto S, et al. Differentiation of angiomyolipoma with minimal fat from clear cell renal cell carcinoma using non-contrast multiparametric magnetic resonance imaging. *In Vivo.* 2022;36(6):2790-2799. [Crossref]
2. Li H, Li A, Zhu H, et al. Whole-tumor quantitative apparent diffusion coefficient histogram and texture analysis to differentiation of minimal fat angiomyolipoma from clear cell renal cell carcinoma. *Acad Radiol.* 2019;26(5):632-639. [Crossref]
3. Lassel EA, Rao R, Schwenke C, Schoenberg SO, Michaely HJ. Diffusion-weighted imaging of focal renal lesions: a meta-analysis. *Eur Radiol*. 2014;24(1):241-249. [Crossref]
4. Cao J, Luo X, Zhou Z, et al. Comparison of diffusion-weighted imaging mono-exponential mode with diffusion kurtosis imaging for predicting pathological grades of clear cell renal cell carcinoma. *Eur J Radiol*. 2020;130:109195. [Crossref]
5. Parada Villavicencio C, Mc Carthy RJ, Miller FH. Can diffusion-weighted magnetic resonance imaging of clear cell renal carcinoma predict low from high nuclear grade tumors. *Abdom Radiol (NY)*. 2017;42(4):1241-1249. [Crossref]
6. Zhang J, Suo S, Liu G, et al. Comparison of monoexponential, biexponential, stretched-exponential, and kurtosis models of diffusion-weighted imaging in differentiation of renal solid masses. *Korean J Radiol*. 2019;20(5):791-800. [Crossref]
7. Fu J, Ye J, Zhu W, Wu J, Chen W, Zhu Q. Magnetic resonance diffusion kurtosis imaging in differential diagnosis of benign and malignant renal tumors. *Cancer Imaging*. 2021;21(1):6. [Crossref]
8. Cheng ZY, Feng YZ, Liu XL, Ye YJ, Hu JJ, Cai XR. Diffusional kurtosis imaging of kidneys in patients with hyperuricemia: initial study. *Acta Radiol*. 2020;61(6):839-847. [Crossref]
9. Novello L, Henriques RN, Ianuş A, Feiweier T, Shemesh N, Jovicich J. In vivo correlation tensor MRI reveals microscopic kurtosis in the human brain on a clinical 3T scanner. *Neuroimage*. 2022;254:119137. [Crossref]
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. [Crossref]
11. Wu G, Zhao Z, Yao Q, et al. The study of clear cell renal cell carcinoma with MR diffusion kurtosis tensor imaging and its histopathologic correlation. *Acad Radiol*. 2018;25(4):430-438. [Crossref].
12. DiBella EVR, Sharma A, Richards L, Prabhakaran V, Majersik JJ, HashemizadehKolowri SK. Beyond diffusion tensor MRI methods for improved characterization of the brain after ischemic stroke: a review. *AJNR Am J Neuroradiol*. 2022;43(5):661-669. [Crossref]
13. Zhu J, Luo X, Gao J, Li S, Li C, Chen M. Application of diffusion kurtosis tensor MR imaging in characterization of renal cell carcinomas with different pathological types and grades. *Cancer Imaging*. 2021;21(1):30. [Crossref]
14. Lanzman RS, Ljimani A, Pentang G, et al. Kidney transplant: functional assessment with diffusion-tensor MR imaging at 3T. *Radiology*. 2013;266(1):218-225. [Crossref]
15. Dai Y, Yao Q, Wu G, et al. Characterization of clear cell renal cell carcinoma with diffusion kurtosis imaging: correlation between diffusion kurtosis parameters and tumor cellularity. *NMR Biomed*. 2016;29(7):873-881. [Crossref]
16. Ding Y, Tan Q, Mao W, et al. Differentiating between malignant and benign renal tumors: do IVIM and diffusion kurtosis imaging perform better than DWI? *Eur Radiol*. 2019;29(12):6930-6939. [Crossref]
17. Notohamiprodjo M, Dietrich O, Horger W, et al. Diffusion tensor imaging (DTI) of the kidney at 3 tesla-feasibility, protocol evaluation and comparison to 1.5 Tesla. *Invest Radiol*. 2010;45(5):245-254. [Crossref]
18. Minervini A, Di Cristofano C, Gacci M, et al. Prognostic role of histological necrosis for nonmetastatic clear cell renal cell carcinoma: correlation with pathological features and molecular markers. *J Urol*. 2008;180(4):1284-1289. [Crossref]
19. Koh MJ, Lim BJ, Choi KH, Kim YH, Jeong HJ. Renal histologic parameters influencing postoperative renal function in renal cell carcinoma patients. *Korean J Pathol*. 2013;47(6):557-562. [Crossref]
20. Huang Y, Chen X, Zhang Z, et al. MRI quantification of non-Gaussian water diffusion in normal human kidney: a diffusional kurtosis imaging study. *NMR Biomed*. 2015;28(2):154-161. [Crossref]
21. Zhang YD, Wu CJ, Wang Q, et al. Comparison of Utility of Histogram Apparent Diffusion Coefficient and R2* for differentiation of low-grade from high-grade clear cell renal cell carcinoma. *AJR Am J Roentgenol*. 2015;205(2):W193-201. [Crossref]
22. Reynolds HM, Parameswaran BK, Finnegan ME, et al. Diffusion weighted and dynamic contrast enhanced MRI as an imaging biomarker for stereotactic ablative body radiotherapy (SABR) of primary renal cell carcinoma. *PLoS One.* 2018;13(8):e0202387. [Crossref]
23. Ding Y, Zeng M, Rao S, Chen C, Fu C, Zhou J. Comparison of biexponential and monoexponential model of diffusion-weighted imaging for distinguishing between common renal cell carcinoma and fat poor angiomyolipoma. *Korean J Radiol*. 2016;17(6):853-863. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253291

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Diagnostic performance of magnetic resonance imaging for lateral pelvic lymph node metastasis in patients with rectal carcinoma: a meta-analysis and systematic review

Xiaolong Liu
Keping Liao
Peng Wang
Yongqiang Gao
Yongxin Du

The Affiliated Hospital of Kunming University of Science and Technology, The People's Hospital of Puer, Department of Medical Imaging, Puer, China

PURPOSE

Accurate identification of lateral pelvic lymph node (LPLN) metastasis is imperative for guiding LPLN dissection to reduce local recurrence in patients with rectal carcinoma. This meta-analysis aimed to investigate the diagnostic performance of magnetic resonance imaging (MRI) for LPLN metastasis in patients with rectal carcinoma.

METHODS

Embase, PubMed, Web of Science, and the Cochrane Library were searched to identify studies related to the diagnostic performance of MRI for LPLN metastasis in patients with rectal carcinoma through June 2024.

RESULTS

This meta-analysis included 12 studies comprising 1,015 patients. The pooled sensitivity [95% confidence interval (CI)] and specificity (95% CI) of MRI for diagnosing LPLN metastasis were 0.66 (0.53, 0.80) and 0.82 (0.76, 0.88), respectively. The pooled positive likelihood ratio (LR) (95% CI) and negative LR (95% CI) were 2.82 (2.14, 3.51) and 0.41 (0.27, 0.55), respectively. The summary receiver operating characteristic curve indicated an area under the curve of 0.824. The quality of the included studies was acceptable according to the Quality Assessment of Diagnostic Accuracy Studies-2 tool. However, publication bias was present, as indicated by Deeks' funnel plot asymmetry test (P = 0.020). Considering that heterogeneity contributed to publication bias, a meta-regression analysis was conducted and revealed that heterogeneity could be influenced by sample size, with sample size negatively associated with sensitivity (coefficient: -0.002, P = 0.009) and positively associated with negative LR (coefficient: 0.002, P = 0.029).

CONCLUSION

Preoperative MRI demonstrates an acceptable ability to identify LPLN metastasis in patients with rectal carcinoma.

CLINICAL SIGNIFICANCE

Clinically, our findings support that preoperative MRI has acceptable diagnostic ability for LPLN metastasis in patients with rectal carcinoma. The preoperative application of MRI may aid in optimizing treatment strategies and improving prognosis in this population.

KEYWORDS

Rectal carcinoma, lateral pelvic lymph node metastasis, magnetic resonance imaging, sensitivity, specificity

Corresponding author: Xiaolong Liu

E-mail: 23749802@qq.com

Received 15 February 2025; accepted 04 April 2025.

Epub: 21.07.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.253291

You may cite this article as: Liu X, Liao K, Wang P, Gao Y, Du Y. Diagnostic performance of magnetic resonance imaging for lateral pelvic lymph node metastasis in patients with rectal carcinoma: a meta-analysis and systematic review. *Diagn Interv Radiol.* 2025;31(5):423-429.

Lateral pelvic lymph node (LPLN) metastasis is considered one of the major causes of local recurrence in patients with rectal carcinoma.[1] In order to reduce local recurrence rates in patients with LPLN metastasis, LPLN dissection should be performed,[2-4] and accurate diagnosis of LPLN metastasis is imperative for guiding this operation.[5-8] Currently, imaging methods such as computed tomography (CT), endorectal ultrasound, and ^{18}F-fluorodeoxyglucose-positron emission tomography (FDG-PET) are used for diagnosing LPLN metastasis, yet each has limitations in sensitivity or specificity.[7,9] Therefore, investigating potential methods for diagnosing LPLN metastasis is essential to improve the management of patients with rectal carcinoma.

Magnetic resonance imaging (MRI), with its outstanding soft tissue contrast resolution, demonstrates good potential for diagnosing LPLN metastasis in patients with rectal carcinoma.[7] Several studies have explored the diagnostic performance of MRI for LPLN metastasis in these patients.[10-21] For instance, one previous study found that when the short-axis cut-off value was 5 mm, the accuracy, sensitivity, and specificity of MRI for diagnosing LPLN metastasis were 77.6%, 68.6%, and 79.7%, respectively; the area under the curve (AUC) was 0.74.[15] Another study applied a 6.8 mm cut-off for the short axis and reported that the sensitivity, specificity, and AUC were 77.8%, 72.1%, and 0.761, respectively.[20] To support the wider application of MRI in patients with rectal carcinoma suspected of LPLN metastasis, it is crucial to conduct a pooled analysis to evaluate the diagnostic performance of MRI for LPLN metastasis in this population. Accordingly, this meta-analysis aimed to provide a comprehensive evaluation of the diagnostic performance of MRI for LPLN metastasis in patients with rectal carcinoma.

Main points

- The ability of magnetic resonance imaging to diagnose lateral pelvic lymph node metastasis was evaluated.
- This meta-analysis included 12 studies with 1,015 patients with rectal carcinoma.
- The pooled sensitivity and specificity were 0.66 and 0.82, respectively.
- The pooled positive and negative likelihood ratios were 2.82 and 0.41, respectively.
- The pooled area under the curve of the summary receiver operating characteristic curve was 0.824.

Methods

The present study is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement and published recommendations. Ethics information and informed consent forms were not required, as systematic reviews typically involve synthesizing and summarizing existing literature rather than directly engaging in human or animal experiments.

Search scheme

Embase, PubMed, Web of Science, and the Cochrane Library were searched to identify studies related to the diagnosis of LPLN metastasis using MRI technology in patients with rectal carcinoma. The keywords used for the search were as follows: "magnetic resonance imaging," "MRI," "MR," "rectal cancer," "rectal carcinoma," and "lateral pelvic lymph node metastasis." The retrieval period was from database inception to June 2024. After excluding duplicate studies, titles and abstracts of the remaining studies were reviewed based on the eligibility criteria. Subsequently, full-text articles were assessed for study eligibility. KL, PW, YG, and YD independently completed this part of the work. In case of disagreement, a decision was made after consultation.

Criteria of the study screen

During the screening process, the inclusion criteria were as follows: i) patients were diagnosed with rectal carcinoma; ii) patients underwent MRI examination for the detection of LPLN metastasis; iii) studies contained complete 2 × 2 contingency tables [including true positive (TP), false positive (FP), false negative (FN), and true negative (TN)] or provided sufficient data to construct 2 × 2 contingency tables for assessing diagnostic efficacy; iv) studies were published in English. The exclusion criteria were as follows: i) case reports, animal experiments, reviews, or meta-analyses; ii) studies lacking or not using histopathological examination as the reference standard; iii) studies by the same authors with overlapping study populations.

Data collection

The first author's name, publication year, study design, sample size, age, gender, and MRI-related information were collected. In addition, 2 × 2 contingency tables were obtained. If the studies did not report direct data on 2 × 2 contingency tables, they were calculated using sensitivity, specificity, positive sample size (PSZ), and negative sample size (NSZ). The formulas used were as follows: $TP = \text{Sensitivity} \times PSZ$; $FN = PSZ - TP$; $TN = \text{Specificity} \times NSZ$; $FP = NSZ - TN$. Data collection was performed independently by KL, PW, YG, and YD. When results were inconsistent, they were resolved through joint discussion.

Statistical analysis

STATA statistical software (version 14.0; StataCorp, College Station, TX, USA) was used for data analyses. Pooled sensitivity, pooled specificity, pooled positive likelihood ratio (LR), and pooled negative LR, each with a 95% confidence interval (CI), were analyzed. Additionally, the summary receiver operating characteristic (SROC) curve was generated. Heterogeneity was assessed using the chi-square test and the I^2 test; $P < 0.05$ indicated significant heterogeneity for the former, and $I^2 \geq 50\%$ for the latter. Deeks' funnel plot was used to evaluate publication bias through Deeks' asymmetry test. Random-effects models were applied in all syntheses. Meta-regression was conducted to further explore sources of heterogeneity. The quality of the included studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool[22] by XL and KL independently. Discrepancies in assessment were resolved through discussion. A $P < 0.05$ was considered statistically significant.

Results

Study flow

A total of 260 studies were identified through database searching. After excluding 58 duplicates, 202 studies were screened based on title and abstract. Subsequently, 184 studies were excluded, and the remaining 18 studies were assessed through full-text review. Finally, 6 studies were excluded, and a total of 12 studies related to the diagnosis of LPLN metastasis using MRI in patients with rectal carcinoma[10-21] were included in this meta-analysis (Figure 1).

Features of enrolled studies

This meta-analysis included 4 prospective studies and 8 retrospective studies. The MRI findings were all preoperative in the included studies. The MRI modality included T2-weighted imaging (T2WI); T1-weighted imaging and T2WI; and T2WI and diffusion-weighted imaging; however, Dev et al.[16] did not report this information. The cut-off value of the short-axis or long-axis diameter of the LPLN used to distinguish positive and negative samples ranged from 4 to 10 mm. The complete features of all studies are presented in Table 1.

Sensitivity and specificity of magnetic resonance imaging for diagnosing lateral pelvic lymph node metastasis

Heterogeneity existed in the sensitivity data ($I^2 = 83.0\%$, $P < 0.001$). The pooled sensitivity (95% CI) was 0.66 (0.53, 0.80; Figure 2a). The specificity data were also heterogeneous ($I^2 = 92.5\%$, $P < 0.001$). The pooled specificity (95% CI) was 0.82 (0.76, 0.88; Figure 2b).

Positive likelihood ratio and negative likelihood ratio of magnetic resonance imaging for diagnosing lateral pelvic lymph node metastasis

Data on the positive LR of MRI showed no significant heterogeneity ($I^2 = 29.6\%$, $P = 0.155$). The pooled positive LR (95% CI) was 2.82 (2.14, 3.51; Figure 3a). Heterogeneity was present in the negative LR data ($I^2 = 74.1\%$, $P < 0.001$). The pooled negative LR (95% CI) was 0.41 (0.27, 0.55; Figure 3b).

Summary receiver operating characteristic curve of magnetic resonance imaging for diagnosing lateral pelvic lymph node metastasis

An SROC curve was constructed to assess the overall ability of MRI to diagnose LPLN metastasis in patients with rectal carcinoma. The AUC of MRI for diagnosing LPLN metastasis was 0.824. The standard error of the AUC was 0.023 (Figure 4).

Quality assessment

All studies had a low risk of bias regarding the reference standard, as well as follow-up and timing. More than 50% of the studies had an unclear risk of bias regarding patient selection and index test, whereas the remaining studies were assessed as having a low risk of bias. All studies had low applicability concerns regarding the reference standard. More than 50% of the studies had low applicability concerns regarding patient selection, and the others were assessed as having unclear applicability concerns. Moreover, more than 50% of the studies had unclear applicability concerns regarding the index test, whereas the remaining studies were assessed as having low applicability concerns (Figure 5a). Detailed information on each study with high, unclear, or low risk of bias or applicability concerns is shown in Figure 5b.

Publication bias and factors related to heterogeneity

Publication bias was present among the included studies ($P = 0.020$; Supplementary Figure 1). Considering that heterogeneity among studies may contribute to publication bias, a meta-regression analysis was conducted to examine factors potentially influencing heterogeneity. It was found that sample size was negatively associated with sensitivity (coefficient: -0.002, $P = 0.009$). Additionally, sample size was positively associated with negative LR (coefficient: 0.002, $P = 0.029$). Study type, cut-off value, and sample size were not significantly associated with specificity or positive LR (all $P > 0.05$; Table 2).

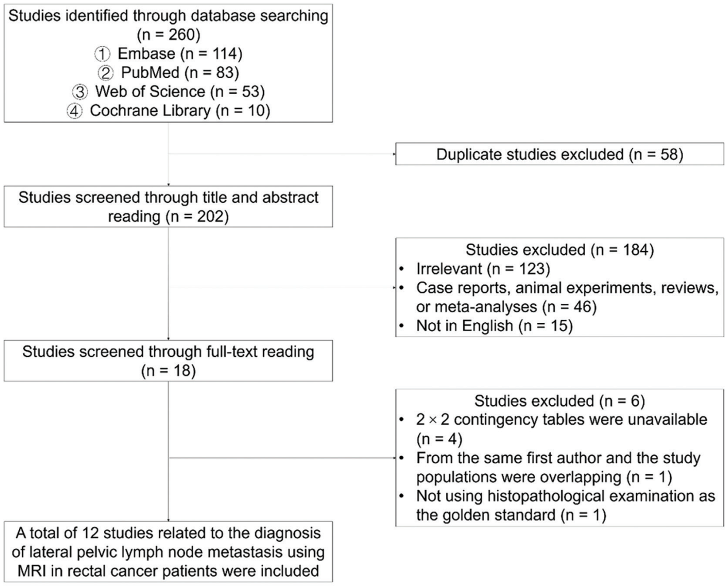

Studies identified through database searching (n = 260)
① Embase (n = 114)
② PubMed (n = 83)
③ Web of Science (n = 53)
④ Cochrane Library (n = 10)
Duplicate studies excluded (n = 58)
Studies screened through title and abstract reading (n = 202)
Studies excluded (n = 184)
• Irrelevant (n = 123)
• Case reports, animal experiments, reviews, or meta-analyses (n = 46)
• Not in English (n = 15)
Studies screened through full-text reading (n = 18)
Studies excluded (n = 6)
• 2 × 2 contingency tables were unavailable (n = 4)
• From the same first author and the study populations were overlapping (n = 1)
• Not using histopathological examination as the golden standard (n = 1)
A total of 12 studies related to the diagnosis of lateral pelvic lymph node metastasis using MRI in rectal cancer patients were included

Figure 1. Study screen. MRI, magnetic resonance imaging.

Table 1. Features of included studies

Study ID	Study type	Sample size	Age (years)	Men (n)	MRI findings	Modality of MRI	Cut-off value[†] (mm)	TP	FP	FN	TN
Matsuoka et al.[10]	Prospective	51	63.0[a]	35	Preoperative	T2WI	5	10	9	5	27
Akasu et al.[11]	Prospective	104	58.0[b]	82	Preoperative	T2WI	4	13	12	2	77
Ogawa et al.[12]	Retrospective	77	(-)	(-)	Preoperative	T1WI and T2WI	5	8	29	2	38
Akiyoshi et al.[13]	Retrospective	77	61.0[b]	55	Preoperative	T2WI	8	21	7	10	39
Ishibe et al.[14]	Prospective	84	62.0[a]	53	Preoperative	T1WI and T2WI	10	12	21	4	47
Ogawa et al.[15]	Retrospective	268	(-)	(-)	Preoperative	T1WI and T2WI	10	14	2	37	215
Dev et al.[16]	Prospective	43	(-)	21	Preoperative	Not mentioned	8	4	3	5	31
Kim et al.[17]	Retrospective	57	57.0[b]	33	Preoperative	T2WI and DWI	7.5	20	10	3	24
Amano et al.[18]	Retrospective	184[‡]	65.0[b]	25	Preoperative	T1WI and T2WI	6	6	5	11	162
Sekido et al.[19]	Retrospective	60	60.0[b]	40	Preoperative	T2WI	7	9	6	3	42
Ishizaki et al.[20]	Retrospective	61	62.0[b]	37	Preoperative	T2WI	6.8	14	12	4	31
Zhang et al.[21]	Retrospective	87	58.7[a]	48	Preoperative	T2WI	7	14	15	7	51

[†] Cut-off value refers to the short-axis or long-axis diameter of lateral pelvic lymph nodes used to distinguish between positive and negative samples.

[‡] Indicates that 184 was the number of regions, not the number of patients.

For age: superscript [a] indicates mean age; superscript [b] indicates median age.

MRI, magnetic resonance imaging; TP, true positive; FP, false positive; FN, false negative; TN, true negative; T2WI, T2-weighted imaging; T1WI, T1-weighted imaging; DWI, diffusion-weighted imaging.

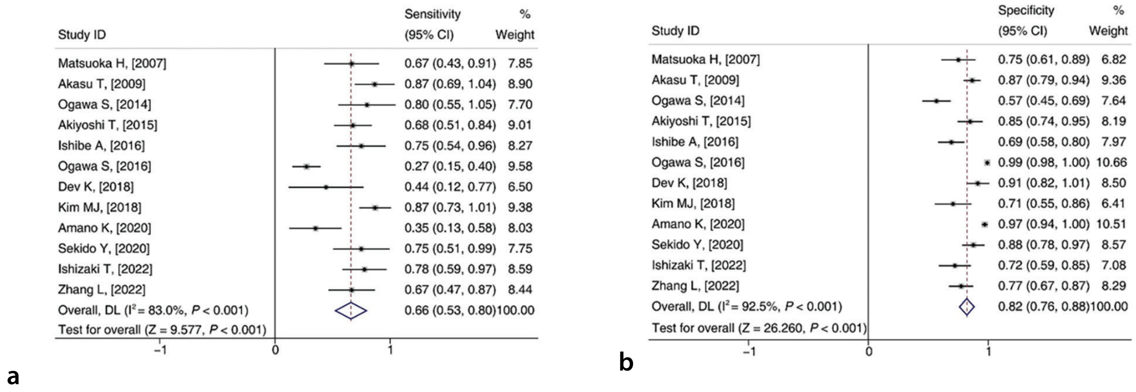


Figure 2. Forest plots of sensitivity and specificity. Pooled sensitivity (**a**) and pooled specificity (**b**) of MRI for diagnosing LPLN metastasis in patients with rectal carcinoma. MRI, magnetic resonance imaging; LPLN, lateral pelvic lymph node, CI, confidence interval.

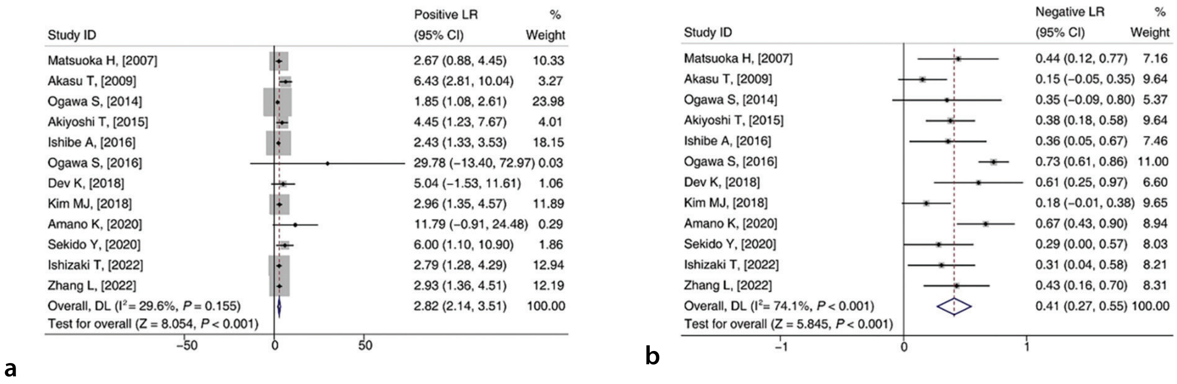


Figure 3. Forest plots of positive and negative likelihood ratios (LRs). Pooled positive LR (**a**) and negative LR (**b**) of MRI for diagnosing LPLN metastasis in patients with rectal carcinoma. MRI, magnetic resonance imaging; LPLN, lateral pelvic lymph node, CI, confidence interval.

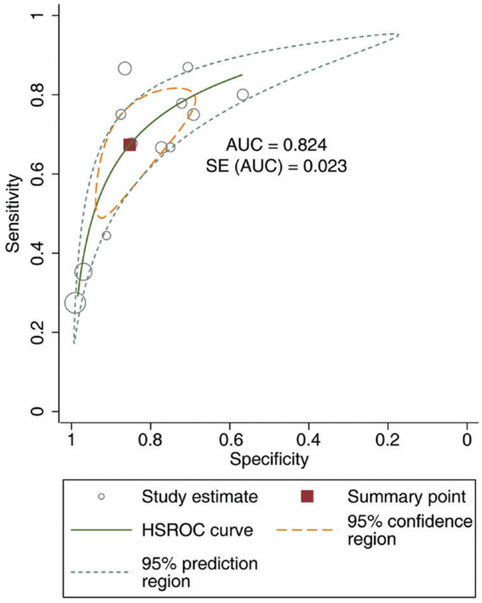
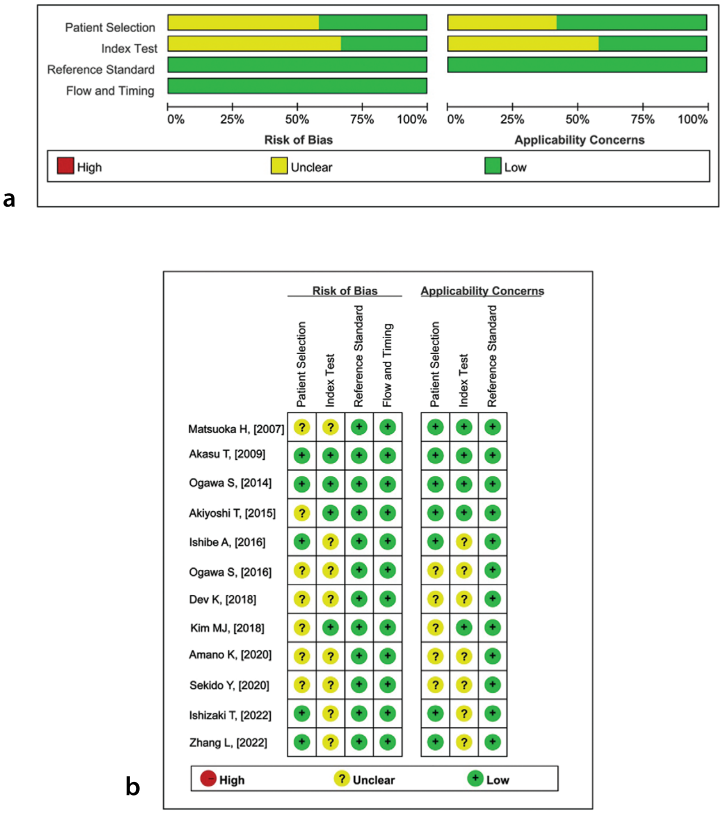


Figure 4. Summary receiver operating characteristic curve of the diagnostic performance of MRI. MRI, magnetic resonance imaging; AUC, area under the curve; HSROC, hierarchical summary receiver operating characteristic.



	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Matsuoka H, [2007]	?	?	+	+	+	+	+
Akasu T, [2009]	+	+	+	+	+	+	+
Ogawa S, [2014]	+	+	+	+	+	+	+
Akiyoshi T, [2015]	?	+	+	+	+	+	+
Ishibe A, [2016]	+	?	+	+	+	?	+
Ogawa S, [2016]	?	?	+	+	?	?	+
Dev K, [2018]	?	?	+	+	?	?	+
Kim MJ, [2018]	?	+	+	+	?	+	+
Amano K, [2020]	?	?	+	+	?	?	+
Sekido Y, [2020]	?	?	+	+	?	?	+
Ishizaki T, [2022]	+	?	+	+	+	?	+
Zhang L, [2022]	+	?	+	+	+	?	+

Figure 5. Quality assessment by QUADAS-2 tools. The proportion of studies with high, unclear, and low risk of bias, as well as applicability concerns (**a**). Detailed information for each study with high, unclear, and low risk of bias, as well as applicability concerns (**b**).

Table 2. Heterogeneity source analysis via meta-regression

Items	Coefficient	Standard error	95% CI	*P* value for t-test
Sensitivity				
Study type	0.010	0.092	(-0.202, 0.222)	0.916
Cut-off value	-0.019	0.024	(-0.073, 0.036)	0.459
Sample size	-0.002	0.198	(-0.003, -0.001)	0.009
P value for F-test	0.018			
Specificity				
Study type	-0.043	0.075	(-0.217, 0.130)	0.579
Cut-off value	0.002	0.019	(-0.043, 0.047)	0.924
Sample size	0.001	0.001	(-0.001, 0.002)	0.061
P value for F-test	0.223			
Positive LR				
Study type	-0.273	1.036	(-2.661, 2.116)	0.799
Cut-off value	-0.100	0.279	(-0.743, 0.542)	0.728
Sample size	0.035	0.027	(-0.029, 0.098)	0.244
P value for F-test	0.650			
Negative LR				
Study type	-0.007	0.107	(-0.253, 0.240)	0.953
Cut-off value	0.025	0.027	(-0.037, 0.088)	0.380
Sample size	0.002	0.001	(0.001, 0.003)	0.029
P value for F-test	0.040			

CI: confidence interval; LR: likelihood ratio.

Discussion

LPLN metastasis occurs in approximately 10% to 25% of patients with rectal carcinoma, which is associated with increased local recurrence rates.[4,23] Of note, two previous meta-analyses found that the pooled sensitivity (95% CI) of MRI for diagnosing LPLN metastasis in patients with rectal carcinoma was 0.72 (0.66, 0.78)[24] and 0.88 (0.85, 0.91)[25]; the pooled specificity (95% CI) was 0.80 (0.73, 0.85)[24] and 0.85 (0.78, 0.90).[25] In the current meta-analysis, we found that the pooled sensitivity (95% CI) and specificity (95% CI) of MRI for diagnosing LPLN metastasis were 0.66 (0.53, 0.80) and 0.82 (0.76, 0.88), respectively, in patients with rectal carcinoma. The pooled sensitivity differed between our meta-analysis and previous meta-analyses.[24,25] A potential reason may be that the cut-off value for lymph node size used to identify LPLN metastasis varied among studies, which contributed to differences in MRI sensitivity and ultimately affected the pooled analysis.

LR refers to the probability ratio of a specific test result between diseased and non-diseased individuals, and the value of LR has important implications.[26-28] In general, a higher positive LR and a lower negative LR suggest superior diagnostic performance of a specific test.[28,29] The present meta-analysis observed that the positive LR and negative LR of MRI for diagnosing LPLN metastasis were 2.82 and 0.41, respectively, in patients with rectal carcinoma. Therefore, our findings suggest that MRI possesses moderate diagnostic performance for LPLN metastasis in patients with rectal carcinoma.

The receiver operating characteristic curve is applied to evaluate the overall diagnostic performance of a test.[30,31] Generally, an AUC value greater than 0.8 indicates good overall diagnostic performance.[30,32] A previous meta-analysis reported that the AUC of MRI for diagnosing LPLN metastasis was 0.88 in patients with rectal carcinoma.[25] Similarly, in our meta-analysis, the AUC was 0.82. Hence, our findings indicate that MRI is useful for diagnosing LPLN metastasis in patients with rectal carcinoma.

Publication bias refers to the tendency for studies with favorable or statistically significant results to be more likely to be published than those with non-substantial results, which may affect the conclusions of a meta-analysis.[33-35] In the current meta-analysis, Deeks' funnel plot asymmetry test showed that publication bias existed regarding the diagnostic performance of MRI for LPLN metastasis in patients with rectal carcinoma. We speculated that a potential contributor to this bias might be heterogeneity among the included studies.[35,36] To further explore the factors influencing heterogeneity, we conducted a meta-regression analysis. It was found that heterogeneity could be influenced by sample size, as sample size was negatively related to sensitivity but positively related to negative LR. Due to the presence of publication bias and heterogeneity in the enrolled studies, our findings should be interpreted with caution. Further rigorous studies are needed to verify the diagnostic performance of MRI for LPLN metastasis in patients with rectal carcinoma.

Several limitations should be noted in this meta-analysis. (1) The cut-off value of the short-axis or long-axis diameter of the LPLN used to distinguish positive and negative samples ranged from 4 to 10 mm in the included studies. Therefore, our meta-analysis could not determine the optimal cut-off value of lymph node size for identifying LPLN metastasis, which should be further investigated. (2) A comparison of the diagnostic performance of MRI with other imaging methods, such as CT and ^{18}F-FDG-PET, could be further explored. (3) Most of the included studies were conducted in Japan, which may limit the generalizability of the findings.

In conclusion, preoperative MRI is recommended for identifying LPLN metastasis in patients with rectal carcinoma, which may further assist in optimizing treatment strategies in this population.

Footnotes

Conflict of interest disclosure

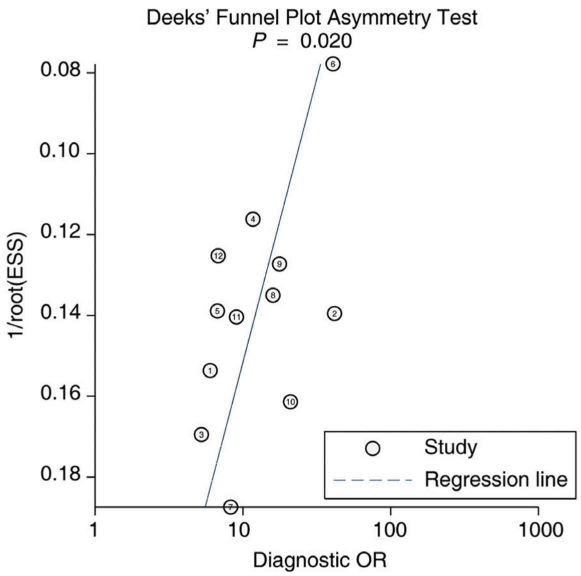
The authors declared no conflicts of interest.

References

1. Choi GS. Lateral pelvic node metastasis in locally advanced rectal cancer: are we exaggerating or ignoring? *Ann Surg Oncol*. 2021;28(11):5803-5804. [Crossref]
2. Fujita S, Mizusawa J, Kanemitsu Y, et al. Mesorectal excision with or without lateral lymph node dissection for clinical stage II/III Lower Rectal Cancer (JCOG0212): a multicenter, randomized controlled, noninferiority trial. *Ann Surg*. 2017;266(2):201-207. [Crossref]
3. Ogura A, Konishi T, Cunningham C, et al. Neoadjuvant (Chemo)radiotherapy with total mesorectal excision only is not sufficient to prevent lateral local recurrence in enlarged nodes: results of the multicenter lateral node study of patients with low cT3/4 Rectal Cancer. *J Clin Oncol*. 2019;37(1):33-43. [Crossref]
4. Chang G, Halabi WJ, Ali F. Management of lateral pelvic lymph nodes in rectal cancer. *J Surg Oncol*. 2023;127(8):1264-1270. [Crossref]
5. Hazen SJA, Sluckin TC, Konishi T, Kusters M. Lateral lymph node dissection in rectal cancer: state of the art review. *Eur J Surg Oncol*. 2022;48(11):2315-2322. [Crossref]
6. Yoo GS, Park HC, Yu JI. Clinical implication and management of rectal cancer with clinically suspicious lateral pelvic lymph node metastasis: a radiation oncologist's perspective. *Front Oncol*. 2022;12:960527. [Crossref]
7. Ogawa S, Itabashi M, Inoue Y, et al. Lateral pelvic lymph nodes for rectal cancer: a review of diagnosis and management. *World J Gastrointest Oncol*. 2021;13(10):1412-1424. [Crossref]
8. Inoue A, Sheedy SP, Wells ML, et al. Rectal cancer pelvic recurrence: imaging patterns and key concepts to guide treatment planning. *Abdom Radiol (NY)*. 2023;48(6):1867-1879. [Crossref]
9. Elhusseini M, Aly EH. Lateral pelvic lymph node dissection in the management of locally advanced low rectal cancer: summary of the current evidence. *Surg Oncol*. 2020;35:418-425. [Crossref]
10. Matsuoka H, Nakamura A, Masaki T, et al. Optimal diagnostic criteria for lateral pelvic lymph node metastasis in rectal carcinoma. *Anticancer Res*. 2007;27(5B):3529-3533. [Crossref]
11. Akasu T, Iinuma G, Takawa M, Yamamoto S, Muramatsu Y, Moriyama N. Accuracy of high-resolution magnetic resonance imaging in preoperative staging of rectal cancer. *Ann Surg Oncol*. 2009;16(10):2787-2794. [Crossref]
12. Ogawa S, Itabashi M, Hirosawa T, Hashimoto T, Bamba Y, Kameoka S. Lateral pelvic lymph node dissection can be omitted in lower rectal cancer in which the longest lateral pelvic and perirectal lymph node is less than 5 mm on MRI. *J Surg Oncol*. 2014;109(3):227-233. [Crossref]
13. Akiyoshi T, Matsueda K, Hiratsuka M, et al. Indications for lateral pelvic lymph node dissection based on magnetic resonance imaging before and after preoperative chemoradiotherapy in patients with advanced low-rectal cancer. *Ann Surg Oncol*. 2015;22(Suppl 3):614-620. [Crossref]
14. Ishibe A, Ota M, Watanabe J, et al. Prediction of lateral pelvic lymph-node metastasis in low rectal cancer by magnetic resonance imaging. *World J Surg*. 2016;40(4):995-1001. [Crossref]
15. Ogawa S, Hida J, Ike H, et al. Selection of lymph node-positive cases based on perirectal and lateral pelvic lymph nodes using magnetic resonance imaging: study of the japanese society for cancer of the colon and rectum. *Ann Surg Oncol*. 2016;23(4):1187-1194. [Crossref]
16. Dev K, Veerenderkumar KV, Krishnamurthy S. Incidence and predictive model for lateral

pelvic lymph node metastasis in lower rectal cancer. *Indian J Surg Oncol*. 2018;9(2):150-156. [Crossref]

17. Kim MJ, Hur BY, Lee ES, et al. Prediction of lateral pelvic lymph node metastasis in patients with locally advanced rectal cancer with preoperative chemoradiotherapy: focus on MR imaging findings. *PLoS One*. 2018;13(4):e0195815. [Crossref]
18. Amano K, Fukuchi M, Kumamoto K, et al. Pre-operative evaluation of lateral pelvic lymph node metastasis in lower rectal cancer: comparison of three different imaging modalities. *J Anus Rectum Colon*. 2020;4(1):34-40. [Crossref]
19. Sekido Y, Nishimura J, Fujino S, et al. Predicting lateral pelvic lymph node metastasis based on magnetic resonance imaging before and after neoadjuvant chemotherapy for patients with locally advanced lower rectal cancer. *Surg Today*. 2020;50(3):292-297. [Crossref]
20. Ishizaki T, Katsumata K, Enomoto M, et al. Predictors of lateral pelvic lymph node metastasis in advanced low rectal cancer treated with neoadjuvant chemotherapy. *Anticancer Res*. 2022;42(4):2113-2121. [Crossref]
21. Zhang L, Shi F, Hu C, et al. Development and external validation of a preoperative nomogram for predicting lateral pelvic lymph node metastasis in patients with advanced lower rectal cancer. *Front Oncol*. 2022;12:930942. [Crossref]
22. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. [Crossref]
23. Lee T, Horvat N, Gollub MJ, Garcia-Aguilar J, Kim TH. Prognostic value of lateral lymph node metastasis in pretreatment MRI for rectal cancer in patients undergoing neoadjuvant chemoradiation followed by surgical resection without lateral lymph node dissection: a systemic review and meta-analysis. *Eur J Radiol*. 2024;178:111601. [Crossref]
24. Hoshino N, Murakami K, Hida K, Sakamoto T, Sakai Y. Diagnostic accuracy of magnetic resonance imaging and computed tomography for lateral lymph node metastasis in rectal cancer: a systematic review and meta-analysis. *Int J Clin Oncol*. 2019;24(1):46-52. [Crossref]
25. Rooney S, Meyer J, Afzal Z, et al. The role of preoperative imaging in the detection of lateral lymph node metastases in rectal cancer: a systematic review and diagnostic test meta-analysis. *Dis Colon Rectum*. 2022;65(12):1436-1446. [Crossref]
26. Parikh R, Parikh S, Arun E, Thomas R. Likelihood ratios: clinical application in day-to-day practice. *Indian J Ophthalmol*. 2009;57(3):217-221. [Crossref]
27. Doi SAR, Kostoulas P, Glasziou P. Likelihood ratio interpretation of the relative risk. *BMJ Evid Based Med*. 2023;28(4):241-243. [Crossref]
28. Elston DM. Likelihood ratios. *J Am Acad Dermatol*. 2022;86(6):1229. [Crossref]
29. McGee S. Simplifying likelihood ratios. *J Gen Intern Med*. 2002;17(8):646-649. [Crossref]
30. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean J Anesthesiol*. 2022;75(1):25-36. [Crossref]
31. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digit Health*. 2022;4(12):853-855. [Crossref]
32. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315-1316. [Crossref]
33. DeVito NJ, Goldacre B. Catalogue of bias: publication bias. *BMJ Evid Based Med*. 2019;24(2):53-54. [Crossref]
34. Rouan J, Velazquez G, Freischlag J, Kibbe MR. Publication bias is the consequence of a lack of diversity, equity, and inclusion. *J Vasc Surg*. 2021;74(2 Suppl):111-117. [Crossref]
35. Lin L, Chu H. Quantifying publication bias in meta-analysis. *Biometrics*. 2018;74(3):785-794. [Crossref]
36. Sun P, Zhao W. Be careful about heterogeneity and publication bias in meta-analysis. *J Clin Anesth*. 2019;53:76. [Crossref]



Supplementary Figure 1. Deeks' funnel plot asymmetry test. EES, expected effect size; OR, odds ratio.

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243100

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Deep learning for named entity recognition in Turkish radiology reports

Abubakar Ahmad Abdullahi[1]*
Murat Can Ganiz[1]*
Ural Koç[2]*
Muhammet Batuhan Gökhan[2]**
Ceren Aydın[2]**
Ali Bahadır Özdemir[2]**

[1]Marmara University Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

[2]Ankara Bilkent City Hospital, Clinic of Radiology, Ankara, Türkiye

***Joint first authors**

****Contributed equally to this work**

Corresponding author: Ural Koç

E-mail: dr_uralkoc@hotmail.com

Received 30 October 2024; revision requested 10 December 2024; last revision received 01 January 2025; accepted 13 January 2025.

Epub: 28.02.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.243100

PURPOSE

The primary objective of this research is to enhance the accuracy and efficiency of information extraction from radiology reports. In addressing this objective, the study aims to develop and evaluate a deep learning framework for named entity recognition (NER).

METHODS

We used a synthetic dataset of 1,056 Turkish radiology reports created and labeled by the radiologists in our research team. Due to privacy concerns, actual patient data could not be used; however, the synthetic reports closely mimic genuine reports in structure and content. We employed the four-stage DYGIE++ model for the experiments. First, we performed token encoding using four bidirectional encoder representations from transformers (BERT) models: BERTurk, BioBERTurk, PubMedBERT, and XLM-RoBERTa. Second, we introduced adaptive span enumeration, considering the word count of a sentence in Turkish. Third, we adopted span graph propagation to generate a multidirectional graph crucial for coreference resolution. Finally, we used a two-layered feed-forward neural network to classify the named entity.

RESULTS

The experiments conducted on the labeled dataset showcase the approach's effectiveness. The study achieved an F1 score of 80.1 for the NER task, with the BioBERTurk model, which is pre-trained on Turkish Wikipedia, radiology reports, and biomedical texts, proving to be the most effective of the four BERT models used in the experiment.

CONCLUSION

We show how different dataset labels affect the model's performance. The results demonstrate the model's ability to handle the intricacies of Turkish radiology reports, providing a detailed analysis of precision, recall, and F1 scores for each label. Additionally, this study compares its findings with related research in other languages.

CLINICAL SIGNIFICANCE

Our approach provides clinicians with more precise and comprehensive insights to improve patient care by extracting relevant information from radiology reports. This innovation in information extraction streamlines the diagnostic process and helps expedite patient treatment decisions.

KEYWORDS

Named entity recognition, radiology reports, bidirectional encoder representations from transformers, Turkish, computed tomography, thorax

Radiology reports are a cornerstone of modern healthcare, capturing intricate diagnostic insights derived from medical images. These unstructured reports encapsulate the clinical context, imaging techniques, findings, and interpretations, which are pivotal in guiding patient care decisions.[1] However, their inherent lack of structure poses challenges for downstream applications that require standardized and structured data, including research, billing, accreditation, and quality improvement.[2] There is a push toward using structured formats instead of free-text radiology reports. Although initiatives such as RadReport[2] and

You may cite this article as: Abdullahi AA, Ganiz MC, Koç U, Gökhan MB, Aydın C, Özdemir AB. Deep learning for named entity recognition in Turkish radiology reports. *Diagn Interv Radiol.* 2025;31(5):430-439.

RadLex[3] have helped standardize radiology reporting, unstructured formats remain the most common format despite the need for standardization. Various research methodologies have been investigated to bridge this gap, including rule-based systems, machine learning, and deep learning.

Our study focuses on applying deep learning to extract named entities from radiology reports written in Turkish. In addition, we developed a new dataset to train the named entity recognition (NER) task and considered the distinctive characteristics of the Turkish language to attain the best possible results. For the NER task, we utilized the DYGIE++ framework[4] and adapted it to the Turkish language. The DYGIE++ framework relies on a bidirectional encoder representations from transformers (BERT)[5] model to extract text embeddings. Therefore, we used the BioBERTurk model,[6] a variant of BERT pre-trained on Turkish biomedical data. This combination allows for the extraction of structured information, which can be used to enhance various medical applications. Our approach builds on previous research and aims to improve the overall effectiveness of information extraction in radiology reporting.

The potential of deep learning applications in Turkish radiology reports has yet to be fully explored. To remedy this, we worked with Ankara Bilkent City Hospital radiologists and hand-labeled a substantial dataset of 1,056 reports. To the best of our knowledge, this is the first dataset in Turkish created for this purpose. These reports have been expertly labeled to include observation and symptom categories, and they serve as a crucial foundation for our experiments.

In this paper, we provide a detailed explanation of our methodology and showcase how using DYGIE++ with various BERT models has been effective for our NER task of extracting observations and symptoms from Turkish radiology reports. Although there are no studies against which we can compare our F1 results (80.1) in Turkish, our results are similar to those in other languages. The implications of our study go beyond Turkish radiology reports; the lessons we learned and the methodologies we established can be applied to multiple languages and medical contexts, leading to improved information extraction practices. We hope to see a future where structured insights can be easily extracted from unstructured reports, leading to a revolution in medical reporting practices.

Main points

- Precise data are extracted from radiology reports to address the challenges of retrieving information from unstructured reports.
- Named entity recognition is used to identify observations and symptoms, even in low-resource languages such as Turkish.
- Diagnostic precision is improved and decision-making expedited to foster improved patient care and healthcare outcomes.

In the following sections, we will present related research and discuss the methodology, results, and conclusions that support our findings. The methodology section will elaborate on the dataset and experimental setup. In the results section, we will showcase the findings of our experiments conducted using varying configurations. In the discussion, we will compare our results with other studies in the field, including those conducted in languages other than Turkish. We hope to contribute to the ongoing dialogue on integrating deep learning into radiology reporting and inspire innovation in healthcare.

Methods

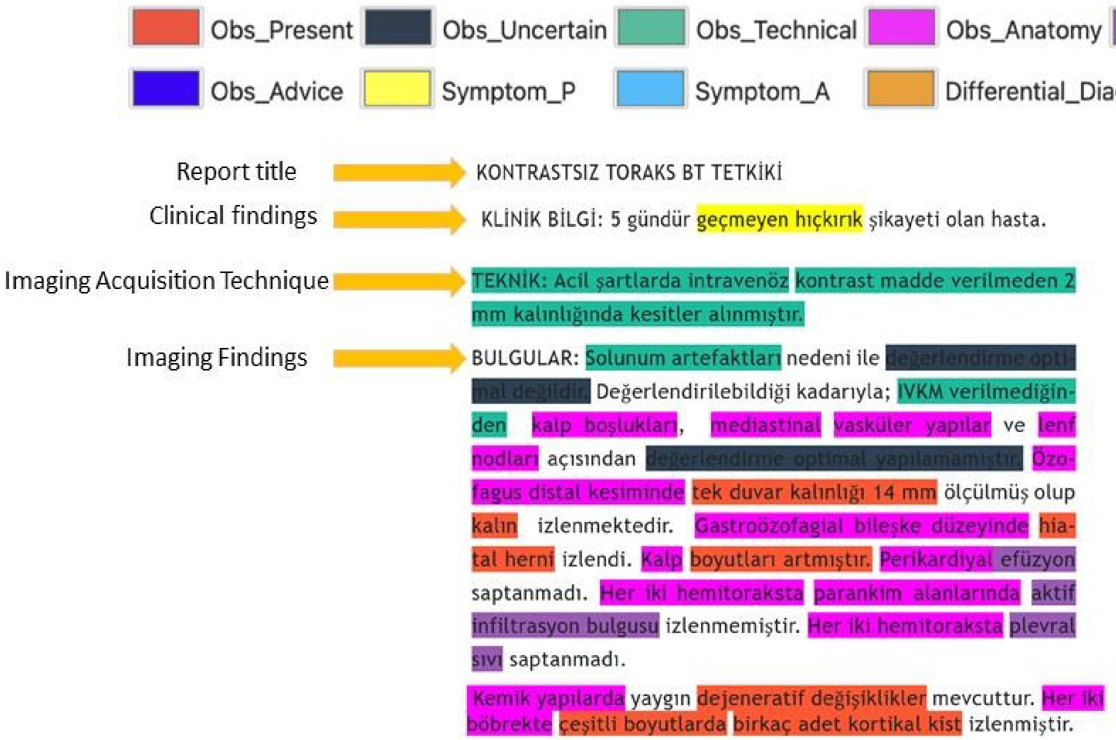
We created a labeled dataset of 1,056 radiology reports produced by the radiologists in our research team. Due to ethical and privacy considerations, it was not feasible to use actual patient data. Therefore, the radiologists drew from their experience of composing authentic radiology reports to generate synthetic reports that resembled the structure and content of genuine ones. This approach ensured that the dataset retained the critical features and complexities of real reports while safeguarding patient confidentiality and data privacy. The reports focused on computed tomography (CT) scans of the thorax area, encompassing the chest, lungs, heart, abdomen, and other vital organs. Figure 1 shows an example of a labeled report. This dataset can be utilized in various medical research projects and assist in developing diagnostic tools and techniques. Table 1 enumerates imaging types and their frequencies. We used the expertise of radiologists to label the data for NER, resulting in nine labels: Obs_Present, Obs_Uncertain, Obs_Technical, Obs_Anatomy, Obs_Absent, Obs_Advice, Symptom_P, Symptom_A, and Differential_Diagnosis. Table 2 enumerates the labels, their descriptions, and their frequencies.

We established a Doccano platform to simplify the labeling of our reports. Doccano is an open-source web-based annotation tool that provides a collaborative environment for annotating text elements such as named entities. It allows users to upload text documents and add annotations to a group of words within the document. Users work in parallel on separate documents that need to be labeled. Due to its user-friendly interface, Doccano was particularly valuable in simplifying the labeling process. An export of the data to the JavaScript object notation lines format became readily available once the labeling was complete. We labeled 1,056 reports by randomly dividing them into three equal parts for three radiologists to label in parallel. Ural Koç, co-author reviewed the labeling results and supervised the entire labeling process.

We named our task entity recognition using the DYGIE++ framework. The DYGIE++ framework is a span-based model for extracting entities, relations, and event triggers. We performed the entity extraction in isolation to accomplish our task. Our approach in the four stages of the DYGIE++ model is detailed as follows:

1. Token encoding: This step uses a BERT model to obtain token representations of the text. It utilizes a sliding window technique, feeding a sentence to the model at each iteration along with 15 surrounding sentences. We experimented with four BERT models: BERTurk, BioBERTurk, PubMedBERT, and XLM-RoBERTa. The BERTurk model was pre-trained on Turkish text sourced from Wikipedia dumps, and we selected it because the model's Turkish language matched our training data. The BioBERTurk model was pre-trained on top of BERTurk with Turkish biomedical texts and radiology theses, making it the most suitable fit for our application domain and language. The PubMedBERT model was pre-trained on English text sourced from the abstracts and articles of academic biomedical publications, and we selected it because its medical data matched our domain. The XLM-RoBERTa model was pre-trained on text from Wikipedia dumps containing 100 languages (including Turkish), and we chose it because medical terms tend to remain consistent across multiple languages.

2. Adaptive span enumeration: A span is a group of adjacent tokens that can be either a single token or a combination of many. We created it by concatenating token representations. The usage of suffixes in the Turkish language results in shorter sentences despite longer word lengths. For instance, the English phrase "the nasogastric tube has been pushed forward" translates to "nazogastrik tüp ileri itildi" or "nazogastrik tüp iler-


🟥 Obs_Present
Obs_Uncertain
Obs_Technical
Obs_Anatomy
Obs_Absent
Obs_Advice
Symptom_P
Symptom_A
Differential_Diagnosis
Report title
Clinical findings
Imaging Acquisition Technique
Imaging Findings

KONTRASTSIZ TORAKS BT TETKİKİ

KLİNİK BİLGİ: 5 gündür geçmeyen hıçkırık şikayeti olan hasta.

TEKNİK: Acil şartlarda intravenöz kontrast madde verilmeden 2 mm kalınlığında kesitler alınmıştır.

BULGULAR: Solunum artefaktları nedeni ile değerlendirme optimal değildir. Değerlendirilebildiği kadarıyla; IVKM verilmediğinden kalp boşlukları, mediastinal vasküler yapılar ve lenf nodları açısından değerlendirme optimal yapılamamıştır. Özofagus distal kesiminde tek duvar kalınlığı 14 mm ölçülmüş olup kalın izlenmektedir. Gastroözofagial bileşke düzeyinde hiatal herni izlendi. Kalp boyutları artmıştır. Perikardiyal efüzyon saptanmadı. Her iki hemitoraksta parankim alanlarında aktif infiltrasyon bulgusu izlenmemiştir. Her iki hemitoraksta plevral sıvı saptanmadı.

Kemik yapılarda yaygın dejeneratif değişiklikler mevcuttur. Her iki böbrekte çeşitli boyutlarda birkaç adet kortikal kist izlenmiştir.

Figure 1. Color-coded example of labeled reports.

Table 1. Imaging types and their frequencies in the labeled dataset of radiology reports

Imaging type	Number of reports	Percentage
Abdominal radiology	363	34.38%
Thorax radiology	224	21.21%
Neuroradiology	187	17.71%
Vascular and thorax radiology	101	9.56%
Musculoskeletal radiology	66	6.25%
Head and neck radiology	45	4.26%
Vascular and thoracoabdominal radiology	25	2.37%
Vascular and neuroradiology	22	2.08%
Vascular and musculoskeletal radiology	15	1.42%
Vascular and abdominal radiology	5	0.47%
Vascular and neck radiology	3	0.28%

letildi" in Turkish, consisting of four- or three-word sentences instead of the seven-word sentence in English. Although DYGIE++ was originally developed using English, we modified our model to accommodate Turkish. We set the maximum number of tokens per span to four instead of the default limit of eight used in English experimentally, as we obtained the best performance using this value.

3. Span graph propagation: This step generates a multidirectional graph by computing the connections between spans. Spans are considered connected if they are likely to be related or refer to the same topic (coreference). We were interested in the coreference propagation in this step, which is crucial for identifying references to an entity throughout the document. Therefore, once we had the entity type of one reference, we could apply it to all the other references in the document.

4. Named entity classification: In the final step, a two-layered feed-forward neural network was used as a scoring function to make predictions for named entities.

For the experiment, we partitioned the 1,056 reports in the labeled dataset into three subsets: 75% for training, 15% for testing, and 10% for development. The training configuration closely followed that of DYGIE++.[4] The training phase spanned 100 epochs and focused on NER; therefore, the loss weights for relation extraction, coreference resolution, and event extraction were set to 0, and the weight for NER was set to 1. We used the AdamW optimizer,[7] with a learning rate of 1e −3 and weight decay of 0.0. The gradient norm was set to 5.0 for stable training with a slanted triangular learning rate scheduler. We used an NVIDIA V100 graphical processing unit as a CUDA device throughout the experiments. The codebase was in Python. We sourced our code from the DYGIE++ GitHub code repository of[4] (github.com/dwadden/dygiepp), which was built on the AllenNLP framework.[8] The loss weights are given as NER: 0.5, relation extraction: 0.5, coreference resolution: 0.0, and event extraction: 1.0.

Statistical analysis

As for the statistical analysis, we used the micro F1 score as the standard to evaluate and compare the performance of our models. Numeric values are given as a number and frequency (%). Cohen's kappa statistic was used to evaluate agreement. A *P* value <0.05 was considered statistically significant. The study did not require ethics committee approval or patient consent.

Results

Our setup comprises four experimental combinations differentiated by the BERT model, as described under "Token encoding" in section 2. Table 3 shows each model's precision, recall, and F1 score. The best perform-

Table 2. Label distribution for the dataset of radiology reports

Code	Name	Description	Frequency	Percentage
Obs_Present	Observations present	Presence of radiological features, identifiable pathophysiological processes, or diagnostic diseases	12,848	34%
Obs_Absent	Absence of observations	Absence of radiological features, identifiable pathophysiological processes, or diagnostic diseases	3,165	8.38%
Obs_Uncertain	Uncertain observations	Lack of certainty about a radiological feature, pathophysiological process, or diagnostic disease	1,102	2.92%
Obs_Technical	Technical observations	Technical situation that describes radiological techniques such as acquisitions	1,546	4.09%
Obs_Anatomy	Anatomical observations	Anatomical parts such as "vertebrae"	16,872	44.65%
Obs_Advice	Observations of advice	Tests and examinations recommended by the radiologist regarding the current diagnosis and treatment process	489	1.29%
Symptom_P	Presence of a symptom	A specific clinical symptom communicated by the clinician to the radiologist	668	1.77%
Symptom_A	Absence of a symptom	Absence of a specific clinical symptom communicated by the clinician to the radiologist	16	0.04%
Differential_ Diagnosis	Differential diagnosis	Differential diagnoses that may occur as a result of the current findings	1,084	2.87%

Table 3. Results for the BERTurk, BioBERTurk, PubMedBert, and XLM-RoBERTa models

BERT model	Precision	Recall	F1
BERTurk	78.3	79.9	79.1
BioBERTurk	**80.0**	**80.1**	**80.1**
PubMedBERT	75.0	76.9	75.9
XLM-RoBERTa	79.5	77.0	78.3

ing model was the BioBERTurk model, with an F1 score of 80.1. The BERTurk, PubMedBERT, and XLM-RoBERTa models scored 79.1, 75.9, and 78.3, respectively.

Figure 2 is a bar chart that displays each label's F1 score for all four BERT models. We report their respective precision, recall, and F1 scores using tables in Appendices 1-4. These tables offer a label-specific perspective, highlighting the strengths and weaknesses of each model. We can see that although the label "Obs_Present" is the most frequent (occurring 50.65% of the time), it does not have the highest F1 score among all the models. This affects the micro average F1 score because labels that occur more frequently contribute more weight to the overall F1 score. Conversely, "Symptom_A" has a 0.0 F1 score for all models because it lacks examples (only 16 occurrences) for the model to learn. Consequently, its effect on the overall F1 score is negligible.

After receiving constructive feedback from the peer reviewers, two radiologists who were not involved in the initial study evaluated the synthetically generated reports using a Likert scale. The Likert scale ranged from 1 to 5, where 1 indicated the least resemblance to real-world reports and 5 indicated the closest resemblance. The responses were analyzed using Cohen's kappa statistic (Cohen's kappa score: 0.92, $P < 0.001$). The evaluation of radiology reports prepared by the study radiologists achieved a high inter-observer agreement among the independent radiologists. Furthermore, the selected categories on the scale indicated that the reports closely resembled real-world radiology reports (Figure 3). After the peer-review process, 25% of the data were randomly re-annotated (UK) to assess intra-annotator agreement. Cohen's kappa statistic was used to evaluate the level of agreement, yielding a kappa value of 0.997 ($P < 0.0001$). This result indicates a high level of agreement and is statistically significant.

The co-occurrence chord diagram and matrix of the nine labels are shown in Figures 4, 5 and Appendices 5, 6.

Discussion

Structured reports have a standardized language and are consistently organized into ordered sections to enable the automated or semi-automated abstraction of reporting data. In recent years, numerous researchers have demonstrated a keen interest in extracting information from unstructured radiology reports, as almost all reports are written in this format. In 2010, Soysal et al.[9] proposed a natural language processing (NLP) system that converts radiology reports into Turkish. The initial medical information extraction system in Turkish, TRIES, follows a three-step conversion process. It begins with a morphological analysis of every word in the sentence, followed by NER and relation extraction. Its purpose is to match the sentence with a set of rule templates. An example is the sentence "The liver is 14 cm in height," which is analyzed as "Liver vertical tall + NESS + POSS3SG 14 cm + COP," later transformed into "[entity: Liver] [attribute: height] + POSS3SG [value: NUMERIC: 14 cm] + COP," and finally converted to "Liver.height = 14cm." The TRIES system has achieved results with a 93% recall and 98% precision rate. However, this method is limited because rule-based systems fail if a relationship cannot be matched to a specific rule.

Little research related to the present study has been conducted in the Turkish language domain. This is a notable shortcoming considering the considerable advancements published in the literature, especially in pre-trained deep learning models. One of the most commonly used pre-trained language models for creating downstream NLP applications via fine-tuning is BERT, which considers the entire context of words by look-

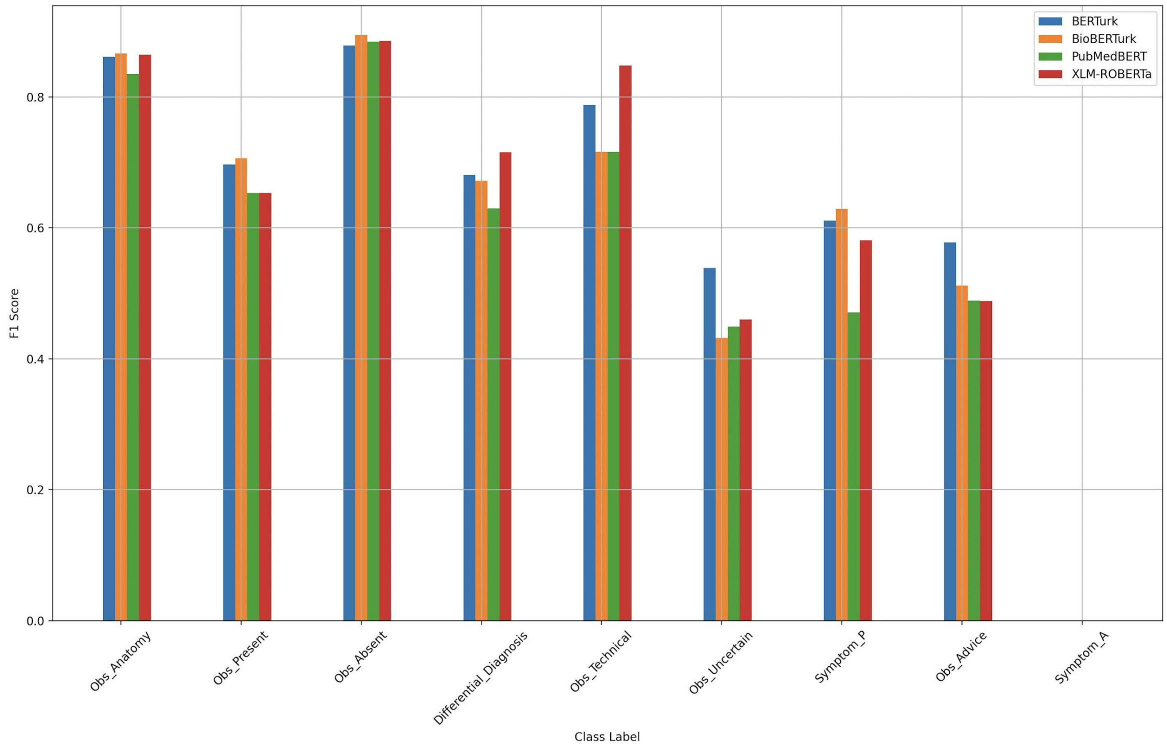


Figure 2. Bar chart of the F1 scores of the BERTurk, BioBERTurk, PubMedBert, and XLM-RoBERTa models.

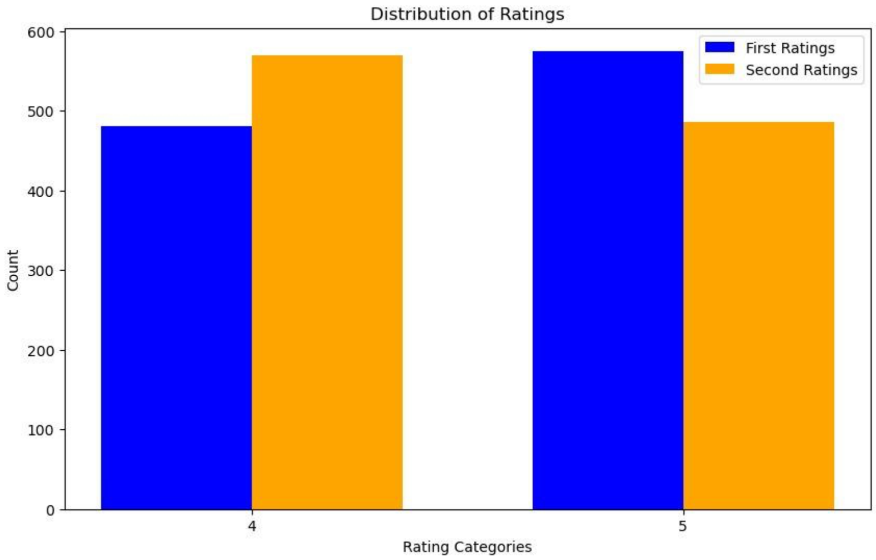


Figure 3. Evaluation of synthetically generated radiology reports by two independent radiologists using a Likert scale (1-5). The analysis showed a high inter-observer agreement (Cohen's kappa score: 0.92, $P < 0.001$), with the majority of scores indicating a strong resemblance to real-world radiology reports.

ing both left and right in a sentence. This model's innovation lies in its pre-training process, which is trained on large amounts of data to perform two tasks. First, masked language modeling (MLM) requires masked words within sentences to be predicted, helping BERT understand the word context and semantics. Second, next sentence prediction (NSP) predicts if one sentence follows another, enabling BERT to grasp sentence relationships. With this bidirectional approach, MLM and NSP allow BERT to capture intricate language relationships. In addition, BERT's architecture allows it to be fine-tuned for specific language-related tasks such as NER. As our task is NER on Turkish data, we experimented with four variations of BERT: BERTurk,[10] BioBERTurk, PubMedBERT,[11] and RoBERTa-XLM.[12]

We found no previous studies related to deep learning in the Turkish language; therefore, we explored other underrepresented languages to gain inspiration to help fill this gap. In a recent study, Jantscher et al.[13] investigated methods for NER and relation extraction from radiology reports in German. To achieve their goal, they fine-tuned a BERT model and used active learning for domain adaptation and training. Three separate datasets were utilized in this study. Reports on head CT were used to fine-tune the GermanMedBERT[14] model, and reports on magnetic resonance imaging (MRI) of the head and pediatric X-rays were used for domain adaptation and training. The researchers aimed to demonstrate that domain adaptation and active learning enhance the effectiveness of NER and relation extraction tasks. The model trained on MRI data performed the best, with an F1 score of 86.0 for NER and 80.0 for relation extraction.

In a similar study,[15] researchers aimed to extract named entities from Polish radiology reports. Using a dataset of 1,200 chest X-ray reports, the study focused on sequence labeling using the inside–outside–beginning annotation schema. This annotation schema consists of 44 tags representing everyday radiological observations while emphasizing generalization for potential application across clinical domains. The experiments involved the use of five BERT models: Pol-

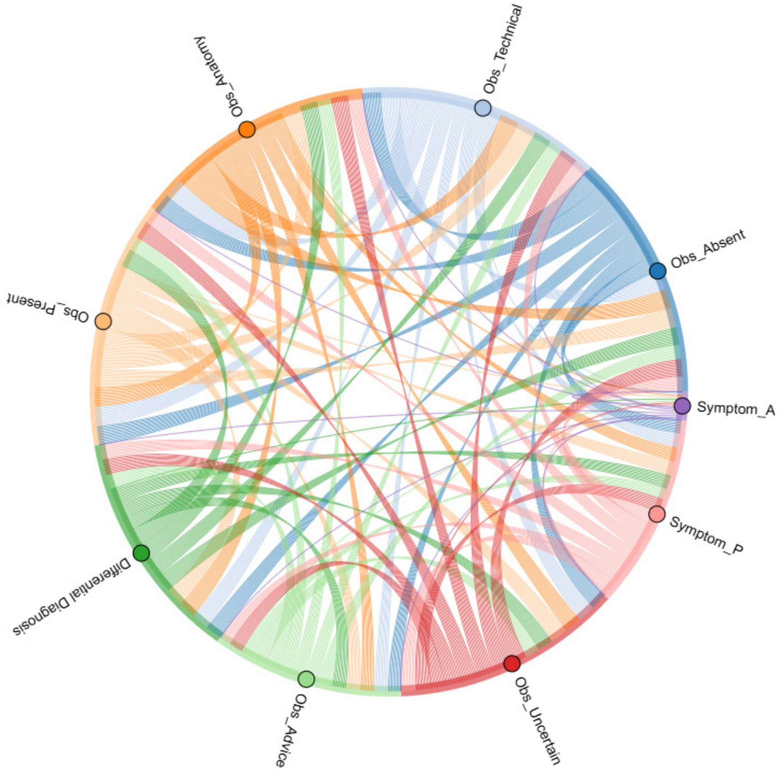


Figure 4. Co-occurrence chord diagram representing the total number of times each label pair appeared together across all reports. In this case, repetitions within the same document are also considered.

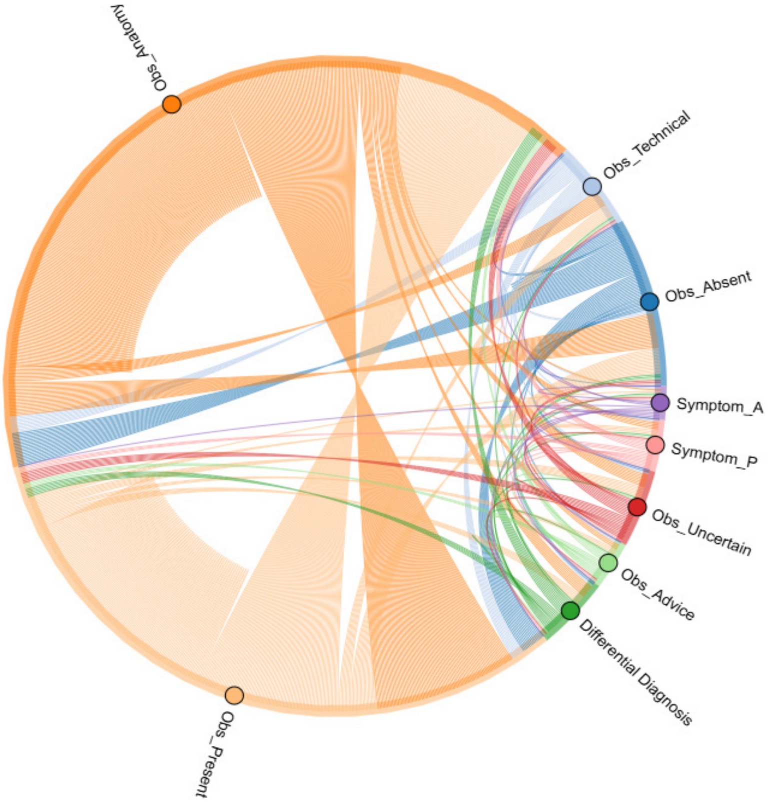


Figure 5. Co-occurrence chord diagram representing the frequency of unique label pairs. Even if the same label pairs appear multiple times, they are counted only once, illustrating the occurrence frequency of these unique combinations.

ish-roberta-base-v2,[16] Polish-distilroberta,[16] Polish-longformer,[16] HerBERT,[17] and mLUKE.[18] The mLUKE model is a multilingual variant of the LUKE model,[19] whereas the rest of the models were pre-trained only on Polish data. The results demonstrated that mLUKE was the most effective model, with an F1 score of 80.9. Its multilingual nature enhanced the domain-specific medical knowledge base across all languages. Certain classes exhibited lower-than-expected scores due to the complexity and variability within those categories. By contrast, others performed well despite limited annotated examples.

Another study[20] focuses on NER applied to chest CT reports in Japanese. The dataset consists of 118,155 reports, 540 of which were annotated by medical experts. Three deep learning models (BiLSTM-CRF, BERT, and BERT-CRF) were used to train NER. Each of the three models was pre-trained on Wikipedia data and CT reports. The labeled dataset was used to evaluate the models, which showed promising results in extracting clinical information from the Japanese chest CT reports. The BiLSTM-CRF model had the highest micro F1 score, with 95.4 for CT and 94.3 for Wikipedia. Higher F1 scores were observed across all models when pre-training with CT reports instead of only Wikipedia. Analysis of the effect of various modifiers on performance shows that the "certainty modifier" entity had a favorable impact, resulting in higher F1 scores. Conversely, the "change modifier" and "characteristics modifier" entities reduced performance, leading to lower F1 scores.

The results of the present study demonstrate that, among the different BERT models, BioBERTurk performed the best. We attribute our model's improved performance to adaptive span enumeration. We ran several iterations to determine the optimal value for the maximum number of tokens per span for the Turkish language. We set it at four instead of the default value of eight in English experimentally, as detailed in the Material and Methods section under "Adaptive span enumeration." This estimation resulted in a 1.5-point increase in BioBERTurk's F1 score. We believe this value to be specific to the Turkish language, and a similar concept can be applied to other languages.

The BERTurk model closely followed BioBERTurk in performance (79.1%) due to its Turkish language embeddings. This is a BERT model that was pre-trained from scratch using only Turkish text. Therefore, we expected it to perform better than multilingual models such as XLM_RoBERTa and English-only models. The XLM-RoBERTa model performed reasonably well (78.3%), but, as expected, it was too generic because it was trained on data from 100 languages. In addition, it is a much larger model, and given its size, we needed a larger dataset for fine-tuning to have a noticeable impact. Finally, PubMedBERT is an English-only model that is pre-trained using English-only medical domain texts. Although the medical terminology in English and Turkish overlaps to a certain degree due to the heavy use of Latin in medicine, as mentioned before, Turkish is very different from English, especially in terms of the heavy use of suffixes that can modify medical concepts in Latin. For example, "appendix" in English can be translated as "Apendiks," "Apendiksin," or "Apendiksinin," with the suffix "-in" indicating possession or a relationship. Similarly, "intubation" in English can be expressed as "Entübasyon," "Entübasyonu," with the suffix "-u" for possession, or "Entübasyonunda," with the locative suffix "-da" to indicate location within a procedure, and so on. There can be a large number of variations with many different suffixes. Due to these profound differences between languages, we observed a significant drop in performance (80.1% vs. 75.9%) when we used PubMedBERT. For PubMedBERT, fine-tuning the model with a large number of Turkish medical texts may increase its performance. A possible solution for PubMedBERT to be considered in future studies is the use of adapters.[21] This method of fine-tuning adds extra layers to the model while retaining the existing ones, which are frozen during training. In this manner, the model preserves its medical knowledge by not updating the frozen weights and incorporates the Turkish context by updating the introduced weights. Our results indicate that it may be difficult to apply deep learning models that have been pre-trained on different languages or even multi-lingual models in domain specific applications such as medicine; however, it is worth using pre-trained models in the target language, adjusting hyper parameters, and applying domain specific fine-tuning.

Our resources, mainly medical data in Turkish, are limited due to the low number of datasets and studies. In fact, our dataset of 1,056 annotated radiology reports is a first in the Turkish medical domain. There are also restrictions for unlabeled data, both in terms of quantity and quality, in the Turkish medical domain compared with the English domain. These restrictions affect our model in several ways. First, we can discuss the domain specialization of large language models such as BERT. Although we used BioBERTurk as a base model that has been fine-tuned for the Turkish medical domain, we might obtain better results by further fine-tuning this model if we had access to a large number of anonymized Turkish radiology reports or related literature in Turkish. Second, we used just 1,056 Turkish radiology reports, which were manually created by radiology experts to mimic actual patient reports. This number can be increased in two ways. One is to involve more experts, which may not be feasible without vital funding and organization, currently beyond our capabilities. The other is to use techniques such as data augmentation,[22] which are useful for increasing the size of the labeled dataset, although the quality would be debatable. Furthermore, these medical text data augmentation methods are devised for English biomedical texts, and applying these directly to Turkish radiology reports may not be feasible due to the key differences between English and Turkish and the agglutinative nature of Turkish, as previously discussed.

We note that the four models exhibit different performance levels for each label. For instance, XLM-RoBERTa performs best for "Obs_Technical," as technical terms are not unique to the Turkish language and were pre-trained in multiple languages. Moreover, BioBERTurk has excellent results for "Symptom_P," as it was trained on the relevant Turkish biomedical data. The "Obs_Uncertain" label posed challenges for all four models because uncertainties usually involve negation-related terms such as "could not be measured" or "evaluation is not optimal." Consequently, most of these predictions tend to be misclassified as "Obs_Absent." The BERTurk model demonstrated the best performance for this specific class label because it is specially trained for the Turkish language. However, the unexpected underperformance of BioBERTurk in predicting the "Obs_Uncertain" label is noteworthy, given its pre-training on Turkish biomedical data. This performance discrepancy warrants a closer examination of pre-training data specificity.

The F1 score of 89.0 for Polish radiology reports in[13] closely aligns with our obtained score of 80.1. The dataset sizes are similar; ours has 1,056 instances, whereas theirs has 1,200. In addition, as in our study, certain classes are high frequency and yield lower F1 scores. We believe that the limited linguistic resources in both the Polish and Turkish languages specific to radiology reporting are the reason for this commonality. The F1 score reported by Sugimoto et al.[20] on Japanese data exceeds ours, and this difference can be attributed to the substantial amount of fine-tuning data they used, totaling over 100,000 reports. In our study, we faced constraints in conducting extensive fine-tuning due to the limited data available. The discrepancy in fine-tuning resources underscores the impact of data volume on model performance and highlights the importance of considering the scale of training data in achieving optimal results. Despite these differences, we see parallel trends in the outcomes of certainty labels.

This study has several limitations that warrant consideration. First, the dataset used in this study was synthetic, created by radiologists to mimic actual Turkish radiology reports. This limitation could affect the generalizability of the findings to real-world applications. Second, a larger dataset, including actual anonymized reports, could enhance the robustness and performance of the models, particularly in identifying less frequent entity labels such as "Symptom_A." Third, although the study focuses on Turkish radiology reports, the findings may not be directly applicable to other low-resource languages without language-specific adaptations. Similar adjustments would be necessary for other languages with unique linguistic features. Fourth, despite the strong performance of the BioBERTurk model, the study was limited to evaluating only four BERT-based models. Exploring additional model architectures or integrating ensemble approaches could potentially yield improved results. Finally, due to resource constraints, fine-tuning was performed with limited training data. Access to a larger corpus of Turkish biomedical texts or radiology reports could further optimize the performance of the deep learning models.

In our future research, we plan to use the mentioned insights to propose a pretrained BERT model for biomedical applications in Turkish. We also plan to develop a language-specific approach to determine optimal token span lengths during adaptive span enumeration. These initiatives will enhance the accuracy and efficiency of information extraction models, as demonstrated in our research. Based on recent developments in artificial intelligence (AI), mainly in large language models, we also plan to experiment with these models, such as GPT-4o and Llama 3, and with different sized models, and compare their performances.

In conclusion, our study highlights the critical role of language-specific adaptations and domain-relevant fine-tuning in enhancing NER for Turkish radiology reports. The introduction of BioBERTurk and the adaptive span enumeration mechanism proved instrumental in achieving the highest performance among the tested models. By experimentally determining an optimal span length tailored to the Turkish language, we demonstrated the necessity of customizing hyperparameters to accommodate linguistic features such as agglutination and complex suffix structures. Furthermore, this research is built on the first-ever NER dataset derived from Turkish radiology reports, a resource labeled by radiology experts. This dataset not only reflects the unique linguistic and domain-specific challenges of Turkish but also lays the groundwork for future advancements in low-resource medical NLP. Our work also underscores the challenges posed by limited annotated datasets and the importance of future efforts in expanding high-quality medical text resources. By leveraging advances in large language models and further fine-tuning with domain-specific data, we aim to push the boundaries of information extraction in low-resource languages. Ultimately, this research contributes to the development of AI tools that streamline clinical workflows, improve diagnostic precision, and enhance patient care. We hope our research contributes to continued innovation that enables healthcare practitioners to access standardized and structured data to improve patient care.

Acknowledgments

We would like to express our gratitude to the two independent radiologists (Yasin Celal Güneş, Mehmet Numan Çolakoğlu) who contributed to this study by evaluating the reports and providing their valuable assessments.

We have provided our data, code, and model files as supplementary links: https://drive.google.com/drive/folders/1VWWRhxd-KkHTvSWD1KreEEAKRcofOxqbo

GitHub: https://github.com/BIGDaTA-Lab-AI/dygiepp-multilingual-radiology

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*. 2016;1937-1945. [CrossRef]
2. Kahn CE Jr, Langlotz CP, Burnside ES, et al. Toward best practices in radiology reporting. *Radiology*. 2009;252(3):852-856. [CrossRef]
3. Langlotz CP. RadLex: a new method for indexing online educational materials. *Radiographics*. 2006;26(6):1595-1597. [CrossRef]
4. Wadden D, Wennberg U, Luan Y, Hajishirzi H. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint. arXiv:1909.03546*, 2019. [CrossRef]
5. Devlin J, Chang MV, Lee K, Toutanova K. Bert: pretraining of deep bidirectional transformers for language understanding. *arXiv preprint. arXiv:1810.04805*, 2018. [CrossRef]
6. Türkmen H, Dikenelli O, Eraslan C, Callı MV, Özbek SS. BioBERTurk: Exploring Turkish Biomedical Language Model¨ Development Strategies in a Low-Resource Setting. *J Healthc Inform Res*. 2023;7(4):433-446. [CrossRef]
7. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint. arXiv:1711.05101*, 2017. [CrossRef]
8. Gardner M, Grus J, Neumann M, et al. Allennlp: a deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*. Published 2018. Accessed December 23, 2024. [CrossRef]
9. Soysal E, Cicekli I, Baykal N. Design and evaluation of an ontology-based information extraction system for radiological reports. *Comput Biol Med*. 2010;40(11-12):900-911. [CrossRef]
10. Schweter S. BERTurk - BERT models for Turkish. *Software*. [CrossRef]
11. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23. [CrossRef]
12. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. *arXiv preprint. arXiv:1911.02116*, 2019. [CrossRef]
13. Jantscher M, Gunzer F, Kern R, Hassler E, Tschauner S, Reishofer G. Information extraction from German radiological reports for general clinical text and language understanding. *Sci Rep*. 2023;13(1):2353. [CrossRef]
14. Manjil Shrestha. Development of a language model for the medical domain. *PhD Thesis*, Hochschule Rhein-Waal, 2021. [CrossRef]
15. Obuchowski A, Klaudel B, Jasik P. Information extraction from Polish radiology reports using Language models. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*. 2023;113-122. [CrossRef]
16. Dadas S, Perelkiewicz M, Poswiata R. Pre-training Polish Transformer-Based Language Models at Scale. *Artificial Intelligence and Soft Computing*, 2020. Springer International Publishing. pages 301-314. ISBN: 9783-030-61534-5. [CrossRef]
17. Mroczkowski R, Rybak P, Wroblewska A, Gawlik I. HerBERT: efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, April 2021, Kiyv, Ukraine. Association for Computational Linguistics, pages 1–10. [CrossRef]
18. Ri R, Yamada I, Tsuruoka Y. mLUKE: the power of entity representations in multilingual pretrained language models. *arXiv preprint. arXiv:2110.08151*, 2021. [CrossRef]
19. Yamada I, Asai A, Shindo H, Takeda H, Matsumoto Y. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020. [CrossRef]
20. Sugimoto K, Takeda T, Oh JH, et al. Extracting clinical terms from radiology reports with deep learning. *J Biomed Inform*. 2021:116:103729. http://doi.org/10.1016/j.jbi.2021.103729. [CrossRef]
21. Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. *Proc Int Conf Mach Learn*. 2019;2790-2799. [CrossRef]
22. Issifu AM, Ganiz MC. A simple data augmentation method to improve the performance of named entity recognition models in the medical domain. *Proc 6th Int Conf Comput Sci Eng (UBMK)*. 2021;763-768. [CrossRef]

Appendix 1. Labels results for BERTurk model

Labels	Precision	Recall	F1
Obs present	0.717	0.697	0.706
Obs absent	0.890	0.899	0.894
Obs uncertain	0.559	0.352	0.432
Obs technical	0.708	0.723	0.716
Obs anatomy	0.849	0.885	0.866
Obs advice	0.478	0.550	0.512
Symptom P	0.688	0.579	0.629
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.580	0.797	0.671

Appendix 2. Labels results for BioBERTurk model

Labels	Precision	Recall	F1
Obs present	0.702	0.691	0.697
Obs absent	0.883	0.874	0.879
Obs uncertain	0.560	0.519	0.538
Obs technical	0.787	0.787	0.787
Obs anatomy	0.846	0.876	0.861
Obs advice	0.520	0.650	0.578
Symptom P	0.647	0.579	0.611
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.585	0.814	0.681

Appendix 3. Labels results for PubMedBERT model

Labels	Precision	Recall	F1
Obs present	0.673	0.635	0.653
Obs absent	0.884	0.884	0.884
Obs uncertain	0.500	0.407	0.449
Obs technical	0.708	0.723	0.716
Obs anatomy	0.805	0.866	0.835
Obs advice	0.440	0.550	0.489
Symptom P	0.533	0.421	0.471
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.536	0.763	0.629

Appendix 4. Labels results for XLM-RoBERTa model

Labels	Precision	Recall	F1
Obs present	0.695	0.616	0.653
Obs absent	0.892	0.879	0.886
Obs uncertain	0.606	0.370	0.459
Obs technical	0.867	0.828	0.848
Obs anatomy	0.853	0.876	0.864
Obs advice	0.476	0.500	0.488
Symptom P	0.750	0.474	0.581
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.628	0.831	0.715

Appendix 5. A co-occurrence matrix showing the total number of times each label pair appeared together across all reports. Repetitions within the same document are included in the calculations

	Obs_ Absent	Obs_ Technical	Obs_ Anatomy	Obs_ Present	Differential diagnosis	Obs_ Advice	Obs_ Uncertain	Symptom_P	Symptom_A
Obs_Absent	32694	14895	167727	125985	10489	4715	10774	7010	164
Obs_Technical	14895	6950	81788	62152	5077	2413	5763	3179	83
Obs_Anatomy	167727	81788	918308	708388	56819	26141	60585	35245	851
Obs_Present	125985	62152	708388	543428	43728	20306	46456	26866	697
Differential diagnosis	10489	5077	56819	43728	3898	1785	3549	2375	65
Obs_Advice	4715	2413	26141	20306	1785	752	1896	972	20
Obs_Uncertain	10774	5763	60585	46456	3549	1896	4258	2332	48
Symptom_P	7010	3179	35245	26866	2375	972	2332	2190	47
Symptom_A	164	83	851	697	65	20	48	47	2

Appendix 6. A co-occurrence matrix showing the frequency of unique label pairs. Each pair is counted only once, regardless of how many times it appears within or across documents

	Obs_ Absent	Obs_ Technical	Obs_ Anatomy	Obs_ Present	Differential diagnosis	Obs_ Advice	Obs_ Uncertain	Symptom_P	Symptom_A
Obs_Absent	0	332	334	334	316	254	304	246	15
Obs_Technical	332	0	332	332	315	253	303	245	15
Obs_Anatomy	334	332	0	334	316	254	304	246	15
Obs_Present	334	332	334	0	316	254	304	246	15
Differential diagnosis	316	315	316	316	0	244	288	234	15
Obs_Advice	254	253	254	254	244	0	232	182	10
Obs_Uncertain	304	303	304	304	288	232	0	223	13
Symptom_P	246	245	246	246	234	182	223	0	14
Symptom_A	15	15	15	15	15	10	13	14	0

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243182

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis

Burak Koçak[1]
Fadime Köse[1]
Ali Keleş[1]
Abdurrezzak Şendur[1]
İsmail Meşe[2]
Mehmet Karagülle[1]

[1]University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

[2]Üsküdar State Hospital, Department of Radiology, İstanbul, Türkiye

PURPOSE

To comprehensively assess Checklist for Artificial Intelligence in Medical Imaging (CLAIM) adherence in medical imaging artificial intelligence (AI) literature by aggregating data from previous systematic and non-systematic reviews.

METHODS

A systematic search of PubMed, Scopus, and Google Scholar identified reviews using the CLAIM to evaluate medical imaging AI studies. Reviews were analyzed at two levels: review level (33 reviews; 1,458 studies) and study level (421 unique studies from 15 reviews). The CLAIM adherence metrics (scores and compliance rates), baseline characteristics, factors influencing adherence, and critiques of the CLAIM were analyzed.

RESULTS

A review-level analysis of 26 reviews (874 studies) found a weighted mean CLAIM score of 25 [standard deviation (SD): 4] and a median of 26 [interquartile range (IQR): 8; 25th–75th percentiles: 20–28]. In a separate review-level analysis involving 18 reviews (993 studies), the weighted mean CLAIM compliance was 63% (SD: 11%), with a median of 66% (IQR: 4%; 25th–75th percentiles: 63%–67%). A study-level analysis of 421 unique studies published between 1997 and 2024 found a median CLAIM score of 26 (IQR: 6; 25th–75th percentiles: 23–29) and a median compliance of 68% (IQR: 16%; 25th–75th percentiles: 59%–75%). Adherence was independently associated with the journal impact factor quartile, publication year, and specific radiology subfields. After guideline publication, CLAIM compliance improved ($P = 0.004$). Multiple readers provided an evaluation in 85% (28/33) of reviews, but only 11% (3/28) included a reliability analysis. An item-wise evaluation identified 11 underreported items (missing in ≥50% of studies). Among the 10 identified critiques, the most common were item inapplicability to diverse study types and subjective interpretations of fulfillment.

CONCLUSION

Our two-level analysis revealed considerable reporting gaps, underreported items, factors related to adherence, and common CLAIM critiques, providing actionable insights for researchers and journals to improve transparency, reproducibility, and reporting quality in AI studies.

CLINICAL SIGNIFICANCE

By combining data from systematic and non-systematic reviews on CLAIM adherence, our comprehensive findings may serve as targets to help researchers and journals improve transparency, reproducibility, and reporting quality in AI studies.

KEYWORDS

Artificial intelligence, machine learning, checklist, diagnostic imaging, radiology

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Received 11 December 2024; revision requested 04 January 2025; last revision received 19 January 2025; accepted 22 January 2025.

Epub: 10.02.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.243182

You may cite this article as: Koçak B, Köse F, Keleş A, Şendur A, Meşe İ, Karagülle M. Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis. *Diagn Interv Radiol.* 2025;31(5):440-455.

With the exponential increase in artificial intelligence (AI) publications related to medical imaging,[1] ensuring transparency and reproducibility has become crucial for advancing the field and integrating AI into clinical practice.[2-4] To address these needs, various AI-focused reporting guidelines have been introduced,[5-7] one of which is the Checklist for Artificial Intelligence in Medical Imaging (CLAIM).[8] Published in March 2020, the CLAIM was designed to improve reporting clarity and scientific communication in medical imaging AI.[8] Inspired by the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines,[9] the original 2020 version of the CLAIM featured a 42-item checklist to help authors and reviewers achieve clear, comprehensive, and reproducible reporting in AI studies. In May 2024, an updated CLAIM was published following a formal Delphi process, refining the checklist to 44 items to address new challenges and developments while retaining the original structure.[10] The update included refinements to terminology and revisions to some items. The CLAIM is part of the EQUATOR network, a central hub for reporting guidelines.[11]

Since its release, the CLAIM has gained widespread attention across multiple medical specialties involving imaging and AI, with over 850 citations in Google Scholar as of January 2025. Despite its popularity, assessments of CLAIM adherence remain highly variable,[12-14] often with particular focus on specific diseases,[15-18] techniques,[19-21] or individual journals.[22] A comprehensive assessment of CLAIM adherence across these diverse studies is notably lacking. Such an analysis, previously applied to frameworks such as the Radiomics Quality Score (RQS),[23] would reveal the CLAIM's overall adherence patterns, highlight underreported items, and provide guidance for future revisions beyond the 2024 CLAIM update,[10] along with the development of new, alternative AI checklists.

Main points

- To our knowledge, no prior research has synthesized data from published reviews on Checklist for Artificial Intelligence in Medical Imaging (CLAIM) adherence, leaving a gap in providing a comprehensive overview independent of disease, technique, or journal.
- Our two-level analysis identified significant reporting gaps in the medical imaging artificial intelligence literature, with a third of CLAIM items omitted, on average.
- Eleven specific CLAIM items were identified as being consistently underreported in the majority of studies, highlighting critical areas for improvement.
- Factors such as the publication year, journal impact quartile, and the radiology subfield influenced CLAIM adherence.
- Reviews assessing CLAIM adherence exhibited variability in their methodologies, with some using scoring and others focusing on compliance, leading to inconsistencies in evaluation and reporting.

This study aims to comprehensively assess CLAIM adherence in the medical imaging AI literature published to date using a two-level approach: review level and study level. The review-level analysis aggregates data from previous systematic and non-systematic reviews, whereas the study-level analysis examines unique individual papers within these reviews, mostly focusing on checklist items. Furthermore, factors influencing high or low CLAIM adherence are examined at the study level. Finally, critiques of the CLAIM guidelines are systematically analyzed across eligible reviews for both levels.

Methods

Literature search and screening

A literature search was conducted through PubMed, Scopus, and Google Scholar to identify reviews on the application of the CLAIM[8] using the syntax "Checklist for Artificial Intelligence in Medical Imaging." The final search was performed on August 6, 2024. Since the search syntax was simple, we did not use advanced database features to target specific fields (e.g., title, abstract, or keywords). Instead, we used the general search box, which typically searches across all fields in the database entries.

For Google Scholar, the first 100 results were screened based on the filter setting "relevance," whereas all entries were reviewed in the other two databases. Google Scholar can provide valuable additions to systematic reviews, even when screening is limited to the top 100 results.[24] Because its "relevance"-based ranking typically prioritizes the most pertinent articles, this approach was chosen to manage the large volume of results often retrieved from Google Scholar, many of which include duplicates or less relevant entries. Notably, Google Scholar was treated as a supplementary source to mitigate the risk of missing key papers, complementing the more comprehensive searches conducted in PubMed and Scopus, where all entries were reviewed.

Three readers (F.K., A.K., and A.S.; all 3rd- or 4th-year radiology residents) initially screened all records to identify review articles evaluating medical imaging AI studies using the CLAIM (2020 version).[8] Records were excluded if they lacked a CLAIM evaluation (2020 version),[8] full-text access, and relevance to medical imaging; relied on self-reported data; or had significant overlap with another study. Each reader cross-checked another reader's results.

Duplicates were removed using Zotero software. The full-text articles and available supplements were downloaded for evaluation by the same three readers, who divided the workload equally. For articles where full-text access was unavailable through our institutional libraries, we tried to reach out directly to the authors to request access.

Eligibility

After the initial screening, articles were evaluated for eligibility by the same three readers under the supervision of a radiology specialist experienced in informatics and AI (B.K.). For the review-level analysis, reviews with adequate adherence data on the 42-item CLAIM were included; those with incomplete or unclear data were excluded. For the study-level analysis, only reviews with 42-item CLAIM data for each study (i.e., a completed checklist for each study) were included. Duplicate and retracted studies, along with the studies with unclear references to their source articles, were removed. Papers using a modified 42-item CLAIM with subsections that retained the main framework were included in the study-level analysis but excluded from the review-level analysis unless CLAIM adherence could be evaluated at that level.

Analyzing data at the individual study level was crucial to gain item-level insights as well as several other baseline characteristics, as this level of granularity could not have been achieved through a review-level-only analysis. Although we acknowledge the potential limitations of using a highly selected sample, this approach was necessary to address the study's objectives and provide meaningful insights at the desired level of detail.

Data extraction

For the review-level analysis, data extraction was initially performed by a radiology specialist experienced in informatics and AI (B.K.) and was subsequently confirmed by another radiology specialist (M.K.). Extracted data included the review's scope, radiology subfield, number of studies (or evaluations) in the reviews, online publication year, number of readers, reader independence, decision-making methods, reproducibility analysis, consideration of non-applicable (n/a) items in the adherence evaluation, CLAIM adherence evaluation method, and source of the CLAIM evaluation.

For the study-level analysis, the three radiology residents independently extracted and cross-checked the data. The cross-checking was performed by having the readers review and validate one another's work. In cases of disagreement, an experienced reader (B.K.) was consulted to resolve the issue. Extracted information included the journal name, publication year, publication type, journal scope and focus, radiology subfield (expanded from the review-level data), journal's h5-index (from Google Scholar Metrics), 2023 impact factor quartile (2024 release; Journal Citation Reports, Clarivate Analytics, Web of Science Group), and CLAIM adherence by item.

Full-text articles, including the text, figures, tables, and supplements, were reviewed to identify adherence data, including item-specific CLAIM data, organized according to the original item order, if necessary. For adherence data sourced from the reviews, only studies with a clear source attribution were included. In cases of multiple rater evaluations, consensus data were prioritized; if unavailable, one evaluation (the first) was selected. In the study-level analysis, only one assessment per study was included when multiple pipelines were assessed, whereas all assessments were considered in the review-level analysis, which are referred to as "studies" in this research. For studies using a modified CLAIM with subsections within a 42-item framework, an item was considered reported if ≥50% of its subitems were positively evaluated. Partially reported items were classified as reported, in alignment with the common standard checklist format (i.e., reported, not reported, and not applicable).

Two radiology specialists with experience in informatics and AI (B.K. and İ.M.) evaluated the review papers in both the review-level and study-level analyses for critiques about the CLAIM. The PDFs were then screened using Google's NotebookLM tool, with various targeted prompts to identify additional critiques and to minimize the risk of missing important ones. The results from this additional screening were double-checked by both readers, verified against their sources, and integrated with the initial human evaluation findings.

Adherence metrics

This study applied two commonly used CLAIM adherence metrics: the CLAIM score and CLAIM compliance. The CLAIM score represents the total number of reported items, whereas CLAIM compliance is calculated as the percentage of reported items relative to the total applicable CLAIM items.

For the study-level analysis, these two metrics were calculated directly from the extracted item-level data. In the review-level analysis, metrics were extracted as a mean and used as reported when directly provided; if not, they were derived from tables, figures, or supplementary files where possible, converted from the median and interquartile range (IQR), if necessary, according to the methods proposed by Luo et al.[25] and Wan et al.[26], or computed as weighted combinations when presented by category.

Statistical analysis

Statistical analysis was conducted using R (main packages: ggstatsplot and Hmisc) and JASP (version 0.19.1; Apple Silicon). Descriptive statistics, including frequency, percentage, mean, standard deviation (SD), median, IQR, and 25th–75th percentiles, were reported based on variable distribution. In the review-level analysis, adherence metrics were weighted by the number of studies or evaluations using the "Hmisc" R package and presented using both the mean and median without considering statistical normality. For the study-level data, normality was tested with the Shapiro–Wilk test, and the associated statistical results are presented accordingly. In addition, differences between continuous variables were assessed using the Mann–Whitney U test or Student's t-test based on distribution. The Kruskal–Wallis test was applied to compare multiple categories, with Dunn's post-hoc tests and the Bonferroni correction. Correlations were assessed with Spearman's rho. Univariable and multivariable logistic regression was performed to identify the potential factors related to high and low CLAIM adherence metrics according to the median. No multiplicity correction was performed in the logistic regression analyses due to the exploratory nature of the study. Statistical significance was set at $P < 0.05$.

Results

Literature search

Figure 1 summarizes the eligibility process. Finally, 33 eligible reviews encompassing 1,458 study evaluations were included in the review-level analysis. For the study-level analysis, 15 reviews (13 from the previous set and 2 additional reviews) were included, covering 421 unique eligible studies. In total, 35 reviews met the eligibility criteria for both levels of analysis (Table 1).[12-22,27-50] The final dataset used in this study is publicly available from the Open Science Framework and can be accessed via the following link: https://osf.io/rx67y/

Baseline characteristics of papers eligible for the review-level analysis

The baseline characteristics of the 33 papers included in the review-level analysis are summarized in Table 2.

Multiple readers conducted CLAIM evaluations in 85% of reviews (28/33), with most assessments (79%, 22/28) performed independently and finalized by consensus (82%, 23/28). A reliability analysis was included in only a few multi-reader studies (11%, 3/28). One study reported an intraclass correlation coefficient (ICC) above 0.87 for inter-observer reliability across task categories.[46] Another study found an ICC of 0.815 for inter-observer reliability, with varying kappa values for individual items.[14] A third study reported an intra-observer repeatability coefficient of 0.22, which was lower and better than that of other checklists evaluated, except one.[31]

Figure 2 highlights the consideration of item applicability in the included reviews, along with the resultant metrics from this study. Regarding CLAIM adherence, 55% (18/33) of reviews considered the applicability of items, allowing for the calculation of a CLAIM compliance metric. For approximately 79% (26/33) of the reviews, appropriate data to calculate CLAIM scores were available, although the origin of the scores varied, with only 36% (12/33) providing direct reports.

Adherence based on the review-level analysis

Among the 26 reviews with available CLAIM scores, encompassing 874 studies, the weighted mean CLAIM score was 25 (SD: 4), and the weighted median was 26 (IQR: 8; 25th–75th percentiles: 20–28). For the 18 reviews providing CLAIM compliance data, covering 993 studies, the weighted mean CLAIM compliance was 63% (SD: 11%), with

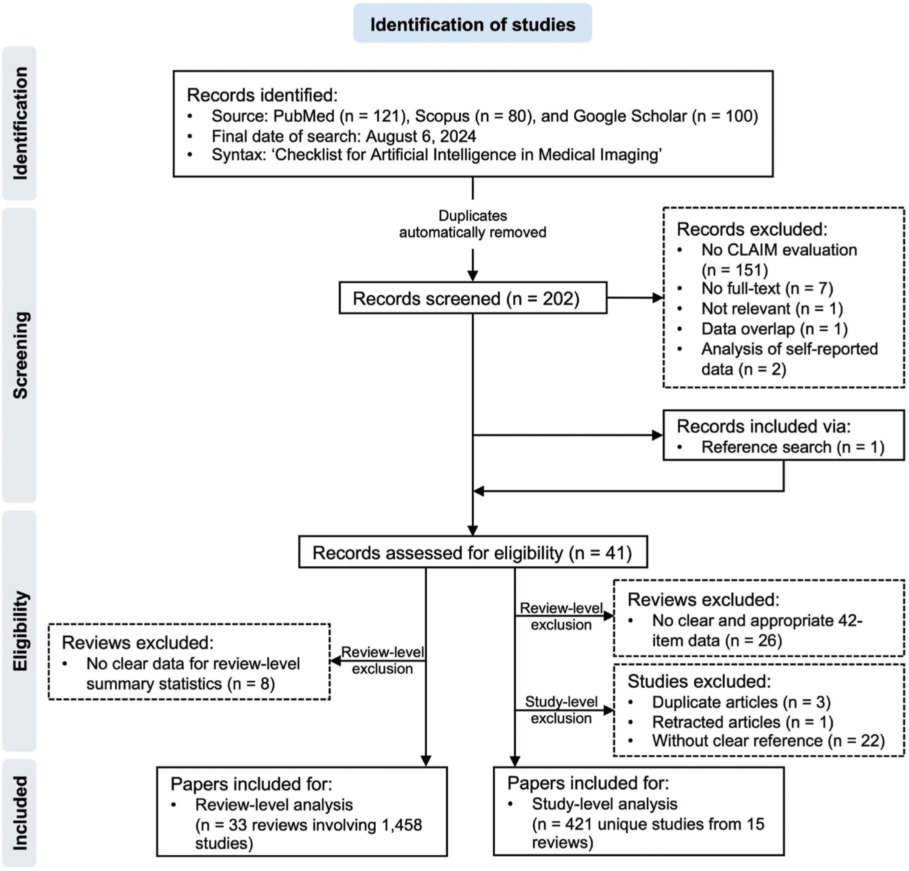

Identification of studies
Identification
Screening
Eligibility
Included
Records identified:
Source: PubMed (n = 121), Scopus (n = 80), and Google Scholar (n = 100)
Final date of search: August 6, 2024
Syntax: 'Checklist for Artificial Intelligence in Medical Imaging'
Duplicates automatically removed
Records screened (n = 202)
Records excluded:
No CLAIM evaluation (n = 151)
No full-text (n = 7)
Not relevant (n = 1)
Data overlap (n = 1)
Analysis of self-reported data (n = 2)
Records included via:
Reference search (n = 1)
Records assessed for eligibility (n = 41)
Review-level exclusion
Reviews excluded:
No clear and appropriate 42-item data (n = 26)
Reviews excluded:
No clear data for review-level summary statistics (n = 8)
Review-level exclusion
Study-level exclusion
Studies excluded:
Duplicate articles (n = 3)
Retracted articles (n = 1)
Without clear reference (n = 22)
Papers included for:
Review-level analysis (n = 33 reviews involving 1,458 studies)
Papers included for:
Study-level analysis (n = 421 unique studies from 15 reviews)

Figure 1. Identification of eligible studies for the review- and study-level analyses. CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

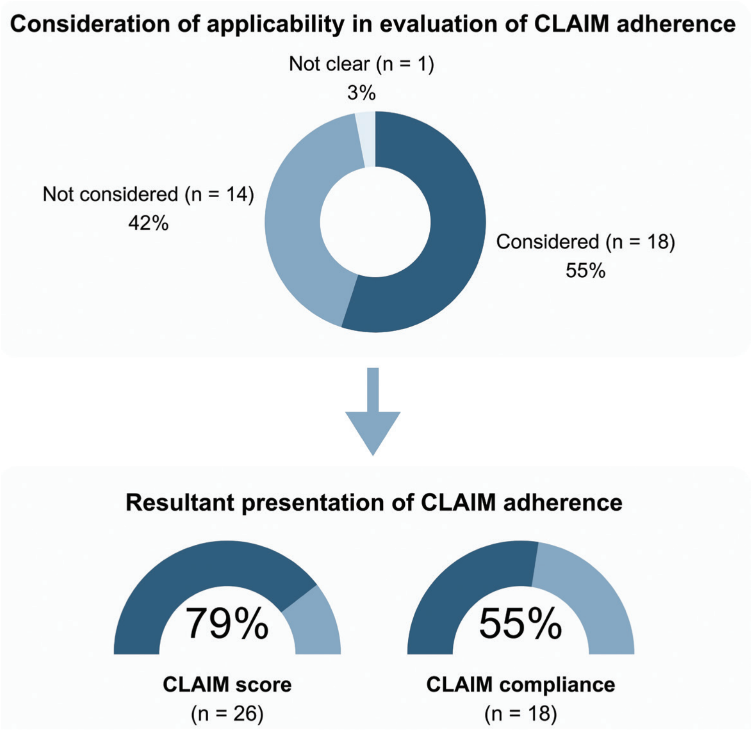


Figure 2. Consideration of item applicability and resultant CLAIM adherence metrics in the review-level analysis, emphasizing the methodological variability among reviews evaluating CLAIM adherence. CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

a weighted median of 66% (IQR: 4%; 25th–75th percentiles: 63%–67%).

Baseline characteristics of papers eligible for the study-level analysis

The baseline characteristics of the papers included in the study-level analysis are summarized in Table 3. Publication dates ranged from 1997 to 2024.

Adherence based on the study-level analysis

In the study-level analysis of 421 unique studies, the median CLAIM score was 26 (IQR: 6; 25th–75th percentiles: 23–29), and the median CLAIM compliance was 68% (IQR: 16%; 25th–75th percentiles: 59%–75%). Notably, 11% of the studies (47/421) had a CLAIM score of <21 (i.e., 50% of 42), whereas 10% (40/421) reported a CLAIM compliance of <50%.

Figure 3 illustrates the median CLAIM scores and compliance by journal and publication volume. Among the top 10 journals by publication volume, *Radiology* had the highest median CLAIM score and compliance rate.

Table 4 presents the results from the univariable and multivariable logistic regression analyses to identify factors linked to high and low CLAIM adherence. In the univariable analysis, the publication year, specific radiology subfields, journal h5-index, and certain impact factor quartiles were associated with the CLAIM score or compliance. In the multivariable analysis, the publication year and impact factor quartile emerged as independent predictors of the CLAIM score and compliance. Specifically, publishing in a first quartile (Q1) journal independently predicted higher CLAIM scores and compliance, whereas second quartile (Q2) journals were associated with higher CLAIM compliance. Certain radiology subfields were additional independent predictors of the CLAIM score.

Figure 4a, b illustrate the correlation between the publication year and CLAIM adherence. Although the CLAIM score did not significantly correlate with the publication year (rho: 0.076, P = 0.117), CLAIM compliance showed a weak but significant positive correlation (rho: 0.119, P = 0.015). Although the CLAIM score did not significantly differ between the pre- and post-CLAIM guideline publication periods (P = 0.153), CLAIM compliance was higher post-publication (P = 0.004) (Figure 4c, d). However, neither the CLAIM score (rho: −0.027, P = 0.697)

Table 1. Reviews included in the analyses, detailing the authors, year, journal abbreviation, radiology subfield, and the number of papers or evaluations included in the review- and study-level analyses

Authors (online publication year)	Journals	Radiology subfield	No. of papers or evaluations[1]	
			Review level	Study level
Abdulaal et al.[15] (2024)	Front Radiol	Chest	5	5
Alabed et al.[19] (2022)	Front Cardiovasc Med	Cardiovascular	209	n/a
Alipour et al.[16] (2023)	Diagnostics (Basel)	Musculoskeletal	8	n/a
Assadi et al.[27] (2022)	Medicina (Kaunas)	Cardiovascular	5	5
Bedrikovetski et al.[28] (2022)	Eur J Radiol	General or multi-system	24	24
Belue et al.[12] (2022)	J Am Coll Radiol	Genitourinary	53	n/a
Belue and Turkbey[29] (2022)	Eur Radiol Exp	Genitourinary	47	n/a
Bhandari et al.[13] (2023)	Neuroradiology	Neuro	138	n/a
Bleker et al.[30] (2022)	Life (Basel)	Genitourinary	4	4
Cerdá-Alberich et al.[31] (2023)	Insights Imaging	General or multi-system	10	9
Dagher et al.[32] (2024)	J Neuroimaging	Neuro	6	n/a
Hardacre et al.[33] (2021)	Br J Radiol	Cardiovascular	3	3
Hickman et al.[34] (2021)	Radiology	Breast	14	n/a
Hu et al.[35] (2022)	Neuroradiology	Neuro	19	n/a
Hwang et al.[36] (2024)	Radiol Artif Intell	Chest	14	n/a
Jia et al.[37] (2022)	Eur J Radiol Open	Chest	19	7
Karabacak et al.[20] (2022)	Acta Radiol	Neuro	5	n/a
Karabacak et al.[38] (2022)	Quant Imaging Med Surg	Neuro	4	n/a
Kim et al.[22] (2023)	Korean J Radiol	General or multi-system	38	n/a
Kouli et al.[21] (2022)	Neurooncol Adv	Neuro	234	222
Lans et al.[39] (2022)	Artif Intell Med	Musculoskeletal	91	n/a
Le et al.[40] (2021)	Appl Sci	Dental	6	6
O'Shea et al.[41] (2021)	Eur Radiol	General or multi-system	186	n/a
Ozkara et al.[18] (2023)	Cancers (Basel)	Neuro	25	n/a
Raj et al.[42] (2024)	Indian J Orthop	Musculoskeletal	5	n/a
Roberts et al.[43] (2021)	Nat Mach Intell	Chest	37	37
Roest et al.[44] (2022)	Life (Basel)	Genitourinary	8	n/a
Si et al.[14] (2021)	Eur Radiol	Musculoskeletal	36	36
Sivanesan et al.[45] (2022)	Can Assoc Radiol J	General or multi-system	100	n/a
Sushentsev et al.[17] (2022)	Insights Imaging	Genitourinary	5	5
Tsang et al.[46] (2023)	Jpn J Radiol	Pediatric	21	21
Wang et al.[48] (2023)	Radiother Oncol	Neuro	42	n/a
Wang et al.[47] (2024)	Radiother Oncol	Chest	37	n/a
Zhong et al.[49] (2022)	Insights Imaging	Musculoskeletal	n/a	28
Zhong et al.[50] (2023)	J Orthop Surg Res	Musculoskeletal	n/a	9

[1] Values represent the total number of studies or evaluations (i.e., pipelines) included in our analysis after applying the eligibility criteria and therefore may not correspond exactly to the total number of studies reported in the respective papers. n/a, not available.

nor compliance (rho: –0.062, $P = 0.365$) was statistically significantly correlated with the publication year after the CLAIM guideline publication in 2020.

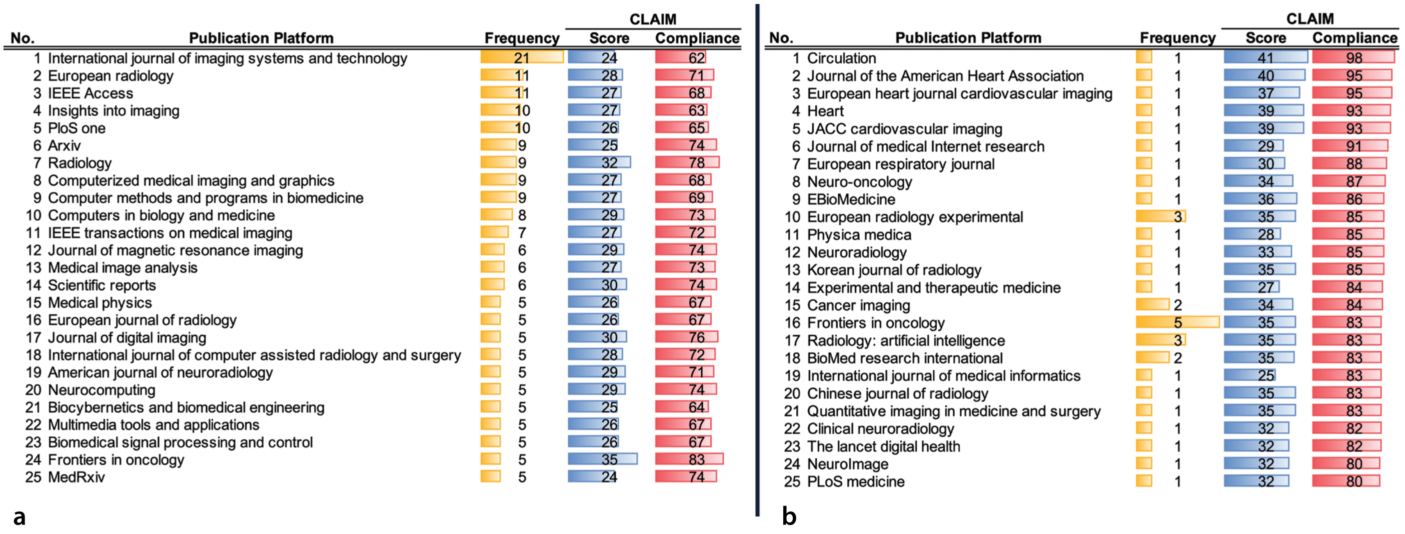
The CLAIM scores and compliance varied significantly across radiology subfields ($P < 0.001$ for both), with post-hoc pairwise comparisons showing that the cardiovascular subfield had consistently distinct results compared with others (Figure 5).

The CLAIM scores and compliance also differed by impact factor quartile ($P < 0.001$ for CLAIM score; $P = 0.002$ for CLAIM compliance) (Figure 6). The post-hoc analysis revealed that journals in Q1 and Q2 had significantly higher CLAIM scores than non-Web of Science indexed journals or publication platforms. However, CLAIM compliance did not show significant pairwise differences across quartiles.

Moreover, the CLAIM scores and compliance were not statistically significantly different among different publication types, such as journal articles, pre-prints, and conference papers ($P > 0.05$).

The item-wise CLAIM adherence is presented in Figure 7. Notably, three items were mostly n/a in ≥50% of the papers: item#10 (selection of data subsets, if applicable), item#21 (the level at which partitions are disjoint, e.g., image, study, patient, institution), and item#27 (ensemble techniques, if applicable).

Considering the applicability of the items, the following 11 items were not reported in ≥50% of the papers (i.e., compliance of <50%): item#12 (de-identification methods), item#13 (how missing data were handled), item#19 (intended sample size and how it was determined), item#29 (statistical measures of significance and uncertainty, e.g., confidence intervals), item#31 (methods for explainability or interpretability and how they were validated), item#33 (flow of participants or cases, using a diagram to indicate



a

No.	Publication Platform	Frequency	CLAIM Score	CLAIM Compliance
1	International journal of imaging systems and technology	21	24	62
2	European radiology	11	28	71
3	IEEE Access	11	27	68
4	Insights into imaging	10	27	63
5	PloS one	10	26	65
6	Arxiv	9	25	74
7	Radiology	9	32	78
8	Computerized medical imaging and graphics	9	27	68
9	Computer methods and programs in biomedicine	9	27	69
10	Computers in biology and medicine	8	29	73
11	IEEE transactions on medical imaging	7	27	72
12	Journal of magnetic resonance imaging	6	29	74
13	Medical image analysis	6	27	73
14	Scientific reports	6	30	74
15	Medical physics	5	26	67
16	European journal of radiology	5	26	67
17	Journal of digital imaging	5	30	76
18	International journal of computer assisted radiology and surgery	5	28	72
19	American journal of neuroradiology	5	29	71
20	Neurocomputing	5	29	74
21	Biocybernetics and biomedical engineering	5	25	64
22	Multimedia tools and applications	5	26	67
23	Biomedical signal processing and control	5	26	67
24	Frontiers in oncology	5	35	83
25	MedRxiv	5	24	74

b

No.	Publication Platform	Frequency	CLAIM Score	CLAIM Compliance
1	Circulation	1	41	98
2	Journal of the American Heart Association	1	40	95
3	European heart journal cardiovascular imaging	1	37	95
4	Heart	1	39	93
5	JACC cardiovascular imaging	1	39	93
6	Journal of medical Internet research	1	29	91
7	European respiratory journal	1	30	88
8	Neuro-oncology	1	34	87
9	EBioMedicine	1	36	86
10	European radiology experimental	3	35	85
11	Physica medica	1	28	85
12	Neuroradiology	1	33	85
13	Korean journal of radiology	1	35	85
14	Experimental and therapeutic medicine	1	27	84
15	Cancer imaging	2	34	84
16	Frontiers in oncology	5	35	83
17	Radiology: artificial intelligence	3	35	83
18	BioMed research international	2	35	83
19	International journal of medical informatics	1	25	83
20	Chinese journal of radiology	1	35	83
21	Quantitative imaging in medicine and surgery	1	35	83
22	Clinical neuroradiology	1	32	82
23	The lancet digital health	1	32	82
24	NeuroImage	1	32	80
25	PLoS medicine	1	32	80

Figure 3. Tabulated bar charts for the study-level analysis of the median CLAIM score and compliance by journal, sorted by publication frequency **(a)** and CLAIM compliance **(b).** CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

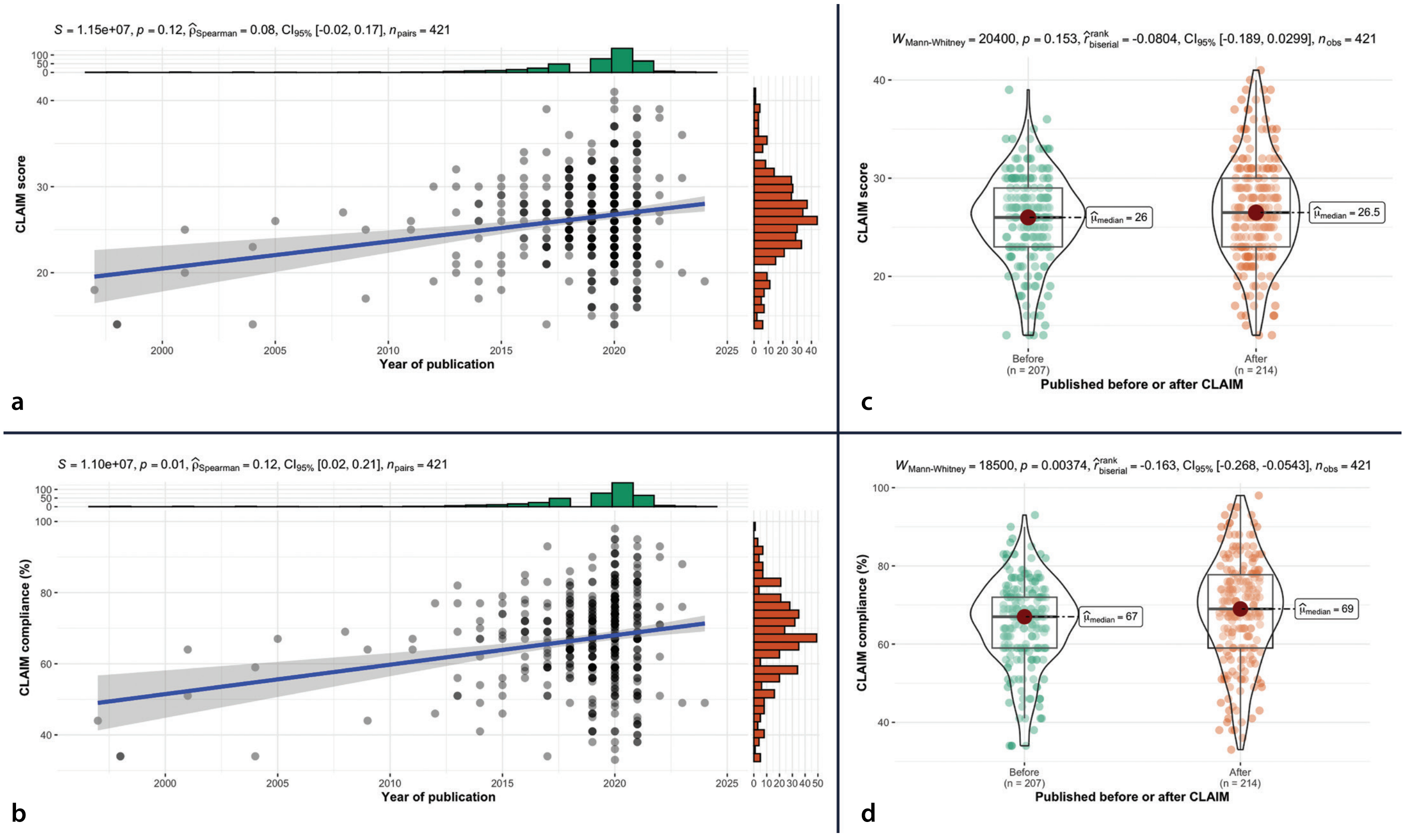


Figure 4. Study-level analysis of the publication year, CLAIM score, and compliance. Scatterplots with marginal distributions showing the correlation between the publication year and CLAIM score **(a)** and compliance **(b)**. Combined box and violin plots illustrating the CLAIM score **(c)** and compliance **(d)** in relation to the release of the CLAIM guidelines in 2020. CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CI, confidence interval.

inclusion and exclusion), item#34 (demographic and clinical characteristics of cases in each partition), item#36 (estimates of diagnostic accuracy and their precision), item#37 (failure analysis of incorrectly classified cases), item#40 (registration number and name of registry), and item#41 (where the full study protocol can be accessed). Figure 8 further highlights the above-mentioned 11 items categorized into three domains: data handling and description, model evaluation and performance, and open science.

The item-wise correlation results for reporting status and year are presented in Table 5, according to pre- and post-publication and post publication of the CLAIM. Considering the entire period, a positive weak-to-moderate and statistically significant reporting trend (rho ≥0.2) was observed for item#19 (intended sample size and how it was determined), item#21 (level at which partitions are disjoint), item#31 (methods for explainability or interpretability and how they were validated), item#33 (flow of participants or cases, using a diagram to indicate inclusion and exclusion), and item#42 (sources of funding and other support; role of funders). Moreover, a negative weak-to-moderate reporting trend (rho ≤−0.2) was observed for item#11 (definitions of data elements, with references to common data elements), item#15 (rationale for choosing the reference standard), item#17 (annotation tools), item#18 (measurement of inter- and intra-rater variability), and item#39 (implications for practice, including the intended use and/or clinical role). Considering the post-publication period, a positive weak-to-moderate reporting trend (rho ≥0.2) was observed in item#10 (selection of data subsets), item#19 (intended sample size and how it was determined), and item#33 (flow of participants or cases, using a diagram to indicate inclusion and exclusion). In addition, a negative weak-to-moderate reporting trend (rho ≤−0.2) was observed for item#9 (data pre-processing steps) and item#39 (implications for practice, including the intended use and/or clinical role).

Table 2. Baseline characteristics of eligible papers included in the review-level analysis

Characteristic	Sub-category	Value
Scope, count (%)	Broad (AI, ML, or deep learning)	22 (67%)
	Deep learning	9 (27%)
	Radiomics	2 (6%)
Radiology subfield, count (%)	Neuro	8 (24%)
	Chest	5 (15%)
	Genitourinary	5 (15%)
	General or multi-system	5 (15%)
	Musculoskeletal	4 (12%)
	Cardiovascular	3 (9%)
	Pediatric	1 (3%)
	Breast	1 (3%)
	Dental	1 (3%)
Number of papers within reviews, median (IQR; 25th–75th percentiles)	-	19 (36; 6–42)
Publication year (online), count (%)	2021	6 (18%)
	2022	15 (45%)
	2023	7 (21%)
	2024	5 (15%)
Number of readers, count (%)	Multiple	28 (85%)
	Single	4 (12%)
	Not clear	1 (3%)
Dependence of reading, count (%)	Independent	22 (67%)
	Not clear	6 (18%)
	Not applicable	5 (15%)
Final decision of reading, count (%)	Consensus	23 (70%)
	Not clear	5 (15%)
	Not applicable	5 (15%)
Reliability analysis, count (%)	No	25 (76%)
	Not applicable	5 (15%)
	Yes	3 (9%)
Source of CLAIM evaluation, count (%)	As reported	12 (36%)
	Calculated from table or figure data	15 (45%)
	As reported + calculated from table or figure	2 (6%)
	As reported with median–mean conversion	3 (9%)
	As reported with a weighted combination of different categories	1 (3%)

Percentages may not total 100% due to rounding. IQR, interquartile range; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; AI, artificial intelligence; ML, machine learning.

Critiques in reviews eligible for the entire study

In analyzing the 35 reviews that applied the CLAIM, we identified 10 key critiques, which we organized into 7 categories: fulfillment, applicability, feasibility and practicality, structure, interpretation, relative importance, and scoring. The most common critique was the inapplicability of certain items to all study types. Another frequent issue was the subjective nature of deciding whether an item was sufficiently reported. Table 6 presents all the critiques along with their representative source articles.

Discussion

Main findings and related implications

This study comprehensively evaluated CLAIM adherence in the medical imaging AI literature through a two-level approach: review- and study-level analyses. Considering both analyses, on average, one-third of CLAIM items were inadequately reported, indicating room for improvement in adhering to reporting guidelines. Since adherence was independently assessed rather than self-reported, efforts to improve compliance should focus on improving awareness and engagement among researchers in terms of transparent reporting practices through guidelines. Notwithstanding their well-known benefits,[51] recent meta-research shows that radiology, nuclear medicine, and medical imaging journals rarely mandate AI-specific guidelines, despite the CLAIM being the most recommended.[52,53] Journals can actively endorse and promote the CLAIM[8] and its updates[10] to improve reporting quality and transparency while ensuring proper checklist usage with auditing practices.[54,55]

Our correlation analysis revealed a very weak but positive trend between CLAIM compliance and publication year. Although compliance was higher in the post-publication period, the trend was not statistically significant. Long-term follow-up studies on checklists such as STARD have demonstrated slow but significant improvements in research reporting quality over time.[56] Although a similar trend was observed in our analysis, more time and data are needed to better understand this progression and assess the CLAIM's true impact.

Table 3. Baseline characteristics of eligible papers included in the study-level analysis

Variable	Category	Value
Radiology subfield, count (%)	Neuro	222 (53%)
	Musculoskeletal	73 (17%)
	Chest	49 (12%)
	General or multi-system	33 (8%)
	Pediatric	21 (5%)
	Genitourinary	9 (2%)
	Cardiovascular	8 (2%)
	Dental	6 (1%)
Publication type, count (%)	Journal article	403 (96%)
	Preprint	14 (3%)
	Conference paper	4 (1%)
Scope of journals, count (%)	Radiology or imaging-related	170 (40%)
	No	251 (60%)
Focus of journals, count (%)	AI-focused	14 (3%)
	No	407 (97%)
h-5 index of journal, median (IQR; 25th–75th percentiles)	-	67 (70; 44–113)
Impact factor quartile, count (%)	Q1	222 (53%)
	Q2	116 (28%)
	Q3	29 (7%)
	Q4	17 (4%)
	No	37 (9%)
Top 10 most frequent publication platform, count (%)	International journal of imaging systems and technology	21 (5%)
	European radiology	11 (3%)
	IEEE access	11 (3%)
	Insights into imaging	10 (2%)
	PloS one	10 (2%)
	Arxiv	9 (2%)
	Radiology	9 (2%)
	Computerized medical imaging and graphics	9 (2%)
	Computer methods and programs in biomedicine	9 (2%)
	Computers in biology and medicine	8 (2%)

IQR, interquartile range; AI, artificial intelligence.

We observed that adherence assessments in reviews often lacked consistency due to the absence of standardized methods. We identified two primary approaches, the CLAIM score and CLAIM compliance (%), differing by item applicability. To improve comparability and fairness in the evaluation of adherence, we strongly recommend prioritizing the CLAIM compliance rate over the CLAIM score in future evaluations. The compliance rate accounts for the applicability of individual items, which can vary between studies, thereby providing a more accurate and equitable assessment. Moreover, this approach could be formally recommended or mandated by the developers in future versions of the CLAIM to ensure consistent and standardized adherence evaluations.

Publication year, impact factor quartile, and radiology subfields were key independent predictors of high or low CLAIM adherence. Studies in higher-impact journals (Q1 and Q2) showed stronger adherence, underscoring their role in setting transparent reporting standards and enabling rigorous peer review. However, it should be acknowledged that high-quality research can also be published in lower-impact journals, and high-impact journals are not immune to poor-quality research. Factors contributing to stronger adherence in higher-impact journals may include stricter editorial and peer-review processes, greater visibility of reporting guidelines in these journals, and, potentially, a higher familiarity of authors with these standards. In this respect, encouraging CLAIM adoption, particularly in lower-impact journals, could help enhance reporting transparency and reproducibility. It is important to note, however, that these observations are based on assumptions and warrant further investigation.

In addition, certain subfields, such as cardiovascular imaging, exhibited unique adherence patterns, reflecting differences in the maturity of AI reporting practices. These findings may indicate the need for specific strategies to improve CLAIM adherence across diverse medical imaging subfields and ensure consistent reporting standards throughout the discipline. Further research may be required to investigate whether unique adherence patterns in certain subfields, such as cardiovascular imaging, could

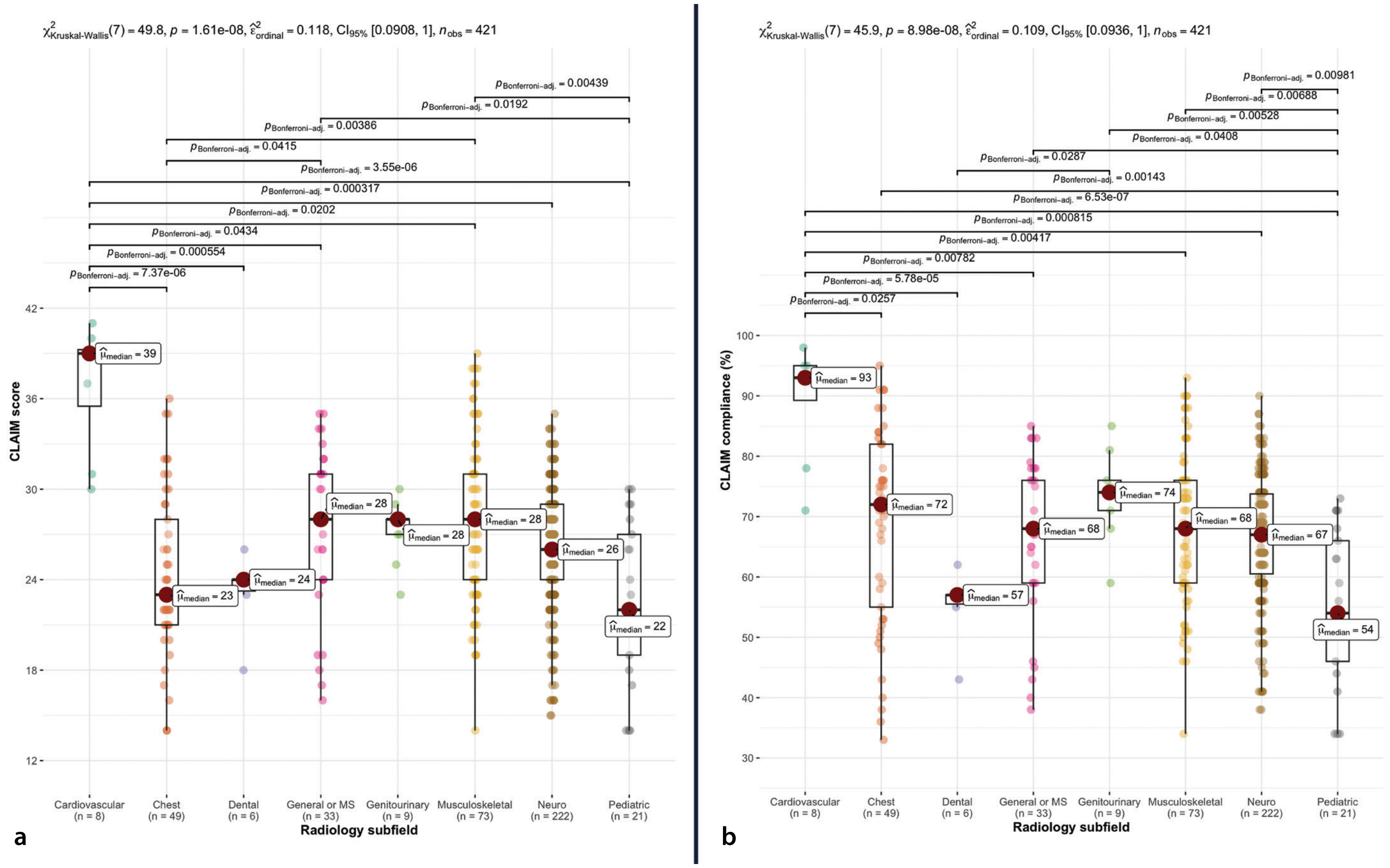


Figure 5. Box plots for the study-level analysis of the CLAIM score **(a)** and compliance **(b)** by radiology subfield, with pairwise comparisons. The Kruskal–Wallis test showed statistically significant differences across all categories in both analyses **(a, b)**. Only statistically significant pairwise comparisons are displayed for clarity. MS, multi-system; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CI, confidence interval.

Table 4. Univariable and multivariable analysis of the study-level data to identify factors related to high and low CLAIM adherence

Variable	Category[1]	Univariable analysis				Multivariable analysis			
		CLAIM score		CLAIM compliance		CLAIM score		CLAIM compliance	
		Estimate	*P*	Estimate	*P*	Estimate	*P*	Estimate	*P*
Publication year	-	0.069	**0.028**	0.092	**0.010**	0.110	**0.007**	0.095	**0.028**
Radiology subfield	Dental	−2.590	**0.026**	−16.748	0.986	−2.672	**0.025**	−16.669	0.986
	Cardiovascular	14.585	0.977	16.384	0.985	15.722	0.984	16.998	0.982
	Genitourinary	0.272	0.760	1.070	0.221	0.542	0.639	0.643	0.474
	Neuro	−0.598	0.149	−0.418	0.265	−0.336	0.483	−0.025	0.953
	Chest	−1.613	**0.001**	0.274	0.548	−1.963	**<0.001**	−0.079	0.876
	Pediatric	−1.466	**0.014**	−1.345	**0.030**	−1.241	0.059	−1.106	0.091
	Musculoskeletal	−0.267	0.565	−0.155	0.713	−0.361	0.487	−0.075	0.869
Publication type	Print	0.387	0.700	1.014	0.382	-	-	-	-
	Preprint	−0.916	0.430	1.686	0.188	-	-	-	-
Scope of journals	Radiology or imaging-related	0.314	0.123	0.278	0.163	-	-	-	-
Focus of journals	AI-focused	0.256	0.652	−0.534	0.346	-	-	-	-
h5 index of journal	-	0.005	**0.013**	0.004	**0.027**	0.003	0.246	0.001	0.771
Impact factor quartile of journal	Q1	1.387	**<0.001**	0.750	**0.040**	1.754	**0.018**	2.577	**0.017**
	Q2	1.154	**0.004**	0.289	0.456	1.414	0.053	2.152	**0.046**
	Q3	0.386	0.454	−0.302	0.565	0.836	0.296	1.547	0.173
	Q4	0.128	0.836	−1.044	0.148	0.277	0.764	1.165	0.350

P values achieving statistical significance are in bold. CLAIM, Checklist for Artificial Intelligence in Medical Imaging; AI, artificial intelligence.

[1] Reference categories not shown.

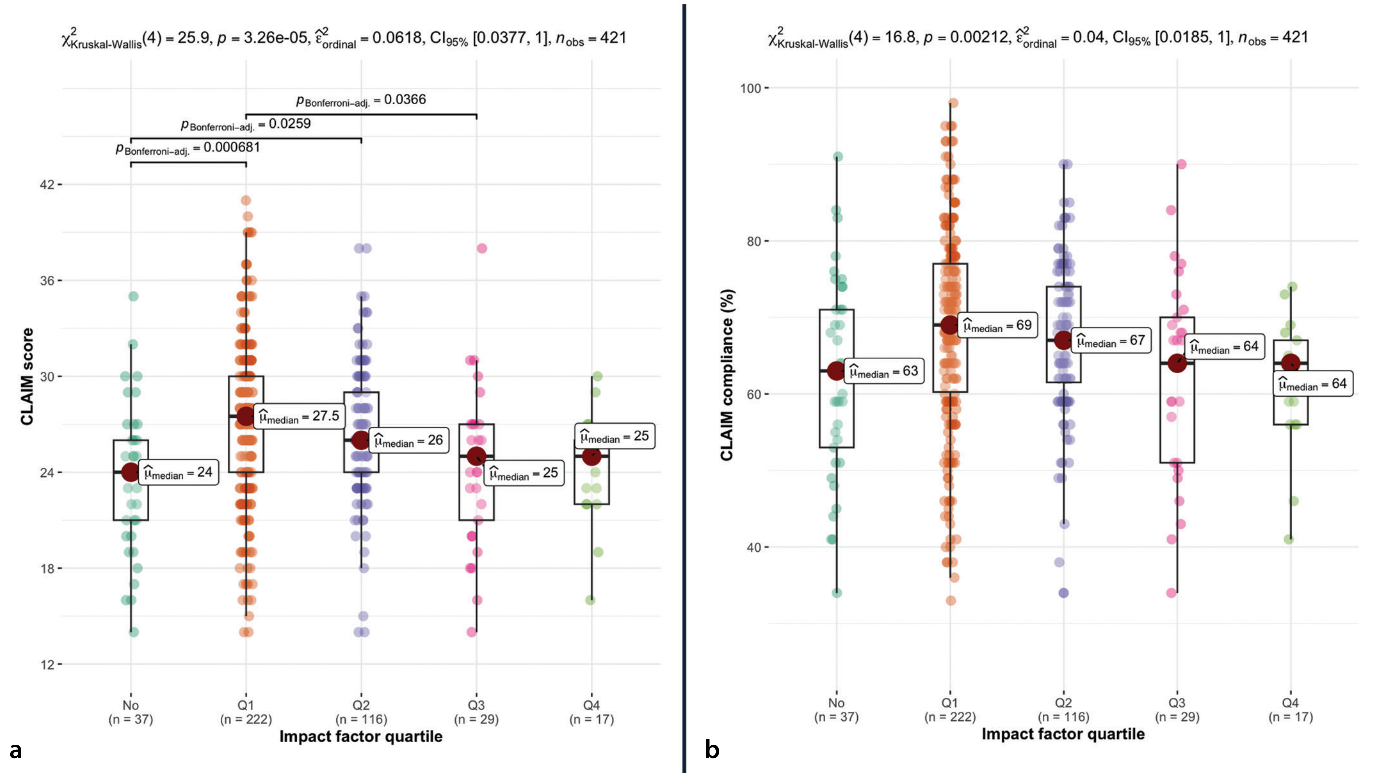


Figure 6. Box plots for the study-level analysis of the CLAIM score **(a)** and compliance **(b)** by impact factor quartile, with pairwise comparisons. The Kruskal–Wallis test showed statistically significant differences across all categories in both analyses **(a, b)**. Only statistically significant pairwise comparisons are displayed for clarity. CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CI, confidence interval.

be influenced by the contribution of specific authors or research groups.

Eleven items were underreported in ≥50% of studies: de-identification methods (item#12), missing data handling (item#13), sample size determination (item#19), statistical significance and uncertainty (item#29), explainability methods (item#31), participant flow (item#33), demographic data (item#34), diagnostic accuracy estimates (item#36), failure analysis (item#37), registration details (item#40), and protocol access (item#41). This suggests challenges in fulfilling the CLAIM requirements, possibly due to inadequate knowledge, training, resource limitations, or the perceived irrelevance of certain items for specific study types. Interestingly, several of these items reflect broader challenges in AI research, such as securing adequate sample sizes, addressing uncertainty, enhancing model explainability to avoid the "black-box" problem, and promoting principles of open science, even if not explicitly stated. These 11 items, therefore, warrant particular attention when preparing AI manuscripts to improve the overall reporting transparency and rigor of AI research in medical imaging.

From the 35 eligible reviews, several key critiques were identified, including concerns about the inapplicability of certain items to all study types and the subjective nature of reporting decisions. Although the CLAIM 2024 update has addressed applicability by introducing three checklist options and leaving judgment to the evaluators,[10] subjective interpretation still remains a significant issue. Notably, our analysis revealed that CLAIM evaluations involved multiple readers in 85% of reviews, but only 11% assessed evaluation reliability, revealing a critical gap. Despite high reported reproducibility, such assessments need improved experimental settings to thoroughly investigate interpretation-related issues, as previously achieved for RQS.[57] Additionally, leveraging automated tools, such as those powered by large language models used for RQS,[58] might have the potential to help reduce subjectivity and improve consistency.

Based on the other critiques identified, future versions of the CLAIM can also be improved by simplifying definitions and improving clarity, removing subjective items based on reproducibility studies with rigorous analysis, and providing holistic guidance for interpreting manuscripts alongside their code. Additional improvements could include prioritizing items by assigning weights through evidence-based voting methods and developing user-friendly online tools, similar to the METhodological RadiomICs Score (METRICS),[59] for an adherence assessment that considers item applicability. These refinements would help streamline CLAIM evaluations and improve their utility for the medical imaging community.

Previous studies

To the best of our knowledge, no research has yet been conducted to evaluate CLAIM adherence by synthesizing data from both systematic and non-systematic reviews, providing a comprehensive overview of the topic. However, similar efforts have been made in the field of radiomics research,[23,60,61] particularly with the RQS,[62] which is widely regarded as the standard for assessing the methodological quality of radiomics studies, although recent alternatives have emerged.[59]

In 2023, Spadarella et al.[60], who first published their research online in 2022, conducted a review-level analysis of 44 reviews. They reported a median RQS of 21%. Later, in late 2024, Kocak et al.[23] deepened the analysis by performing a study-level analysis of 1,574 unique papers from 89 reviews, finding a median RQS of 31%. In 2025, in another very recent coincidental and independent study, Barry et al.[61] conducted a multi-lev-

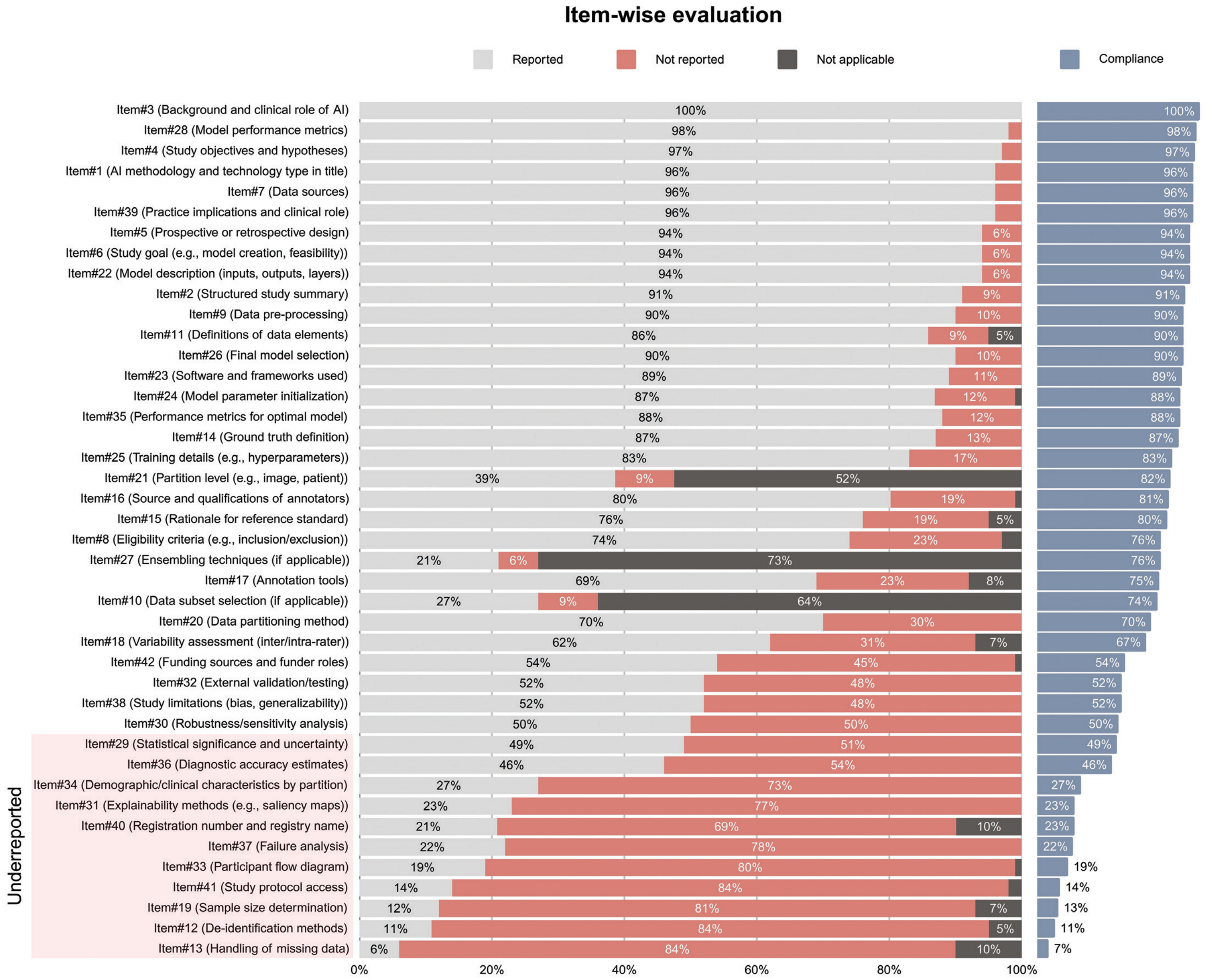


Figure 7. Item-wise analysis of the study-level data, ranked by compliance rates [calculated as follows: reported / (reported + not reported) × 100], considering the applicability of items. The compliance rates are based on the actual number of publications that reported or did not report each item. Note that item names have been abbreviated.

el meta-analysis of 3,258 RQS assessments from 130 systematic reviews as a continuation of the earlier study by Spadarella et al.[60], reporting an overall mean RQS of 9.4 ± 6.4 (95% confidence interval, 9.1–9.6) [26.1% ± 17.8% (25.3%–26.7%)]. It is important to note, however, that these RQS scores are not directly comparable to CLAIM adherence, as the two tools serve different purposes: RQS assesses the methodological quality of radiomics research, whereas the CLAIM focuses on reporting the quality of medical imaging AI research.

Furthermore, our results can be compared with those reported in the studies synthesized for this research.[12-22,27-50] In the review articles evaluated in the review-level analysis, the raw CLAIM scores ranged from 20 to 40, whereas the CLAIM adherence rates differed widely between 41% and 81%. This considerable variability underscores the inconsistent adherence to the CLAIM observed across the literature, highlighting the critical importance of our study in addressing these gaps.

Strengths and limitations

This study provides several strengths with notable implications for evaluating AI reporting quality in medical imaging. First, integrating data from multiple reviews offers a comprehensive assessment, unlike topic-specific studies, and provides a generalizable understanding of reporting practices. Second, our two-step analysis delivers both a broad overview and detailed insights, enabling item-wise evaluation to pinpoint areas needing particular improvement. Third, we identified factors associated with CLAIM adherence, offering actionable insights for enhancing reporting standards. Fourth, we presented two adherence metrics (the CLAIM score and compliance), facilitating comparability with other studies and setting a benchmark for future research. Finally, our analysis of critiques from eligible reviews offers valuable feedback to guide future updates to the CLAIM guidelines beyond 2024 and new alternative AI checklists.[10]

Our study has several limitations that should be carefully considered when inter-

Table 5. Item-wise correlation between reporting status and online publication year

CLAIM items[1]	Pre- and post-publication of CLAIM			Post-publication of CLAIM		
	rho	*P*	flag[2]	rho	*P*	flag[2]
Item#1 (AI methodology and technology type in title)	−0.097	0.046	*	−0.074	0.281	
Item#2 (Structured study summary)	0.034	0.491		0.022	0.748	
Item#3 (Background and clinical role of AI)	−0.038	0.435		0.071	0.300	
Item#4 (Study objectives and hypotheses)	−0.131	0.007	**	−0.162	0.018	*
Item#5 (Prospective or retrospective design)	0.092	0.060		0.017	0.806	
Item#6 (Study goal, e.g., model creation, feasibility)	−0.098	0.045	*	0.046	0.502	
Item#7 (Data sources)	0.024	0.626		−0.009	0.899	
Item#8 (Eligibility criteria, e.g., inclusion/exclusion)	−0.055	0.261		−0.014	0.842	
Item#9 (Data pre-processing)	−0.086	0.078		−0.217	0.001	**
Item#10 (Data subset selection, if applicable)	0.191	<0.001	***	0.225	<0.001	***
Item#11 (Definitions of data elements)	−0.220	<0.001	***	−0.057	0.405	
Item#12 (De-identification methods)	0.099	0.042	*	0.068	0.322	
Item#13 (Handling of missing data)	0.134	0.006	**	0.110	0.110	
Item#14 (Ground truth definition)	−0.057	0.240		−0.137	0.045	*
Item#15 (Rationale for reference standard)	−0.205	<0.001	***	−0.069	0.312	
Item#16 (Source and qualifications of annotators)	−0.078	0.111		−0.153	0.025	*
Item#17 (Annotation tools)	−0.244	<0.001	***	−0.092	0.180	
Item#18 [Variability assessment (inter/intra-rater)]	−0.211	<0.001	***	−0.121	0.078	
Item#19 (Sample size determination)	0.220	<0.001	***	0.250	<0.001	***
Item#20 (Data partitioning method)	0.140	0.004	**	−0.112	0.102	
Item#21 (Partition level, e.g., image, patient)	0.345	<0.001	***	0.116	0.091	
Item#22 [Model description (inputs, outputs, layers)]	0.036	0.462		−0.109	0.110	
Item#23 (Software and frameworks used)	−0.127	0.009	**	−0.134	0.050	
Item#24 (Model parameter initialization)	−0.124	0.011	*	−0.176	0.010	*
Item#25 (Training details, e.g., augmentation, hyperparameters)	0.141	0.004	**	−0.123	0.073	
Item#26 (Final model selection)	−0.057	0.246		−0.117	0.088	
Item#27 (Ensemble techniques, if applicable)	0.186	<0.001	***	0.167	0.014	*
Item#28 (Model performance metrics)	−0.076	0.119		−0.129	0.060	
Item#29 (Statistical significance and uncertainty)	0.026	0.594		0.060	0.386	
Item#30 (Robustness/sensitivity analysis)	0.022	0.656		0.015	0.831	
Item#31 (Explainability methods, e.g., saliency maps)	0.222	<0.001	***	−0.002	0.982	
Item#32 (External validation/testing)	0.009	0.846		0.098	0.152	
Item#33 (Participant flow diagram)	0.356	<0.001	***	0.202	0.003	**
Item#34 (Demographic/clinical characteristics by partition)	0.195	<0.001	***	0.126	0.065	
Item#35 (Performance metrics for optimal model)	−0.020	0.684		−0.097	0.158	
Item#36 (Diagnostic accuracy estimates)	0.101	0.039	*	0.172	0.012	*
Item#37 (Failure analysis)	0.075	0.125		−0.009	0.892	
Item#38 [Study limitations (bias, uncertainty, generalizability)]	0.173	<0.001	***	0.107	0.119	
Item#39 (Practice implications and clinical role)	−0.270	<0.001	***	−0.298	<0.001	***
Item#40 (Registration number and registry name)	−0.141	0.004	**	−0.093	0.174	
Item#41 (Study protocol access)	0.160	0.001	**	−0.075	0.275	
Item#42 (Funding sources and funder roles)	0.207	<0.001	***	0.092	0.178	

[1] Note that item names have been abbreviated; [2]* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

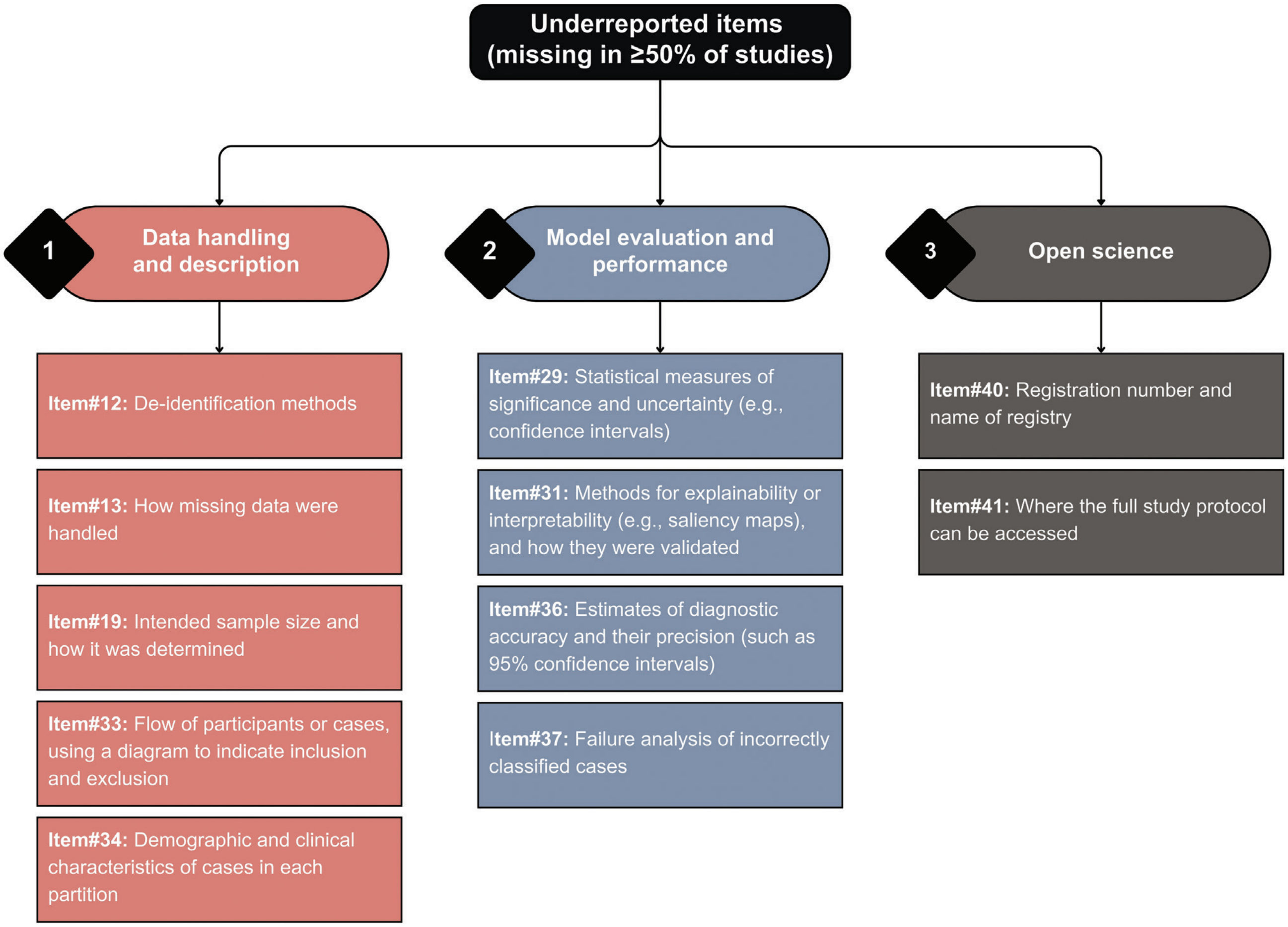

Underreported items (missing in ≥50% of studies)
1
Data handling and description
2
Model evaluation and performance
3
Open science
Item#12: De-identification methods
Item#13: How missing data were handled
Item#19: Intended sample size and how it was determined
Item#33: Flow of participants or cases, using a diagram to indicate inclusion and exclusion
Item#34: Demographic and clinical characteristics of cases in each partition
Item#29: Statistical measures of significance and uncertainty (e.g., confidence intervals)
Item#31: Methods for explainability or interpretability (e.g., saliency maps), and how they were validated
Item#36: Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
Item#37: Failure analysis of incorrectly classified cases
Item#40: Registration number and name of registry
Item#41: Where the full study protocol can be accessed

Figure 8. Eleven underreported items (i.e., missing in ≥50% of studies), categorized by relevant domains.

Table 6. Critiques identified in the analysis of the 35 review papers eligible for review- or study-level analyses

Category	Critique identified about the CLAIM with representative source articles
Fulfillment	Certain items may be viewed as overly strict or difficult to meet[43]
	Certain items are too technical, requiring advanced engineering or statistical knowledge[14]
Applicability	Some items are not applicable to all study types[12-14,30,39]
Feasibility and practicality	Some items may be impractical or infeasible in real-world settings[22]
Structure	Dividing the checklist into distinct sections sometimes complicates quality assessment[39]
Interpretation	Deciding if an item is sufficiently reported is subjective[13,39,44]
	Certain items may be viewed as vague or lack clarity in their current form[22]
	Certain items provide limited guidance on holistically interpreting a manuscript alongside its code[45]
Relative importance	Certain items may be more crucial than others but are currently weighted equally[13,39]
Scoring	Lack of standardized score or compliance calculation strategy[44]

CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

preting the results. First, this study was not registered (e.g., in PROSPERO). This decision was due to the unique nature of conducting a collective review of previous reviews of the CLAIM. Given the limited number of studies employing a similar strategy, and despite our group's experience with other guidelines, the methodology required adaptations based on the challenges and limitations encountered during data collection and analysis. These evolving methodological adjustments made it difficult to provide a fully transparent outline of the approach at the outset. Second, this research was limited to three databases, PubMed, Scopus, and Google Scholar, which we selected based on their broad coverage and relevance to the field, according to our experience. However, we acknowledge that the inclusion of additional databases, such as Embase and Web of Science, could further improve the comprehensiveness of the search. Third, the assessment of reporting quality was based solely on the CLAIM (2020 version). In the future, other AI-specific reporting guidelines, such as CONSORT-AI and TRIPOD-AI, could be considered to provide a more comprehensive evaluation of reporting standards.[63] Fourth, many articles were published before the CLAIM guidelines were introduced in 2020. However, the goal of this study was to highlight the overall state of reporting quality in the field, with some analyses covering both pre- and post-guideline periods. Fifth, our analysis focused solely on reporting quality and did not include evaluating the studies' actual impact, such as citation counts; there may not yet have been sufficient time for recent studies to have accumulated citations for meaningful comparisons. Additionally, the scope of our study is limited to exploring other factors that could affect the clinical translation of AI, such as methodological quality. Evaluating these factors may require supplementary tools, such as METRICS.[59] Sixth, this study was conducted after the CLAIM 2024 update.[10] Although the main framework of the original CLAIM was preserved,[8] earlier findings might have better informed the current update but could still aid future revisions and new guidelines. Seventh, the results of this study rely on prior systematic and non-systematic reviews as well as the expertise of the evaluators involved in those studies. The potential limited familiarity with certain aspects of the CLAIM in those articles and inconsistencies may have influenced the findings of this study. Eighth, due to the lack of a standard checkbox format in the initial CLAIM, consideration of item applicability may vary among reviews, potentially influencing adherence results, although both the CLAIM score and CLAIM compliance were assessed in the two-level analysis. Ninth, extracting data from systematic reviews can be subjective and may vary depending on the readers' experience. To minimize potential errors, we implemented a rigorous process involving the cross-checking of extracted data and resolving disagreements through consensus or by consulting an experienced reader, when necessary, at different stages of the study. Finally, the number of studies included in the study-level analysis was smaller than the number of studies represented in the review articles analyzed at the review level. However, to gain item-level insights, it was essential to conduct the analysis at the individual study level, as this granularity could not have been achieved at the review level. The sample size for the study-level analysis was determined merely by the availability of data in the existing literature, which may have introduced some degree of bias. Therefore, the findings should be interpreted with this limitation in mind.

In conclusion, this study provides a comprehensive evaluation of CLAIM adherence in the medical imaging AI literature, revealing significant variability and highlighting areas for improvement. Our two-level analysis, encompassing review- and study-level data, identified substantial reporting gaps, with a third of checklist items often omitted. Factors such as publication year, journal impact quartiles, and subfield-specific differences emerged as key independent predictors of adherence, underscoring the role of high-impact journals and tailored strategies for different subfields. The CLAIM compliance rate was highlighted as a more objective and fairer metric for adherence assessment. Additionally, several important critiques of the CLAIM were identified, providing valuable insights for researchers and developers. We hope these findings serve as actionable guidance for the scientific community to enhance transparency, reproducibility, and reporting quality in AI studies.

Acknowledgements

The authors wish to express their sincere gratitude to the anonymous reviewers for their insightful comments and constructive feedback, which have significantly improved the quality of this paper.

The language of this manuscript was checked and improved by ChatGPT (4o). This tool was used solely to improve the clarity and quality of the content originally written by the authors. The authors conducted strict supervision after using this tool.

Footnotes

Conflict of Interest

Burak Koçak, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

1. Kocak B, Baessler B, Cuocolo R, Mercaldo N, Pinto Dos Santos D. Trends and statistics of artificial intelligence and radiomics research in radiology, nuclear medicine, and medical imaging: bibliometric analysis. *Eur Radiol*. 2023;33(11):7542-7555. [CrossRef]
2. Nensa F, Pinto Dos Santos D, Dietzel M. Beyond accuracy: reproducibility must lead AI advances in radiology. *Eur J Radiol*. 2024;180:111703. [CrossRef]
3. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. *JAMA*. 2020;323(4):305-306. [CrossRef]
4. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J Med Internet Res*. 2023;25:e48763. [CrossRef]
5. Vasey B, Novak A, Ather S, Ibrahim M, McCulloch P. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. *Clin Radiol*. 2023;78(2):130-136. [CrossRef]
6. 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(9):1364-1374. [CrossRef]
7. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. [CrossRef]
8. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029. [CrossRef]
9. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. [CrossRef]
10. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical

Imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. [CrossRef]

11. Pandis N, Fedorowicz Z. The international EQUATOR network: enhancing the quality and transparency of health care research. *J Appl Oral Sci*. 2011;19(5):0. [CrossRef]
12. Belue MJ, Harmon SA, Lay NS, et al. The low rate of adherence to Checklist for Artificial Intelligence in Medical Imaging criteria among published prostate MRI artificial intelligence algorithms. *J Am Coll Radiol*. 2023;20(2):134-145. [CrossRef]
13. Bhandari A, Scott L, Weilbach M, Marwah R, Lasocki A. Assessment of artificial intelligence (AI) reporting methodology in glioma MRI studies using the Checklist for AI in Medical Imaging (CLAIM). *Neuroradiology*. 2023;65(5):907-913. [CrossRef]
14. Si L, Zhong J, Huo J, et al. Deep learning in knee imaging: a systematic review utilizing a Checklist for Artificial Intelligence in Medical Imaging (CLAIM). *Eur Radiol*. 2022;32(2):1353-1361. [CrossRef]
15. Abdulaal L, Maiter A, Salehi M, et al. A systematic review of artificial intelligence tools for chronic pulmonary embolism on CT pulmonary angiography. *Front Radiol*. 2024;4:1335349. [CrossRef]
16. Alipour E, Pooyan A, Shomal Zadeh F, Darbandi AD, Bonaffini PA, Chalian M. Current status and future of artificial intelligence in MM imaging: a systematic review. *Diagnostics (Basel)*. 2023;13(21):3372. [CrossRef]
17. Sushentsev N, Moreira Da Silva N, Yeung M, et al. Comparative performance of fully-automated and semi-automated artificial intelligence methods for the detection of clinically significant prostate cancer on MRI: a systematic review. *Insights Imaging*. 2022;13(1):59. [CrossRef]
18. Ozkara BB, Chen MM, Federau C, et al. Deep Learning for Detecting Brain Metastases on MRI: a systematic review and meta-analysis. *Cancers*. 2023;15(2):334. [CrossRef]
19. Alabed S, Maiter A, Salehi M, et al. Quality of reporting in AI cardiac MRI segmentation studies - a systematic review and recommendations for future studies. *Front Cardiovasc Med*. 2022;9:956811. [CrossRef]
20. Karabacak M, Ozkara BB, Ozturk A, et al. Radiomics-based machine learning models for prediction of medulloblastoma subgroups: a systematic review and meta-analysis of the diagnostic test performance. *Acta Radiol*. 2023;64(5):1994-2003. [CrossRef]
21. Kouli O, Hassane A, Badran D, Kouli T, Hossain-Ibrahim K, Steele JD. Automated brain tumor identification using magnetic resonance imaging: A systematic review and meta-analysis. *Neurooncol Adv*. 2022;4(1):vdac081. [CrossRef]
22. Kim DY, Oh HW, Suh CH. Reporting Quality of Research Studies on AI applications in medical images according to the CLAIM guidelines in a radiology journal with a strong prominence in Asia. *Korean J Radiol*. 2023;24(12):1179-1189. [CrossRef]
23. Kocak B, Keles A, Kose F, Sendur A. Quality of radiomics research: comprehensive analysis of 1574 unique publications from 89 reviews. *Eur Radiol*. 2024. [CrossRef]
24. Briscoe S, Abbott R, Lawal H, Shaw L, Coon JT. Feasibility and desirability of screening search results from Google Search exhaustively for systematic reviews: a cross-case analysis. *Res Synth Methods*. 2023;14(3):427-437. [CrossRef]
25. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Stat Methods Med Res*. 2018;27(6):1785-1805. [CrossRef]
26. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol*. 2014;14(1):135. [CrossRef]
27. Assadi H, Alabed S, Maiter A, et al. The role of artificial intelligence in predicting outcomes by cardiovascular magnetic resonance: a comprehensive systematic review. *Medicina (Kaunas)*. 2022;58(8):1087. [CrossRef]
28. Bedrikovetski S, Seow W, Kroon HM, Traeger L, Moore JW, Sammour T. Artificial intelligence for body composition and sarcopenia evaluation on computed tomography: a systematic review and meta-analysis. *Eur J Radiol*. 2022;149:110218. [CrossRef]
29. Belue MJ, Turkbey B. Tasks for artificial intelligence in prostate MRI. *Eur Radiol Exp*. 2022;6(1):33. [CrossRef]
30. Bleker J, Kwee TC, Yakar D. Quality of multicenter studies using MRI radiomics for diagnosing clinically significant prostate cancer: a systematic review. *Life (Basel)*. 2022;12(7):946. [CrossRef]
31. Cerdá-Alberich L, Solana J, Mallol P, et al. MAIC-10 brief quality checklist for publications using artificial intelligence and medical images. *Insights Imaging*. 2023;14(1):11. [CrossRef]
32. Dagher R, Ozkara BB, Karabacak M, et al. Artificial intelligence/machine learning for neuroimaging to predict hemorrhagic transformation: Systematic review/meta-analysis. *J Neuroimaging*. 2024;34(5):505-514. [CrossRef]
33. Hardacre CJ, Robertshaw JA, Barratt SL, et al. Diagnostic test accuracy of artificial intelligence analysis of cross-sectional imaging in pulmonary hypertension: a systematic literature review. *Br J Radiol*. 2021;94(1128):20210332. [CrossRef]
34. Hickman SE, Woitek R, Le EPV, et al. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. *Radiology*. 2022;302(1):88-104. [CrossRef]
35. Hu J, Wang Y, Guo D, et al. Diagnostic performance of magnetic resonance imaging-based machine learning in Alzheimer's disease detection: a meta-analysis. *Neuroradiology*. 2023;65(3):513-527. [CrossRef]
36. Hwang EJ, Jeong WG, David PM, Arentz M, Ruhwald M, Yoon SH. AI for detection of tuberculosis: implications for global health. *Radiol Artif Intell*. 2024;6(2):e230327. [CrossRef]
37. Jia LL, Zhao JX, Pan NN, et al. Artificial intelligence model on chest imaging to diagnose COVID-19 and other pneumonias: a systematic review and meta-analysis. *Eur J Radiol Open*. 2022;9:100438. [CrossRef]
38. Karabacak M, Ozkara BB, Mordag S, Bisdas S. Deep learning for prediction of isocitrate dehydrogenase mutation in gliomas: a critical approach, systematic review and meta-analysis of the diagnostic test performance using a Bayesian approach. *Quant Imaging Med Surg*. 2022;12(8):4033-4046. [CrossRef]
39. Lans A, Pierik RJB, Bales JR, et al. Quality assessment of machine learning models for diagnostic imaging in orthopaedics: A systematic review. *Artif Intell Med*. 2022;132:102396. [CrossRef]
40. Le VNT, Kim JG, Yang YM, Lee DW. Evaluating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)-based quality of reports using convolutional neural network for odontogenic cyst and tumor detection. *Appl Sci*. 2021;11(20):9688. [CrossRef]
41. O'Shea RJ, Sharkey AR, Cook GJR, Goh V. Systematic review of research design and reporting of imaging studies applying convolutional neural networks for radiological cancer diagnosis. *Eur Radiol*. 2021;31(10):7969-7983. [CrossRef]
42. Raj M, Ayub A, Pal AK, et al. Diagnostic accuracy of artificial intelligence-based algorithms in automated detection of neck of femur fracture on a plain radiograph: a systematic review and meta-analysis. *Indian J Orthop*. 2024;58(5):457-469. [CrossRef]
43. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3(3):199-217. [CrossRef]
44. Roest C, Fransen SJ, Kwee TC, Yakar D. Comparative performance of deep learning and radiologists for the diagnosis and localization of clinically significant prostate cancer at MRI: a systematic review. *Life (Basel)*. 2022;12(10):1490. [CrossRef]
45. Sivanesan U, Wu K, McInnes MDF, Dhindsa K, Salehi F, van der Pol CB. Checklist for Artificial Intelligence in Medical Imaging Reporting adherence in peer-reviewed and preprint manuscripts with the highest altmetric attention scores: a meta-research study. *Can Assoc Radiol J*. 2023;74(2):334-342. [CrossRef]

46. Tsang B, Gupta A, Takahashi MS, Baffi H, Ola T, Doria AS. Applications of artificial intelligence in magnetic resonance imaging of primary pediatric cancers: a scoping review and CLAIM score assessment. *Jpn J Radiol*. 2023;41(10):1127-1147. [CrossRef]

47. Wang TW, Hong JS, Huang JW, Liao CY, Lu CF, Wu YT. Systematic review and meta-analysis of deep learning applications in computed tomography lung cancer segmentation. *Radiother Oncol*. 2024;197:110344. [CrossRef]

48. Wang TW, Hsu MS, Lee WK, et al. Brain metastasis tumor segmentation and detection using deep learning algorithms: a systematic review and meta-analysis. *Radiother Oncol*. 2024;190:110007. [CrossRef]

49. Zhong J, Hu Y, Zhang G, et al. An updated systematic review of radiomics in osteosarcoma: utilizing CLAIM to adapt the increasing trend of deep learning application in radiomics. *Insights Imaging*. 2022;13(1):138. [CrossRef]

50. Zhong J, Xing Y, Zhang G, et al. A systematic review of radiomics in giant cell tumor of bone (GCTB): the potential of analysis on individual radiomics feature for identifying genuine promising imaging biomarkers. *J Orthop Surg Res*. 2023;18(1):414. [CrossRef]

51. Agha RA, Fowler AJ, Limb C, et al. Impact of the mandatory implementation of reporting guidelines on reporting quality in a surgical journal: a before and after study. *Int J Surg*. 2016;30:169-172. [CrossRef]

52. Koçak B, Keleş A, Köse F. Meta-research on reporting guidelines for artificial intelligence: are authors and reviewers encouraged enough in radiology, nuclear medicine, and medical imaging journals? *Diagn Interv Radiol*. 2024;30(5):291-298. [CrossRef]

53. Zhong J, Xing Y, Lu J, et al. The endorsement of general and artificial intelligence reporting guidelines in radiological journals: a meta-research study. *BMC Med Res Methodol*. 2023;23(1):292. [CrossRef]

54. Kocak B, Keles A, Akinci D'Antonoli T. Self-reporting with checklists in artificial intelligence research on medical imaging: a systematic review based on citations of CLAIM. *Eur Radiol*. 2024;34(4):2805-2815. [CrossRef]

55. Kocak B, Ponsiglione A, Stanzione A, et al. CLEAR guideline for radiomics: early insights into current reporting practices endorsed by EuSoMII. *Eur J Radiol*. 2024;181:111788. [CrossRef]

56. Korevaar DA, Wang J, van Enst WA, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology*. 2015;274(3):781-789. [CrossRef]

57. Akinci D'Antonoli T, Cavallo AU, Vernuccio F, et al. Reproducibility of radiomics quality score: an intra- and inter-rater reliability study. *Eur Radiol*. 2024;34(4):2791-2804. [CrossRef]

58. Mese I, Kocak B. ChatGPT as an effective tool for quality evaluation of radiomics research. *Eur Radiol*. 2024. [CrossRef]

59. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging*. 2024;15(1):8. [CrossRef]

60. Spadarella G, Stanzione A, Akinci D'Antonoli T, et al. Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. *Eur Radiol*. 2023;33(3):1884-1894. [CrossRef]

61. Barry N, Kendrick J, Molin K, et al. Evaluating the impact of the Radiomics Quality Score: a systematic review and meta-analysis. *Eur Radiol*. 2025. [CrossRef]

62. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. [CrossRef]

63. Park SH, Suh CH. Reporting Guidelines for Artificial Intelligence Studies in Healthcare (for Both Conventional and Large Language Models): what's new in 2024. *Korean J Radiol*. 2024;25(8):687-690. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.242999

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Automatic bone age assessment: a Turkish population study

Samet Öztürk[1]
Murat Yüce[2]
Gül Gizem Pamuk[3]
Candan Varlık[3]
Ahmet Tan Cimilli[3]
Musa Atay[3]

[1]Esenler Obstetrics & Gynecology and Pediatrics Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]Icahn School of Medicine at Mount Sinai Biomedical Engineering and Imaging Institute, New York, USA

[3]University of Health Sciences Türkiye, Bağcılar Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

Corresponding author: Samet Öztürk

E-mail: drozturksamet@gmail.com

Received 15 October 2024; revision requested 26 November 2024; last revision received 13 January 2025; accepted 22 January 2025

Epub: 17.03.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.242999

PURPOSE

Established methods for bone age assessment (BAA), such as the Greulich and Pyle atlas, suffer from variability due to population differences and observer discrepancies. Although automated BAA offers speed and consistency, limited research exists on its performance across different populations using deep learning. This study examines deep learning algorithms on the Turkish population to enhance bone age models by understanding demographic influences.

METHODS

We analyzed reports from Bağcılar Hospital's Health Information Management System between April 2012 and September 2023 using "bone age" as a keyword. Patient images were re-evaluated by an experienced radiologist and anonymized. A total of 2,730 hand radiographs from Bağcılar Hospital (Turkish population), 12,572 from the Radiological Society of North America (RSNA), and 6,185 from the Radiological Hand Pose Estimation (RHPE) public datasets were collected, along with corresponding bone ages and gender information. A random set of 546 radiographs (273 from Bağcılar, 273 from public datasets) was initially randomly split for an internal test set with bone age stratification; the remaining data were used for training and validation. BAAs were generated using a modified InceptionV3 model on 500 × 500-pixel images, selecting the model with the lowest mean absolute error (MAE) on the validation set.

RESULTS

Three models were trained and tested based on dataset origin: Bağcılar (Turkish), public (RSNA–RHPE), and a Combined model. Internal test set predictions of the Combined model estimated bone age within less than 6, 12, 18, and 24 months at rates of 44%, 73%, 87%, and 94%, respectively. The MAE was 9.2 months in the overall internal test set, 7 months on the public test set, and 11.5 months on the Bağcılar internal test data. The Bağcılar-only model had an MAE of 12.7 months on the Bağcılar internal test data. Despite less training data, there was no significant difference between the combined and Bağcılar models on the Bağcılar dataset ($P > 0.05$). The public model showed an MAE of 16.5 months on the Bağcılar dataset, significantly worse than the other models ($P < 0.05$).

CONCLUSION

We developed an automatic BAA model including the Turkish population, one of the few such studies using deep learning. Despite challenges from population differences and data heterogeneity, these models can be effectively used in various clinical settings. Model accuracy can improve over time with cumulative data, and publicly available datasets may further refine them. Our approach enables more accurate and efficient BAAs, supporting healthcare professionals where traditional methods are time-consuming and variable.

CLINICAL SIGNIFICANCE

The developed automated BAA model for the Turkish population offers a reliable and efficient alternative to traditional methods. By utilizing deep learning with diverse datasets from Bağcılar Hospital and publicly available sources, the model minimizes assessment time and reduces variability. This advancement enhances clinical decision-making, supports standardized BAA practices, and improves patient care in various healthcare settings.

KEYWORDS

Bone age assessment, deep learning, artificial intelligence, convolutional neural network, InceptionV3

You may cite this article as: Öztürk S, Yüce M, Pamuk GG, Varlık C, Cimilli AT, Atay M. Automatic bone age assessment: a Turkish population study. *Diagn Interv Radiol.* 2025;31(5):456-464.

Children's growth is characterized by non-linear progression, typically advancing in a sequential manner. Although metrics such as height and weight are useful for monitoring growth, bone development often provides the closest approximation to chronological age. The Greulich and Pyle (GP) and Tanner and Whitehouse (TW) methods are commonly employed for bone age assessment (BAA).[1,2] However, these methods rely on the expertise of radiologists and are subject to interpretation biases.[3] To address this, automatic BAA models have been developed, offering enhanced accuracy, repeatability, and efficiency.[4] Our study aims to evaluate the performance of deep learning algorithms within the Turkish population and enhance model efficacy at a population level. Additionally, we seek to demonstrate that establishing a model "from scratch" is feasible for a medium-sized hospital without relying on funds, grants, or dedicated commercial software.

Methods

Ethics approval

Approval was granted by the Non-Interventional Clinical Research Ethics Committee of University of Health Sciences Türkiye, Bağcılar Training and Research Hospital, with the ethics committee decision numbered 2023/09/08/051 and dated September 22, 2023. Informed consent was waived due to the retrospective nature of the study. All procedures in the present study involving human participants were performed in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Main points

- Population-specific deep learning model: We developed an automated model using the YOLOv8m architecture for hand detection and modified InceptionV3 for bone age assessment (BAA) tailored for the Turkish population by integrating Bağcılar Hospital, Radiological Society of North America, and Radiological Hand Pose Estimation datasets.
- Improved accuracy with combined data: The Combined model achieved a mean absolute error of 9.2 months and a 96% correlation with the reference standard, outperforming our single-source models.
- Clinical application and future prospects: This provides a consistent and efficient BAA tool, reducing radiologists' workload and variability. We aim to enhance accuracy with more diverse data and validate the model through broader clinical studies.

Data collection and dataset creation

Wrist and hand radiographs, bone age reports, and gender information for patients aged 0–18 years were collected from the Picture Archiving and Communication System without interpretational hindrances. A total of 2,933 radiographs conforming to Turkish standards were acquired from hospital records. Patients aged >18 years, images with severe artifacts or inappropriate field of view, and reports without BAA were excluded; 2,730 X-rays were found to be eligible (Figure 1). While integrating X-rays from Bağcılar into the dataset, evaluations by S.Ö. (who had 6 years of radiology experience) were compared with the clinical reading report. When the difference was ≤6 months, the report was deemed accurate. In cases where the difference was >6 months, A.T.C. (who had 32 years of radiology experience) and S.Ö. reevaluated images together, and a reference standard was obtained with a consensus decision. Additionally, two different open-source public datasets were incorporated [Radiological Society of North America (RSNA): https://www.rsna.org/rsnai/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017 and Radiological Hand Pose Estimation (RHPE): https://www.kaggle.com/datasets/ipythonx/rhpe-bone-age] with the filtering age range set to 0–18 years, resulting in a hybrid dataset sourced from various devices, vendors, and populations.[5,6] After filtering, the RSNA dataset consisted of 12,572 labeled radiographs, while the RHPE dataset included 6,185 labeled radiographs, and these datasets were further concatenated with the Bağcılar dataset. From the combination of all three datasets, an internal test dataset (n = 546) was created by randomly selecting 10% of Bağcılar data (n = 273) and an equal amount of public data (n = 273). The remaining data were used to create three distinct training and validation splits (Bağcılar, Public, and Combined), maintaining a 9:1 training-to-validation ratio (Figure 1). Bone age-based stratification was applied during the random splitting of each dataset using the train_test_split function from the scikit-learn Python library.

Model structure

In 2017, an RSNA BAA competition was held. The structure of the models used by the competitors and their error rates were published by Halabi et al.[5] The winner of the competition was a commercial company that profited from this work. The authors do not have any collaboration, partnership, or

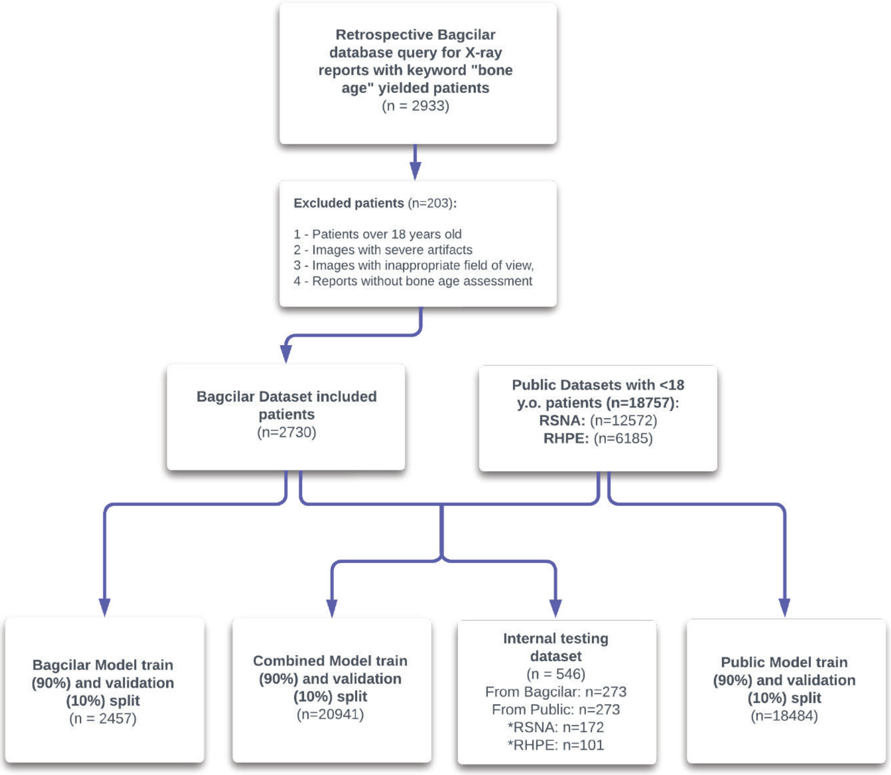

Retrospective Bagcilar database query for X-ray reports with keyword "bone age" yielded patients
(n = 2933)
Excluded patients (n=203):
1 - Patients over 18 years old
2 - Images with severe artifacts
3 - Images with inappropriate field of view,
4 - Reports without bone age assessment
Bagcilar Dataset included patients
(n=2730)
Public Datasets with <18 y.o. patients (n=18757):
RSNA: (n=12572)
RHPE: (n=6185)
Bagcilar Model train (90%) and validation (10%) split
(n = 2457)
Combined Model train (90%) and validation (10%) split
(n=20941)
Internal testing dataset
(n = 546)
From Bagcilar: n=273
From Public: n=273
*RSNA: n=172
*RHPE: n=101
Public Model train (90%) and validation (10%) split
(n=18484)

Figure 1. Flowchart illustrating the data collection process, inclusion and exclusion criteria, and dataset splitting methodology for the bone age prediction study. RSNA, Radiological Society of North America; RHPE, Radiological Hand Pose Estimation.

funding agreement with this company. The authors' models were built and trained "from scratch" using published architectures. As in the competition, a custom InceptionV3 model proved to be more suitable for this study. As a custom preprocessing step to improve model performance, a hand-detection model was also added using the YOLOv8m architecture. In the hospital's routine radiography acquisitions, some of the X-rays had a field of view large enough to include the elbow, whereas in others, phalanges were not included. The goal was to crop and adjust only the hand and wrist portion using YOLOv8m. Due to the heterogeneous nature of the hospital's dataset, encompassing images from diverse regions, a YOLOv8m model was initially trained for hand detection. Images were cropped with detected bounding boxes of the hand area before training the InceptionV3 model. All images were resized to 500 × 500 pixels, and an InceptionV3-based deep convolutional neural network (CNN) was constructed to process pixel information. Binary gender data (0 for female, 1 for male) were incorporated to account for gender effects via a densely connected layer with 32 neurons. Gender and pixel information were merged into a single network, followed by two densely connected layers with rectified linear unit activation, each containing 1,000 neurons, facilitating complex pattern learning. The output layer utilized mean absolute error (MAE) loss for regression simplification (Figure 2). A consistent model architecture was used throughout the study. It was trained and tested on three distinct datasets: the Bağcılar dataset, the public datasets (RSNA and RHPE), and a combined dataset consisting of both. For clarity, references to the "Bağcılar model," "Public model," and "Combined model" datasets pertain to the data used during model training and testing, not to distinct model architectures.

Model training process

The study utilized Keras 3.02, TensorFlow 2.15, and Python 3.9 for training, using an Nvidia RTX 3090 24GB graphics card. Data augmentation techniques, including rotation (up to 20 degrees), horizontal/vertical shifting (up to 20%), zooming (up to 20%), and horizontal flipping, were applied across the entire dataset to encourage the learning of patient-specific features. The final model was trained using Adam optimization with a batch size of 32 for 500 epochs. Learning rate adjustments and early stopping mechanisms were implemented. Models were trained and validated (90% training, 10% validation) and tested on an initially separated internal testing dataset, which was composed of an equally distributed number of images from both local and public sources (Figure 1).

Statistical analysis

Normality analysis was conducted using the Kolmogorov–Smirnov test. For comparisons between variables showing normal distribution, t-tests were employed, while one-way analysis of variance (ANOVA) was utilized for multiple variable comparisons. The Mann–Whitney U test and Kruskal–Wallis analysis were employed for variables that were not normally distributed. Post-hoc analyses were conducted using Bonferroni-corrected Mann–Whitney U and Tukey tests. A significance threshold of $P < 0.05$ was applied. Python version 3.9 was utilized for statistical analyses and plot generation.

Results

A total of 21,487 patients were included in the study, with a mean bone age of 10.4 ± 3.5 years, and 51% were female. A total of 18,757 cases were from public datasets (RSNA and RHPE), with a mean bone age of 10.5 ± 3.4 years and 50% female representation (Figure 3). The Bağcılar dataset had a mean bone age of 9.8 ± 3.9 years, with 38% female patients. Table 1 shows demographic data and information regarding the referring departments and International Classification of Diseases-10 (ICD-10) diagnosis codes for the Bağcılar dataset. The primary referring departments were general pediatrics (48%) and pediatric endocrinology (47.5%). The majority of cases (81%) were referred with preliminary diagnoses under the ICD-10 main category "endocrine, nutritional, and metabolic diseases."

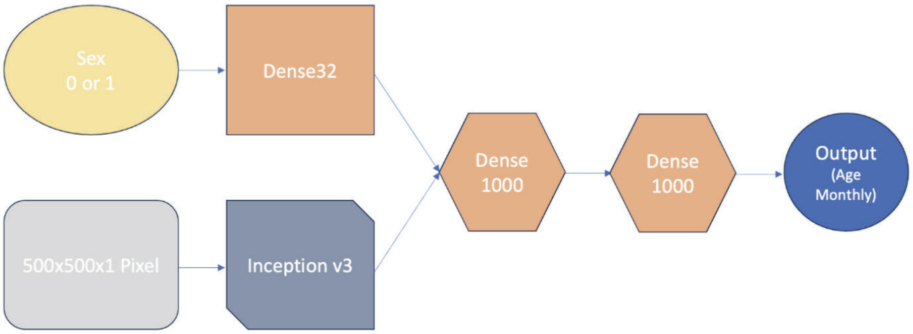


Figure 2. Architecture of the bone age prediction model: Combining sex input (encoded as 0 or 1) through a Dense32 layer and image input (500 × 500 × 1 pixels) processed via InceptionV3, followed by two dense layers (1,000 neurons each) to predict age in months.

The performance metrics for the models, evaluated in the internal testing dataset, showed that the Public model had an MAE of 11.3 months, with a mean squared error (MSE) of 302.1 and a root MSE (RMSE) of 17.4. The Bağcılar model (BM) showed a slightly higher MAE of 12.6 months but improved MSE and RMSE values of 260.3 and 16.1, respectively. The Combined model demonstrated the best overall performance, achieving the lowest MAE of 9.2 months, along with an MSE of 170.7 and an RMSE of 13.1, highlighting its superior accuracy compared with the other models.

Based on the internal testing dataset, the BM achieved bone age predictions within absolute differences of ≤6, ≤12, ≤18, and ≤24 months for 31%, 57%, 77%, and 88% of cases, respectively, with a Pearson correlation of 93%. The public dataset model (PM) achieved predictions within the same ranges for 45%, 69%, 81%, and 89% of cases, also with a Pearson correlation of 93%. The combined dataset model (CM) demonstrated the best performance, with predictions within ≤6, ≤12, ≤18, and ≤24 months for 44%, 73%, 87%, and 94% of cases, respectively, and a Pearson correlation of 96%, highlighting its superior accuracy and clinical utility.

Comparison of bone age predictions from the PM, BM, and CM models with the reference standard in the internal testing dataset revealed no statistically significant differences for any model, as determined by independent t-tests ($P > 0.05$).

The distribution of patients by age group and gender across the training, validation, and internal testing datasets is presented in Table 2. The mean and standard deviation of predicted bone ages alongside the reference standard for each age group and across models are shown in Table 3. Analyses of variance conducted for each age group between the three model assessments, and the reference standard revealed significant differences in the 0–3, 3–6, 6–9, 9–12, 12–15, and 15–18

Table 1. Clinical characteristics of patients in the Bağcılar dataset

	n	%
Gender		
Male	1,693	62
Female	1,037	38
Referring department		
General pediatrics	1,310	48
Pediatric endocrinology	1,296	47.5
Orthopedics	40	1.5
Other*	84	3
ICD-10 category **		
Endocrine, nutritional, and metabolic diseases	2,211	81
Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	164	6
Factors influencing health status and contact with health services	104	3.8
Diseases of the blood and blood-forming organ and certain disorders involving the immune mechanism	65	2.4
Diseases of the respiratory system	46	1.7
Diseases of the genitourinary system	43	1.6
Diseases of the musculoskeletal system and connective tissue	33	1.2
Other	64	2.3

Average bone age was 9.8 years (standard deviation: 3.9). *Including mainly health board, family medicine, emergency department referrals; **ICD-10: International Classification of Diseases, tenth revision.

Table 2. Age and gender distribution of training and validation datasets for each model and internal testing dataset

			Age groups (years)					
	Gender	Split	0–3	3–6	6–9	9–12	12–15	15–18
Bağcılar model	**Female**	**Train**	64	184	363	462	165	131
		Val	9	19	46	52	18	13
	Male	**Train**	97	139	143	227	168	68
		Val	8	18	9	26	20	8
Public model	**Female**	**Train**	273	716	2,242	3,344	1,477	201
		Val	33	94	249	347	168	17
	Male	**Train**	251	945	1,265	1,807	3,445	669
		Val	27	91	146	208	387	82
Combined model	**Female**	**Train**	338	923	2,624	3,785	1,661	321
		Val	41	90	276	420	167	41
	Male	**Train**	346	1,065	1,395	2,030	3,615	743
		Val	37	128	168	238	405	84
Internal testing set	**Female**	**Test**	9	30	76	121	46	17
	Male	**Test**	21	37	36	57	71	25

groups (*P* < 0.001). Subsequently, a Tukey post-hoc analysis was carried out to elucidate the differences between models and the reference standard for each respective age group where significance was observed in the ANOVA:

- 0–3 and 3–6 years: The PM differed significantly from both the reference standard and the other models (P < 0.001).
- 6–9 years: Significant differences were observed between the BM and both the reference standard (*P* = 0.016) and the CM (*P* = 0.042). The PM also differed significantly from the reference standard (*P* = 0.0003) and the CM (*P* = 0.001).
- 9–12 years: Significant differences were found between the reference standard and the BM (*P* = 0.034) as well as between the BM and PM (*P* = 0.002).
- 12–15 years: The BM differed significantly from the reference standard (*P* = 0.005) and the CM (*P* = 0.002).
- 15–18 years: The BM showed significant differences compared with the reference standard (*P* = 0.011) and the CM (*P* = 0.003).

These findings are also shown with boxplots in Figure 4 and indicate that while significant differences exist among certain models and age groups, the degree of discrepancy varies, emphasizing the variability in model performance across age groups. Notably, no significant difference was found between the reference standard and the CM across all age groups.

The MAEs of the models in the Public internal testing data were 6.2, 6.9, and 12.5 months for the PM, CM, and BM, respectively. The ANOVA conducted on the absolute error differences of the three model predictions in the public internal testing dataset revealed a statistically significant difference (*s* = 60.01, *P* < 0.001). Tukey post-hoc analysis of the model assessments in the Public internal testing dataset showed that the BM had a statistically significantly lower performance compared with the PM and CM, with MAE differences of 6.3 and 5.5 months, respectively (*P* < 0.05). There was no significant difference between the PM and CM (*P* > 0.05) (Table 4).

The MAEs of the models in the Bağcılar internal testing dataset were 16.5, 11.4, and 12.7 months for the PM, CM, and BM respectively. Analyses of variance among the absolute error differences of the three models in the Bağcılar internal testing dataset found a statistically significant difference (*s* = 11.19, *P* < 0.001). In the Tukey post-hoc analysis conducted among the model assessments in the Bağcılar test dataset, the PM showed a statistically significantly lower performance compared with the BM and CM, with MAE differences of 3.8 and 5 months, respectively (*P* < 0.05). However, no significant difference was observed between the BM and CM (*P* > 0.05) (Table 5).

Bland–Altman plots were generated to display the differences between the BM, PM, CM, and the reference standard in months,

Table 3. Mean and standard deviation of reference standard and predicted bone ages for different age groups in the internal testing dataset

(Months)	Public model predictions	Bağcılar model predictions	Combined model predictions	Reference standard
0–36	46.8 ± 27.3	30.6 ± 11.2	31.6 ± 10.8	29.9 ± 6.4
36–72	71.6 ± 17.6	64.2 ± 19.5	57.8 ± 14.1	60.6 ± 9.8
72–108	105.9 ± 19.9	103.3 ± 18.2	97.3 ± 18.1	96.6 ± 9.2
108–144	131.8 ± 16.1	126.4 ± 15.9	128.5 ± 14.7	130.5 ± 9.1
144–180	162.2 ± 16.5	155.5 ± 19.5	163.2 ± 17.9	162.6 ± 9.2
180–216	194.8 ± 14.4	189.0 ± 16.2	199.8 ± 15.2	198.6 ± 9.3

Table 4. Post-hoc Tukey test for Public internal testing data. In the analysis of mean absolute error (MAE) for Public internal testing data, no significant difference was observed between the Combined model (CM) and the Public model (PM)

Post-hoc, Tukey test for Public test data						
Group 1	Group 2	MAE difference (months)	*P* value	Lower (months)	Upper (months)	Significant
BM MAE	CM MAE	−5.50	<0.001	−6.97	−4.03	Yes
BM MAE	PM MAE	−6.28	<0.001	−7.75	−4.81	Yes
CM MAE	PM MAE	−0.78	0.43	−2.24	0.68	No

Table 5. Post-hoc Tukey test for Bağcılar internal testing data. In the analysis of mean absolute error (MAE) for Bağcılar internal testing data, no significant difference was observed between the Bağcılar model (BM) and the Combined model (CM)

Post-hoc, Tukey test for Bağcılar internal testing data						
Group 1	Group 2	Mean difference (months)	*P* value	Lower (months)	Upper (months)	Significant
BM MAE	CM MAE	−1.24	0.496	−3.84	1.35	No
BM MAE	PM MAE	3.77	0.002	1.18	6.37	Yes
CM MAE	PM MAE	5.02	<0.001	2.43	7.62	Yes

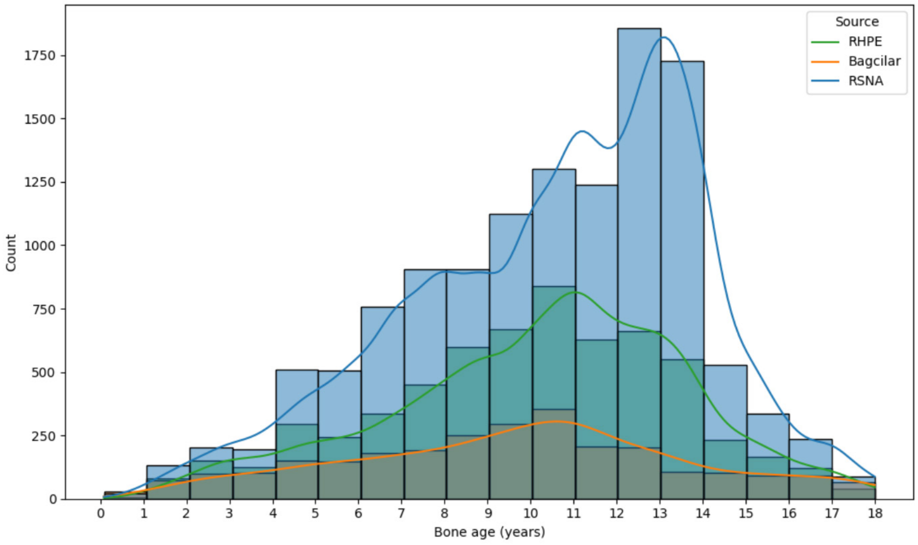


Figure 3. Bone age distribution across datasets: Histogram plots showing the distribution of bone ages (in years) for the Radiological Hand Pose Estimation, Bağcılar, and Radiological Society of North America datasets.

RHPE, Radiological Hand Pose Estimation; RSNA, Radiological Society of North America.

highlighting the variance between the model assessments and the reference standard within the internal testing dataset (Figure 5). Additionally, scatter plots with linear regression lines were created for each model to provide a clearer understanding of their performance across different internal testing datasets (Figure 6).

Discussion

The accuracy of BAA is largely dependent on the experience of the physician, as traditional evaluation is often a subjective estimation. Traditional assessment studies are typically conducted by experienced physicians through visual inspection and manual marking. This process requires significant time and effort, and different physicians may have varying standards when evaluating the same radiograph. Therefore, automated assessment approaches for BAA are increasingly gaining interest.

On average, an experienced radiologist spends approximately 1.4 minutes using the GP method and 7.9 minutes using the TW method to assess a patient.[7] Moreover, both methods are associated with high intra- and inter-observer variability. The reported range of BAA results averages 0.96 years (11.5 months) for GP and 0.74 years (8.9 months) for TW.[8] In some stages of child development, changes can be very subtle, especially after the age of 14 years, and the sensitivity perceivable by the human eye through radiological examination may be lacking.[9] The absence of significant differences between the predicted bone age using our proposed models and those obtained using the GP and TW methods enhances the reliability of our approach.

Our model utilized upper extremity radiographs containing hand and wrist regions with bone age reports sourced from Bağcılar. The images were taken with different presets and exposures, resulting in an inhomogeneous dataset. Some radiographs did not include joints prioritized in

BAA. Others were not captured at the correct position or angle. In some cases, bones were superimposed, and the parent's hand was often visible in infant radiographs. Considering this data heterogeneity, our model better reflects daily clinical practice compared to similar studies.

In this study, public datasets and data from Bağcılar Hospital were used as sources. Racial differences, imaging parameters, and image quality may have influenced the results. However, we believe that a model trained with these parameters could be more consistent than the inter- and intra-observer variability associated with the GP and TW methods.[8] Further prospective studies are needed to assess the added value of such models in daily clinical practice.

Recently, many deep learning methods have been developed for BAA, and RSNA even organized a competition for this purpose.[5] With the developed methods, the timing and pattern of ossification centers according to age can be extracted from images using deep learning techniques for BAA. Thus, this process, which is time-consuming and subjective, with differences between evaluators and even variations within the evaluator, can be carried out on more solid foundations.[10] In recent studies, we see models where various ensemble techniques are employed, combining multiple models into one.[11] Liu et al.[12] suggested that ranking learning may be a more suitable approach for the BAA task than classification and regression. In their study, they achieved accurate BAAs with an MAE of approximately 6 months using a proposed method based on a rank-monotonic enhanced ranking CNN.[12] Li et al.[13] developed a two-stage, fully automatic model

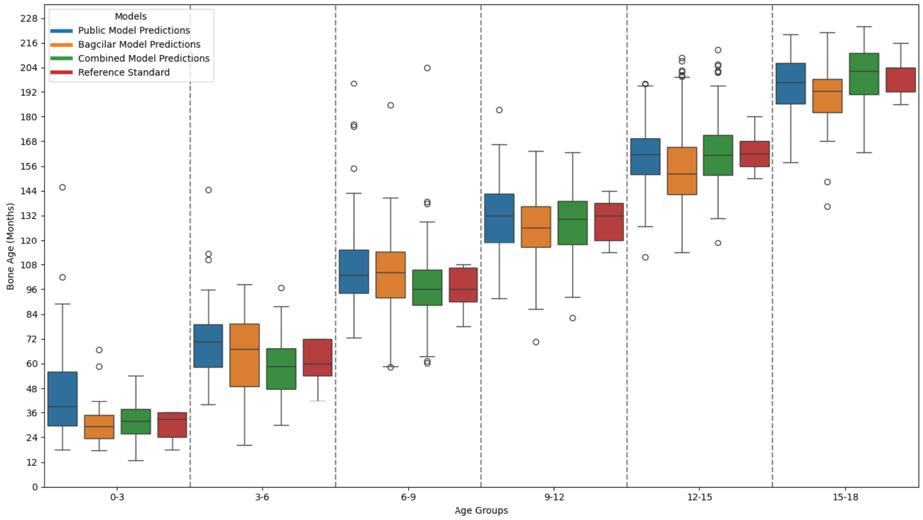


Figure 4. Boxplot of bone age predictions for each model across age groups.

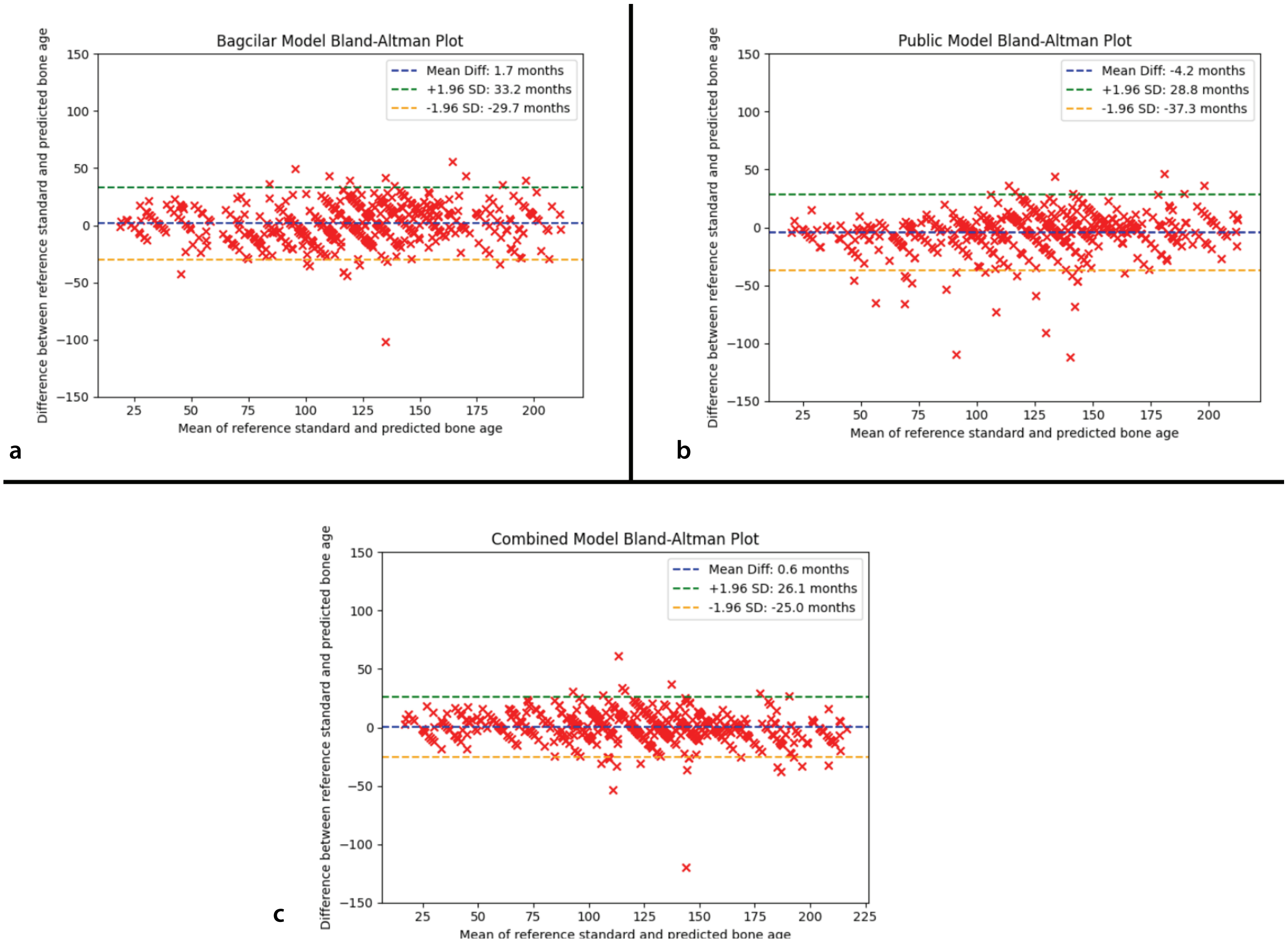


Figure 5. Bland–Altman plots showing the difference between the Bağcılar model, Public model, Combined model, and the reference standard in months, illustrating the variance between the reference standard and the model assessment in the internal testing dataset.

that does not require manual annotation. They demonstrated MAEs of 5.45 months on the RSNA dataset and 3.34 months on a specific dataset.[13] Similarly, our model does not involve annotation. It is an end-to-end model where the bone age is directly assessed by using the cropped hand portion of the X-ray alongside gender information as input. Since our main goal is to show the effect of population differences on model performances, we preferred a validated method that produced the best performance in the RSNA 2017 bone age prediction challenge.

Kim et al.[14] developed a model based on a completely Korean, healthy population, assuming chronological age as the real bone age, such as an atlas study. The developed deep learning model followed a rigorous preprocessing process for estimating chronological age from hand radiographic images. Background removal and transformation networks were applied using manual annotations from an experienced musculoskeletal radiologist. ResNet-50 was used as the basic architecture for age estimation. Compared with their GP-based model, the Korean model showed a lower MAE (8.2 vs. 10.5 months; *P* = 0.002). Additionally, the rate of BAAs within 6 months of chronological age was higher (44.5% vs. 36.4%; *P* = 0.04) with the Korean model. Similarly, our study is also a population-specific model study. In their model, many radiographs were not used as it was based on a non-patient population, such as an atlas. Consequently, there were 21,036 training sets left, and separate test datasets were obtained from two institutions, consisting of 343 and 321 data sets, respectively. Manual annotations were used in creating the model, which is generally time-consuming and cumbersome. Our developed model demonstrated performance comparable with existing models. Utilizing heterogeneous datasets plays a critical role in enhancing model generalizability by exposing the algorithm to a wider range of population and imaging variations. This diversity allows the model to better identify under-represented patterns and reduces the risk of overfitting to specific subsets. The improved performance of the Combined model compared with the locally trained BM underscores the importance of incorporating data from heterogeneous sources in achieving better generalization. Furthermore, increasing the diversity of the included population and imaging modalities can further enhance these models by enabling them to capture relevant information from under-represent-

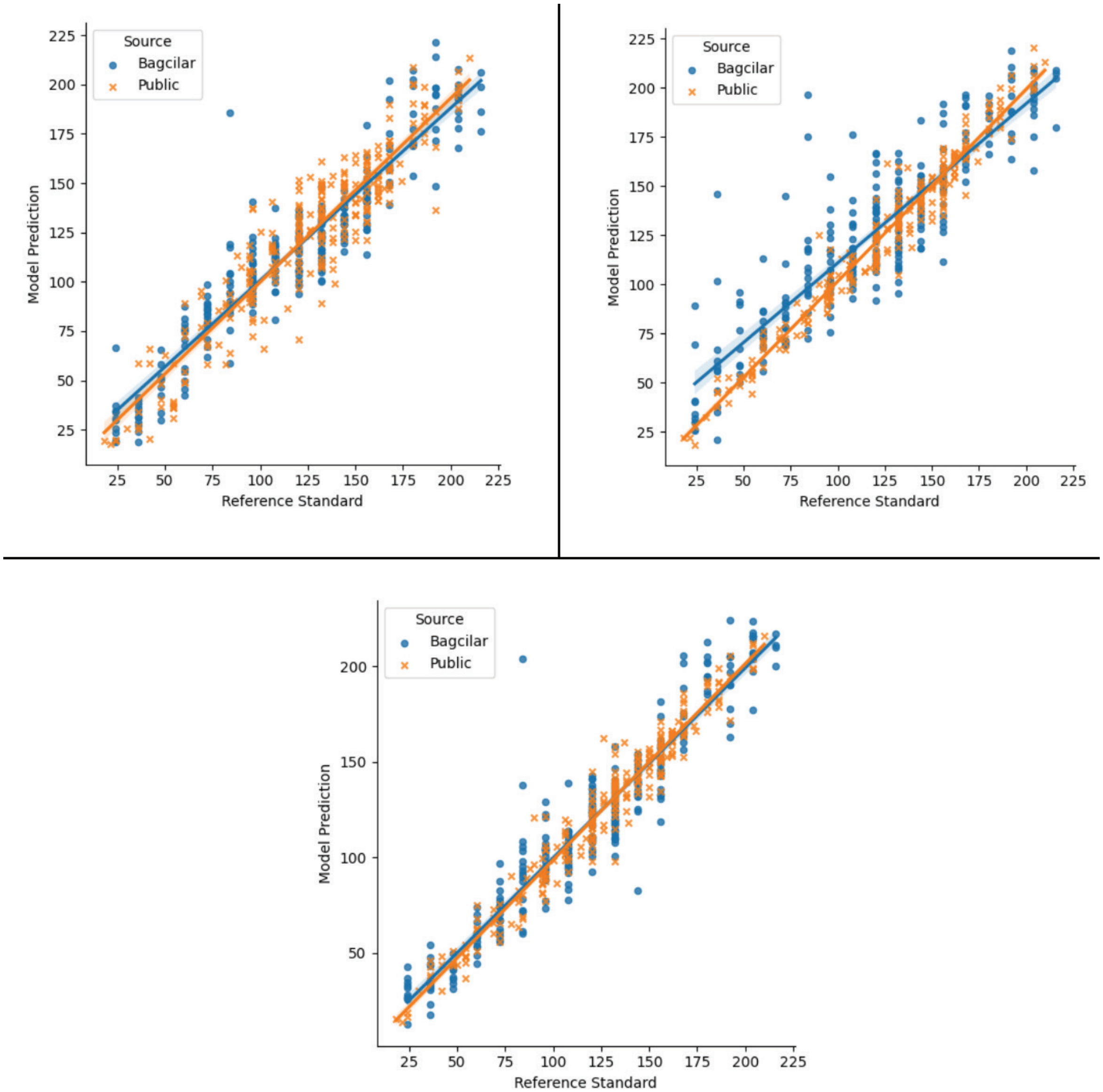


Figure 6. Bone age assessments of the **(a)** Bağcılar model, **(b)** Public model, and **(c)** Combined model on the internal testing dataset in months. Translucent bands around the regression line represent confidence intervals.

ed portions of the data. Greater diversity ultimately strengthens model robustness and improves its capacity to extract meaningful insights. In our model, X-rays requested for BAA and previously reported by radiologists were used. Even though the demographic and diagnostic information was not extensively available in public datasets, models developed using these sources performed worse on local data, indicating important population differences alongside data curation-related information loss.

Spampinato et al.[15] achieved an MAE of 9.6 months using Bonet and the RSNA dataset. Larson et al.[16] achieved an MAE of 6.24 months on the RSNA dataset with a deep residual network structure based on the GP mapping method using ResNet50. Pan et al.[17] used a U-Net model to segment hand mask images from raw X-ray images, employing a deep active learning technique that reduces annotation burden, achieving an MAE of 8.59 months on the RSNA dataset. In our developed Combined model, the MAE value for all data was 9.2 months, 6.9 months for the public dataset, and 11.4 months for the Bağcılar dataset. The described methods, similar to our model, do not involve annotation. Annotation-based methods involve using processed images and adding manual bounding box annotations to these images. These strategies can extract features from specific regions based on prior knowledge and then generate age estimates. Annotation-based methods, which involve additional manual annotations, generally exhibit better performance and higher accuracy compared with annotation-free methods. However, manual annotation is time-consuming and has made it difficult for experimental methods to transition to clinical applications.

Unlike many previous studies that rely on homogeneous datasets, our model was trained and validated using a heterogeneous dataset that includes radiographs from both Bağcılar and public datasets (RSNA and RHPE). This dataset reflects a wide range of imaging conditions, patient demographics, and ethnic backgrounds, thereby increasing the model's robustness and generalizability to real-world clinical settings. The inclusion of such diverse data sources is crucial, as it enables the model to handle a broader spectrum of clinical scenarios, making it more applicable across different populations. The results of our study are promising and highlight the potential of automated BAA models. The Combined model, which integrated data from both Bağcılar and public datasets, demonstrated a high Pearson correlation of 96% with the reference standard, indicating strong predictive accuracy. Specifically, the Public model achieved an MAE of 11.3 months when tested across all test data, while the BM had a higher MAE of 12.6 months. However, when data from both sources were combined, the MAE improved to 9.2 months, highlighting the advantage of integrating diverse datasets to enhance model performance. This improvement could be attributed to the increase in the number of data and the model's increased focus on significant areas due to heterogeneity, enabling the model to account for these differences more effectively, resulting in more accurate and reliable assessments.

The importance of data diversity is further emphasized when examining the model's performance across different age groups. The Combined model showed consistent accuracy across various age ranges, particularly during the critical growth periods of 9–12 and 12–15 years. In contrast, the BM alone exhibited significant deviations from the reference standard in these age groups. This consistency across age groups is crucial for clinical application, as it ensures that the model can be reliably used across a broad patient demographic, minimizing the risk of misclassification and improving overall patient care.

The study has several limitations. Primarily, the limited data quantity has been a key factor, particularly with a small number of radiographs for children under 3 years and a considerably low amount of high-quality data. Another limitation is the absence of a study demonstrating inter-observer differences in our Bağcılar dataset. However, there are many studies in the literature addressing this issue. Additionally, the bone ages in our data were determined using manual methods, such as GP and TW, which, despite having their own limitations, are commonly used in daily practice. Nevertheless, there was no statistically significant difference found between the bone ages obtained with our Combined model and those obtained with clinical methods. Furthermore, the model's performance in older adolescents (aged 15–18 years) showed higher MAEs compared with younger age groups. This could be due to the increased complexity of bone maturation patterns in these age ranges, where small differences in ossification can lead to significant variations in BAA. Addressing this issue may require the development of more specialized models or the inclusion of additional features, such as hormonal markers or elbow and shoulder X-rays, which could provide further insights into bone development in these populations.

In conclusion, this study presents the development of an automatic BAA model using data from Bağcılar, RSNA, and RHPE, making it one of the few studies to incorporate a Turkish population in deep learning-based BAA research. Our model is particularly notable for its ability to integrate heterogeneous data, demonstrating that the inclusion of diverse datasets can significantly enhance model performance. The proposed models offer the advantage of automated analysis without any need for annotation.

Despite the challenges posed by population-level differences, heterogeneous data, and image quality issues, these models can be effectively adopted in various clinical environments, and accuracies can be increased over time with prospectively cumulating data. By enabling more accurate and efficient BAAs, our approach offers valuable support to healthcare professionals, particularly in settings where traditional methods are time-consuming and subject to variability.

Future research should aim to expand the dataset, particularly for younger and older age groups, to improve the model's accuracy and generalizability. Additionally, exploring the incorporation of other clinical parameters, such as hormonal levels, could provide a more comprehensive assessment of bone age, particularly in complex cases. Finally, further validation studies, including prospective trials and cross-institutional collaborations, will be crucial for ensuring the widespread adoption and clinical utility of automated BAA models.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. *Am J Med Sci*. 1959;238(3). [Crossref]
2. Tanner JM. Assessement of skeletal maturity and predicting of adult height (TW2 method). Prediction of adult height. Published online 1983:22-37. [Crossref]
3. Kim JR, Shim WH, Yoon HM, et al. Computerized bone age estimation using deep learning based program: evaluation of the accuracy and efficiency. *AJR Am J Roentgenol*. 2017;209(6):1374-1380. [Crossref]

4. Nadeem MW, Goh HG, Ali A, Hussain M, Khan MA, Ponnusamy VA. Bone age assessment empowered with deep learning: a survey, open research challenges and future directions. *Diagnostics (Basel)*. 2020;10(10):781. [Crossref]

5. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. *Radiology*. 2019;290(2):498-503. [Crossref]

6. Escobar M, González C, Torres F, Daza L, Triana G, Arbeláez P. Hand pose estimation for pediatric bone age assessment. In: Shen D, ed. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science). *Springer, Cham*; 2019. [Crossref]

7. Christoforidis A, Badouraki M, Katzos G, Athanassiou-Metaxa M. Bone age estimation and prediction of final height in patients with beta-thalassaemia major: a comparison between the two most common methods. *Pediatr Radiol*. 2007;37(12):1241-1246. [Crossref]

8. King DG, Steventon DM, O'Sullivan MP, et al. Reproducibility of bone ages when performed by radiology registrars: an audit of Tanner and Whitehouse II versus Greulich and Pyle methods. *Br J Radiol*. 1994;67(801):848-851. [Crossref]

9. Yekeler E. Kemik Yaşı Atlası. First edition; 2021.

10. Lee BD, Lee MS. Automated bone age assessment using artificial intelligence: the future of bone age assessment. *Korean J Radiol*. 2021;22(5):792-800. [Crossref]

11. Pan I, Thodberg HH, Halabi SS, Kalpathy-Cramer J, Larson DB. Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge. *Radiol Artif Intell*. 2019;1(6):e190053. [Crossref]

12. Liu B, Zhang Y, Chu M, Bai X, Zhou F. Bone age assessment based on rank-monotonicity enhanced ranking CNN. *IEEE Access*. 2019;7:120976-120983. [Crossref]

13. Li Z, Chen W, Ju Y, et al. Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. *Front Artif Intell*. 2023;6:1142895. [Crossref]

14. Kim PH, Yoon HM, Kim JR, et al. Bone age assessment using artificial intelligence in Korean pediatric population: a comparison of deep-learning models trained with healthy chronological and Greulich-Pyle ages as labels. *Korean J Radiol*. 2023;24(11):1151-1163. [Crossref]

15. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. *Med Image Anal*. 2017;36:41-51. [Crossref]

16. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. 2018;287(1):313-322. [Crossref]

17. Pan X, Zhao Y, Chen H, Wei D, Zhao C, Wei Z. Fully automated bone age assessment on large-scale hand X-ray dataset. *Int J Biomed Imaging*. 2020;2020:8460493. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253134

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Gastrointestinal bleeding detection on digital subtraction angiography using convolutional neural networks with and without temporal information

Derek Smetanick[1]
Sailendra Naidu[2]
Alex Wallace[2]
M-Grace Knuttinen[2]
Indravadan Patel[2]
Sadeer Alzubaidi[2]

[1]The University of Arizona College of Medicine, Department of Interventional Radiology, Tucson, USA

[2]Mayo Clinic College of Medicine and Science, Department of Radiology, Phoenix, USA

Corresponding author: Derek Smetanick

E-mail: dereksmetanick@arizona.edu

Received 17 February 2025; revision requested 08 April 2025; last revision received 14 June 2025; accepted 22 June 2025.

Epub: 07.08.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.253134

PURPOSE

Digital subtraction angiography (DSA) offers a real-time approach to locating lower gastrointestinal (GI) bleeding. However, many sources of bleeding are not easily visible on angiograms. This investigation aims to develop a machine learning tool that can locate GI bleeding on DSA prior to transarterial embolization.

METHODS

All mesenteric artery angiograms and arterial embolization DSA images obtained in the interventional radiology department between January 1, 2007, and December 31, 2021, were analyzed. These images were acquired using fluoroscopy imaging systems (Siemens Healthineers, USA). Thirty-nine unique series of bleeding images were augmented to train two-dimensional (2D) and three-dimensional (3D) residual neural networks (ResUNet++) for image segmentation. The 2D ResUNet++ network was trained on 3,548 images and tested on 394 images, whereas the 3D ResUNet++ network was trained on 316 3D objects and tested on 35 objects. For each case, both manually cropped images focused on the GI bleed and uncropped images were evaluated, with a superimposition post-processing (SIPP) technique applied to both image types.

RESULTS

Based on both quantitative and qualitative analyses, the 2D ResUNet++ network significantly outperformed the 3D ResUNet++ model. In the qualitative evaluation, the 2D ResUNet++ model achieved the highest accuracy across both 128 × 128 and 256 × 256 input resolutions when enhanced with the SIPP technique, reaching accuracy rates between 95% and 97%. However, despite the improved detection consistency provided by SIPP, a reduction in Dice similarity coefficients was observed compared with models without post-processing. Specifically, the 2D ResUNet++ model combined with SIPP achieved a Dice accuracy of only 80%. This decline is primarily attributed to an increase in false positive predictions introduced by the temporal propagation of segmentation masks across frames.

CONCLUSION

Both 2D and 3D ResUNet++ networks can be trained to locate GI bleeding on DSA images prior to transarterial embolization. However, further research and refinement are needed before this technology can be implemented in DSA for real-time prediction.

CLINICAL SIGNIFICANCE

Automated detection of GI bleeding in DSA may reduce time to embolization, thereby improving patient outcomes.

KEYWORDS

Convolutional neural networks, digital subtraction angiography, gastrointestinal bleeding, image segmentation, interventional radiology, machine learning

You may cite this article as: Smetanick D, Naidu S, Wallace A, Knuttinen MG, Patel I, Alzubaidi S. Gastrointestinal bleeding detection on digital subtraction angiography using convolutional neural networks with and without temporal information. *Diagn Interv Radiol.* 2025;31(5):465-473.

Gastrointestinal (GI) bleeding involves active hemorrhaging from blood vessels within the GI tract. In 5%–10% of cases, patients require either surgery or transcatheter arterial embolization.[1] To perform transcatheter embolization, interventional radiologists often use digital subtraction angiography (DSA) to image the hemorrhage in real time. DSA works by visualizing contrast-opacified vessels and subtracting surrounding anatomical structures, such as soft tissues and bone, to provide a clearer view of the vascular system. The resulting images reveal areas where contrast "pools," indicating the site of bleeding to the interventional radiologist.[2] Although DSA offers a real-time method for locating bleeding, some sources may not be easily visible on angiograms. A neural network used as a decision support tool may assist radiologists in identifying bleeding sites prior to transcatheter arterial embolization.

Convolutional neural networks (CNNs) have demonstrated both accuracy and efficiency in object detection within images.[3] Ronneberger et al.[4] pioneered the U-Net architecture, an extension of the fully convolutional network, which includes a contracting path to capture image context and an expanding path to enable precise localization for segmentation. Neural networks based on the ResUNet architecture have addressed the high computational demands of three-dimensional (3D) convolutional networks.[5] Zhang et al.[6] implemented this design for road detection using a combination of upsampling and downsampling residual blocks. This model was further developed by Jha et al.[7], who proposed the residual neural networks (ResUNet++) architecture and tested it on a segmentation task to identify polyps in two-dimensional (2D) colonoscopy images. Given that ResUNet++ outperformed both the original ResUNet and U-Net models in image segmentation,[7] this architecture serves as the foundation for our model, which aims to segment GI bleeding on DSA images.

Main points

- Automated image segmentation may play a beneficial role in detecting gastrointestinal (GI) bleeding in real time in digital subtraction angiography (DSA) prior to transarterial embolization.
- The three-dimensional (3D) residual neural networks use the temporal resolution from the DSA sequence to predict the bleeding location.
- The two-dimensional neural network outperformed the 3D neural network in segmenting GI bleeding on images.
- Increasing image resolution and using a graphics processing unit may improve both the accuracy of image segmentation and the processing speed, respectively.

This study aims to investigate the utility of a deep learning approach for the automated detection of GI bleeding on DSA images, specifically by comparing 2D and 3D ResUNet++ architectures. We hypothesized that both models could identify bleeding sites, but that one may outperform the other. Our rationale for using a deep learning approach stems from the temporal variability and subtlety of GI bleeds, which may evade human detection on sequential angiographic images. Automated segmentation could assist radiologists by identifying bleeding pixels in real time, potentially reducing time to embolization. This study also evaluates a novel temporal consistency algorithm–superimposition post-processing (SIPP)–to determine whether incorporating temporal bleed memory improves segmentation performance across sequences. We address the following research questions. (1) Can deep learning accurately identify bleeding on DSA? (2) How does a 2D model compare with a 3D model in this context? (3) Does temporal information improve performance when integrated through post-processing?

It is also critical to consider the clinical impact of GI bleeding segmentation in DSA without introducing workflow delays. In practice, a supportive model must identify bleeding sites faster than the interventional radiologist to improve procedural outcomes. Earlier identification could reduce contrast volume, lower radiation exposure, and shorten procedure times.

Methods

Image datasets for training and testing

Mayo Clinic Phoenix approved this study as exempt on 01/31/2024 due to its retrospective nature (IRB application #: 24-000309). Between 2007 and 2021, a total of 96 patients underwent mesenteric artery angiography or arterial embolization DSA procedures for suspected GI bleeding. Of these, 70 patients showed no active extravasation on angiography and were excluded. The remaining 26 patients, who demonstrated confirmed active hemorrhage, were included in the study, as shown in Figure 1. No images were excluded based on patient age, motion artifacts, or image corruption. From the 26 patients, 39 unique image series positive for active hemorrhage were identified by an interventional radiologist and selected for neural network training. These cases involved hemorrhaging in the small and large intestines. On average, each series contained 11 bleeding images. To avoid inflated model performance, data were split at the patient level for training and testing. The bleeding images were cropped to highlight the hemorrhage in higher resolution. The dataset was then augmented by replicating each image nine times, systematically shifting the bleed location to the following regions: upper-left, upper-center, upper-right, middle-left, cen-

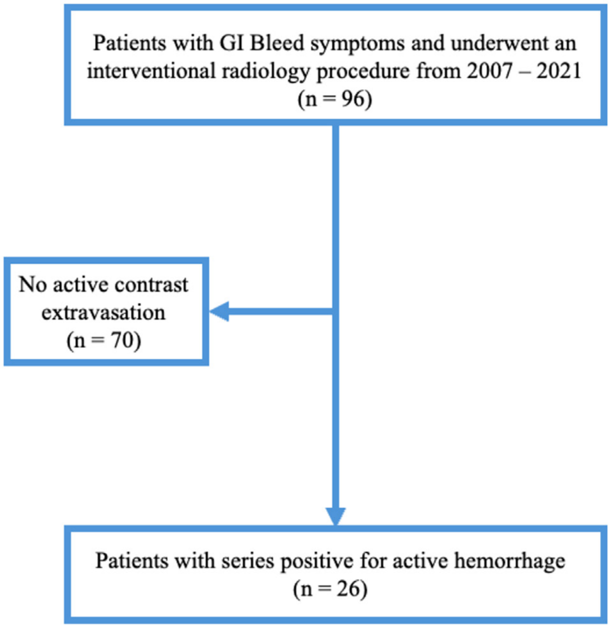

Patients with GI Bleed symptoms and underwent an interventional radiology procedure from 2007 – 2021 (n = 96)
No active contrast extravasation (n = 70)
Patients with series positive for active hemorrhage (n = 26)

Figure 1. Criteria and number of patients from initial retrieval to the final study cohort. GI, gastrointestinal.

ter, middle-right, lower-left, lower-center, and lower-right. This approach increased the dataset size by 900%. Segmentation masks were created manually using Photoshop (Adobe Inc., San Jose, CA, USA) with a thresholding tool to isolate the bleeding. The segmentations displayed bleeding areas in white on a solid black background to produce binary images. The same augmentation technique was applied to the segmentation masks to ensure proper pixel alignment with the original images. Table 1 summarizes the number of GI bleeding-positive and-negative images in the test set after augmentation.

Although 70 patients had no visible extravasation, including all of their image sequences as negative controls would have created a heavily imbalanced dataset. Instead, non-bleeding frames from within the same DSA sequences of the 26 bleeding-positive patients were used. These frames provided sufficient negative control data for training and testing while preserving representative angiographic conditions and avoiding overrepresentation of non-bleeding cases. Moreover, the model's task was to identify where bleeding occurred, rather than whether bleeding was present. In this context, even within bleeding-positive images, the majority of pixels are negative for bleeding.

Both 128 × 128-pixel and 256 × 256-pixel images were used to train separate 2D CNNs, whereas only 128 × 128-pixel images were used to train a 3D CNN for image segmentation. A post-processing technique–superimposing all masks within a series into a single mask for final image segmentation–was applied to both 2D and 3D segmentations. In total, these three networks were evaluated across four distinct testing scenarios: (1) uncropped images from the DSA sequence, (2) cropped images focusing on the bleed, (3) uncropped images with the SIPP technique applied, and (4) cropped images with the SIPP technique.

Superimposition post-processing technique

The SIPP technique algorithm was developed to address the temporal inconsistency of GI bleeding predictions across angiographic image sequences. Bleeding may not be clearly visible in every frame. To mitigate this, SIPP enforces temporal continuity by propagating the presence of bleeding pixels forward through the predicted image sequence. For each frame in the sequence, the model produces a binary segmentation mask M_t, where each pixel is labeled either as 1 (bleeding present) or 0 (no bleeding). The mask $M_t \in \{0,1\}^{\{H \times W\}}$ is a 2D grid with the same height (H) and width (W) as the original image and represents the classification of each pixel. SIPP modifies these predictions by updating each new mask M_t to include any pixel that was previously marked as bleeding. This is defined as:

$$\widehat{M}_t = M_t \vee \widehat{M}_{\{t-1\}}$$

Where ∨ represents a logical OR operation performed on all pixels between the current prediction M_t and the accumulated mask from the previous frame $\widehat{M}_{t-1}$. This rule ensures that once a pixel is marked as bleeding, it remains labeled as such in all following frames of the DSA. This effectively preserves prior bleeding evidence even if the current frame is less confident. This simple yet effective mechanism improves temporal consistency and reduces missed detections due to frame-level variability. A flowchart is provided in Figure 2 to further explain the SIPP technique.

Deep neural network architecture

Both a 3D and a 2D ResUNet++ were constructed based on the architecture shown in Figure 3. A single DSA frame served as the input image for the 2D network, whereas the entire DSA series served as the input for the 3D network. After entering the network, the image passed through a series of convolutional layers with a 3 × 3 kernel size and increasing numbers of filters (16, 32, 48, and 64), referred to as the encoding phase. Each convolutional layer was followed by batch normalization to improve training speed and

Table 1. The number of images positive for gastrointestinal bleeding and the number of control images negative for gastrointestinal bleeding were tabulated for each of the different models

	GI bleed images	Control images
2D uncropped 128 × 128	343	95
2D cropped 128 × 128	321	73
2D uncropped 256 × 256	354	84
2D cropped 256 × 256	321	73
3D uncropped 128 × 128	343	281
3D cropped 128 × 128	273	287
2D uncropped 128 × 128 w/SIPP	343	281
2D cropped 128 × 128 w/SIPP	3195	2421
2D uncropped 256 × 256 w/SIPP	354	270
2D cropped 256 × 256 w/SIPP	3196	2420
3D uncropped 128 × 128 w/SIPP	319	241
3D cropped 128 × 128 w/SIPP	273	287

GI, gastrointestinal; SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional.

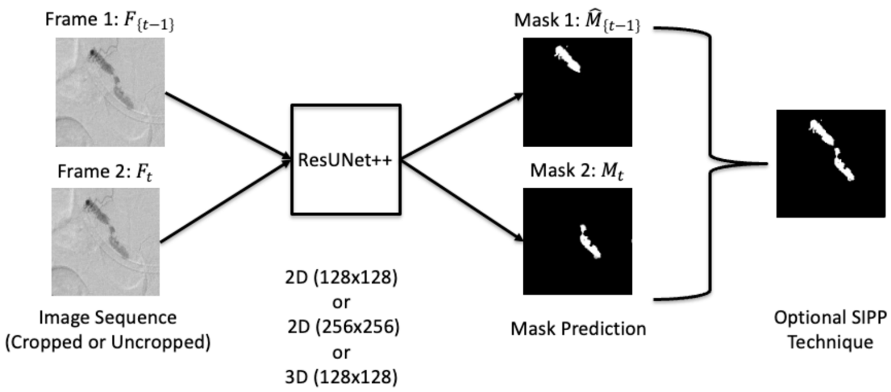


Figure 2. Schematic of the image segmentation pipeline with the optional superimposition post-processing technique. Each image sequence (cropped or uncropped) is passed through a ResUNet++ model configured as either 2D (128 × 128 or 256 × 256) or 3D (128 × 128) to generate frame-wise predicted masks. If applied, the SIPP technique performs a logical OR operation between the current and previous masks to enhance temporal consistency in bleeding detection. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

stability by standardizing the inputs. A rectified linear unit (ReLU) activation function was then applied to introduce non-linearity, enabling the network to learn complex patterns and shapes. The spatial dimensions of the feature maps were reduced through 2D max pooling after each convolutional layer, allowing the network to retain the most important features. After the encoding phase, the features were upsampled back to the original image size using transposed convolutions with a 3 × 3 kernel size and decreasing numbers of filters (64, 48, 32, and 16). Each layer was again followed by batch normalization and ReLU activation. At each step of the decoding path, the feature maps were concatenated with the corresponding feature maps from the encoding phase, allowing the network to leverage both low-level and high-level features for more accurate segmentation. The final output layer consisted of a 1 × 1 convolutional layer with a single filter and sigmoid activation. The resulting segmentation map assigned each pixel a predicted class. Both the 2D and 3D ResUNet++ models described in this study were deep learning architectures designed for semantic segmentation tasks. Although implemented as machine learning models during training and inference, their structural design–comprising convolutional layers, encoding–decoding paths, and feature concatenations–was fundamentally that of deep learning architectures.

The convolutional ResUNet++ networks were implemented using the Keras framework[8] with a TensorFlow backend (Google, Inc.),[9] using Python version 3.9. All experiments were performed on a computer with an Intel Core i7-8700 central processing unit (CPU) @ 3.20 GHz (Intel). To prevent overfitting, a smaller learning rate of 1.0×10^{-4} was used during training to avoid issues such as model instability or failure to converge. Data augmentation was also applied to artificially increase dataset variability, further helping to mitigate overfitting. The architecture was optimized using the Adam optimizer. A batch size of 20 and 20 training epochs were used for each experiment to maintain consistency. Binary cross-entropy loss was employed to optimize the segmentation task.

Quantitative evaluation

The MATLAB software (MathWorks, Natick, MA, USA) was used to quantify the results from predicted and actual masks by measuring mask overlap. A pixel-by-pixel analysis identified true positive pixels (TPP), true negative pixels (TNP), false positive pixels (FPP), and false negative pixels (FNP). TP and TN values were calculated by dividing TPP and TNP by the respective numbers of positive and negative pixels in the ground truth. FP and FN values were calculated by dividing FPP and FNP by the total number of pixels in the ground truth, respectively. These scores were computed for each of the 12 experiments. Dice similarity coefficients (DSCs) were calculated to quantitatively assess the spatial overlap between the predicted segmentation masks and the ground truth annotations. For each model and imaging configuration, the Dice coefficients were computed on a per-sample basis and summarized as mean values with corresponding 95% confidence intervals (CIs). All Dice analysis was performed as part of the quantitative evaluation.

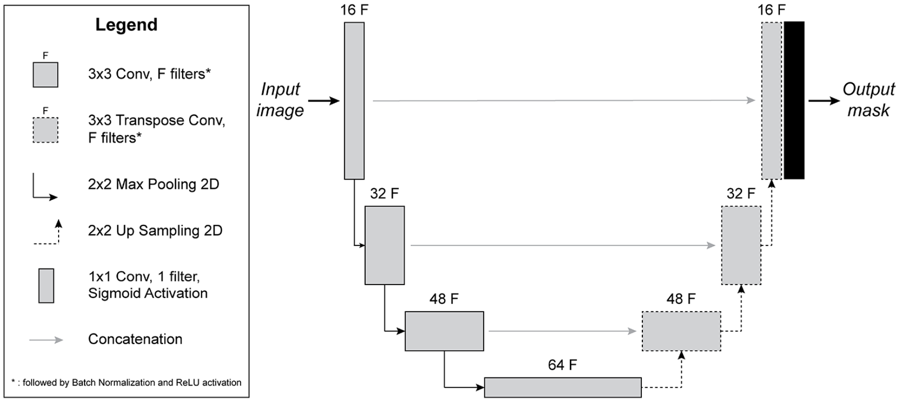


Figure 3. Neural network architecture used in both the 2D ResUNet++ and 3D ResUNet++ models. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

Qualitative evaluation

Although quantitative metrics provided objective measures of segmentation accuracy, a qualitative evaluation was also conducted to assess clinical relevance. This evaluation was performed by a single evaluator–a medical student–who visually compared the predicted segmentation masks with both the ground truth masks and the original DSA images. Each image was classified as TP, TN, FP, or FN using the same definitions applied in the quantitative evaluation. To aid in the classification process, a MATLAB script was used to help identify TN, FP, and FN images. A prediction was considered a TP if the white pixels in the predicted mask overlapped with those in the ground truth mask. This overlap was initially assessed visually and subsequently verified to ensure at least one pixel of overlap, which served as a safeguard to minimize human error in classification. This minimal overlap threshold was intentionally selected based on the model's intended clinical use: to serve as a real-time assistive tool during embolization procedures. In such settings, even a small correctly flagged area could be sufficient to prompt further investigation by an interventional radiologist. The model is not intended to deliver volumetric precision but rather to alert clinicians to potential regions of bleeding. Cases where the white pixels of the predicted and ground truth masks overlapped but also included some FP areas were generally classified as TPs, unless the FP region exceeded 10% of the image area. All ground truth segmentation masks were manually created using a thresholding method and validated by a team of fellowship-trained interventional radiologists to ensure accuracy before comparison.

Statistical analysis

A one-way analysis of variance single-factor test was conducted in MATLAB to determine the statistical significance within the TP results of the quantitative evaluation. An α value of 0.05 was selected, with the null hypothesis stating that there is no statistically significant difference among the various networks. If the *P* values obtained from the analysis were less than α, the null hypothesis was rejected, indicating a statistically significant difference between the networks. In cases where such a difference was detected, a Tukey–Kramer post-hoc test was performed to identify which networks exhibited this disparity.

Results

Quantitative evaluation

The accuracy, intersection over union (IoU), loss, and precision obtained during the initial training and testing of the 2D 128

× 128 ResUNet++, 2D 256 × 256 ResUNet++, and 3D 128 × 128 ResUNet++ models are presented in Table 2. The accuracy and precision scores were comparable across all three networks. The 3D 128 × 128 ResUNet++ exhibited the lowest IoU at 0.06.

Depicted in Figure 4 are the accuracy scores for the 12 different 2D and 3D ResUNet++ structures based on a DSA framewise basis. These results are summarized in Table 3, whereas outcomes of the statistical analysis are presented in Table 4. A statistically significant improvement in the accuracy score was observed using the SIPP technique for all six different ResUNet++ structures: 2D uncropped 128 × 128, 2D cropped 128 × 128, 2D uncropped 256 × 256, 2D cropped 256 × 256, 3D uncropped 128 × 128, and 3D cropped 128 × 128 compared with the control trial. Notably, there was no statistical significance between the 2D uncropped 128 × 128 model and the 2D uncropped 256 × 256 model, the 2D uncropped 128 × 128 model and the 3D uncropped 128 × 128 model, the 2D cropped 128 × 128 model and the 2D cropped 256 × 256 model, and the 2D uncropped 256 × 256 and the 3D uncropped 128 × 128 model when the SIPP method was not used. The largest mean accuracy values were 0.961 and 0.956 for the 2D cropped 128 × 128 with SIPP model and the 2D cropped 256 × 256 with SIPP model, respectively. There was no statistically significant difference between the accuracy values for these two different networks. Both models had a statistically significantly higher accuracy than the 3D cropped 128 × 128 model with SIPP. The 2D cropped 128 × 128 and the 2D cropped 256 × 256 models also maintained the highest accuracy for models without SIPP, with accuracy scores of 0.853 and 0.812, respectively. There was no statistically significant difference between these two models. These models had a statistically significantly higher accuracy than the 3D cropped 128 × 128 model. The 2D uncropped 256 × 256 with SIPP model had a statistically significantly higher accuracy than the 2D uncropped 128 × 128 with SIPP model and the 3D uncropped 128 × 128 with SIPP model. Meanwhile, the 2D uncropped 128 × 128 with SIPP model had a statistically significantly higher accuracy than the 3D uncropped 128 × 128 with SIPP model.

DSCs for each model configuration, with and without SIPP, are summarized in Table 5. Compared with their corresponding original models, the use of SIPP resulted in statistically significant reductions in Dice coefficients for the 2D uncropped 128 × 128 model [from 0.042 (95% CI: 0.0264–0.0575) to 0.019 (95% CI: 0.0133–0.0248)] and the 2D cropped 128 × 128 model [from 0.798 (95% CI: 0.7720–0.8232) to 0.190 (95% CI: 0.1839–0.1959)]. Similarly, the 2D cropped 256 × 256 model exhibited a substantial decrease in Dice score when SIPP was applied [from 0.797 (95% CI: 0.7708–0.8223) to 0.278

Table 2. The results from training the 2D ResUNet++ on 128 × 128-pixel and 256 × 256-pixel images, as well as the 3D ResUNet++ on 128 × 128-pixel images, are tabulated. The metrics of accuracy, intersection-over-union, and precision were included for all three neural networks

Method	Accuracy	IoU	Precision
2D 128 × 128 ResUNet++	**0.95**	0.62	0.99
2D 256 × 256 ResUNet++	**0.96**	0.61	0.98
3D 128 × 128 ResUNet++	**0.96**	0.06	0.95

2D, two-dimensional; 3D, three-dimensional; IoU, intersection-over-union; ResUNet++, residual neural networks.

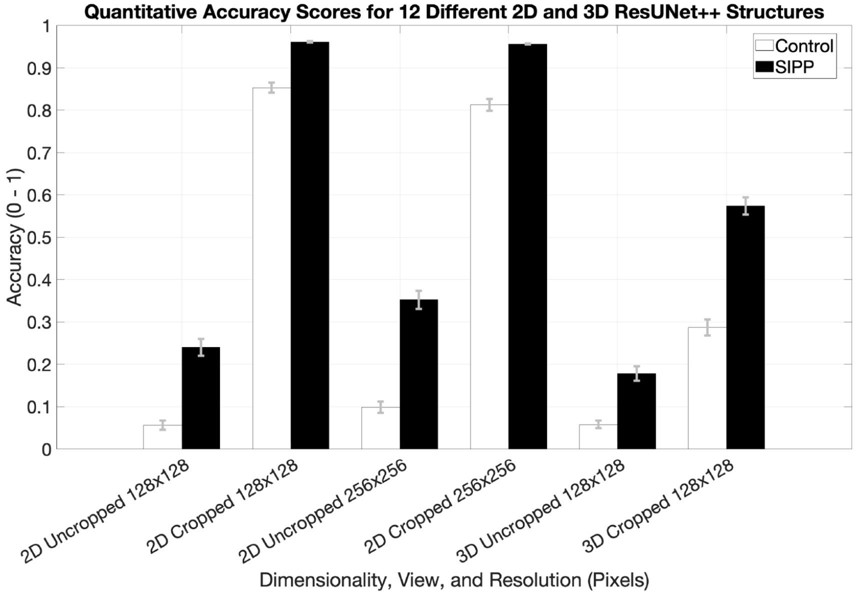


Figure 4. Bar graph showing differences in the quantitative accuracy of segmentation results for the 12 testing scenarios. The control represents cases without post-processing, whereas the other cases used the superimposition post-processing technique. Error bars indicate one standard deviation. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

Table 3. The true positive, true negative, false positive, and false negative rates were tabulated for the 12 different cases for the quantitative results

	True positive	True negative	False positive	False negative
2D uncropped 128 × 128	0.056	0.966	0.034	0.001
2D cropped 128 × 128	0.853	0.996	0.003	0.002
2D uncropped 256 × 256	0.099	0.978	0.022	0.001
2D cropped 256 × 256	0.812	0.997	0.002	0.003
3D uncropped 128 × 128	0.058	0.998	0.002	0.001
3D cropped 128 × 128	0.287	0.999	0.001	0.006
2D uncropped 128 × 128 w/SIPP	0.240	0.919	0.081	0.001
2D cropped 128 × 128 w/SIPP	0.961	0.898	0.098	0
2D uncropped 256 × 256 w/SIPP	0.352	0.972	0.028	0.001
2D cropped 256 × 256 w/SIPP	0.956	0.946	0.051	0.001
3D uncropped 128 × 128 w/SIPP	0.178	0.985	0.015	0.001
3D cropped 128 × 128 w/SIPP	0.573	0.982	0.018	0.003

SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional.

(95% CI: 0.2703–0.2858)]. In contrast, for the 2D uncropped 256 × 256, 3D uncropped 128 × 128, and 3D cropped 128 × 128 models, although minor changes in Dice coefficients were observed, the corresponding 95% CIs overlapped. Therefore, these changes are not statistically significant based on CI analysis. Overall, these results indicate that although SIPP altered segmentation performance, its effects were not uniformly beneficial across all models, and in some cases, it led to considerable declines in segmentation accuracy. These results are visually represented in Figure 5.

Qualitative evaluation

An example image from the 2D cropped 128 × 128 model, the 2D uncropped 256 × 256 model, and the 3D cropped 128 × 128 model is shown in Figure 6. The original image is on the left, the ground truth is in the middle, and the predicted image is on the right. Each image was reviewed manually for quality control to compare the ground truth with the predicted image. The results from the qualitative evaluation are displayed in Table 6 and plotted in Figure 7.

From the highest TP accuracy count to the lowest TP count, the twelve networks ranked as follows for the qualitative results: 2D cropped 256 × 256 with SIPP, 2D cropped 128 × 128 with SIPP, 2D cropped 128 × 128, 2D cropped 256 × 256, 3D cropped 128 × 128 with SIPP, 3D cropped 128 × 128, 2D uncropped 256 × 256 with SIPP, 2D uncropped 128 × 128 with SIPP, 3D uncropped 128 × 128 with SIPP, 2D uncropped 256 × 256, 3D uncropped 128 × 128, and 2D uncropped 128 × 128. The range of TP accuracy was from 0.999 to 0.122. The models using the SIPP technique had higher accuracy rates than their control counterparts. The ranking order was similar to the TP accuracies from the quantitative section. The main differences in the qualitative list compared with the quantitative list are that 2D cropped 256 × 256 with SIPP marginally outperformed 2D cropped 128 × 128 with SIPP, and 3D cropped 128 × 128 marginally outperformed 2D uncropped 256 × 256 with SIPP.

Discussion

The widely used U-Net architecture for medical image segmentation is leveraged in this study through the ResUNet++ variant. ResUNet preserves input dimensions and minimizes information loss, as described by Yousef et al.[10], whereas U-Net++ incorporates nested skip connections to enhance semantic segmentation, as detailed by Zhou et al.[11] The effectiveness of ResUNet++ has been validated by Jha et al.[7], supporting its use in segmentation tasks. This study evaluates segmentation accuracy using standard analyses similar to those employed in cone-beam CT acquisitions for prostate treatments.[12]

Using 2D ResUNet++ for DSA images offers distinct advantages over 3D ResUNet++. Although 3D ResUNet++ benefits from incorporating temporal information across image sequences, it did not outperform the 2D model. For uncropped DSA images, 3D ResUNet++ performed similarly to 2D ResUNet++, likely because downscaling the original 1064 × 1064-pixel images to 128 × 128 or 256 × 256 pixels led to a loss of crucial spatial detail. This limitation was addressed by manually cropping the images to focus specifically on bleeding regions, allowing the bleed to occupy approximately 5% of the image area and substantially improving training and testing resolution. This process improved segmentation accuracy for both 2D and 3D ResUNet++ models, emphasizing the importance of image resolution for accurate GI bleeding localization and favoring 2D model performance. These quantitative findings were further supported by qualitative assessments.

The Keras framework[8] was used to evaluate accuracy, IoU, loss, and precision metrics during the training of both 2D and 3D ResUNet++ models on cropped images. Across

Table 4. A one-way analysis of variance with a Tukey–Kramer post-hoc test was conducted, and the resulting *P* values were tabulated to compare different models. The significance level (α) was set at 0.05. Statistical differences in segmentation accuracy were observed for models with *P* values less than α

Model 1	Model 2	*P* value
2D uncropped 128 × 128	2D uncropped 128 × 128 w/SIPP	<0.001
2D cropped 128 × 128	2D cropped 128 × 128 w/SIPP	<0.001
2D uncropped 256 × 256	2D uncropped 256 × 256 w/SIPP	<0.001
2D cropped 256 × 256	2D cropped 256 × 256 w/SIPP	<0.001
3D uncropped 128 × 128	3D uncropped 128 × 128 w/SIPP	<0.001
3D cropped 128 × 128	3D cropped 128 × 128 w/SIPP	<0.001
2D uncropped 128 × 128	2D uncropped 256 × 256	0.098
2D uncropped 128 × 128	3D uncropped 128 × 128	1.000
2D cropped 128 × 128	2D cropped 256 × 256	0.192
2D uncropped 256 × 256	3D uncropped 128 × 128	0.143
2D cropped 128 × 128 w/SIPP	2D cropped 256 × 256 w/SIPP	0.995
2D cropped 256 × 256 w/SIPP	3D cropped 128 × 128 w/SIPP	<0.001
2D cropped 128 × 128 w/SIPP	3D cropped 128 × 128 w/SIPP	<0.001
2D cropped 256 × 256	3D cropped 128 × 128	<0.001
2D cropped 128 × 128	3D cropped 128 × 128	<0.001
2D uncropped 256 × 256 w/SIPP	2D uncropped 128 × 128 w/SIPP	<0.001
2D uncropped 256 × 256 w/SIPP	3D uncropped 128 × 128 w/SIPP	<0.001
2D uncropped 128 × 128 w/SIPP	3D uncropped 128 × 128 w/SIPP	0.001

SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional.

Table 5. Mean Dice similarity coefficients and corresponding 95% confidence intervals are reported for each model configuration, comparing results with and without superimposition post-processing. Statistically significant differences, identified by non-overlapping confidence intervals, are indicated in bold

Model	Original mean (95% CI)	SIPP mean (95% CI)
2D uncropped 128 × 128	0.042 [0.0264, 0.0575]	0.019 [0.0133, 0.0248]
2D cropped 128 × 128	0.798 [0.7720, 0.8232]	0.190 [0.1839, 0.1959]
2D uncropped 256 × 256	0.069 [0.0493, 0.0893]	0.065 [0.0533, 0.0757]
2D cropped 256 × 256	0.797 [0.7708, 0.8223]	0.278 [0.2703, 0.2858]
3D uncropped 128 × 128	0.054 [0.0381, 0.0694]	0.064 [0.0479, 0.0795]
3D cropped 128 × 128	0.334 [0.2955, 0.3731]	0.281 [0.2523, 0.3104]

CI, confidence interval; SIPP, superimposition post-processing.

all metrics, the 2D ResUNet++ outperformed its 3D counterpart. Higher IoU indicates superior segmentation, and although 3D ResUNet++ had a lower IoU, its performance improved following the application of the SIPP technique. SIPP accumulates bleeding-positive pixels across sequential frames, enhancing temporal consistency. Originally applied to 3D ResUNet++ to address intermittent bleeding visibility, SIPP also improved segmentation performance for 2D ResUNet++ models. However, quantitative analysis revealed that SIPP increased FP rates, as errors persisted across frames, whereas FN rates remained relatively unaffected by post-processing. The increase in FP rates resulting from the SIPP technique also contributed to a decrease in DSCs across most models. Since the Dice coefficient is sensitive to both FPs and FNs, the propagation of errors across sequential frames reduced overall spatial overlap precision, despite improvements in bleeding pixel continuity. This tradeoff highlights an important limitation of SIPP: although it enhances temporal consistency and bleed detection sensitivity, it may compromise segmentation specificity, as reflected in Dice score reductions.

Since transarterial embolization is performed in real time under fluoroscopy, model inference speed is critical. Doubling image resolution from 128 × 128 to 256 × 256 pixels nearly quadrupled the model runtime. Interestingly, there was no statistically significant difference in runtime between 2D ResUNet++ trained on 256 × 256 images and 3D ResUNet++ trained on 128 × 128 images, indicating that 3D models also demand substantial computational resources. Prior studies using graphics processing unit (GPU) hardware have demonstrated that 512 × 512-pixel images can be segmented in less than 1 second,[4] suggesting that GPU acceleration could greatly enhance model performance and enable the training of higher-resolution 3D networks. Although training on a GPU would have considerably expedited model development, cost constraints and limited institutional access to dedicated GPU hardware necessitated CPU-based training in this study. For future real-time deployment, GPU acceleration will be critical to support high-throughput inference and maintain clinical usability.

Ground truth segmentation quality greatly impacts machine learning model performance. To ensure reliable labeling, ground truth masks underwent rigorous validation. A chart review was conducted to confirm each bleeding episode's anatomical site, with

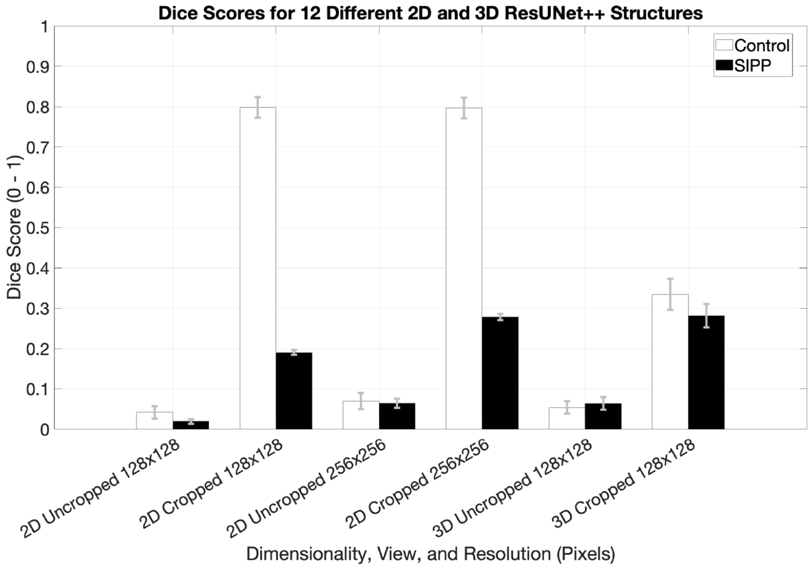


Figure 5. Mean Dice similarity coefficients and 95% confidence intervals for twelve different 2D and 3D ResUNet++ segmentation models, evaluated with and without SIPP. Bars indicate the mean DSC values, and error bars represent the corresponding 95% confidence intervals. Dice coefficients range from 0 (no overlap) to 1 (perfect overlap). SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks; DSC, Dice similarity coefficients.

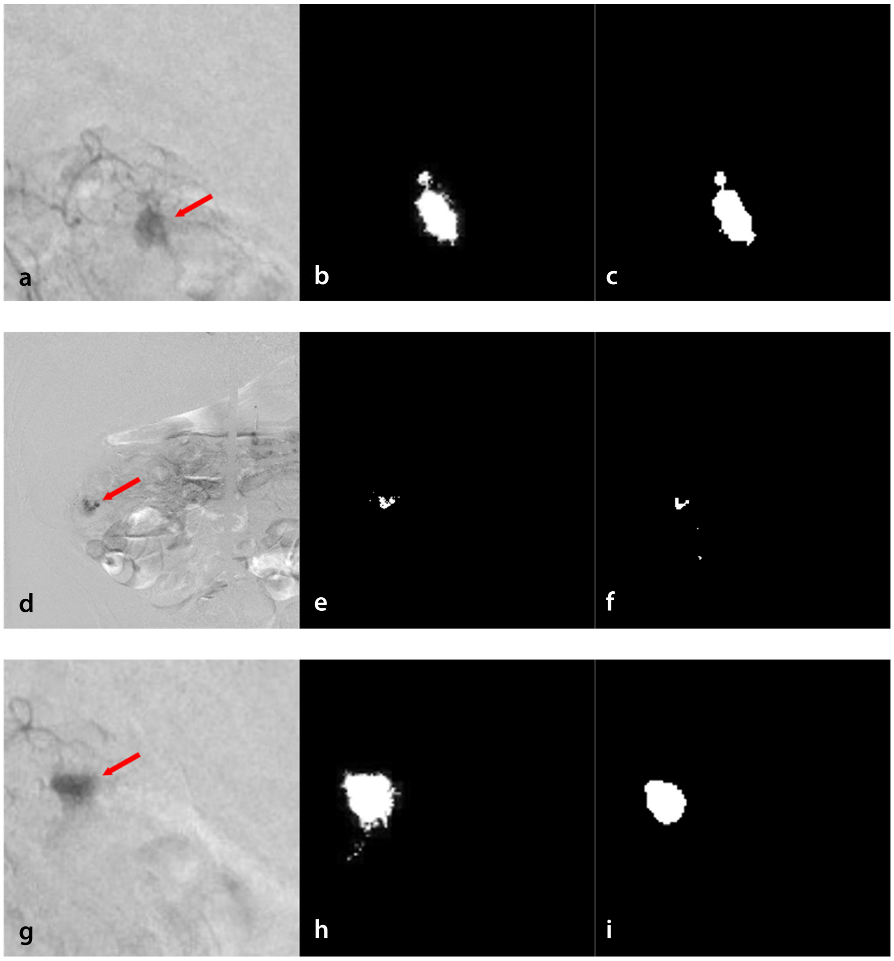


Figure 6. Results for the 2D and 3D ResUNet++ models: **(a)** original image tested on the 2D cropped 128 × 128 model; **(b)** ground truth; **(c)** predicted image; **(d)** image tested on the 2D uncropped 256 × 256 model; **(e)** corresponding ground truth; **(f)** predicted image; **(g)** image tested on the 3D cropped 128 × 128 model; **(h)** ground truth; **(i)** predicted image. Red arrows in **(a)**, **(d)**, and **(g)** point toward the GI bleed. GI, gastrointestinal; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

ambiguities resolved in consultation with fellowship-trained interventional radiologists. Manual image thresholding using MATLAB assigned white pixel values (255) to bleeding regions and black (0) to background areas, creating binary masks. Although manual segmentation is labor-intensive, it remains the gold standard for validation, as emphasized by Yepes-Calderon et al.[13] Potential operator bias was minimized by having a single individual perform all segmentations. Data augmentation techniques, including cropping and translation, as described by Shorten et al.,[14] expanded the training dataset. Cropping enhanced effective resolution, and systematic translations increased the dataset size by 900%. Due to image series grouping for 3D training, fewer images were available for the 3D models compared with the 2D models. Further research is needed to assess how expanded 3D datasets could impact model performance.

When comparing Tables 2 and 4, an apparent contradiction emerges because models such as 2D cropped 128 × 128 with SIPP and 2D cropped 256 × 256 with SIPP show high true positive rates (TPRs) and true negative rates (TNRs) in Table 2 yet exhibit a notable decrease in DSCs in Table 4. This discrepancy stems from fundamental differences in how these metrics are calculated. TPRs and TNRs incorporate TNs, which dominate pixel-based segmentation tasks and can inflate performance metrics, particularly when background regions vastly outnumber bleeding pixels. In contrast, the Dice coefficient is a spatial overlap metric that does not consider TNs and is highly sensitive to both FPs and FNs. Since the SIPP technique propagates predictions across frames, it can increase FPs and lead to reduced Dice scores despite stable or improved TPRs and TNRs. This tradeoff underscores a central tension in medical image segmentation: balancing sensitivity and temporal consistency with spatial specificity. Given the model's intended role as an assistive tool during real-time embolization, the slight increase in FPs introduced by SIPP may be clinically acceptable if it ensures that critical bleeding regions are not missed. Both methods were incorporated in this study for transparency.

In Table 6, some models display a TNR of 1.0 while still reporting a nonzero FPR. This discrepancy stems from differences in denominator definitions: TPs and TNs were calculated relative to the number of positive and negative pixels in the ground truth, whereas FPs and FNs were normalized over the total number of pixels in the image. As a result, even a small number of FPPs yields a measurable FPR despite a perfect TNR. This normalization strategy was chosen to consistently reflect prediction error impacts across images of varying sizes and class balances.

Recent studies have further demonstrated the potential of machine learning for DSA-based bleeding detection. Barash et al.[15] utilized a CNN to classify DSA images as either normal or containing active bleeding, achieving an area under the curve of 85.0% and an accuracy of 77.43%. Similarly, Liu et al.[16] introduced a method using parametric color imaging to enhance DSA sequences and better localize bleeding points. Additionally, Min et al.[17] developed a two-stage deep learning model, "InterNet," to detect active abdominal arterial bleeding on emergency DSA images. Their model considerably improved workflow efficiency, reducing radiologist interpretation time from 84.88 to 43.78 seconds. This highlights the potential of artificial intelligence tools to expedite bleeding detection during high-stakes procedures. Compared with these classification-based approaches, the present study focuses on semantic segmentation to directly identify and localize bleeding regions at the pixel level. Furthermore, our study introduces the SIPP technique to enhance temporal consistency.

Table 6. The true positive, true negative, false positive, and false negative rates were tabulated for the 12 different cases in the qualitative results

	True positive rate	True negative rate	False positive rate	False negative rate
2D uncropped 128 × 128	0.122	1	0.402	0.285
2D cropped 128 × 128	0.969	1	0.003	0.023
2D uncropped 256 × 256	0.22	1	0.338	0.292
2D cropped 256 × 256	0.953	1	0	0.038
3D uncropped 128 × 128	0.163	1	0.149	0.311
3D cropped 128 × 128	0.597	1	0.014	0.182
2D uncropped 128 × 128 w/SIPP	0.408	1	0.325	0
2D cropped 128 × 128 w/SIPP	0.997	1	0.002	0
2D uncropped 256 × 256 w/SIPP	0.571	1	0.244	0
2D cropped 256 × 256 w/SIPP	0.999	1	0	0
3D uncropped 128 × 128 w/SIPP	0.276	1	0.234	0.179
3D cropped 128 × 128 w/SIPP	0.853	1	0.018	0.054

SIPP, superimposition post-processing.

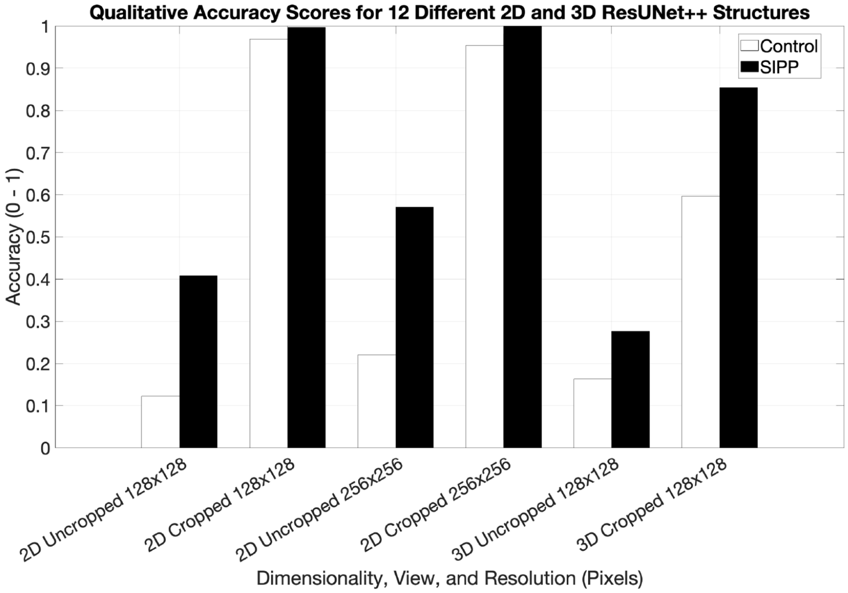


Figure 7. Bar graph showing differences in the qualitative accuracy of segmentation results for the 12 testing scenarios. The "control" represents cases without post-processing, whereas the other cases used the SIPP technique. SIPP, superimposition post-processing; 2D, two-dimensional; 3D, three-dimensional; ResUNet++, residual neural networks.

Limitations

Several limitations must be acknowledged. First, the sample size was relatively small (26 patients), limiting statistical power and generalizability. Second, no external validation set from a separate institution was used, raising concerns about model robustness across different imaging protocols and vendors. Third, training was performed on a CPU rather than a GPU, which constrained image resolution, limited model complexity, slowed inference speeds, and necessitated manual cropping of bleeding regions to preserve resolution for training. Although necessary under computational constraints, manual cropping introduces potential bias and is not feasible for clinical deployment. In future work, GPU-accelerated training and inference will be pursued to allow the processing of entire uncropped DSA images at full resolution. Alternatively, a sliding window approach could be implemented, whereby the model systematically analyzes overlapping regions of the full image to detect bleeding without manual preselection. Fourth, the dataset included only bleeding-positive cases, limiting the ability to fully assess FPRs and overall specificity. Future studies can address these limitations by expanding datasets, incorporating external validation cohorts, utilizing GPU acceleration, and including negative control cases to better assess real-world model performance.

In conclusion, this study investigated the use of 2D ResUNet++ and 3D ResUNet++ neural network models to segment GI bleeding in DSA prior to transarterial embolization. Most notably, the 2D ResUNet++ outperformed the 3D ResUNet++ model. In qualitative analysis, the 2D ResUNet++ model achieved the highest accuracy, ranging from 95% to 97%, when enhanced with the SIPP technique. The highest DSC observed was 80% for the same model. Both quantitative and qualitative analyses highlight the potential feasibility of this model for real-time bleeding segmentation in the interventional radiology suite. Furthermore, training and testing with more 3D data are recommended to further refine the performance of the 3D ResUNet++ model. Incorporating GPU acceleration is also advised for faster processing. Future studies should evaluate the impact of these tools on DSA images in real time.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Shin JH. Refractory gastrointestinal bleeding: role of angiographic intervention. *Clin Endosc*. 2013;46(5):486-491. [Crossref]
2. Taslakian B, Ingber R, Aaltonen E, Horn J, Hickey R. Interventional radiology suite: a primer for trainees. *J Clin Med*. 2019;8(9):1347. [Crossref]
3. Ajit A, Acharya K, Samanta A. A review of convolutional neural networks. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE). IEEE; 2020:1-5. [Crossref]
4. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *Arxiv*. [Crossref]
5. Qiu Z, Yao T, Mei T. Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Trans Multimedia*. 2018;20(4):939-949. [Crossref]
6. Zhang Z, Liu Q, Wang Y. Road extraction by deep residual U-Net. *Arxiv*. [Crossref]
7. Jha D, Smedsrud PH, Riegler MA, et al. ResUNet++: an advanced architecture for medical image segmentation. *Arxiv*. [Crossref]
8. Chollet F. 2015. Keras. GitHub. [Crossref]
9. Keeton K. Proceedings of the 12th USENIX conference on operating systems design and implementation. USENIX Association; 2016. [Crossref]
10. Yousef R, Khan S, Gupta G, et al. U-Net-based models towards optimal MR brain image segmentation. *Diagnostics*. 2023;13(9):1624. [Crossref]
11. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a Nested U-Net architecture for medical image segmentation. deep learn med image anal multimodal learn clin decis support. 2018;11045:3-11. [Crossref]
12. Tegtmeier RC, Kutyreff CJ, Smetanick JL, et al. Custom-trained deep learning-based auto-segmentation for male pelvic iterative CBCT on C-arm linear accelerators. *Pract Radiat Oncol*. 2024;14(5):e383-e394. [Crossref]
13. Yepes-Calderon F, McComb JG. Eliminating the need for manual segmentation to determine size and volume from MRI. A proof of concept on segmenting the lateral ventricles. *PLoS One*. 2023;18(5):e0285414. [Crossref]
14. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60. [Crossref]
15. Barash Y, Livne A, Klang E, et al. Artificial intelligence for identification of images with active bleeding in mesenteric and celiac arteries angiography. *Cardiovasc Intervent Radiol*. 2024;47(6):785-792. [Crossref]
16. Liu J, Zhou X, Guan W, Gong S, Liu J. Research on detection method of bleeding point in two-dimensional DSA image based on parametric color imaging. *Comput Biol Med*. 2022;146:105496. [Crossref]
17. Min X, Feng Z, Gao J, et al. InterNet: detection of active abdominal arterial bleeding using emergency digital subtraction angiography imaging with two-stage deep learning. *Front Med (Lausanne)*. 2022;9:762091. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.242401

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

BREAST IMAGING

ORIGINAL ARTICLE

The role of the Kaiser score system in uncertain malignant potential (B3) breast lesions: a pilot study

Fatma Çelik Yabul
Hafize Otçu Temur
Bahar Atasoy
Serdar Balsak
Alpay Alkan
Şeyma Yıldız

Bezmialem Vakıf University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

PURPOSE

This study aims to evaluate the effectiveness of the Kaiser score (KS) system in assessing breast lesions with uncertain malignant potential (B3).

METHODS

Breast magnetic resonance imaging (MRI) scans from a total of 76 patients with histologically proven B3 lesions were included in this study. The KS was recorded for each MRI scan. The patients were classified based on biopsy results, and upgraded lesions were identified. Statistical analysis was conducted to evaluate the association between high KS values and upgraded lesions.

RESULTS

The mean age of the 76 patients was calculated as 49.6 ± 10.1. A significant association was observed between the KS system and the prediction of malignancy upgrade ($P < 0.001$). Furthermore, among the descriptors, spiculation, margin, and upgrading prediction demonstrated a statistically significant difference ($P < 0.001$). Additionally, the specificity improved when the accepted KS cut-off value was set at seven instead of five. A significant association was also observed between the KS system and the papilloma upgrade rate within the B3 lesion subgroups ($P < 0.001$).

CONCLUSION

Breast radiology plays a crucial role in the diagnosis of B3 lesions. Our findings suggest that the KS system holds promise as a tool for predicting the upgrade potential of B3 lesions.

CLINICAL SIGNIFICANCE

This study demonstrated that the KS system may assist in predicting the upgrade potential of B3 breast lesions. It also demonstrated that spiculation and margin descriptors within the KS system possess a high positive predictive value for upgrade prediction. Additionally, we believe that the KS system can help prevent unnecessary surgeries in patients with B3 lesions.

KEYWORDS

B3 breast lesion, Kaiser score system, breast magnetic resonance imaging, breast, magnetic resonance imaging

Corresponding author: Fatma Çelik Yabul

E-mail: fatmayabul@gmail.com

Received 21 July 2023; revision requested 18 September 2023; last revision received 18 January 2024; accepted 04 February 2025.

Epub: 26.03.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.242401

Uncertain malignant potential lesions (B3) of the breast can be classified as atypical ductal hyperplasia (ADH), radial scar, papillary lesions, lobular neoplasia (LN), and flat epithelial hyperplasia (FEH). These lesions are commonly characterized by an increased lifetime risk of breast cancer in women.[1-3] Due to the heterogeneity of high-risk lesion groups, upgrade rates for high-risk breast lesions have varied in the literature, ranging from 6% to 32%.[4-6]

The management of B3 lesions is determined by pathological findings, patient age, risk factors, and the type of biopsy performed. Radiological-pathological discordance remains one of the key criteria for excision.[7-9]

You may cite this article as: Çelik Yabul F, Otçu Temur H, Atasoy B, Balsak S, Alkan A, Yıldız Ş. The role of the Kaiser score system in uncertain malignant potential (B3) breast lesions: a pilot study. *Diagn Interv Radiol.* 2025;31(5):474-479.

The Kaiser score (KS) system is a decision-making tool in parallel with the Breast Imaging-Reporting and Data System (BI-RADS) classification, considering the morphological and dynamic features in breast magnetic resonance imaging (MRI).[10,11] Dietzel and Baltzer[10] developed a clinical decision tool originally referred to as the Tree flowchart and later renamed the KS after Werner A. Kaiser's contributions to its development. Additionally, they published an essay that included a practical guide for the interpretation of breast MRI examinations using the KS.[10] Their contribution has been further extended with a recently published article in which they emphasized that the KS served as an evidence-based decision-making tool to objectively differentiate between benign and malignant breast lesions.[12]

This study aims to evaluate the upgrade potential of high-risk breast lesions and to determine the role of the KS in avoiding potentially unnecessary surgical excisions.

Methods

This retrospective study was approved according to the principles of the Declaration of Helsinki by the Ethics Committee of the Bezmialem Vakıf University (approval no: E-54022451-050.01.04-3208, date: 20.10.2021), and all participants signed a written informed consent form.

Patient selection

Patients' core biopsy-proven B3 lesions, collected between 2016 and 2021, were retrieved from the archives. Initially, 130 patients were reviewed. Among these, only 81 had MRI scans available in the system. Of the 81 patients, 5 were excluded due to the absence of pathological contrast enhancement on MRI. Patients with excision results or a follow-up period of at least 2 years (24–72 months) were included in this study. Based on these criteria, a total of 76 patients were considered eligible for this study. The age, risk status, and complaints of the patients were recorded.

Main points

- The Kaiser score (KS) system may help predict uncertain malignant potential (B3) breast lesions upgrade.
- Spiculation and margin identifiers in the KS system have a high positive predictive value in upgrade prediction.
- Unnecessary surgeries can be avoided in cases diagnosed with B3 lesions by using the KS system.

Magnetic resonance imaging acquisition and image interpretation

All breast MRI scans were conducted using a 1.5 T scanner (Siemens Magnetom Avanto Fit, Siemens Healthineers; Erlangen, Germany) with a bilateral 16-channel breast coil in the prone position. Apparent diffusion coefficient (ADC) maps, subtraction, and maximum intensity projection images were acquired. Axial T2-weighted fat-suppressed imaging [repetition time (TR)/echo time (TE): 4560/59 ms; slice thickness: 4 mm, matrix: 340 × 512], axial T1-weighted imaging (TR/TE: 571/11 ms; slice thickness: 4 mm, matrix: 340 × 512), one precontrast and five postcontrast 3D T1 turbo spin-echo imaging (TR/TE: 5.16/2.38 ms; flip angle: 100, slice thickness: 1 mm), and diffusion-weighted imaging (b-values: 0–800 s/mm^2) series were obtained. The gadolinium-based contrast agent was administered at 0.1 mmol/kg using a mechanical power injector, followed by a 15–20 cm^3 saline flush.

Two breast radiologists evaluated all the images using the Siemens Syngo Via (Erlangen, Germany) workstation. They were blinded to clinical data and histopathology results. The KS was assigned via the online version to the patients after reaching consensus. Descriptors evaluated in the KS were spiculation, dynamic enhancement curves, margins, internal enhancement, and edema around the lesion. Using the KS, the patients were scored from 1 to 11. A score of 5–7 was categorized as BI-RADS 4, and a score of 8–11 was categorized as BI-RADS 5 and considered positive. Optional moderators were noted, such as evidence of microcalcification overlapping the area of contrast and ADC values. The cut-off value was $>1.4 \times 10^{-3}$ mm^2/s as recommended in the KS.

Histopathological evaluation

The patients' diagnoses were obtained using one of the following methods: tru-cut biopsy under ultrasonographic guidance (n = 59), vacuum-assisted biopsy (VAB) under mammographic guidance (n = 10), or biopsy under MRI guidance (n = 7). On average, 3–4 samples were obtained for tru-cut biopsies using a 14-gauge needle. The results of core biopsy, surgical excision, or follow-up evaluations were analyzed. Cases with an upgrade to ductal carcinoma *in situ* (DCIS) following excision, including those with progression detected during follow-up, were considered positive.

Statistical analysis

Statistical analysis was performed using SPSS software (IBM Corp. Released 2021. IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY, USA). In addition to descriptive statistics [mean ± standard deviation for continuous variables, frequencies with percentages for categorical variables, and area under the receiver operating characteristic (ROC) curve (AUC) with standard error], the Shapiro–Wilk test was used to assess the distribution of the data. Comparisons of KS descriptors and upgrade rates were performed using Fisher's exact test and the Fisher–Freeman–Halton test. ROC analysis was performed using MedCalc version 12 to assess the overall diagnostic performance of KS in predicting progression, and the optimal cut-off value was determined using Youden's J index. Differences in KS and upgrade rates among high-risk lesion subgroups (ADH, radial scar, atypical papillomas, LCIS, LN, and FEH) were evaluated. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. A type 1 error rate of $\alpha = 0.05$ was considered statistically significant.

Results

A total of 76 patients were evaluated. Thirty patients were classified as high-risk due to a personal history of breast cancer (n = 8) or a family history of breast cancer in immediate relatives (n = 22). Clinical findings such as pain (n = 12), palpable mass (n = 7), and bloody nipple discharge (n = 6) were present in 33% of the patients. The mean age of the patients was calculated as 49.6 ± 10.1.

The histopathological distribution of lesions is presented in Table 1. Among the 76

Table 1. Histopathological distribution of the B3 lesions

Histopathologic results	n (%)
Papilloma	36 (47.4%)
Flat epithelial hyperplasia	18 (23.7%)
Radial scar	12 (15.8%)
Lobular neoplasia	5 (6.6%)
Atypic ductal hyperplasia	5 (6.6%)
Total	76

B3 lesions, papilloma was the most common diagnosis (47.4%), followed by FEH (23.7%), radial scar (15.8%), LN (6.6%), and ADH (6.6%).

In the follow-up cases (n = 40), no progression was observed. During follow-up, the lesions remained stable in 33 patients (82.5%) and regressed in size in 7 patients (17.5%). Surgical excision was performed in 36 patients (47.3%), and DCIS was detected in 12 of 76 patients following excision. No upgrade to invasive cancer was identified. The overall upgrade rate in patients with B3 lesions was 16%.

Based on MRI results, non-mass enhancement was observed in 34 cases (44.7%), whereas mass enhancement was observed in 42 cases (55.2%). There was no statistically significant difference between upgraded lesions and MRI enhancement patterns ($P > 0.050$).

In the evaluation of optional moderators (suspicious microcalcifications and high ADC values), seven patients had microcalcifications overlapping with the contrast-enhanced area on MRI. Based on the KS, two points were added to these patients. An upgrade was detected in four of them.

No high ADC values were identified in the evaluation of lesions that would warrant a four-point reduction in the KS. All lesion ADC values were below 1.4×10^{-3} mm^2/s. The KS and MRI findings for the upgraded lesions are presented in Table 2.

A positive KS was a significant predictor of lesion upgrade status ($P < 0.001$). In patients with a KS exceeding 5, the sensitivity and specificity for predicting an upgrade were 81.25% and 83.33%, respectively. The NPV, PPV, and overall accuracy were 94.34%, 56.52%, and 82.89%, respectively (Table 3).

When the KS cut-off value was set at 7, the sensitivity, specificity, and accuracy for predicting an upgrade were 68.75%, 98.3%, and 80.2%, respectively, with an AUC of 0.86 and a standard error of 0.07 (95% confidence interval: 0.76–0.93, $P < 0.001$). Additionally, a cut-off value of >6 was identified (Figure 1).

Evaluation of MRI findings showed that spiculation was a significant predictor of lesion upgrade ($P < 0.001$). Furthermore, the subgroups of B3 lesions were analyzed. Among these subgroups, the KS was a significant predictor of the upgrade rate for papilloma ($P < 0.001$).

Discussion

B3 of the breast are commonly encountered in needle biopsies. Due to the potential for malignancy, the management of these lesions following needle biopsy remains controversial, with no universally accepted standard recommendation. Although surgical biopsy is widely recommended for ADH, the management of other B3 lesions should be determined on a patient basis through a multidisciplinary approach. Criteria for surgical excision may include sampling adequacy (e.g., needle gauge, number of samples, and accurate targeting), lesion size, and radiology–pathology concordance.[7-9]

The literature has studied the role of MRI in managing high-risk lesions. Londero et al.[13] reported that the absence of enhancement on breast MRI effectively eliminated the risk of invasive cancer and served as a reliable indicator for excluding surgery in B3 lesions. Similarly, in our study, no progress was seen during follow-up in cases that were di-

Table 2. MRI findings of the upgraded lesions

No	Kaiser score	Spiculation	Margin	Contrast	Edema	Internal enhancement	Pathology
1	3	Negative	Irregular	Type 1	Negative	Homogeneous	FEH
2	3	Negative	Circumscribed	Type 1	Negative	Homogeneous	ADH
3	7	Positive	Irregular	Type 2	Negative	Homogeneous	ADH
4	7	Positive	Irregular	Type 2	Negative	Homogeneous	LN
5	7	Positive	Irregular	Type 2	Negative	Homogeneous	LN
6	2	Negative	Circumscribed	Type 2	Negative	Homogeneous	LN
7	11	Positive	Irregular	Type 3	Positive	Homogeneous	Papilloma
8	5	Negative	Circumscribed	Type 2	Negative	Homogeneous	Papilloma
9	8	Negative	Irregular	Type 3	Negative	Inhomogeneous	Papilloma
10	8	Negative	Irregular	Type 3	Negative	Inhomogeneous	Radial scar
11	8	Negative	Irregular	Type 3	Negative	Inhomogeneous	Radial scar
12	7	Positive	Irregular	Type 2	Negative	Homogeneous	Papilloma
13	11	Positive	Irregular	Type 3	Positive	Inhomogeneous	Papilloma
14	9	Positive	Irregular	Type 3	Negative	Inhomogeneous	Papilloma
15	5	Negative	Circumscribed	Type 2	Negative	Homogeneous	FEH
16	11	Positive	Irregular	Type 3	Positive	Inhomogeneous	Papilloma

MRI, magnetic resonans imaging; FEH, flat epithelial hyperplasia; ADH, atypical ductal hyperplasia; LN, lobular neoplasia.

Table 3. Kaiser score positivity and upgrade ratio

	Upgrade (+)	Upgrade (–)	Total
Kaiser (+)	13	10	23
Kaiser (–)	3	50	53
Total	16	60	76

agnosed as B3 lesions but were not included in the study because MRI did not show any contrast enhancement.

The KS system is a decision-making tool that integrates five morphology and kinetic criteria, along with two optional modifiers (microcalcifications and ADC values), to differentiate benign from malignant breast tumors. The KS system offers a standardized approach to breast MRI evaluation, enhancing its utility in clinical practice. In recent years, there has been a rapid increase in studies employing this flowchart.[14-20] Studies have demonstrated that inter-reader agreement is high and that the KS enhances the diagnostic performance of MRI.[14-18] Wang et al.[19] has also showed that the KS is a useful diagnostic tool that helps radiologists with different levels of breast MRI experiences make more accurate diagnoses. According to Zhang et al.[20], KS is a better way to diagnose breast lesions than BI-RADS, whether the lesions show non-mass enhancement or are evaluated on their own. Furthermore, Wengert et al.[21] gave useful supporting data and pushed for the use of KS to eliminate BI-RADS 4 mammography calcifications. However, no studies to date have specifically evaluated the application of KS in B3 lesions.

In our study, the upgrade rates were comparable with those reported in the literature. However, the excision rates were higher than those documented in previous studies.[4-6] This can be attributed to the large proportion of high-risk patients and the limited availability of VAB in our country.

There were three false-negative cases in our study. Two cases (LN and FEH) were upgraded to low-grade DCIS following excision (Figure 2). It is well-established that MRI has low specificity for detecting low-grade DCIS,[22] which may explain these false-negative results. In one of these cases (ADH), microcalcifications led to an increased score, highlighting the importance of incorporating optional moderators in the KS system.

We observed false-positive results in 10 cases. Four patients had papillomas, and three had radial scars. In the false-positive papilloma cases, the type 2 contrast enhancement pattern contributed to the increased scores (Figure 3). The literature indicates that papillomas are a heterogeneous group that may exhibit varying enhancement patterns,[23] which we believe contributes to the higher false-positive rate. Additionally, contour irregularity and spiculation positivity increased the scores in cases of radial scars. Radial scars were present in six of the patients with false-positive results. Radial scars are inherently characterized by irregular contours.[24] In the KS system, scoring begins at six points due to the spiculation positivity commonly observed in radial scars, leading to false-positive outcomes.

Evaluation of the KS descriptors revealed that 11 patients exhibited positive spiculation and contour irregularity. The KS values of all 11 cases ranged from 6 to 11, and 8 of them were upgraded (Figure 4). These findings show that spiculation positivity and contour irregularity are significantly associated with lesion upgrade.

Our study identified three cases with edema, all of which underwent an upgrade. Recent studies have shown that peritumoral edema is associated with poor prognosis.[25] Consequently, the presence of edema may have a high positive value for predicting B3 lesions upgrade.

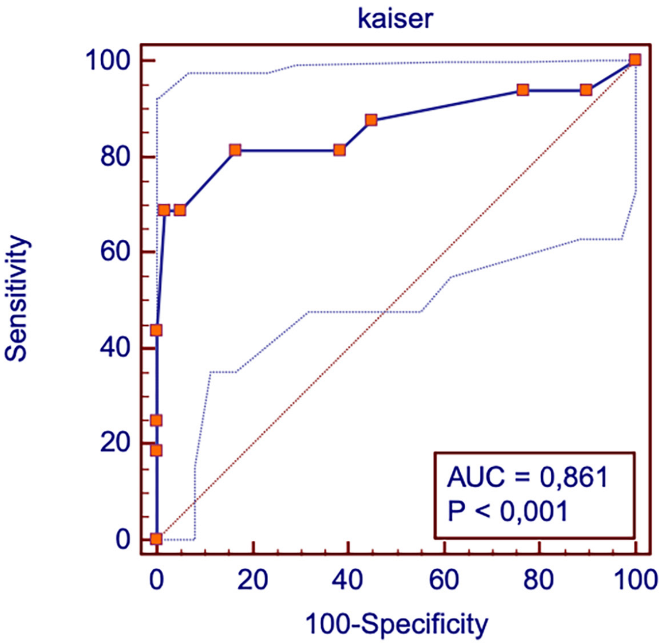


Figure 1. Sensitivity and specificity ratio of KS 7. KS, Kaiser score; AUC, area under the receiver operating characteristic curve.

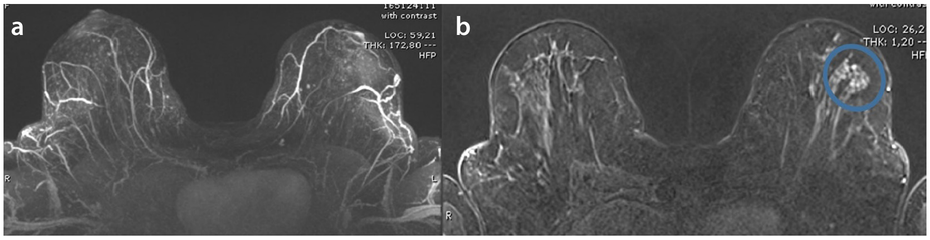


Figure 2. False-negative case, MIP series **(a)** early postcontrast series **(b),** non-mass enhancement, root sign absent, type 2, circumscribed lesion, Kaiser score 2, and BI-RADS 2/3. MIP, maximum intensity projection; BI-RADS, Breast Imaging-Reporting and Data System.

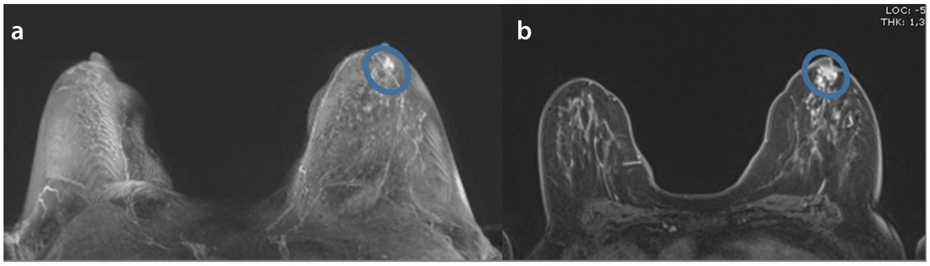


Figure 3. True-positive case, MIP series **(a)**, early postcontrast series **(b)** mass, inhomogeneous enhancement, root sign absent, type 3, Kaiser score 8, and BI-RADS 5. MIP, maximum intensity projection; BI-RADS, Breast Imaging-Reporting and Data System

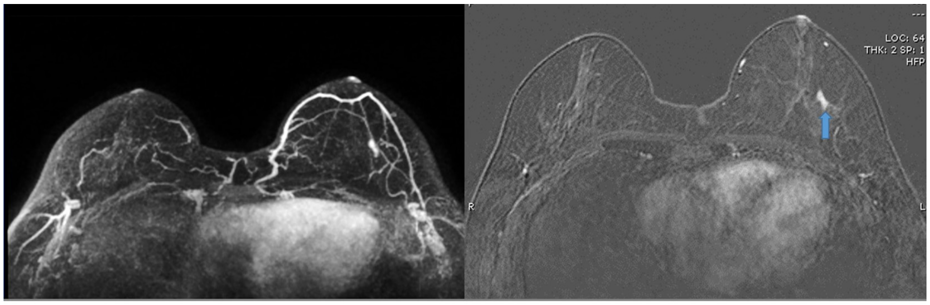


Figure 4. Root sign present, mass lesion, type 2 enhancement, edema absent, Kaiser score 7, and BI-RADS4. BI-RADS, Breast Imaging-Reporting and Data System.

Additionally, the internal enhancement pattern may significantly influence the prediction of lesion upgrade. The inhomogeneous enhancing pattern increased the lesion score from 4 to 8. Significantly, three of these lesions are confirmed true positives.

The acceptance of a 5 KS value for differentiating malignant from benign tumors resulted in an accuracy of 82.89%. Nevertheless, when the cut-off value was set at 7, the specificity (98.3%) improved without significantly reducing accuracy.

The limitations of our study include the heterogeneity of B3 lesion pathologies. The study included a small number of ADH lesions because their exclusion would not have made a statistical difference. The high number of papillomas may be due to the broad MRI indication, which aimed to reduce the risk of papillomatosis and detect cancer in the ipsilateral breast.[26] Furthermore, our study is single-centered and retrospective in design. Additionally, two breast radiologists conducted the KS assessment; however, another limitation is the absence of statistical analysis for inter-reader agreement. This study can be conducted prospectively on specific B3 lesion subgroups.

In conclusion, we speculate that increasing the KS threshold value in future studies with larger sample sizes could help avoid unnecessary surgeries. In conclusion, the KS system demonstrates the ability to predict B3 lesion upgrade accurately.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. *Ann Oncol*. 2008;19(4):614-622. [Crossref]
2. Lakhani SREI, Schnitt SJ, Tan PH, van de Vijver MJ. WHO classification of tumours of the breast, 4th ed. International Agency for Research on Cancer, Lyon; 2012;50-54. [Crossref]
3. Morrow M, Schnitt SJ, Norton L. Current management of lesions associated with an increased risk of breast cancer. *Nat Rev Clin Oncol*. 2015;12(4):227-238. [Crossref]
4. Mooney KL, Bassett LW, Apple SK. Upgrade rates of high-risk breast lesions diagnosed on core needle biopsy: a single-institution experience and literature review. *Mod Pathol*. 2016;29(12):1471-1484. [Crossref]
5. Polat DS, Schopp JG, Arjmandi F, et al. Performance of a clinical and imaging-based multivariate model as decision support tool to help save unnecessary surgeries for high-risk breast lesions. *Breast Cancer Res Treat*. 2021;185:479-494. [Crossref]
6. Oktay A, Aslan Ö, Taşkın F, et al. Outcomes of high-risk breast lesions diagnosed using image-guided core needle biopsy: results from a multicenter retrospective study. *Diagn Interv Radiol*. 2023;20;29(4):579-587. [Crossref]
7. Rageth CJ, O'Flynn EA, Comstock C, et al. First International Consensus Conference on lesions of uncertain malignant potential in the breast (B3 lesions). *Breast Cancer Res Treat*. 2016;159(2):203-213. [Crossref]
8. Bahl M. Management of high-risk breast lesions. *Radiol Clin North Am*. 2021;59(1):29-40. [Crossref]
9. Rageth CJ, O'Flynn EAM, Pinker K, et al. Second International Consensus Conference on lesions of uncertain malignant potential in the breast (B3 lesions). *Breast Cancer Res Treat*. 2019;174(2):279-296. [Crossref]
10. Dietzel M, Baltzer PAT. How to use the Kaiser score as a clinical decision rule for diagnosis in multiparametric breast MRI: a pictorial essay. *Insights Imaging*. 2018;9(3):325-335. [Crossref]
11. Baltzer PA, Dietzel M, Kaiser WA. A simple and robust classification tree for differentiation between benign and malignant lesions in MR-mammography. *Eur Radiol*. 2013;23(8):2051-2060. [Crossref]
12. Baltzer PAT, Krug KB, Dietzel M. Evidence-based and structured diagnosis in breast MRI using the Kaiser Score. *Rofo*. 2022;194(11):1216-1228. [Crossref]
13. Londero V, Zuiani C, Linda A, Girometti R, Bazzocchi M, Sardanelli F. High-risk breast lesions at imaging-guided needle biopsy: usefulness of MRI for treatment decision. *AJR Am J Roentgenol*. 2012;199(2):240-250. [Crossref]
14. Marino MA, Clauser P, Woitek R, et al. A simple scoring system for breast MRI interpretation: does it compensate for reader experience? *Eur Radiol*. 2016;26(8):2529-2537. [Crossref]
15. Istomin A, Masarwah A, Vanninen R, Okuma H, Sudah M. Diagnostic performance of the Kaiser score for characterizing lesions on breast MRI with comparison to a multiparametric classification system. *Eur J Radiol*. 2021;138:109659. [Crossref]
16. Cloete DJ, Minne C, Schoub PK, Becker JHR. Magnetic resonance imaging of fibroadenoma-like lesions and correlation with breast imaging-reporting and data system and Kaiser scoring system. *SA J Radiol*. 2018;22(2):1532. [Crossref]
17. Milos RI, Pipan F, Kalovidouri A, et al. The Kaiser score reliably excludes malignancy in benign contrast-enhancing lesions classified as BI-RADS 4 on breast MRI high-risk screening exams. *Eur Radiol*. 2020;30(11):6052-6061. [Crossref]
18. Jajodia A, Sindhwani G, Pasricha S, et al. Application of the Kaiser score to increase diagnostic accuracy in equivocal lesions on diagnostic mammograms referred for MR mammography. *Eur J Radiol*. 2021;134:109413. [Crossref]
19. Wang Q, Fu F, Chen Y, et al. Application of the Kaiser score by MRI in patients with breast lesions by ultrasound and mammography. *Diagn Interv Radiol*. 2022;28(4):322-328. [Crossref]
20. Zhang B, Feng L, Wang L, Chen X, Li X, Yang Q. Kaiser score for diagnosis of breast lesions presenting as non-mass enhancement on MRI. *Nan Fang Yi Ke Da Xue Xue Bao*. 2020;40(4):562-566. [Crossref]
21. Wengert GJ, Pipan F, Almohanna J, et al. Impact of the Kaiser score on clinical decision-making in BI-RADS 4 mammographic calcifications examined with breast MRI. *Eur Radiol*. 2020;30(3):1451-1459. [Crossref]
22. Taskin F, Kalayci CB, Tuncbilek N, et al. The value of MRI contrast enhancement in biopsy decision of suspicious mammographic microcalcifications: a prospective multicenter study. *Eur Radiol*. 2021;31(3):1718-1726. [Crossref]
23. Yılmaz R, Bender Ö, ÇelikYabul F, Dursun M, Tunacı M, Acunas G. Diagnosis of nipple

discharge: value of magnetic resonance imaging and ultrasonography in comparison with ductoscopy. *Balkan Med J*. 2017;34(2):119-126. [Crossref]

24. Bargallo X, Ubeda B, Ganau S, et al. Magnetic resonance imaging assessment of radial scars/complex sclerosing lesions of the breast. *Curr Med Imaging*. 2022;18(2):242-248. [Crossref]

25. Panzironi G, Moffa G, Galati F, Marzocca F, Rizzo V, Pediconi F. Peritumoral edema as a biomarker of the aggressiveness of breast cancer: results of a retrospective study on a 3 T scanner. *Breast Cancer Res Treat*. 2020;181(1):53-60. [Crossref]

26. Gültekin MA, Yabul FÇ, Temur HO, et al. Papillary lesions of the breast: addition of DWI and TIRM sequences to routine breast MRI could help in differentiation benign from malignant. *Curr Med Imaging*. 2022;18(9):962-969. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253412

Copyright© Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Letter to editor: dual-energy computed tomography-based volumetric thyroid iodine quantification: correlation with thyroid hormonal status, pathologic diagnosis, and phantom validation

Ahmet Gürkan Erdemir

Hacettepe University Faculty of Medicine,
Department of Radiology, Ankara, Türkiye

Dear Editor,

I read with great interest the article by Lee[1], which presents a promising non-invasive method for quantifying intrathyroidal iodine concentration using dual-energy computed tomography (DECT). Their study demonstrates that DECT-derived iodine maps can effectively distinguish between thyroid functional states and detect diffuse thyroid disease without the need for contrast enhancement. This approach holds particular promise in the peri-radioactive iodine (RAI) therapy setting, especially when applied both before and after treatment to monitor iodine organification.

However, a key translational challenge remains: The optimal timing of DECT imaging in relation to RAI therapy and contrast exposure is still poorly defined. As noted in the American College of Radiology Manual on Contrast Media[2], iodinated contrast agents are contraindicated during active RAI treatment phases due to the risk of competitive inhibition. Although the guideline recommends a conservative delay of several months, emerging evidence suggests that the kinetics of iodine organification and clearance may not support such prolonged avoidance windows.

Nimmons et al.[3] conducted a prospective study assessing urinary iodine clearance after intravenous contrast administration. In their cohort, the median time to return to baseline urinary iodine levels was 43 days, with 75% of patients normalizing within 59 days and 90% within 74 days. This study not only highlights the interindividual variability in iodine kinetics but also raises the question of whether personalized biomarkers–such as serial urinary iodine levels or DECT-based iodine density–could better guide the safe reinitiation of RAI planning.

Given this, DECT could potentially evolve from a diagnostic modality into a monitoring tool for individualized iodine readiness. It may be employed to quantify residual iodine load following contrast exposure to help determine the optimal timing for RAI therapy, to longitudinally track thyroidal iodine washout without relying on urinary measurements, to identify iodine-induced dysregulation–such as prolonged retention or the Wolff-Chaikoff effect–in elderly patients or those with renal impairment, and to implement contrast-deferred DECT protocols that help avoid unnecessary delays in oncologic management.

Still, standardization is needed. Future DECT protocols should be prospectively validated against urinary iodine and RAI uptake metrics. Additionally, combining DECT with functional nuclear imaging (e.g., single photon emission computed tomography/computed tomography) may enhance clinical decision-making by simultaneously capturing iodine content and tracer uptake. These integrations could ultimately reduce uncertainty in post-contrast scenarios where thyroid nodules are incidentally discovered.

In conclusion, the work of Lee[1] offers a valuable step forward, but its clinical integration–particularly in the nuanced post-contrast period–demands further clarification. DECT's capacity to measure thyroidal iodine *in vivo* opens a pathway toward more tailored and efficient RAI planning, provided it is used with informed caution and in concert with evolving evidence on iodine kinetics.

KEYWORDS
Agent, contrast, opaque, radionuclide, thyroid, thyroid gland

Corresponding author: Ahmet Gürkan Erdemir

E-mail: a.gurkan.erdemir@gmail.com

Received 16 April 2025; accepted 05 May 2025.

Epub: 18.08.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.253412

You may cite this article as: Erdemir AG. Letter to editor: dual-energy computed tomography-based volumetric thyroid iodine quantification: correlation with thyroid hormonal status, pathologic diagnosis, and phantom validation. *Diagn Interv Radiol.* 2025;31(5):480-481.

Footnotes

Conflict of interest disclosure

The author declared no conflicts of interest.

References

1. Lee Y. Dual-energy computed tomography-based volumetric thyroid iodine quantification: correlation with thyroid hormonal status, pathologic diagnosis, and phantom validation. *Diagn Interv Radiol*. 2025;31(3):226-233. [CrossRef]
2. ACR Committee on Drugs and Contrast Media. ACR Manual on Contrast Media. Version 2024. American College of Radiology. 2024. [CrossRef]
3. Nimmons GL, Funk GF, Graham MM, Pagedar NA. Urinary iodine excretion after contrast computed tomography scan: implications for radioactive iodine use. *JAMA Otolaryngol Head Neck Surg*. 2013;139(5):479-482. [CrossRef]

DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242833

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Factors effecting the success of retrograde tibiopedal access and recanalization in infrapopliteal artery occlusions

Cemal Aydın Gündoğmuş
Hande Özen Atalay
Vugar Samadli
Levent Oğuzkurt

Koç University Hospital, Department of Radiology, İstanbul, Türkiye

Corresponding author: Cemal Aydın Gündoğmuş

E-mail: cagundogmus@gmail.com

Received 16 May 2024; revision requested 09 June 2024; last revision received 11 July 2024; accepted 07 August 2024.

Epub: 09.09.2024

Publication date: 08.09.2025

DOI: 10.4274/dir.2024.242833

PURPOSE

Peripheral arterial disease (PAD) is increasingly prevalent, particularly among the aging population. Retrograde tibiopedal access (RTPA) has emerged as a useful endovascular treatment for PAD. However, there is limited research examining factors that influence the efficacy of RTPA. To investigate factors affecting the access, crossing, and recanalization success rates of RTPA for infrapopliteal PAD treatment.

METHODS

A retrospective study was conducted on 720 patients who underwent endovascular treatment for PAD. Of these, 104 patients (mean age: 65.5 ± 16.2; 89 men) with 131 RTPA trials were included in the final evaluation. The disease and its duration, Rutherford score, smoking status, access site, and its occlusion status, access, crossing, and recanalization success were noted. Data were analyzed using Pearson's chi-square and Mann–Whitney U tests and multivariate logistic regression to evaluate the impact of various factors on success rates.

RESULTS

The access success rate was 82.6%, the crossing success rate was 95.4%, and the recanalization success rate was 74%. Access success was significantly higher when the dorsal pedal artery (DPA) was the access artery compared with the posterior tibial artery (91.3% vs. 74.2%, $P = 0.009$). Access success was notably lower in patients with thromboangiitis obliterans compared with patients with diabetes mellitus (DM) and non-DM atherosclerosis (68.6% vs. 90.3% and 80.3%, $P = 0.019$). Recanalization success was higher when the puncture site was non-occluded (76.7% vs. 53.5%, $P = 0.023$).

CONCLUSION

The study suggests that RTPA is a generally effective and safe technique for infrapopliteal PAD treatment. The most favorable outcomes are observed in individuals with DM who have a non-occluded DPA at the puncture site. Recanalization success is only affected by the patency of the artery at the puncture site.

CLINICAL SIGNIFICANCE

These findings offer targeted guidance for clinicians and highlight areas requiring further investigation.

KEYWORDS

Angiography, atherosclerosis, diabetes, retrograde tibiopedal access, thromboangiitis obliterans

Peripheral arterial disease (PAD) is a cardiovascular disorder distinguished by a stenosis or occlusion of peripheral arteries, typically impacting the lower extremities.[1,2] Recent studies highlight that PAD is a burgeoning concern in contemporary hospital admissions, particularly among the aging population.[3-5] It has been estimated that up to 20% of individuals aged 80 and above suffer from PAD, reflecting a substantial clinical and public health concern.[6]

You may cite this article as: Gündoğmuş CA, Özen Atalay H, Samadli V, Oğuzkurt L. Factors effecting the success of retrograde tibiopedal access and recanalization in infrapopliteal artery occlusions. *Diagn Interv Radiol.* 2025;31(5):482-488.

Among the therapeutic options for PAD, endovascular interventions have been steadily rising in prominence.[7,8] These minimally invasive procedures serve as an alternative to open surgical approaches, often offering advantages in terms of shorter hospital stays, reduced morbidity, and quicker recovery times. Endovascular treatments have become an initial treatment modality of choice for many clinicians dealing with patients with PAD, especially those at high surgical risk or those who have failed other treatment options.[9]

One of the more recent advances in the realm of endovascular interventions for PAD is the utilization of retrograde tibiopedal access (RTPA). This method has proven particularly useful in cases where antegrade access is not feasible or the occluded segment of the artery is not easily traversable through standard methods.[2,10] The technique can facilitate the crossing of long intraluminal complex lesions and may provide additional options for limb salvage in otherwise challenging scenarios. Despite the growing body of evidence supporting the benefits of RTPA, there is a notable paucity of research exploring the variables that influence its efficacy.[10-13] Most studies have primarily focused on technical success and safety profiles, with limited attention to how patient-specific factors and the anatomical characteristics of occlusions may affect the procedure's outcome. Furthermore, there is no clear data on the effect of the occluded access artery on recanalization success, while the success of RTPA in treating infrapopliteal arteries is not well established.

Therefore, the present study aims to address this gap by investigating various factors that may have an impact on the effectiveness of RTPA, such as patient demographics, underlying diseases, and the access artery and its condition. By contributing to this underexplored area of research, more targeted guidance for clinicians is offered, thereby potentially improving patient outcomes in the management of PAD.

Methods

The present retrospective study was conducted in accordance with the ethical standards outlined by the World Medical Association in the Declaration of Helsinki. Approval for the study was obtained from the Ethics Committee of Koç University Ethical Board (reference number/date: 2023.131.IRB.043/12.04.2023). Prior to the procedure, written informed consent was obtained from all patients.

Of the 720 patients who had endovascular treatment for PAD in a tertiary referral center between November 2015 and February 2023, 129 patients with 158 RTPA trials were included in this retrospective study. A total of 26 patients were excluded from the study, with 13 patients undergoing RTPA for the treatment of acute thromboembolism on top of chronic atherosclerotic occlusions, and the remaining 13 patients undergoing RTPA specifically for occluded suprapopliteal arteries (Figure 1). A total of 131 access trials in 104 patients (89 men and 15 women; mean age: 65.5 ± 16.2) with infrapopliteal artery disease were evaluated using procedural images and reports. Patients' diagnoses, disease duration, Rutherford scores, and smoking status were collected from the hospital records.

Retrograde tibiopedal access technique

All endovascular treatment procedures were performed either with sedation or with ultrasound (US)-guided sciatic nerve blockage in addition to local anesthesia. The access sites, including the femoral and ipsilateral ankle, were prepared in a sterile fashion prior to the procedure in all patients. All patients in whom there was an attempt to use RTPA were first approached in an antegrade way via ipsilateral common femoral or superficial femoral artery access. If the infrapopliteal artery occlusion could not be crossed by way of an antegrade approach, RTPA was performed. The patients were placed in a supine position on the angiography table. To obtain access to the dorsal pedal artery (DPA), the foot was held in a neutral position with minimum flexion. On the other hand, access to the posterior tibial artery (PTA) was achieved by rotating the foot laterally and gently bending the knee. All RTPA's were conducted under US guidance by a single interventional radiologist with over 20 years of expertise in performing procedures that necessitate image-guided vascular access.

A transverse placement of a linear 9–15 MHz transducer (Logiq S8, GE HealthCare Technologies, Inc., Chicago, Illinois) was performed to visualize and identify the most suitable access site for the target artery (Figure 2). Subsequently, a small skin wheal was induced using 1 mL of 1% prilocaine (Citanest 10 mg/mL, AstraZeneca). In this procedure, a 4-cm 21G micropuncture needle (Micropuncture Introducer Set, Cook Medical) is carefully inserted into the artery's anterior wall, ensuring avoidance of the posterior wall, before a 200-cm-long, 0.018-inch

Main points

- There is very limited data on the effect of the occluded access artery on recanalization success.
- Retrograde tibiopedal access (RTPA) success in treating infrapopliteal arteries is not well-established.
- The access success rate was 100% in 30 cases in which the access artery was patent.
- The target vessel at the puncture site was occluded in 101 (77.1%) RTPA trials. The access success rate was 82.6% (109/131), the crossing success rate was 95.4% (104/109), and the recanalization success rate was 74% (77/104).
- The most favorable outcomes were observed in individuals with diabetes mellitus who had a non-occluded dorsal pedal artery at the puncture site.

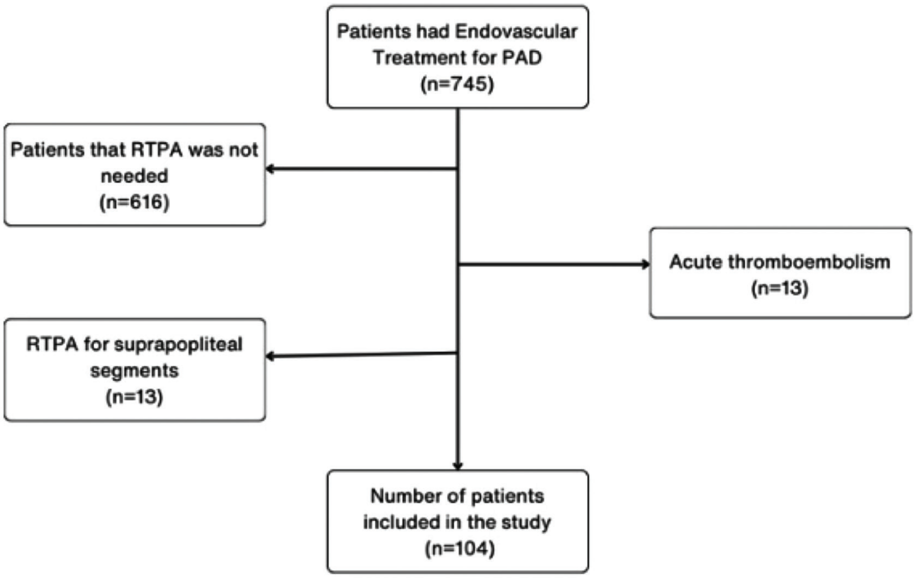

Patients had Endovascular Treatment for PAD (n=745)
Patients that RTPA was not needed (n=616)
Acute thromboembolism (n=13)
RTPA for suprapopliteal segments (n=13)
Number of patients included in the study (n=104)

Figure 1. Flowchart of patient selection. PAD, peripheral arterial disease; RTPA, retrograde tibiopedal access.

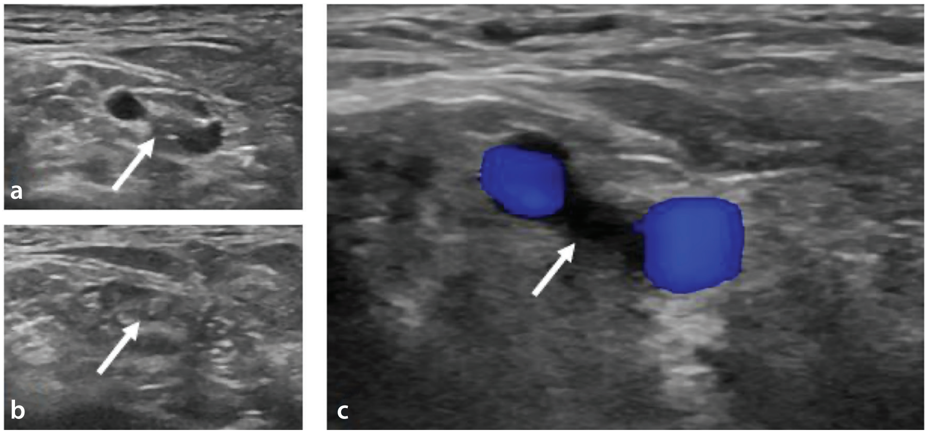


Figure 2. Ultrasound visualization of the occluded posterior tibial artery (PTA) with a linear high-frequency transducer placed transversely in a 46-year-old male patient with thromboangiitis obliterans. **(a)** The PTA (arrow) is seen at the center of the image between the posterior tibial veins. **(b)** The posterior tibial veins are compressed because of pressure applied with the transducer, whereas the PTA (arrow) cannot be compressed. **(c)** In color Doppler imaging, the venous flow can be observed, but there is no arterial flow, which reveals the presence of occlusion.

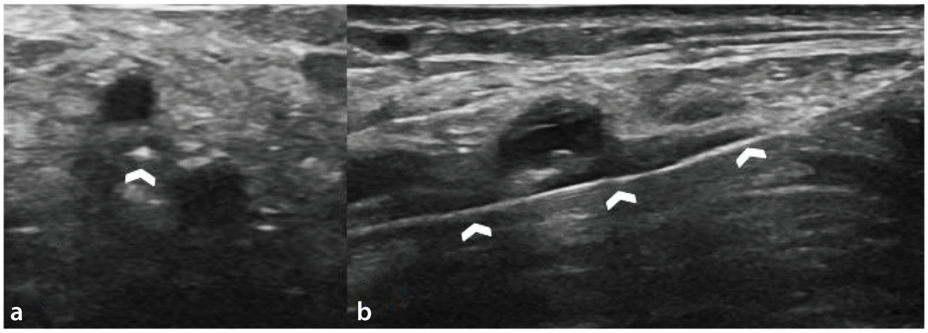


Figure 3. Retrograde tibiopedal access procedure. Using real-time ultrasound (US) imaging, a 21G needle is inserted, and the posterior wall or veins are carefully observed. Transvers **(a)** and longitudinal **(b)** US images indicate 0.018-inch guidewire (arrowheads) with a hydrophilic tip being inserted into the distally occluded posterior tibial artery. Using US to visualize the needle and the guidewire's tactile feedback, the arterial access is confirmed.

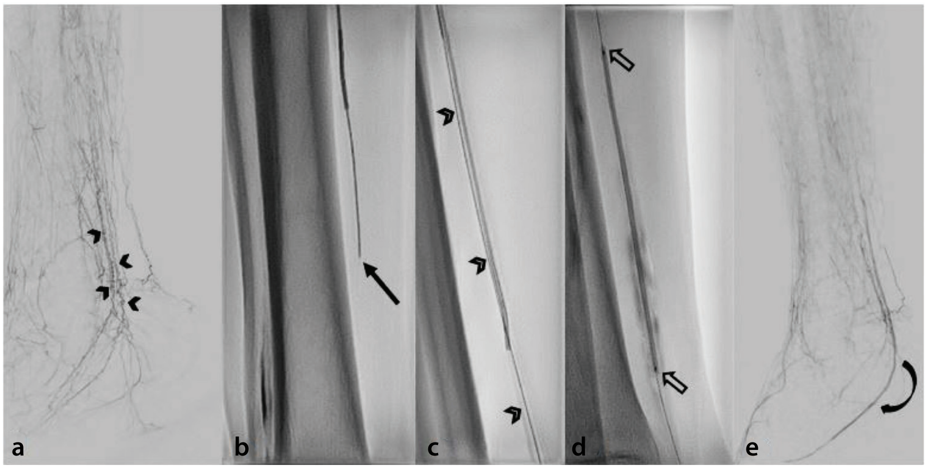


Figure 4. Initial angiogram of the patient and successful recanalization of infrapopliteal arteries. **(a)** The angiogram indicates corkscrew-shaped collaterals (arrowheads) compatible with thromboangiitis obliterans. (**b**) The occluded segment of the posterior tibial artery (PTA) is crossed with the anterograde approach (arrow) using a 014-inch guidewire. **(c)** When the anterograde crossing is unsuccessful, retrograde tibiopedal access (RTPA) is performed, and a 0.018-inch guidewire (arrowhead) is advanced within the occluded PTA lumen. **(d)** Balloon angioplasty is performed with an over-the-wire 2-mm balloon (between arrows). **(e)** The final angiogram demonstrates patent tibial artery flow (curved arrow) and successful recanalization with RTPA.

hydrophilic tip guidewire (V-18 ControlWire, Boston Scientific) is then advanced (Figure 3). Since the targeted arterial segments might be occluded, the verification of arterial access relies on other factors, including direct sonographic observation of the needle within the middle section of the occlusion, tactile feedback obtained from the guidewire, and fluoroscopic or sonographic visualization of the guidewire. A 90-cm support catheter with a straight tip and a diameter of 2.6 F (CXI, Cook Medical) was inserted in cases requiring additional support. Following the successful crossing of the guidewire through the obstruction, the support catheter, or a low-profile balloon catheter was advanced. To ensure accurate placement within the patent lumen, a contrast injection was administered. Subsequently, the procedure involves the utilization of bareback pre-dilatations utilizing balloons with a diameter ranging from 2 to 3 mm. This facilitates the advancement of a guidewire in an antegrade manner through the occluded segments. In most cases, the use of a snare or flossing was not needed following pre-dilatations. Nevertheless, methodologies such as Controlled Antegrade and Retrograde subintimal Tracking (CART) can be employed if deemed required. The subsequent course of treatment involved the continuation of standard endovascular procedures in an antegrade manner. This included the use of routine angiographic imaging, administration of standard medication dosages, angioplasty using balloons ranging from 1.5 to 3 mm for the infrapopliteal and required inframalleolar arteries, and angioplasty using either plain or drug-eluting balloons (Figure 4). The placement of an introducer sheath in the RTPA site was avoided, and a temporary hemostasis measure in the form of a 4F dilator in the 21G introducer set was placed at the RTPA site following pre-dilatation.

Technical success parameters

The evaluation of technical success parameters included the following criteria. The ability to achieve percutaneous access to a distal artery, and successful insertion of a support or a balloon catheter over a wire was considered as achieving access success. The ability to pass the wire to the proximal patent segment of the occlusion was regarded as crossing success. Finally, the successful restoration of flow in the occluded segment as confirmed by angiography was defined as recanalization success. The puncture site complications were assessed via US prior to discharge and on the 7th-day clinical follow-up.

Statistical analysis

Categorical variables were presented as counts and percentages. Successful and unsuccessful RTPA and recanalization attempts were compared in terms of patients' gender, diagnoses, and smoking status using Pearson's chi-square test. A Mann–Whitney U test was performed to compare continuous data, such as age, disease duration, and Rutherford's score, between successful and unsuccessful access and recanalization attempts. Multivariate logistic regression analyses were used to calculate the effects of confounders on RTPA and recanalization success. Two logistic regression statistical models were employed to analyze the access and recanalization success rates of RTPA. In Model A, the effects of all confounders were evaluated. Confounders that did not have a significant effect on the regression model were removed from Model B to obtain optimal results.

A confidence level of 95% was selected, and $P < 0.05$ was considered statistically significant. The data analysis was preformed using IBM SPSS Statistics 22 software.

Results

A total of 87% of the patients had chronic limb-threatening ischemia. The remaining had severe claudication. Fifty-seven (54.8%) patients had diabetes mellitus (DM), 26 (25%) had thromboangiitis obliterans (TAO), and 21 (20.2%) had non-DM atherosclerosis. Fifty-seven (54.8%) patients were active smokers. The DPA, or distal anterior tibial artery, was the access artery in 69 (52.3%) RTPA trials, whereas the PTA was the access artery in 62 (47.7%). The target vessel at the puncture site was occluded in 101 (77.1%) RTPA trials. The access success rate was 82.6% (109/131), the crossing success rate was 95.4% (104/109), and the recanalization success rate was 74% (77/104). Five crossing failures were due to extravasation of the wire in three cases and the inability to traverse the occlusion in two cases. The access success rate was 100% in 30 cases in which the access artery was patent.

A snare was used in four cases from the antegrade access to create an intraluminal through-and-through guidewire. The CART process was required in two cases and was performed successfully in one. Reverse CART was never used.

Patients' age, gender, smoking status, Rutherford scores, and disease duration were not found to be different between successful and failed RTPA trials. Access was significantly more successful when the access artery was DPA when compared with PTA (91.3%, 74.2%, $P = 0.009$, respectively). The access success rate was significantly lower in patients with TAO compared with those with DM and non-DM atherosclerosis (68.6%, 90.3%, and 80.3%, $P = 0.019$, respectively) (Table 1). On the other hand, the recanalization success rate was found to be associated with only the occlusion of the entry site. The recanalization success rate was higher when the puncture site was non-occluded (76.7%, 53.5%, $P = 0.023$) (Table 2).

In the study, two logistic regression statistical models were employed to analyze the access and recanalization success rates of RTPA. A 1.025-fold increase in access success was associated with each unit increase in age for Model A, as measured by an odds ratio (OR) of 1.025 and a 95% confidence interval (CI) ranging from 1.011 to 1.040. Furthermore, compared with PTA access, DPA access increased the successful access rate by 3.185 times, as indicated by an OR of 3.185 and a 95% CI of 1.120 to 9.057. Model B followed a comparable structure, wherein a 1.033-fold increase in access success was observed for every unit increase in age (OR: 1.033, 95% CI: 1.020–1.045). Moreover, DPA access increased access success rates by 2.773 times compared with PTA access in this model, with an OR of 2.773 and a 95% CI ranging from 1.028 to 7.482 (Table 3). The non-occluded access artery increased recanalization success rates by 2,760 times compared with cases with an occluded access artery, with an OR of 2.773 and a 95% CI ranging from 1.117 to 6.817.

Vasospasm at the puncture site was seen in 11 (10.5%) patients. A self-limiting hematoma was seen in two (1.9%) patients. A pseudoaneurysm, or arteriovenous fistula, was not seen in any patients.

Discussion

The findings of this study reveal insights into the outcomes of RTPA trials in patients with infrapopliteal artery involvement but different underlying conditions. Most prominent of all, the access success rate was highest among patients with DM who had a non-occluded DPA at the puncture site. However, the recanalization success rate was broadly influenced only by the occlusion status of the puncture site, regardless of other patient-specific factors or underlying conditions.

Table 1. Comparison of patient-related factors in successful and unsuccessful retrograde tibiopedal access attempts

	Successful RTPA	Unsuccessful RTPA	*P*
Sex			
Male	90	19	
Female	19	3	0.471[a]
Age*	68 (26–93)	64.5 (39–89)	0.344[b]
Current smoking status			
Smoker	58 (84.1)	11 (15.9)	
Non-smoker	51 (82.3)	11 (17.7)	0.783[c]
Access artery			
DPA	63 (91.3)	6 (8.7)	
PTA	46 (74.2)	16 (25.8)	0.009[c]
Access site			
Occluded	80 (79.2)	21 (20.8)	
Non-occluded	29 (96.7)	1 (3.3)	0.025[c]
Diagnosis			
DM	65 (90.3)	7 (9.7)	
TAO	24 (68.6)	11 (31.4)	
AS	20 (80.3)	4 (16.7)	0.019[c]
Disease duration*	15 (2–40)	20 (8–30)	0.391[b]
Rutherford score*	5 (3–6)	4 (3–6)	0.508[b]

*, Median (min-max); [a], Fisher's exact test; [b], Mann–Whitney U test; [c], Pearson's chi-square test. RTPA, retrograde tibiopedal access; DPA, dorsal pedal artery; PTA, posterior tibial artery; DM, diabetes mellitus; TAO, thromboangiitis obliterans; AS, non-diabetic atherosclerosis.

The study found that a high percentage of patients (87%) had chronic limb-threatening ischemia, suggesting that this intervention is often considered for severe cases and as limb salvage. Over half of the patients had DM, aligning with the high prevalence of vascular complications in this patient group. Notably, a significant number of patients (54.8%) were active smokers, further accentuating the comorbid factors often seen in patients with vascular disease. This observation aligns with prior research that has demonstrated a notable prevalence of smoking among individuals with PAD.[12,14]

This study shows an overall access success rate of 82.6%, a crossing success rate of 95.4%, and a recanalization success rate of 74%, indicating that RTPA is generally a reliable technique. The study by Montero-Baker et al.[12] examined the application of RTPA in the treatment of 51 infrapopliteal segment occlusions. The authors reported that the overall success rate of this approach was 86.3%, which is slightly higher than the success rate observed in the present study. However, in Montero-Baker's study, the puncture artery was patent in RTPA, and the guidance was performed using a C-arm, not US.

Access success varied significantly among different underlying conditions. To the best of our knowledge, this study is the first to examine the predictive value of the diagnosis of DM or TAO, access sites, including DPA and PTA, and access artery occlusion status in relation to the technical success achieved in the occlusion of an infrapopliteal artery through RTPA. Notably, patients with TAO had a significantly lower success rate in comparison with those with DM and non-DM atherosclerosis. This could indicate that the etiological factors underlying TAO may present unique challenges to successful vascular access, warranting further investigation.

Table 2. Comparison of patient-related factors in successful and unsuccessful recanalization attempts

	Successful recanalization	Unsuccessful recanalization	*P*
Sex			
Male	90	19	
Female	19	3	0.471[a]
Age*	68 (30–93)	65.5 (26–91)	0.420[b]
Current smoking status			
Smoker	41 (59.4)	28 (40.6)	
Non-smoker	36 (58.1)	26 (41.9)	0.875[a]
Access artery			
DPA	45 (65.2)	24 (34.8)	
PTA	32 (51.6)	30 (48.4)	0.114[a]
Access site			
Occluded	54 (53.5)	47 (46.5)	
Non-occluded	23 (76.7)	7 (23.3)	0.023[a]
Diagnosis			
DM	43 (59.7)	29 (40.3)	
TAO	16 (45.7)	19 (54.3)	
AS	18 (75)	6 (25)	0.078[a]
Disease duration*	15 (2–40)	19 (3–30)	0.804[b]
Rutherford score*	5 (3–6)	5 (3–6)	0.860[b]

*, Median (min-max); [a], Fisher's exact test; [b], Mann–Whitney U test; [c], DPA, dorsal pedal artery; PTA, posterior tibial artery; DM, diabetes mellitus; TAO, thromboangiitis obliterans; AS, non-diabetic atherosclerosis.

Interestingly, the DPA was a more successful access route compared with the PTA, with success rates of 91.3% and 74.2%, respectively. In a previous study conducted by Grözinger et al.[11], which examined the parameters influencing the recanalization success of the superficial femoral artery and infrapopliteal artery using RTPA, the impact of the access artery on technical success did not yield any statistically significant results. Furthermore, the study by Grözinger et al.[11] categorized the access artery into two categories: infrapopliteal arteries and superficial femoral-popliteal arteries. In the present study, the access artery was evaluated in terms of two categories, DPA and PTA, which are both located below the knee (around the ankle), and this provides a more precise anatomical delineation. The study's results imply that clinicians should carefully evaluate the selection of the access artery as a crucial element in the planning of these operations.

The recanalization success rate was shown to be influenced by the occlusion status of the entry site, notwithstanding the high success rate seen in terms of access. The results indicated that recanalization was more effective in cases where the puncture site was patent, hence supporting the significance of maintaining vascular patency at the puncture site and the selection of the access artery to achieve good outcomes.

Table 3. Logistic regression analysis of factors affecting success of retrograde tibiopedal access attempts

	Model A				Model B			
	OR	%95 CI Lower	%95 CI Upper	*P*	OR	%95 CI Lower	%95 CI Upper	*P*
Age	1.025	1.011	1.040	0.001	1.033	1.020	1.045	0.000
Access artery (DPA)	3.185	1.120	9.057	0.030	2.773	1.028	7.482	0.044
Access site (non-occluded)	2.962	0.620	14.149	0.174	3.401	0.762	15.185	0.109
Sex (male)	2.898	0.718	11.696	0.135				
Diagnosis (ref: TAO)				0.462				
Diagnosis (DM)	2.292	0.617	8.516	0.216				
Diagnosis (AS)	1.386	0.371	5.185	0.628				

DPA, dorsal pedal artery; TAO, thromboangiitis obliterans; DM, diabetes mellitus; AS, non-diabetic atherosclerosis; OR, odds ratio; CI, confidence interval.

The research findings indicated that there were no statistically significant variations in outcomes when considering factors such as patients' age, gender, smoking status, Rutherford score, and disease duration. The study conducted by Okuno et al.[1] examined the potential impact of gender, age, and current smoking status on the risk of restenosis following endovascular therapy. The results indicated that none of these factors demonstrated a statistically significant association with the risk of restenosis. Another study conducted by Grözinger et al.[11] did not yield any statistically significant evidence, indicating that the Rutherford score has an impact on the technical success of endovascular treatments with RTPA. Similarly, the present study suggested that these factors lack statistical significance in relation to technical success. This observation suggests that the efficacy of the technique may not be greatly impacted by these variables.

The present investigation documented a rather modest incidence of complications, with vasospasm observed in 10.5% of the patient cohort and a self-resolving hematoma in 1.9% of cases. No major complications were noted related to RTPA. In the multicenter prospective study performed by Walker et al.[13] involving 197 patients, in which the researchers included all occlusions in the infra-inguinal region, no major complications related to RTPA were observed; the overall rate of minor complications remained below 6% and consisted of local pain, infection, ecchymosis, bleeding, and acute vessel dissection. In another study conducted by Goltz et al.[10], significant complications were not detected. However, minor complications consisting of hematoma and vasospasm were observed in 12.5% of the patients, aligning closely with the findings of the present study. Significantly, the absence of more serious complications such as pseudoaneurysms and arteriovenous fistulas suggests that RTPA is generally safe when performed with proficiency and accuracy.

The present study is subject to certain limitations, including the limited number of participants, the retrospective methodology, and the single institution setting. Furthermore, the restricted sample size may potentially limit the generalizability of its findings to wider groups. Further research is warranted to validate these findings and to explore the enduring effects of RTPA, including the inclusion of a broader range of patients. Since the technical aspects and determinants of successful RTPA were the main objective of this study, patency periods and long-term patency rates were not included in the results. However, the success of a method cannot be measured by its technical success alone. Due to the study group's heterogeneity, any assessments of the clinical severity of PAD, such as the WIfi Classification,[15] were excluded from the analysis. However, the primary goal of the study was to compare the success of RTPA in various diseases and clinical circumstances to guide clinicians toward the best decision when contemplating RTPA.

One further limitation of the study pertains to the fact that the retrograde access, crossing, and recanalization procedures were conducted exclusively by a proficient interventional radiologist with expertise in this domain. Achieving access, crossing, or recanalization success and the results of the present study can vary among procedures conducted by various professionals.

In conclusion, this study elucidates the determinants impacting the efficacy of RTPA, emphasizing that the most favorable outcomes were observed in individuals with DM who had a non-occluded DPA at the puncture site. The success rates often exhibit a high level of efficacy; however, it is important to consider that several factors, including the selection of the access artery and the underlying medical condition, might exert an influence on the resulting outcomes. The findings provide valuable insights for clinicians in customizing their strategy based on the unique qualities and situation of each patient. Additional research is needed to further elucidate these observations and formulate more precise clinical recommendations.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Okuno S, Iida O, Shiraki T, et al. Impact of calcification on clinical outcomes after endovascular therapy for superficial femoral artery disease: assessment using the peripheral artery calcification scoring system. *J Endovasc Ther*. 2016;23(5):731-737. [CrossRef]
2. Htun WW, Kyaw H, Aung YL, Maw M, Kwan T. Primary retrograde tibio-pedal approach for endovascular intervention of femoropopliteal disease with chronic total occlusion. *Cardiovasc Revasc Med*. 2020;21(2):171-175. [CrossRef]
3. Malyar NM, Freisinger E, Meyborg M, et al. Low rates of revascularization and high in-hospital mortality in patients with ischemic lower limb amputation: morbidity and mortality of ischemic amputation. *Angiology*. 2016;67(9):860-869. [CrossRef]
4. Schmidt A, Bakker OJ, Bausback Y, Scheinert D. The tibiopedal retrograde vascular access for challenging popliteal and below-the-knee chronic total occlusions: literature review and description of the technique. *J Cardiovasc Surg (Torino)*. 2017;58(3):371-382. [CrossRef]
5. Kersting J, Kersting J, Kamper L, Das M, Haage P. Guideline-oriented therapy of lower extremity peripheral artery disease (PAD) - current data and perspectives. *Rofo*. 2019;191(4):311-322. English, German. [CrossRef]
6. Ali NMOS, Alsaffar MHAAA. Transpedal and tibiopedal retrograde revascularization for peripheral vascular disease. *Cureus*. 2022;14(2):e22082. [CrossRef]
7. Ortiz D, Jahangir A, Singh M, Allaqaband S, Bajwa TK, Mewissen MW. Access site complications after peripheral vascular interventions: incidence, predictors, and outcomes. *Circ Cardiovasc Interv*. 2014;7(6):821-828. [CrossRef]
8. Mustapha JA, Saab F, McGoff T, et al. Tibio-pedal arterial minimally invasive retrograde revascularization in patients with advanced peripheral vascular disease: the TAMI technique, original case series. *Catheter Cardiovasc Interv*. 2014;83(6):987-994. [CrossRef]
9. Agarwal S, Sud K, Shishehbor MH. Nationwide trends of hospital admission and outcomes among critical limb ischemia patients: from 2003-2011. *J Am Coll Cardiol*. 2016;67(16):1901-1913. [CrossRef]
10. Goltz JP, Planert M, Horn M, et al. Retrograde transpedal access for revascularization of below-the-knee arteries in patients with critical limb ischemia after an unsuccessful antegrade transfemoral approach. *Rofo*. 2016;188(10):940-948. English. [CrossRef]
11. Grözinger G, Hallecker J, Grosse U, et al. Tibiopedal and distal femoral retrograde vascular access for challenging chronic total occlusions: predictors for technical success, and complication rates in a large single-center cohort. *EurRadiol*. 2021;31(1):535-542. [CrossRef]
12. Montero-Baker M, Schmidt A, Bräunlich S, et al. Retrograde approach for complex popliteal and tibioperoneal occlusions. *J Endovasc Ther*. 2008;15(5):594-604. [CrossRef]
13. Walker CM, Mustapha J, Zeller T, et al. Tibiopedal access for crossing of infrainguinal artery occlusions: a prospective multicenter observational study. *J Endovasc Ther*. 2016;23(6):839-846. [CrossRef]

14. Aygun MS, Tureli D, Deniz S, Oguzkurt L. Ultrasound-guided retrograde tibial access through chronically occluded tibial arteries: a last resort recanalization technique. *Diagn Interv Radiol*. 2022;28(6):621-626. [CrossRef]

15. de Athayde Soares R, Matielo MF, Brochado Neto FC, et al. WIfI classification versus angiosome concept: a change in the infrapopliteal angioplasties paradigm. *Ann Vasc Surg*. 2021;71:338-345. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242946

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Flow-diverting stents in the management of extracranial carotid artery aneurysms

Celal Cinar[1]
Erol Akgul[2]
Alperen Elek[1]
Mahmut Kusbeci[1]
Egemen Ozturk[3]
Hasan Bilen Onan[4]
Irem Islek[2]
Mohammad Naim Forogh[1]
Mohammad Nawas Nasiri[1]
Ismail Oran[1]

[1] Ege University Faculty of Medicine, Department of Interventional Radiology, İzmir, Türkiye

[2] İstanbul Medipol University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

[3] Uşak Training and Research Hospital, Clinic of Radiology, Uşak, Türkiye

[4] Çukurova University Faculty of Medicine, Department of Radiology, Adana, Türkiye

PURPOSE

This study aims to investigate the indications and therapeutic efficacy of flow-diverting stents (FDSs) in the management of extracranial carotid artery aneurysms (ECAAs) and dissections.

METHODS

A retrospective analysis was conducted on 18 patients treated for ECAAs with an FDS between 2010 and 2024. Patient demographics, aneurysm characteristics, procedural details, and clinical and radiologic follow-up outcomes were extracted from medical records. Procedures were performed under general anesthesia using standard endovascular techniques. Patients received preoperative and postoperative antiplatelet therapy and were fully anticoagulated during the procedure. Follow-up assessments included digital subtraction angiography or computed tomography angiography at 6–12 months and clinical evaluations to monitor symptom resolution and complications.

RESULTS

Eighteen patients, with an average age of 46.44 ± 17.54 years, underwent 19 endovascular interventions. Technical success was achieved in all cases. Single stent deployment was used in 15 aneurysms, and telescopic stent deployment in 7. Total occlusion of the aneurysm was achieved in 94.4% of cases. One patient required retreatment due to the separation of two overlapped telescopic stents. All patients were discharged within 2 days post-procedure, with symptomatic patients experiencing the complete resolution of symptoms. No complications or adverse events were reported during the follow-up period.

CONCLUSION

The endovascular treatment of ECAAs with FDSs appears to be a safe and effective alternative, achieving high technical success and positive clinical outcomes.

CLINICAL SIGNIFICANCE

The use of FDSs for treating ECAAs significantly improves patient outcomes with minimal complications.

KEYWORDS

Carotid artery, stenting, flow diverter, aneurysm, neuroendovascular treatments

Corresponding author: İsmail Oran

E-mail: ismailoran@gmail.com

Received 19 July 2024; revision requested 13 September 2024; accepted 26 November 2024.

Epub: 30.12.2024

Publication date: 08.09.2025

DOI: 10.4274/dir.2024.242946

Extracranial carotid artery aneurysms (ECAAs) account for <1% of all peripheral arterial aneurysms.[1] The most common etiologies of ECAAs include atherosclerosis and dissection with or without trauma.[2] These aneurysms are often diagnosed incidentally during examinations for other pathologic processes and are mostly asymptomatic.[3] Although the risk of ECAA rupture and exsanguination is minimal, complications such as thrombosis, embolization, and nerve compression frequently indicate the need for repair.[4,5]

In cases where ECAAs are located more distally in the internal carotid artery (ICA) and near the base of the skull, endovascular therapy is recommended. Despite the lack of consensus, various types of stents are available for the endovascular treatment of ECAAs. Coated stents

You may cite this article as: Cinar C, Akgul E, Elek A, et al. Flow-diverting stents in the management of extracranial carotid artery aneurysms. *Diagn Interv Radiol.* 2025;31(5):489-495.

are often avoided in tortuous carotid arteries due to their stiffness and lack of maneuverability during the procedure.[3] However, flow-diverting stents (FDSs) have proven to be more effective in treating extracranial aneurysms and dissections.[4,6]

This study aims to investigate the indications and therapeutic efficacy of FDSs in the management of ECAAs and dissections.

Methods

The Institutional Review Board of Ege University Faculty of Medicine approved this retrospective study (protocol number: 24-8T/23, date: 26.06.2024). Informed consent was not required due to this study's retrospective and observational nature. All identifiable details were anonymized during data collection and analysis to ensure patient confidentiality.

We conducted a retrospective analysis on a cohort of 18 patients treated in two institutions for ECAAs using FDSs between 2010 and 2024. Patient demographics, aneurysm characteristics, procedural details, and clinical and radiologic follow-up outcomes were extracted from medical records. These cases were confirmed angiographically using computed tomography (CT) and magnetic resonance imaging. Inclusion criteria encompassed patients diagnosed with cervical ICA aneurysms, irrespective of aneurysm etiology and presentation. Patients with aneurysms located outside the cervical ICA were excluded.

Main points

- The study achieved a 100% technical success rate in treating extracranial carotid artery aneurysms (ECAAs) with flow-diverting stents (FDSs), as all 19 endovascular interventions in 18 patients were successfully performed without peri-procedural complications.
- Follow-up imaging indicated a 94.4% total occlusion rate of aneurysms, with 17 out of 18 patients showing complete occlusion. Only one patient required retreatment due to the separation of two overlapped telescopic stents, which was successfully addressed with a third stent.
- All symptomatic patients experienced the resolution of their symptoms post-treatment.
- The study reported no complications or adverse events, such as transient ischemic attack or stroke, during the follow-up period, indicating a safe profile for FDSs in ECAA treatment.

All procedures were performed under general anesthesia using standard endovascular techniques. The choice between single or telescopic stent deployment was based on aneurysm morphology, size, and the presence of associated vascular lesions. Antiplatelet therapy was administered preoperatively and continued postoperatively in accordance with institutional protocols. Patients were pre-loaded for 5 days with antiplatelet medication (300 mg/day of aspirin and 75 mg/day of clopidogrel). In cases of resistance to clopidogrel, 10 mg/day of prasugrel was used. Platelet function was measured using multiple electrode aggregometry (Multiplate® Analyzer; Roche Diagnostics, Munich, Germany). All tests were undertaken 1 day before the endovascular procedure. According to the consensus opinion of the Working Group on High On-Treatment Platelet Reactivity, platelet aggregation (adenosine diphosphate) values >47 U (the normal range in the absence of an antiaggregant is 57–113 U, as reported by the manufacturer) is considered indicative of nonresponsiveness or hyporesponsiveness (resistance).[7]

All patients were fully anticoagulated with intravenous heparin during the procedure. Post-procedure, dual antiplatelet therapy was continued for 6–12 months, and aspirin was continued for the patient's lifetime.

A 6 or 7 Fr introducer was placed in the groin region for the vascular intervention, followed by navigation into the common carotid artery proximal to the dissection. A microwire inside a microcatheter was then crossed through the dissection segment. An FDS of the appropriate diameter and length was selected according to the measurements made from three-dimensional angiography. After the microcatheter was placed in the lesion, Pipeline (Medtronic, Irvine, CA, USA), Derivo (Acandis, Pforzheim, Germany), and Surpass Evolve (Stryker Neurovascular, Kalamazoo, MI) stents were used.

Technical success was defined as the accurate placement and deployment of the FDS in the targeted segment of the cervical ICA without peri-procedural complications. Digital subtraction angiography or CT angiography was performed routinely at 6 and 12 months after stent deployment. Total occlusion of the aneurysm on imaging was defined as the absence of residual filling. Clinical follow-up assessments were performed to monitor symptom resolution and potential complications.

No statistical comparisons were made in this descriptive study. Summary statistics are reported as median and range for continuous variables or frequency counts and percentages for categorical variables.

Result

A total of 18 patients, comprising 8 men (45%) and 10 women (55.5%), underwent 19 endovascular interventions. The average age was 46.44 ± 17.54 years, ranging from 8 to 68 years. Six cases were discovered incidentally during imaging investigations for other pathological processes, whereas the other patients presented with various symptoms (Table 1).

Clopidogrel resistance was detected in three patients; they were re-loaded with prasugrel. All patients were treated with FDSs. Technical success was achieved in all cases (100%) (Figures 1-4). Single stent deployment was utilized in 15 locations, whereas telescopic (dual) stent deployment was employed in 7 aneurysms. Imaging follow-up indicated that the total occlusion of the aneurysm was achieved in 17 out of 18 patients (94.4%). One patient required retreatment (patient 16) due to the separation of two overlapped telescopic stents, resulting in residual filling. When evaluated retrospectively, it was thought that the stent separation was caused by insufficient manipulation during initial stent deployment and leaving the short overlapped segment. This was successfully addressed with a third stent. Notably, aneurysm occlusion persisted in subsequent follow-ups.

All patients were discharged on postoperative day 1 or 2. Clinically, symptom resolution was observed in symptomatic patients, including the complete disappearance of neck pain and swallowing difficulties. No complications or adverse events (transient ischemic attack or stroke) were reported during the follow-up period.

Discussion

Our study found that FDSs are highly effective in treating ECAAs. Among the 18 patients who underwent 19 endovascular interventions, technical success was achieved in all cases, with 94.4% (17 out of 18) of aneurysms showing total occlusion on follow-up imaging. One patient required retreatment due to the separation of two telescopic stents. Clinical outcomes were positive, with symptomatic patients experiencing the resolution of symptoms, and no complications

Table 1. Summary of patients with ECAAs treated with an FDS

No/age (year)/ sex	Etiology	Presentation	Aneurysm: side/ geometry/length/ diameter/neck	Associated vascular lesions	FDS name/size	Radiologic follow-up	Clinical follow-up
1/8/F	Fall	Swallowing difficulties, neck mass	R/saccular/6 cm/3 cm/0.5 cm	None	Pipeline/5 × 30 mm (two telescopic)	1-year CTA: total occlusion	Disappearance of symptoms
2/45/M	Unknown	TIA	L/fusiform/2 cm/0.9 cm/1.5 cm	None	Pipeline/5 × 30 mm	1- year CTA: total occlusion	No complaint
3/21/M	Fall (suicide)	Incidental (polytrauma)	R/saccular/1.5 cm/1.5 cm/0.5 cm L/fusiform/3.5 cm/2 cm/2.5 cm	Aortic transection treated with stent graft	Pipeline/5 × 30 mm each	6-month CTA: total occlusion	No complaint
4/54/F	Unknown	Neck mass	R/saccular/2.8 cm/2.2 cm/1 cm	None	Pipeline/5 × 30 mm (two telescopic)	6-month CTA: total occlusion	No complaint
5/ 65/M	Unknown	Incidental	L/saccular/1.6 cm/0.5 cm/1 cm	None	Pipeline/5 × 30 mm	1-year CTA: total occlusion	No complaint
6/68/F	Unknown	Neck mass	R/saccular/3 cm/3 cm/1.3 cm	None	Pipeline/5 × 30 mm	6-month CTA: total occlusion	No complaint
7/35/F	Unknown	Acute neck pain	L/fusiform/2.5 cm/1.5 cm/2 cm	R narrowing of cervical ICA due to long segment dissection	Pipeline/5 × 30 mm	12-month CTA: total occlusion	No complaint
8/45/M	Unknown	TIA	L/two saccular/2.5 cm/1.5 cm/1 cm and 1.2 cm/1.2 cm/0.7 cm	None	Pipeline/5 × 30 mm (two telescopic)	24-month CTA: total occlusion	No complaint
9/64/F	Unknown	Incidental	L/saccular/0.8 cm/0.8 cm/0.4 cm	L cavernous ICA aneurysm 15 mm in diameter	Pipeline/5 × 30 mm	15-month CTA: total occlusion	No complications
10/65/F	Fibromuscular dysplasia	Incidental	R/Saccular/0.3 cm/0.3 cm/0.3 cm	Two intracranial aneurysms 6 mm and 4 mm in diameter	Pipeline/5 × 30 mm	6-month CTA: total occlusion	No complications
11/51/F	Unknown	Syncope	R/saccular/0.5 cm/0.3 cm/0.4 cm L/fusiform/1.5 cm/0.8 cm/0.5 cm	Two intracranial aneurysms 1.3 cm and 0.5 cm in diameter	Pipeline/4.5 × 20 mm (R) Surpass/ 5 × 50 mm (L)	38-month DSA: total occlusion	No complaint
12/16/M	Unknown	Right transient hemiparesis 8 months before	L/fusiform/4 cm/1.5 cm/ 3 cm	None	Derivo/5.5 × 50 mm (two telescopic)	6-month DSA: apparently diminished aneurysm. 36-month CTA: almost total occlusion with minimal neck filling	No complaint
13/33/M	Unknown	Acute stroke, mechanical thrombectomy 4 weeks before	L/fusiform/1.5 cm/1 cm/1.5 cm	Dissecting stenosis involving whole prepetrous ICA segment	Surpass/5 × 20 mm and Derivo/ 5 × 50 mm (telescopic)	30-month DSA: aneurysm occlusion with 50% long segment stenosis	No complaint
14/48/F	Unknown	Incidental	R/saccular/0.9 cm/0.9 cm/0.5 cm L/fusiform/3 cm/1 cm/2.5 cm	Cavernoma in cervical spinal cord	Pipeline/5 × 30 mm in each	6-month DSA: total occlusion	No complaint
15/52/F	Unknown	Incidental	R/saccular/1.5 cm/1.5 cm/0.5 cm	Occluded L ICA	Pipeline/5 × 30 mm	12-month Doppler US: total occlusion	No complaint
16/61/M	Unknown	Neck pain	R/fusiform/2.5 cm/1.5 cm/0.6 cm	Three intracranial aneurysms 5, 6, and 15 mm in diameter	Derivo/4.5 × 30 and 5.5 × 30 mm (telescopic) Derivo/5.5 × 50 mm (retreatment)	6-month DSA: residual filling due to separation of overlapped stents. Retreatment with third stent. 18-month DSA: total occlusion	No complaint
17/54/F	Connective tissue disorder	Neck pain	R/saccular/1 cm/1 cm/0.4 cm and 0.9 cm/0.8 cm/0.4 cm L/saccular/2 cm/1 cm/0.5 cm	None	Pipeline/4.75 × 30 mm in each.	6-month CTA: total occlusion	Neck pain subsides
18/51/M	Fibromuscular dysplasia	Neck pain	R/fusiform/1 cm/0.8 cm/0.8 cm L/fusiform/1 cm/0.9 cm/0.7 cm	None	Derivo/5 × 30 mm (two telescopic in R) Derivo/5 × 30 mm (L)	6-month DSA: total occlusion	Neck pain subsides

F, female; M, male; ECAAs, extracranial carotid artery aneurysms; FDS, flow-diverting stent; TIA, transient ischaemic attack; L, left; R, right; ICA, internal carotid artery; CTA, computed tomography angiography; DSA, digital subtraction angiography; US, ultrasound.

ovascular methods, surgical treatment was preferred for Attigah type I ECAAs at their institution (64.0% vs. 40.0%, $P = 0.09$), and both surgical and endovascular treatments were deemed safe.

A meta-analysis by Galyfos et al.[20] involving 374 patients with 383 ECAAs (220 were treated with open surgery and 81 with endovascular methods) found similar 30-day mortality rates for open surgery and endovascular treatments [4% vs. 0%; pooled odds ratio (OR), 2.67; 95% confidence interval (CI), 0.291–24.451]. Stroke and transient ischemic attack rates were also comparable (5.5% vs. 1.2%; pooled OR, 1.42; 95% CI, 0.412–4.886), but cranial injury was more common in open surgery (14.5% vs. 0%; OR, 3.98; 95% CI, 1.178–13.471).[20] The literature also shows that the perioperative stroke rate for endovascular treatment ranges from 2% to 3.1%.[9,10] Similarly, Ni et al.[21] demonstrated in a study with a 2-year follow-up that no deaths or neurological adverse events occurred.

Endovascular modalities described in the literature include covered stenting,[12] bare metal stenting,[2] multiple stent techniques (telescoping stenting, overlapped stenting),[22] and stent-assisted coiling[14] for treating ECAAs. Self-expanding carotid stents have traditionally been used for treating carotid atherosclerosis in high-risk patients due to their positive effects on coronary atherosclerosis. Recently, these stents have also been employed to address spontaneous dissections or those caused by trauma or angioplasty.[4,8,23] However, mechanical tests reveal that self-expanding carotid stents tend to stiffen, with bending stiffness increasing non-linearly as deflection rises.[24] This stiffness makes these stents less suitable for use in a distal cervical or petrous ICA, where sharp angulation at the skull base occurs. Furthermore, carotid stents with large cell designs are highly porous and may lack sufficient radial force to seal a false lumen or induce thrombosis in a pseudoaneurysm.[25]

In our study, we used FDSs for all cases. Although there are limited reports on the use of FDSs,[4,5,25-28] they offer several advantages over traditional closed- and open-cell stents. Kurre et al.[28] reported their experience with stent placement for acute ICA dissections in 73 patients presenting with acute ischemia, using FDSs in approximately 30% of cases. They reported excellent success rates (100%) for justified reconstructions of the cervical ICA and a low complication rate (8%), with no new ischemic symptoms in treated dis-

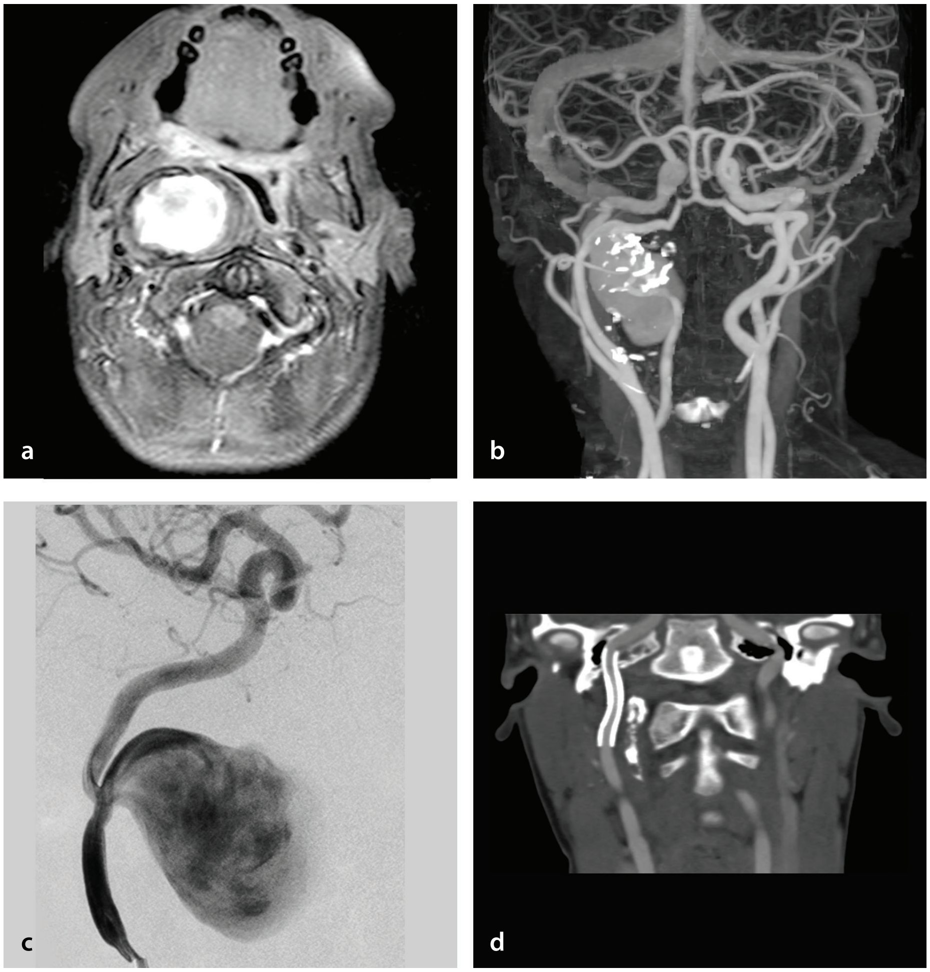


Figure 1. Patient 1: Chronic traumatic aneurysm **(a)**. Magnetic resonance imaging shows a huge mass compressing the esophagus and laryngeal air passage. **(b)** Subtracted computed tomography (CT) angiography reveals the aneurysm with extensive wall calcification. **(c)** Angiography demonstrates a huge aneurysm arising from the pre-petrous segment of the right internal carotid artery. **(d)** One-year follow-up CT angiography confirms the disappearance of the aneurysm and shrinkage of the mass.

or adverse events were reported during the follow-up period.

ECAAs can arise from various etiologies, including atherosclerosis, trauma, infections, and inflammatory conditions.[2] Many cases, similar to those in our study, can be idiopathic.[8] Giannopoulos et al.'s[9] systematic review found trauma to be the cause in 54.3% (38 out of 70) of cases. Similar to our study, the literature indicates that neurological symptoms occur in approximately 42% to 51% of ECAA cases.[10-12] Given the high morbidity associated with ECAAs, treatment is recommended upon diagnosis, especially if symptomatic.[13,14] Untreated ECAAs can lead to distal embolization (particularly in true aneurysms) or exert a mass effect on adjacent structures (particularly in false aneurysms).[15,16]

Several therapeutic strategies have been proposed for managing ECAAs, including surgical, endovascular, and conservative therapies. In addition, there is a case report of a complex ECAA treated with both endovascular and open surgical approaches.[17] However, the optimal treatment modality remains controversial due to the lack of established guidelines. Open surgery for ECAA treatment has a 2.6% peri-procedural mortality rate, and cranial nerve injury occurs in 11.8% to 26% of cases.[2,9] Moreover, open surgery can be risky if the aneurysm's location and patient suitability are not optimal. Attigah et al.[18] classified aneurysms into high (type I) and very low (type V) positions, with these positions being more suitable for endovascular treatment. In an observational study by Choi et al.[19] involving 41 patients treated with surgical, conservative, or end-

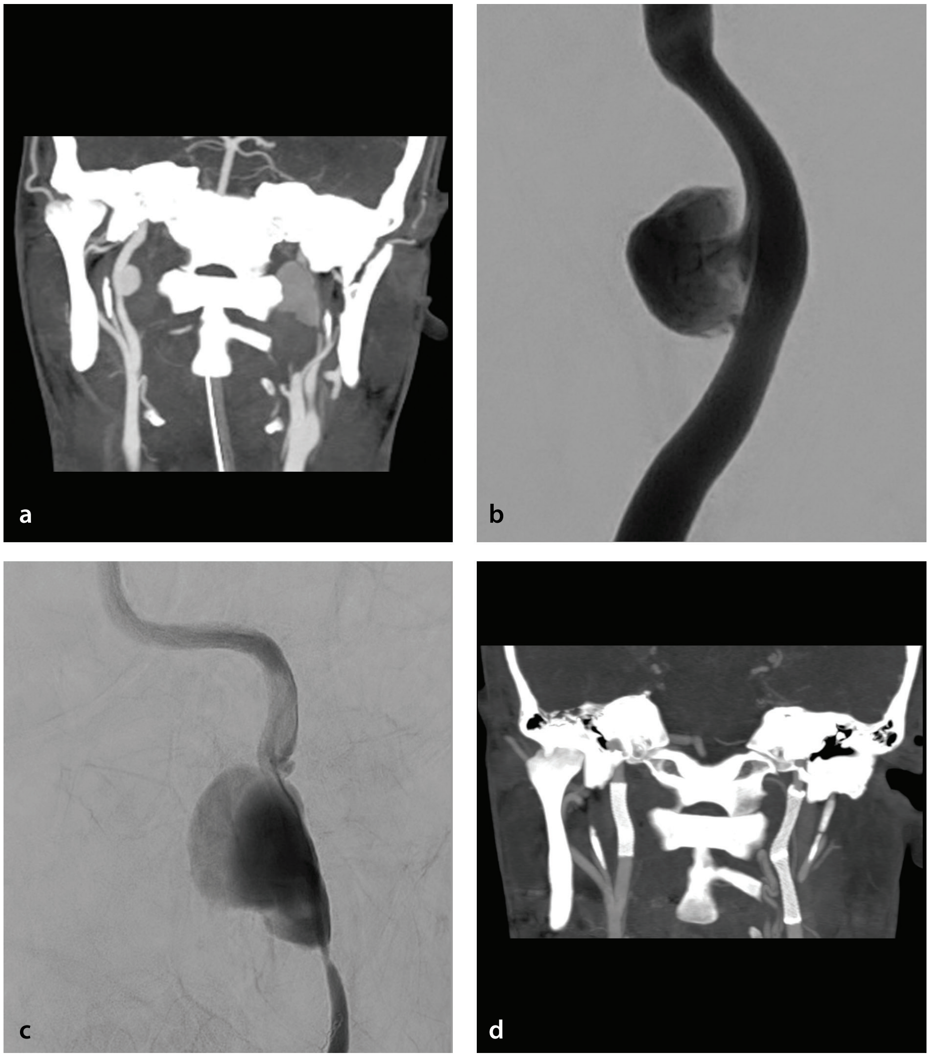


Figure 2. Patient 3: Bilateral acute traumatic dissecting aneurysms. **(a)** Computed tomography (CT) angiography reveals a bilateral high-cervical internal carotid artery (ICA) dissection accompanied by saccular (R) and fusiform (L) aneurysms. **(b, c)** Angiograms show more clearly the anatomy of the right **(b)** and left **(c)** cervical ICA aneurysms. **(d)** Six-month control CT angiography confirms the disappearance of both aneurysms and the normal calibration of the dissected segments.

sections.[28] Similarly, Hilditch et al.[5] treated seven young patients with symptomatic extracranial ICA dissection using FDSs, with no serious perioperative complications. None of the patients experienced recurrent ischemic events following ICA reconstruction, and no postprocedural in-stent stenosis was observed.

FDSs are approved for the treatment of wide-necked intracranial aneurysms and are potentially suitable for treating dissections with or without aneurysms at the skull base due to several unique features. The softer and more flexible characteristics of FDSs provide greater durability against stent fracture in the highly mobile high-cervical ICA transition at the skull base. FDSs are low-porosity woven tubes, offering three times the vessel wall coverage compared with traditional intracranial stents.[25] The higher metal coverage of the parent vessels (30%–50%) of FDSs can improve the closure of a dissection flap or pseudoaneurysm and reduce continued blood flow into a false lumen. This can also reduce recurrent embolic events, providing an advantage over braided stents, which generally offer less metal coverage.[29] In addition, FDSs facilitate the neo-intimal remodeling of the parent artery. Another significant feature of the FDS is its greater flexibility and adaptable radial force compared with traditional self-expanding carotid stents, allowing easier accommodation to sharp angulation at the skull base.

Despite the advantageous properties of FDSs in treating ECAAs with or without dissection, certain factors may limit their future use. Notably, some features of the extracranial cervical vessel structure, such as high lumen pressures and frequent positional changes due to neck movement, pose a higher risk of stent migration compared with intracranial vessels.[30] Both proximal migration in the anterior and posterior circulation and the spontaneous shortening of FDSs have been reported.[31] In our series, we encountered the separation of two overlapped telescopic stents in only one patient. Another concern is the need for dual antiplatelet agents 6–12 months following FDS placement, complicating the management of any medical conditions requiring surgery.[26]

Another significant limitation of FDSs in cervical segments is the parent artery diameter, as current flow diverters are recommended for vessel diameters of up to 5.2–5.75 mm, designed to open approximately 0.25 mm above their nominal diameter, with the largest available size being approximately 5.25 mm. For arteries measuring wider than 5.75 mm, other adjunctive endovascular techniques should be considered. For example, Amuluru et al.[4] and Rahal et al.[32] reported concurrent anchoring strategies with FDS deployment in cases where the distal cervical ICA measured ≥5.25 mm. To ensure adequate coverage of the aneurysm neck and to cover long segment dissection, if any, we intentionally used multiple FDSs in a telescoping configuration in six patients. Tsang et al.[26] also highlighted the use of the telescoping method in six of the seven cases in their series.

This study has several limitations that should be considered when interpreting the findings. The retrospective, non-randomized design may introduce recall bias and limit the establishment of causal relationships. The small sample size because of the low incidence of ECAAs limits the study's power. The etiology and aneurysm type also differed in each case. Additionally, some patients did not have a long enough follow-up period. Although these limitations require cautious interpretation, they also point to opportunities for future research to address these constraints and enhance our understanding of the subject.

In conclusion, considering the patient's condition and the characteristics of the aneurysm, the endovascular treatment of ECAAs with FDSs appears to be a safe and feasible alternative.

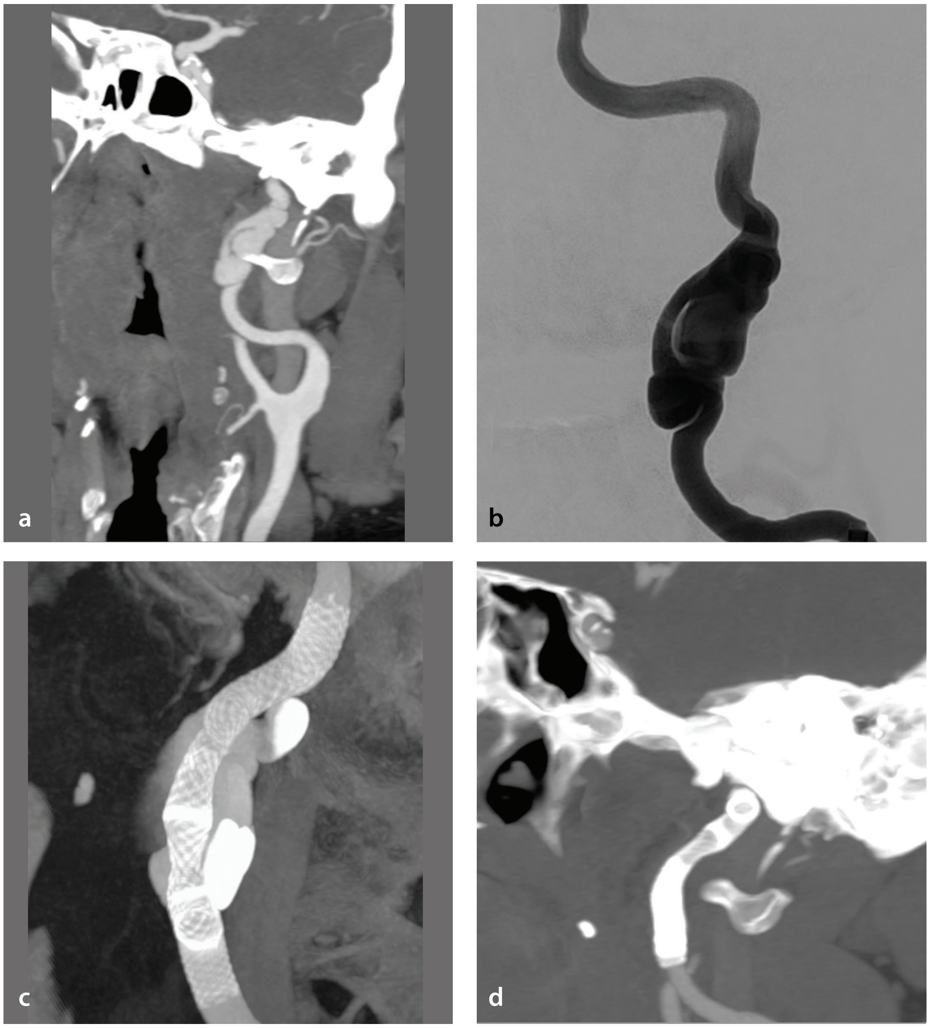


Figure 3. Patient 8: Chronic dissecting aneurysms in the high-cervical segment of the left internal carotid artery. Computed tomography (CT) angiography **(a)** and digital subtraction angiography **(b)**. **(c)** Cone-beam CT after telescopic flow-diverting stent implantation in the dissected segment. **(d)** Twenty-four-month control CT angiography confirms aneurysm occlusion and the reconstruction of the dissected segment.

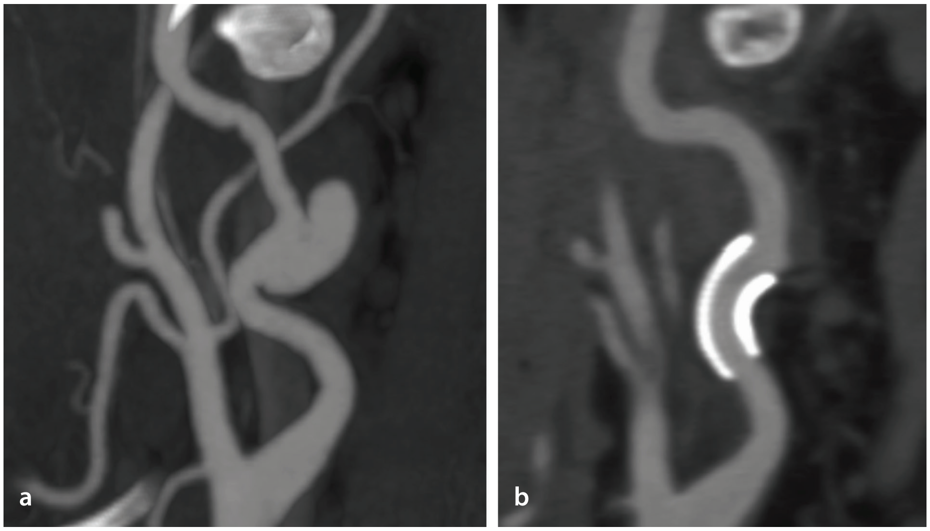


Figure 4. Patient 2: Sausage-shaped, presumably dissecting aneurysm located in the left cervical internal carotid artery (ICA). **(a)** Computed tomography (CT) angiography shows an aneurysm located in the low cervical ICA. **(b)** One-year control CT angiography confirms occlusion of the aneurysm.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. El-Sabrout R, Cooley DA. Extracranial carotid artery aneurysms: Texas Heart Institute experience. *J Vasc Surg*. 2000;31(4):702-712. [CrossRef]
2. Welleweerd JC, Nelissen BG, Koole D, et al. Histological analysis of extracranial carotid artery aneurysms. *PLoS One*. 2015;10(1):e0117915. [CrossRef]
3. Robijn SMM, Welleweerd JC, Lo RTH, Moll FL, de Borst GJ. Treatment of an extracranial internal carotid artery aneurysm with a flow-diverting stent. *J Vasc Surg Cases*. 2015;1(2):191-193. [CrossRef]
4. Amuluru K, Al-Mufti F, Roth W, Prestigiacomo CJ, Gandhi CD. Anchoring pipeline flow diverter construct in the treatment of traumatic distal cervical carotid artery injury. *Interv Neurol*. 2017;6(3-4):153-162. [CrossRef]
5. Hilditch CA, Brinjikji W, Schaafsma J, et al. Flow-diverter stents for internal carotid artery reconstruction following spontaneous dissection: a technical report. *Clin Neuroradiol*. 2019;29:707-715. [CrossRef]
6. Martínez-Galdámez M, Rodríguez C, Hermosín A, et al. Internal carotid artery reconstruction with a "mega flow diverter": first experience with the 6×50 mm DERIVO embolization device. *Neurointervention*. 2018;13(2):133-137. [CrossRef]
7. Tantry US, Bonello L, Aradi D, et al. Consensus and update on the definition of on-treatment platelet reactivity to adenosine diphosphate associated with ischemia and bleeding. *J Am Coll Cardiol*. 2013;62(24):2261-2273. [CrossRef]
8. Gao P, Qi J, Wang M, et al. Endovascular treatment of extracranial carotid artery aneurysms using self-expandable covered stent grafts: a single center retrospective study. *Vascular*. 2022;30(1):14-20. [CrossRef]
9. Giannopoulos S, Trinidad E, Aronow H, Soukas P, Armstrong EJ. Endovascular repair of extracranial carotid artery aneurysms: a systematic review. *Vasc Endovascular Surg*. 2020;54(3):254-263. [CrossRef]
10. Li Z, Chang G, Yao C, et al. Endovascular stenting of extracranial carotid artery aneurysm: a systematic review. *Eur J Vasc Endovasc Surg*. 2011;42(4):419-426. [CrossRef]
11. Srivastava SD, Eagleton MJ, O'Hara P, Kashyap VS, Sarac T, Clair D. Surgical repair of carotid artery aneurysms: a 10-year, single-center experience. *Ann Vasc Surg*. 2010;24(1):100-105. [CrossRef]
12. Zhou W, Lin PH, Bush RL, et al. Carotid artery aneurysm: evolution of management over

two decades. *J Vasc Surg*. 2006;43(3):493-496. [CrossRef]

13. Kaupp HA, Haid SP, Jurayj MN, Bergan JJ, Trippel OH. Aneurysms of the extracranial carotid artery. *Surgery*. 1972;72(6):946-952. [CrossRef]
14. Welleweerd JC, Moll FL, de Borst GJ. Technical options for the treatment of extracranial carotid aneurysms. Expert Rev Cardiovasc Ther. 2012;10(7):925-931. [CrossRef]
15. Mohan IV, Stephen MS. Peripheral arterial aneurysms: open or endovascular surgery? *Progress in cardiovascular diseases*. 2013;56(1):36-56. [CrossRef]
16. Pulli R, Dorigo W, Alessi Innocenti A, Pratesi G, Fargion A, Pratesi C. A 20-year experience with surgical management of true and false internal carotid artery aneurysms. *Eur J Vasc Endovasc Surgery*. 2013;45(1):1-6. [CrossRef]
17. Montanari F, Venturini L, Valente I, Minucci M, Donati T, Tshomba Y. Hybrid treatment of large extracranial carotid artery aneurysm. *J Vasc Surg Cases Innov Tech*. 2023;9(2):101117. [CrossRef]
18. Attigah N, Külkens S, Zausig N, et al. Surgical therapy of extracranial carotid artery aneurysms: long-term results over a 24-year period. *Eur J Vasc Endovasc Surg*. 2009;37(2):127-133. [CrossRef]
19. Choi E, Gwon JG, Kwon SU, Lee DH, Kwon TW, Cho YP. Management strategy for extracranial carotid artery aneurysms: A single-center experience. *Medicine (Baltimore)*. 2022;101(19):e29327. [CrossRef]
20. Galyfos G, Eleftheriou M, Theodoropoulos C, et al. Open versus endovascular repair for extracranial carotid aneurysms. *J Vasc Surg*. 2021;74(3):1017-1023. [CrossRef]
21. Ni L, Pu Z, Zeng R, et al. Endovascular stenting for extracranial carotid artery aneurysms: experiences and mid-term results. *Medicine (Baltimore)*. 2016;95(46):e5442. [CrossRef]
22. Cornwall JW, Png CYM, Han DK, Tadros RO, Marin ML, Faries PL. Endovascular techniques in the treatment of extracranial carotid artery aneurysms. *J Vasc Surg*. 2021;73(6):2031-2035. [CrossRef]
23. Yadav JS, Roubin GS, King P, Iyer S, Vitek J. Angioplasty and stenting for restenosis after carotid endarterectomy. Initial experience. *Stroke*. 1996;27(11):2075-2079. [CrossRef]
24. Carnelli D, Pennati G, Villa T, Baglioni L, Reimers B, Migliavacca F. Mechanical properties of open-cell, self-expandable shape memory alloy carotid stents. *Artif Organs*. 2011;35(1):74-80. [CrossRef]
25. Brzezicki G, Rivet DJ, Reavey-Cantwell J. Pipeline embolization device for treatment of high cervical and skull base carotid artery dissections: clinical case series. *J Neurointerv Surg*. 2016;8(7):722-728. [CrossRef]
26. Tsang AC, Leung KM, Lee R, Lui WM, Leung GK. Primary endovascular treatment of post-irradiated carotid pseudoaneurysm at the skull base with the Pipeline embolization device. *J Neurointerv Surg*. 2015;7(8):603-607. [CrossRef]
27. Fischer S, Perez MA, Kurre W, Albes G, Bäzner H, Henkes H. Pipeline embolization device for the treatment of intra- and extracranial fusiform and dissecting aneurysms: initial experience and long-term follow-up. *Neurosurgery*. 2014;75(4):364-374. [CrossRef]
28. Kurre W, Bansemir K, Aguilar Pérez M, et al. Endovascular treatment of acute internal carotid artery dissections: technical considerations, clinical and angiographic outcome. *Neuroradiology*. 2016;58(12):1167-1179. [CrossRef]
29. Nerva JD, Morton RP, Levitt MR, et al. Pipeline embolization device as primary treatment for blister aneurysms and iatrogenic pseudoaneurysms of the internal carotid artery. *J Neurointerv Surg*. 2015;7(3):210-216. [CrossRef]
30. Malek AM, Higashida RT, Phatouros CC, et al. Endovascular management of extracranial carotid artery dissection achieved using stent angioplasty. *AJNR Am J Neuroradiol*. 2000;21(7):1280-1292. [CrossRef]
31. Chalouhi N, Tjoumakaris SI, Gonzalez LF, et al. Spontaneous delayed migration/shortening of the pipeline embolization device: report of 5 cases. *AJNR Am J Neuroradiol*. 2013;34(12):2326-2330. [CrossRef]
32. Rahal JP, Dandamudi VS, Heller RS, Safain MG, Malek AM. Use of concentric solitaire stent to anchor pipeline flow diverter constructs in treatment of shallow cervical carotid dissecting pseudoaneurysms. *J Clin Neurosci*. 2014;21(6):1024-1028. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243198

INTERVENTIONAL RADIOLOGY

TECHNICAL NOTE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Utilization of a steerable microcatheter and adjunctive techniques for prostatic artery embolization in anatomically challenging vesicoprostatic trunks

Hippocrates Moschouris[1]
Çağın Şentürk[2]
Konstantinos Stamatiou[3]

[1]Tzanio General Hospital, Clinic of Radiology, Piraeus, Greece

[2]İzmir Tınaztepe University Galen Hospital, Department of Interventional and Neuroendovascular Radiology, İzmir, Türkiye

[3]Tzanio General Hospital, Clinic of Urology, Piraeus, Greece

ABSTRACT

Prostatic artery (PA) origination from a common trunk with the superior vesical artery (SVA) is a frequent cause of technical difficulties in PA catheterization for PA embolization (PAE). These difficulties, which substantially increase the operative time, radiation dose, cost, and technical failure rate of PAE, can often be overcome by the utilization of a steerable microcatheter (MC) with a tip that can be manually adjusted at an angle that optimally conforms to the shape and origin of the common vesicoprostatic trunk. Adjunctive techniques that can be applied when the steerable MC fails to engage the PA include: 1) the protective temporary embolization of the SVA so that a permanent embolic can be redirected into the PA; 2) PAE via collaterals between superior vesical branches and the PA; and 3) embolization from a proximal position of the MC near the PA orifice to exploit preferential flow to the PA. In the authors' recent experience, the utilization of a steerable MC with and without adjunctive techniques (in 12 and 23 patients, respectively) resulted in a 35% increase in the technically successful embolization of PAs originating from vesicoprostatic trunks with no significant complications. Familiarization with alternative devices and techniques may substantially improve the technical outcome of PAE in cases with challenging arterial anatomy.

KEYWORDS

Angiography, microcatheter, prostatic artery, prostatic artery embolization, vesicoprostatic trunk

Prostatic artery (PA) origination from a common *vesicoprostatic* trunk with the superior vesical artery (SVA) (referred to as type 1 PA origination, according to a widely used angiographic classification)[1] is the most prevalent (or second most, depending on the population in question) variety of PA origination.[1,2] Compared with other types of PA origination, type 1 is associated with significantly more technical difficulties during attempted PA catheterization for PA embolization (PAE). These difficulties, mainly caused by a short and cranially oriented vesicoprostatic trunk, with or without an unfavorable angle of origin of the PA from this common trunk, may lead to prolonged operative times, increased radiation doses for staff and patients, and the need for additional microcatheters (MCs) and microguidewires (MGWs).[2,3] The rate of unsuccessful attempts at superselective catheterization of the PA is also significantly higher for type 1 origination compared with all other types combined,[2] and this eventually translates into unilateral (rather than bilateral) PAE procedures with suboptimal clinical outcomes.

Among other approaches, the utilization of a steerable MC with a tip manually adjustable by the operator to angles of 0°–180°, has been briefly reported in the literature[3-5] as an option to address the challenges associated with vesicoprostatic trunks during PAE. There is likely room for refinement and the application of adjunctive techniques in the deployment of the steerable MC to further improve the technical outcomes of PAEs. The aim of this work is therefore to describe the utilization of a steerable MC and of adjunctive techniques for PAE in the context of anatomically challenging vesicoprostatic trunks. The clinical efficacy and safety of the described techniques is also briefly evaluated.

Corresponding author: Hippocrates Moschouris

E-mail: hipmosch@gmail.com

Received 24 December 2024; revision requested 6 January 2025; accepted 17 January 2025.

Epub: 07.04.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.243198

You may cite this article as: Moschouris H, Şentürk Ç, Stamatiou K. Utilization of a steerable microcatheter and adjunctive techniques for prostatic artery embolization in anatomically challenging vesicoprostatic trunks. *Diagn Interv Radiol.* 2025;31(5):496-501.

Methods

In the centers of the authors, a steerable 2.4 French (Fr) MC (SwiftNINJA, Merit Medical Systems, Inc., South Jordan, UT, USA) was deployed when catheterization of the common vesicoprostatic trunk proved impossible after attempts of approximately 3 min duration with the initial standard combination of devices. This combination included a 2 Fr MC (Parkway Soft-Asahi Intecc Co., Japan) and a double-angled MGW (0.016" Meister, Asahi Intecc Co.). Compared with the standard device, the tip of the steerable MC can not only be adjusted so that it is optimally oriented to the orifice of the common vesicoprostatic trunk but also provides better support and a steadier angle that is not prone to straightening upon insertion of the MGW. The tip of the steerable MC was locked at the desired angle and catheterization was attempted. After initial engagement of the orifice of the common vesicoprostatic trunk with the MGW, the MGW tended to advance more easily into the SVA than into the PA. After distal advancement of the MGW into the SVA, the angled tip of the steerable MC was "unlocked," and the device was advanced distally over the MGW and into the SVA to secure this first step of catheterization. Subsequently, the MC–MGW combination was slowly retracted under fluoroscopy until it reached the origin of the PA. Catheterization of the latter was then attempted with appropriate rotation of the double-angled MGW and, as necessary, with a new adjustment of the tip of the MC so that it was angled and "locked" in the direction of PA origination (Supplementary Figures 1, 2). The same double-angled MGW used in the initial attempts was combined with the steerable MC. Occasionally, however, if appropriately angled, the steerable MC alone (without the MGW) could be advanced to engage first the common vesicoprostatic trunk and then the PA (Supplementary Videos 1, 2). A high-quality "roadmap" image greatly facilitated the identification of the origins of the target arteries.

Main points

- Prostatic artery (PA) origination from the anterior division of the internal iliac artery in the form of a common trunk with the superior vesical artery (i.e., a vesicoprostatic trunk) is frequently encountered.
- This vesicoprostatic trunk may be clinically relevant to PA embolization because of the frequently associated difficulties arising from it during the procedure.
- These difficulties may result in increased operative times, radiation doses, and costs and even in technical failure of the procedure.
- A steerable microcatheter with a tip that can be manually adjusted to an angle of 0°–180° can be employed with or without adjunctive techniques to overcome the aforementioned difficulties.

When attempts with a steerable MC and MGW proved fruitless, the following techniques were applied:

1. Protective temporary embolization of the SVA. This can be applied when the MC easily advances into the SVA but cannot be redirected into the PA. In the current work, a temporary embolic agent (EmboCube Gelatin, Merit Medical) with particles of a hydrated size of 2.5 mm was injected through the MC into the SVA. Care was taken to avoid a too-distal or too-proximal occlusion, including at the PA origin. After angiographic confirmation of an SVA occlusion, the MC was withdrawn just proximal to the PA origin. A permanent embolic (composed of microspheres) was then slowly injected, and its redirection into the PA was fluoroscopically documented (Figure 1). Care is required to avoid backflow into the anterior division of the internal iliac artery. Occasionally, resistance to the advancement of the MGW in the SVA after its occlusion forces the MGW to enter an otherwise non-selectable PA. At this point, superselective catheterization of the PA can be accomplished. Finally, although gelatin is considered a temporary embolic, the authors confirm the patency of the contralateral SVA prior to the protective occlusion of the ipsilateral SVA.

2. PAE via collaterals. Descending SVA branches with rich anastomoses to the PA may occasionally be encountered. Owing to their usually obtuse angle of origin from the SVA, superselective catheterization and

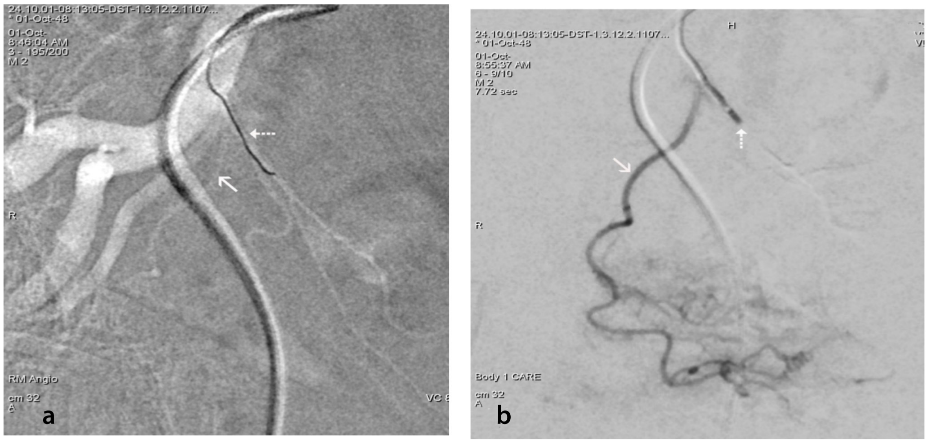


Figure 1. Protective superior vesical artery (SVA) embolization. **(a)** Fluoroscopic image with the "roadmap" technique shows advancement of the microguidewire (MGW) into the SVA (dotted arrow). Redirection of the MGW to the prostatic artery (PA) (arrow) was impossible. **(b)** Angiographic image after protective embolization and proximal occlusion of the SVA (dotted arrow) shows good opacification of the PA (arrow) and of the right hemiprostate. Embolization was safely performed from this position of the microcatheter.

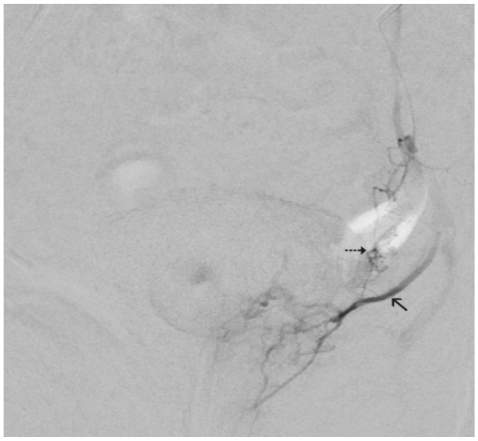


Figure 2. Embolization through collaterals to the prostatic artery (PA). Angiographic image shows the tip of the microcatheter (MC) (dotted arrow) in the distal part of a branch of the superior vesical artery. Contrast injection in this branch opacifies the left hemiprostate and, retrogradely, the left PA (arrow). No significant perfusion of the bladder wall is appreciated. Embolization was safely performed from this position of the MC.

the distal advancement of the MC into these branches is much easier than catheterization of the PA. If a slow manual contrast injection in these branches reveals substantial prostatic opacification and negligible perfusion of the bladder wall, PAE can be safely performed through these descending SVA branches (Figure 2).

3. Embolization from a proximal position of the MC near the PA orifice that exploits preferential flow to the PA. This can be applied when all previous options have failed and when the tip of the steerable MC can only reach the orifice of the PA without engaging it. In the current work, with appropriate rotation of the MC and/or adjustments of the angle of its tip, the latter was positioned against the PA orifice. Angiograms were acquired during slow manual contrast injections to ensure that only the PA was opacified (Figure 3). Embolization was then slowly performed and stopped when a sub-stasis of flow was observed in the PA.

Of a total of 157 patients with benign prostatic hyperplasia (314 pelvic sides) treated with PAE in the centers of the authors during the last 3 years (Table 1), a common vesicoprostatic trunk was observed in 101 pelvic sides (32.2%). In cases with a double or triple PA per pelvic side, only the most prominent PA was registered. All patients were informed in detail of both the standard and the adjunctive PAE techniques and provided written informed consent prior to the procedure. Of the 101 cases with vesicoprostatic trunks, PAE was accomplished with a standard MC in 59 cases, PA catheterization required additional utilization of steerable MC with no adjunctive technique in 23 cases, and a steerable MC and adjunctive techniques were eventually employed in 12 cases (Table 2). In the remaining 7 cases, utilization of one of the aforementioned approaches either failed or adjunctive techniques were contraindicated, and the patients underwent unilateral PAE. All 35 patients who were successfully treated with a steerable MC, with or without adjunctive techniques, underwent bilateral PAE, and the clinical success rate 1 year post-PAE was 87%. No major complications were observed. Minor complications were observed in 6 of the 35 patients. The technical success rate within the entire cohort of 157 patients was 96.8% (bilateral PAE in 128 patients, unilateral PAE in 24 patients, and technical failure/no PAE in 5 patients). The clinical success rate 1 year post-PAE was 83.7%, and complications (minor only) were encountered in 25 of the 157 patients (Supplementary Figure 3, Supplementary Table 1). Finally, imaging from and the clinical outcomes of the 35 patients who were treated with the steerable MC, with or without adjunctive techniques, were comparable with a previous series from the same centers within which only conventional catheterization techniques were applied.[6]

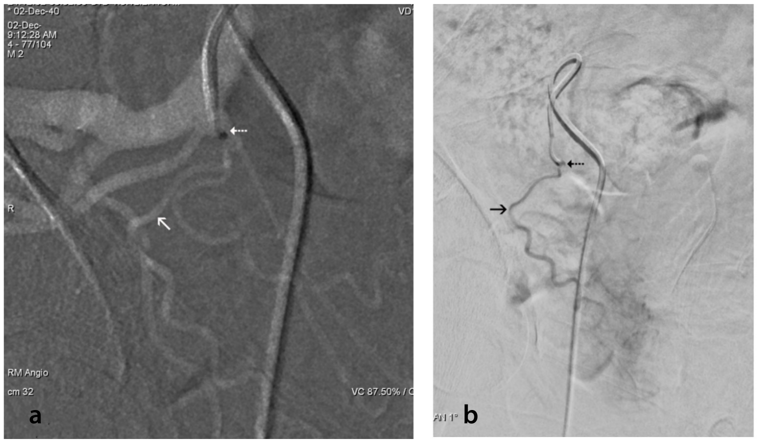


Figure 3. Exploitation of preferential flow to the prostatic artery (PA). **(a)** Fluoroscopic image with the "roadmap" technique shows the tip of the steerable microcatheter (MC) (dotted arrow) in the vesicoprostatic trunk. The PA (arrow) has an acutely angled and tortuous origin from the trunk, and superselective catheterization of the PA was impossible. **(b)** Angiographic image after adjusting the tip of the MC to face the PA ostium (dotted arrow) shows good opacification of the PA (arrow) and of the right hemiprostate and no reflux in the superior vesical artery or in the anterior division of the internal iliac. Embolization was safely performed from this position of the MC.

Table 1. Demographic and clinical features of the patients

Variable	Value for all patients with PAE treated in the two centers (n = 157)	Value for the subgroup treated with steerable MC ± adjunctive techniques (n = 35)
Age (y; mean ± SD)	71.2 ± 10.3	73.1 ± 11.2
BMI (mean ± SD)	26.6 ± 2.6	27.1 ± 3.1
PV (mL; mean ± SD)	87.1 ± 48.2	77.4 ± 41.2
Indication for PAE (proportion of patients)		
Moderate LUTS	54/157	11/35
Severe LUTS	57/157	14/35
Indwelling bladder catheter	40/157	9/35
Hemorrhage of prostatic origin	6/157	1/35

PAE, prostatic artery embolization; MC, microcatheter; SD, standard deviation; BMI, body mass index; PV, prostate volume; LUTS, lower urinary tract symptoms; y, years.

Table 2. Additional data about the adjunctive techniques for PAE

Adjunctive technique	Number of patients	Embolic material	Percentage of prostatic infarction of the treated lobe*	Clinical success 1 year post-PAE (proportion of patients)
Protective SVA embolization	6	Embosphere (100–300 μm)	5%–55%	6/6
PA embolization via collaterals	3	Embosphere (100–300 μm)	31%–39%	3/3
Proximal PA embolization and exploitation of preferential flow to the PA	3	Embosphere (300–500 μm)	0%–26%	2/3

*Percentage of prostatic infarction = the volume of infarcts in the treated lobe/the total volume of the treated lobe (infarcts were evaluated with contrast-enhanced ultrasound). PAE, prostatic artery embolization; SVA, superior vesical artery; PA, prostatic artery.

Discussion

According to a practical and widely accepted approach,[1] PA origination can be angiographically classified into five types. In type 1, the PA and the SVA share a common origin (trunk) from the anterior division of the internal iliac artery. In type 2, the PA originates from the anterior division of the internal iliac artery (separately from the SVA). Type 3 describes PA origination from the obturator artery, while type 4 indicates origination from the internal pudendal artery. Finally, type 5 includes rare PA origins, such as origination from the accessory pudendal or aberrant obturator artery. Among other vasculo-anatomical factors, PA origination type affects the technical outcome of PAE, with type 1 most often associated with difficult or failed PA catheterizations.

Compared with standard MCs, the utilization of a steerable MC appears to significantly increase the chances of the successful catheterization of type 1 PA originations—particularly in cases of short, cranially oriented vesicoprostatic trunks with an acute angle of PA origination[2-5]—by up to approximately 35%, according to the experience presented herein. The steerable MC can also address additional challenges that often coexist with type I PA origin, such as a tortuous and ectatic anterior division of the internal iliac artery or a base catheter facing posterolaterally instead of anteromedially.[7] When preinterventional computed tomographic angiography reveals such a challenging anatomy, it may be more practical to begin with a steerable rather than a standard MC; however, the financial aspects of this approach should be further investigated.

With the adjunctive techniques described herein, PAE can be performed even when superselective catheterization of the PA is impossible. The following technique-specific comments can also be made: 1) protective coiling of the SVA has been described before[7] and, despite its permanent nature, is considered a safe technique. However, since protective occlusion of the SVA is only needed during the few minutes of the injection of microspheres into the PA, the authors prefer a temporary embolic agent for the protection of the SVA, with the potential for complete SVA recanalization in the following days or weeks. Moreover, uniformly cut gelatin particles with a standardized hydrated size probably ensure a more controlled and precise occlusion compared with the gelfoam slurry prepared by the operator;[6] 2) collaterals between vesical and prostatic arteries are not uncommon,[8] but they can only rarely serve as pathways for safe and effective PAE. Distal advancement of the MC in the collaterals and slow, controlled manual contrast injections[6] are required to confirm abundant flow to the prostate and the absence of bladder wall opacification; and, finally 3) it should be acknowledged that the exploitation of preferential flow to the PA is a suboptimal PAE technique that should be applied only when previous options have failed. Relatively larger microspheres with a diameter of 300–500 microns (rather than 100–300 microns) are preferred for this last option to minimize the risk of ischemic complications in the case of reflux to the SVA or to the more distal branches of the anterior division of the internal iliac artery.

Other options to address the aforementioned difficulties are either not widely available—such as utilization of a robotic c atheter[7]—or are more complex and invasive—such as the combination of a larger sheath and a "buddy wire".[5] The adjunctive techniques presented herein appear to be simpler, more widely available, and affordable, as they can be applied not only with steerable but also with standard MCs.

In conclusion, familiarization with the application of a steerable MC and with the adjunctive techniques described herein may substantially improve the technical outcome of PAE in cases of anatomically challenging vesicoprostatic trunks.

Acknowledgements

The authors are grateful to Mr. Konstantinos Chouchos, MSc, and Mrs Maria-Theodora Moschouri (medical student) for their valuable assistance in the preparation of this manuscript.

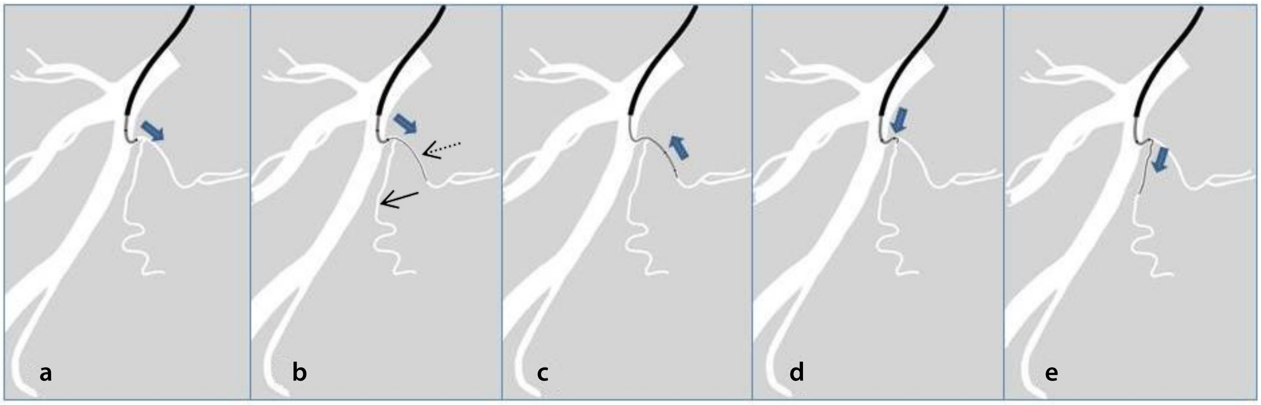
Footnotes

Conflict of interest disclosure

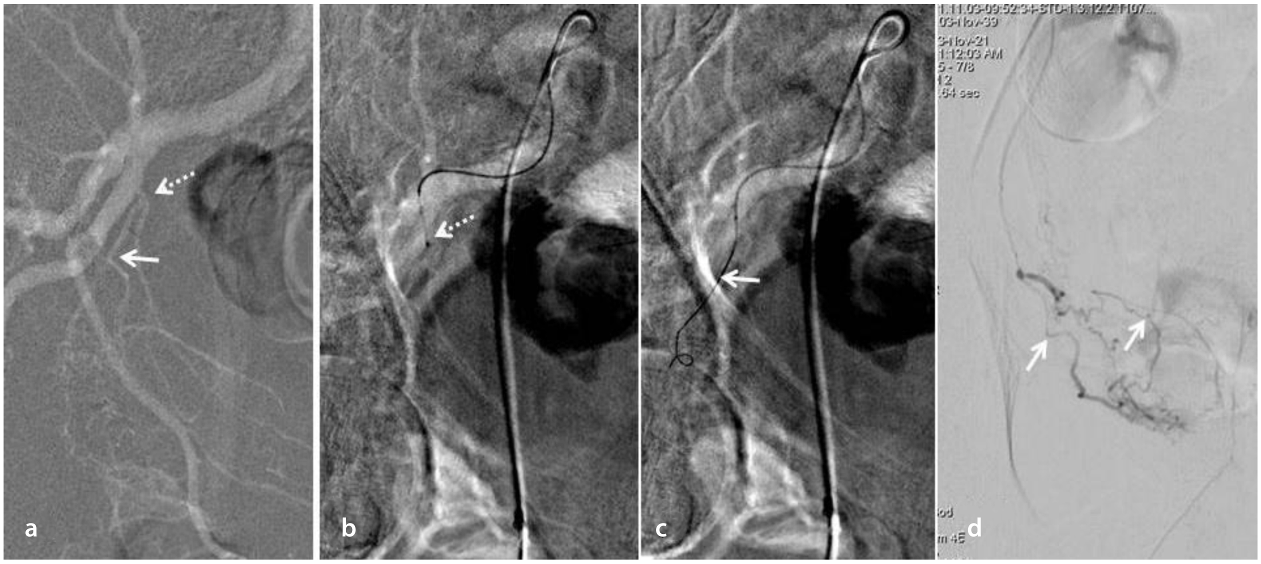
The authors declared no conflicts of interest.

References

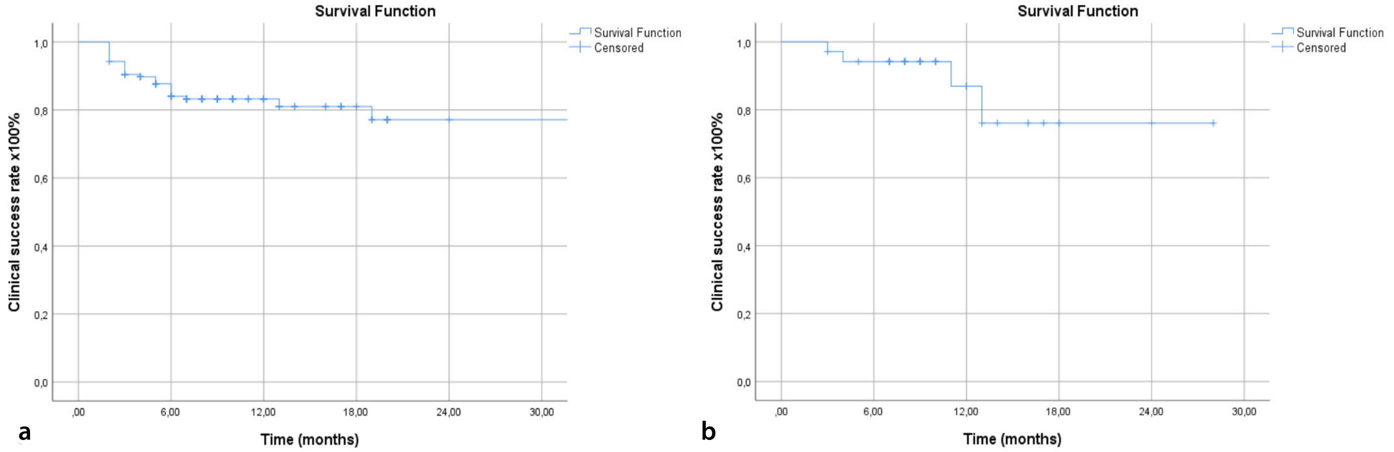
1. de Assis AM, Moreira AM, de Paula Rodrigues VC, et al. Pelvic arterial anatomy relevant to prostatic artery embolisation and proposal for angiographic classification. *Cardiovasc Intervent Radiol*. 2015;38(4):855-861. [Crossref]
2. Moschouris H, Stamatiou K, Tzamarias S, et al. Angiographic imaging of prostatic artery origin in a greek population and correlation with technical and clinical aspects of prostatic artery embolization. *Cureus*. 2023;15(9):e45941. [Crossref]
3. Boeken T, Gautier A, Moussa N, et al. Impact of anatomy type of prostatic artery on the number of catheters needed for prostatic artery embolization. *Diagn Interv Imaging*. 2021;102(3):147-152. [Crossref]
4. Hoffmann JC, Minkin J, Primiano N, Yun J, Eweka A. Use of a steerable microcatheter during superselective angiography: impact on radiation exposure and procedural efficiency. *CVIR Endovasc*. 2019;2(1):35. [Crossref]
5. Mouli S, Hohlastos E, Salem R. Prostate artery embolization. *Semin Intervent Radiol*. 2019;36(2):142-148. [Crossref]
6. Moschouris H, Stamatiou K, Malagari K, et al. The value of contrast-enhanced ultrasonography in detection of prostatic infarction after prostatic artery embolization for the treatment of symptomatic benign prostatic hyperplasia. *Diagn Interv Radiol*. 2019;25(2):134-143. [Crossref]
7. Bagla S, Isaacson AJ. Tips and tricks for difficult prostatic artery embolization. *Semin Intervent Radiol*. 2016;33(3):236-239. [Crossref]
8. Richardson AJ, Acharya V, Kably I, Bhatia S. Prostatic artery embolization: Variant origins and collaterals. *Tech Vasc Interv Radiol*. 2020;23(3):100690. [Crossref]



Supplementary Figure 1. Schematic drawing showing usual maneuvers for prostatic artery (PA) catheterization in challenging case of vesicoprostatic (VP) trunk. **(a)** The tip of the steerable microcatheter (MC) is manually angulated towards the orifice of the VP trunk and locked. **(b)** The microguidewire (MGW) is inserted; more often than not, it advances into the superior vesical artery (black dotted arrow) instead of the PA (black arrow). **(c)** The tip of the MC is unlocked and the latter is inserted into the superior vesical artery over the MGW. **(d)** The MC and MGW are slowly withdrawn close to the origin of the PA; the latter is catheterized with appropriate rotation of the MGW, and, if required, additional angulation and locking of the MC. **(e)** The MGW is advanced into the PA. All drawings correspond to ipsilateral oblique angiographic projections. Blue arrow indicates the direction of advancement (or withdrawal) of MC and MGW.



Supplementary Figure 2. Representative case of utilization of steerable microcatheter (MC) in the context of challenging vesicoprostatic (VP) trunk. Ipsilateral oblique roadmap image **(a)**, shows a short VP trunk (dotted arrow) which originates at 90° angle from the anterior division of the internal iliac artery. The angle of prostatic artery (PA, arrow) origin from the VP trunk is also unfavorable (less than 90°). Similar projections during attempt of catheterization with steerable MC **(b, c)**, show the tip of the MC (dotted arrow, **b**) angulated towards the orifice of the VP trunk. Despite suboptimal support from the base catheter, the microguidewire (arrow, **c**) can be directed into the PA. Anteroposterior angiogram **(d)** after distal advancement of the MC into the PA shows distal prostatic branches (arrows).



Supplementary Figure 3. Kaplan-Meier curves showing the clinical success rates for the entire patient cohort (n = 157) who unterwent PAE in the 2 centers of the authors during the last 3 years **(a)**, and for the subgroup of patients (n = 35) who underwent PAE with steerable MC ± adjunctive techniques **(b)**. PAE, prostatic artery embolization; MC, microcatheter.

Supplementary Table 1. Complications encountered in the patients of this work

Complication*	Number of patients (%)	
	For the entire patient cohort (n = 157)	For the subgroup treated with steerable MC ± adjunctive techniques (n = 35)
Haematospermia	3 (1.9)	-
Haematuria	2 (1.3)	1 (2.8)
Penile ulcers (small, ischemic)	2 (1.3)	-
Acute urinary retention	9 (5.7)	4 (11.5)
Prostatic tissue expulsion	1 (0.6)	-
Rectal bleeding	2 (1.3)	-
Urinary tract infection	2 (1.3)	-
Inguinal haematoma	4 (2.5)	1(2.8)
Total	25 (15.9)	6 (17.1)

*All complications were considered minor, because they required no hospitalization and were self-limiting, or resolved with conservative treatment. MC, microcatheter.

Supplementary Video 1 link: https://youtu.be/JSOz3imIncg

Supplementary Video 1. Despite suboptimal support from and unfavorable rotation of the base catheter, the steerable MC can be appropriately angled and directed into the vesicoprostatic trunk without a MGW. MC, microcatheter, MGW, microguidewire.

Supplementary Video 2 link: https://youtu.be/gU1DBLcCZOk

Supplementary Video 2. Despite unfavorable rotation of the base catheter, appropriate angulation of the steerable MC can help the operator to direct the MGW into the vesicoprostatic trunk and then into the PA. MC, microcatheter, MGW, microguidewire, PA, prostatic artery.

DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.243005

INTERVENTIONAL RADIOLOGY

LETTER TO THE EDITOR

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Proposal for training: the educational value of a musculoskeletal embolization patellar tendinopathy model

Emeric Gremen[1,2]
Julien Ghelfi[1,2]
Marylène Bacle[3]
Julien Frandon[4,5]

[1]Grenoble-Alpes University Faculty of Medicine, Department of Radiology, Grenoble, France

[2]Grenoble Alpes University Hospital, Department of Radiology, Grenoble, France

[3]Montpellier Nîmes University Faculty of Medicine, Department of Radiology, Nîmes, France

[4]Nîmes University Hospital, Department of Medical Imaging, Nîmes, France

[5]University of Montpellier, Medical Imaging Group Nîmes, Nîmes, France

KEYWORDS
Animal model, embolization, MSK, musculoskeletal, neovessels, training

Corresponding author: Emeric Gremen

E-mail: egremen@chu-grenoble.fr

Received 29 August 2024; revision requested 29 September 2024; last revision received 17 October 2024; accepted 31 October 2024.

Epub: 09.12.2024

Publication date: 08.09.2025

DOI: DOI: 10.4274/dir.2024.243005

Dear Editor,

Model training presents significant advantages for training purposes in several key areas of musculoskeletal embolization. The model allows trainees to gain practical experience with the procedures, techniques, and equipment used in real-world scenarios. This hands-on practice is invaluable for building confidence and competence to find and treat neovessels. As we have seen recently,[1] detecting neovessels, which is a direct marker of the technical success of this treatment, is a key but challenging step in performing embolization and must be mastered to avoid inappropriate patient management.

The patellar tendinopathy pig animal model[2] provides a realistic environment for trainees to learn how to diagnose and treat neovascularization. This is particularly important for understanding the complexities and nuances of different degrees of neovessel visualization. In this study, we improved our understanding of the model and developed a classification for the evaluation of these neovessels during pre- and post-embolization angiographies. A Likert scale was used to evaluate the neovessels in our model (grade 0: no neovessels; grade 1: slight tumor blush equivalent to muscle enhancement; grade 2: marked tumor blush greater than muscle enhancement; grade 3: very marked tumor blush). Additionally, we examined associated vascular anomalies (arterial anastomosis, early venous return). These observations are detailed and illustrated in the portfolio (Figures 1 and 2) for the reader's reference. Based on this classification, this model allows for an expert and reliable angiographic evaluation of the effectiveness of the embolization agent used by comparing the neovessels before and after embolization.

During the past year, from 2023 to 2024, 24 angiographies were conducted on our pig patellar tendinopathy model, involving 12 pigs and 24 tendons. Tendinopathy induction involved injecting a total of 50 mg of type 1 collagenase per tendon under ultrasound guidance, using two syringes, each containing 25 mg at a concentration of 25 mg/mL. The injections were performed using 1-mL insulin syringes (G25 with attached needles) for precise and controlled delivery. On all the angiographies performed at D7, the presence of neovessels was noted. Across all the series performed, 5 had grade 1 neovessels, 8 had grade 2, and 11 had grade 3. Moreover, six had multiple arterial anastomoses and seven had early venous return (often in association). For reference, these described vascular anomalies can be associated with any grade of neovessels, sometimes masking their detection. This highlights the value of this model for training in identifying neovessels and then practice in treating them.

One of the key benefits of using an animal model lies in the tactile feedback it offers, which is crucial for learning how to manage resistance and understand the importance of quality injections, as we previously described,[1] in musculoskeletal embolization. This tactile feedback also aids in assessing reflux and making real-time adjustments, skills that are challenging to replicate in purely virtual or silicone-based environments. This hands-on experience may help to develop not only theoretical knowledge but also essential technical competence.

Although simulators represent the future of medical training, and some affordable options do exist, those specifically designed for high-fidelity embolization procedures remain expensive. Moreover, one of the issues with current simulators is their lack of fidelity to the

You may cite this article as: Gremen E, Ghelfi J, Bacle M, Frandon J. Proposal for training: the educational value of a musculoskeletal embolization patellar tendinopathy model. *Diagn Interv Radiol.* 2025;31(5):502-504.

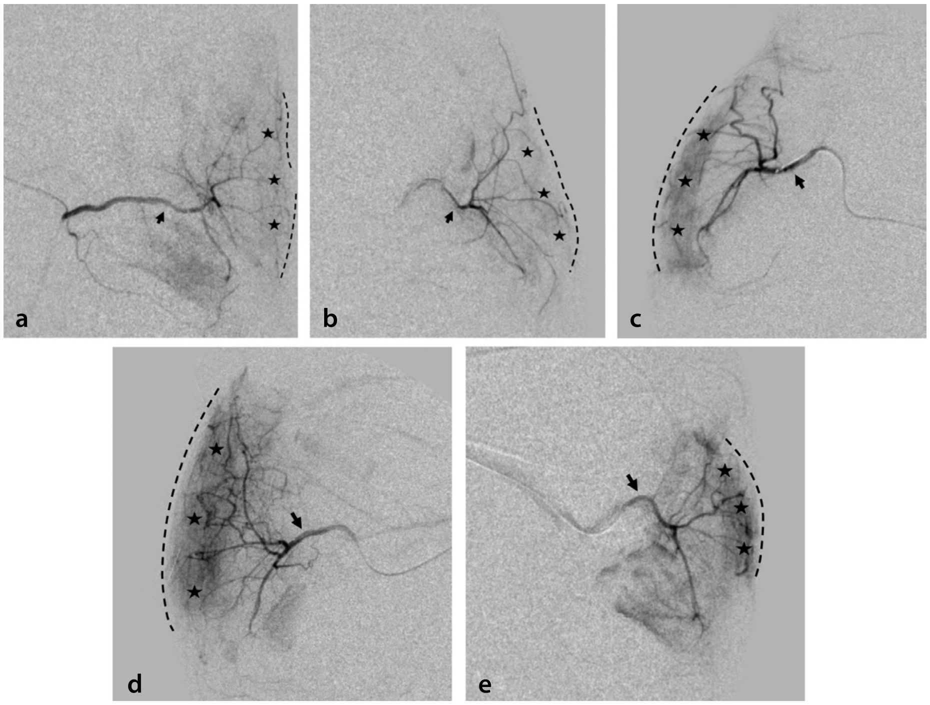


Figure 1. Evaluation of neovessels according to a Likert scale, grade 1 to 3. The dotted lines represent the anterior surface of the inflamed patellar tendon in these profile views from arteriographies performed 7 days after tendinopathy induction. The black arrows represent the genicular artery. Pictures **(a)** and **(b)** show grade 1 neovessels (black stars), which correspond to a slight tumoral blush equivalent to muscle enhancement. It is important to note the difference in enhancement compared with grade 2 **(c)** and grade 3 **(d, e)** neovessels, which exhibit significantly more pronounced enhancement.

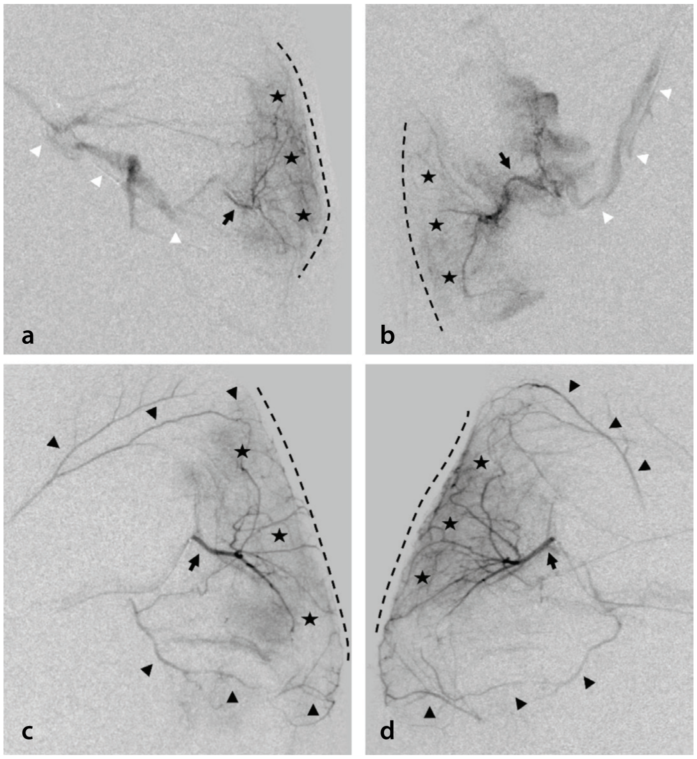


Figure 2. Evaluation of vascular anomalies associated with neovessels. Under the same conditions, images **a** and **b** represent an early venous return (white arrowheads). This early venous return can sometimes mask the appearance of neovessels, as seen in image **b** (grade 1) compared with image **a** (grade 2). Early venous return results in contrast agent leakage, making it difficult to adequately impregnate the arterial vascular network during injection. Images **c** (grade 1) and **d** (grade 2) illustrate arterial collaterals (black arrowheads), which are also important to identify. If not occluded, they continue to supply the neovessel bed, leading to treatment failure.

actual procedure, which limits their effectiveness. This makes purchasing a simulator a significant financial investment that may not deliver the expected training outcomes, further exacerbating the financial barriers for many programs.[3] By contrast, although using an animal model also incurs significant costs, it offers a highly realistic, high-fidelity experience that closely mimics human procedures, making it an invaluable tool for training.

Nevertheless, implementing an animal model from scratch involves considerable costs. For example, acquiring a refurbished C-arm capable of fluoroscopy and digital subtraction angiography for use with animals costs approximately €60,000. Additional expenses for model induction include ultrasound equipment (approximately €10,000) and needles for intrapatellar injections. Supplies for embolization, which are single use for each pig (introducer, guidewire, 4Fr catheters, 2.0Fr microcatheters, embolization particles), cost approximately €2,000 per procedure. The cost of housing for the animals for 1 week is approximately €1,000 per pig, including pharmacy costs (particularly for anesthesia coverage) and the purchase of operating room supplies (compresses, sterile drapes), which are approximately €300 per pig. Radiation shielding for the facility could add another €100,000. Although these costs may seem high, they are indicative and vary depending on the existing infrastructure at each institution.

However, the tactile feedback and intra-procedural experience provided by the animal model remain unparalleled, offering a training fidelity that current simulators cannot achieve. This model allows trainees to practice in a realistic environment, gaining practical skills in embolization that are directly translatable to clinical practice. As such, despite the costs, the animal model remains, until today, an invaluable tool for advanced training in musculoskeletal embolization.

Additionally, this animal model facilitates the continuous creation of new e-learning content (anatomical education, the classification of neovessels, and the inclusion of video recordings of angiograms) and can be employed for various educational purposes, such as anatomical dissections or managing hemorrhages in surgical training just before the sacrifice of the animal. Furthermore, an additional comparative study between traditional methods and this animal model could be useful to more objectively assess its educational effectiveness.

Standardized training is essential for maintaining high medical standards and ensuring the best possible outcomes for patients. We believe that the combination of theoretical e-learning with practical sessions on the animal model provides a balanced solution for acquiring in-depth knowledge and essential technical skills. This blended approach not only enhances the learning experience but also ensures the development and refinement of new techniques, ultimately leading to optimal patient outcomes. Overall, this model serves as an invaluable resource for medical training, providing a comprehensive and safe environment for learning in the field of musculoskeletal embolization.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This study was funded by a research scholarship from the “missions d’enseignement, de recherche, de référence et d’innovation” (MERRI) for manuscript writing.

References

1. Gremen E, Gremen E, Ghelfi J, Bacle M, Frandon J. Musculoskeletal embolization innovation: keys to highlighting neovessels and advanced perspectives. *Cardiovasc Intervent Radiol*. 2024;47(5):680-682. [Crossref]
2. Ghelfi J, Bacle M, Stephanov O, et al. Collagenase-induced patellar tendinopathy with neovascularization: first results towards a piglet model of musculoskeletal embolization. *Biomedicines*. 2021;10(1):2. [Crossref]
3. Mandal I, Ojha U. Training in interventional radiology: a simulation-based approach. *J Med Educ Curric Dev*. 2020;7:2382120520912744. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242926

MODALITY-BASED (UC, CT, MRI, PET-CT) IMAGING

INVITED REVIEW

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

New imaging techniques and trends in radiology

Mecit Kantarcı[1]
Sonay Aydın[2]
Hayri Oğul[3]
Volkan Kızılgöz[2]

[1]Atatürk University Faculty of Medicine, Department of Radiology, Erzurum, Türkiye

[2]Erzincan Binali Yıldırım University Faculty of Medicine, Department of Radiology, Erzincan, Türkiye

[3]Medipol University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

ABSTRACT

Radiography is a field of medicine inherently intertwined with technology. The dependency on technology is very high for obtaining images in ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI). Although the reduction in radiation dose is not applicable in US and MRI, advancements in technology have made it possible in CT, with ongoing studies aimed at further optimization. The resolution and diagnostic quality of images obtained through advancements in each modality are steadily improving. Additionally, technological progress has significantly shortened acquisition times for CT and MRI. The use of artificial intelligence (AI), which is becoming increasingly widespread worldwide, has also been incorporated into radiography. This technology can produce more accurate and reproducible results in US examinations. Machine learning offers great potential for improving image quality, creating more distinct and useful images, and even developing new US imaging modalities. Furthermore, AI technologies are increasingly prevalent in CT and MRI for image evaluation, image generation, and enhanced image quality.

KEYWORDS

Arthrographic applications, cerebrospinal fluid flowmetry, imaging techniques, magnetic resonance spectroscopy, magnetic resonance imaging techniques

Medical imaging is the process of generating visual representations of the body's tissues and organs to examine their structure and function for clinical and scientific purposes. These techniques allow the evaluation of internal structures beneath the skin and bones, facilitating the diagnosis of abnormalities and the treatment of diseases. Medical imaging has become an essential component of healthcare, research, and biological imaging.[1,2]

Imaging technologies play a critical role in diagnosing abnormalities and supporting therapy, providing medical personnel with essential information about their patients' conditions. Techniques such as electroencephalography (EEG), magnetoencephalography (MEG), and electrocardiography capture and quantify data rather than generate visuals. They present information as parameter graphs over time or maps with varying levels of detail. Although these technologies have limitations, they can be considered a form of medical imaging on a smaller scale. By 2010, more than 5 billion medical imaging studies had been completed worldwide.[3]

Medical imaging accounts for approximately 50% of the overall ionizing radiation exposure in the United States. These technologies are crucial for the diagnosis, management, treatment, and prevention of various disorders. Imaging techniques are now essential for diagnosing nearly all major medical conditions, including trauma, malignancies, cardiovascular diseases, neurological disorders, and numerous other health issues. These techniques are operated by highly skilled technicians and medical specialists, such as oncologists and internists.[1]

Medical imaging technologies are predominantly used for medical diagnostics. Diagnosis refers to the systematic identification of a patient's condition and associated symptoms. The process involves gathering data from the patient's medical history, physical examinations, or

Corresponding author: Mecit Kantarcı

E-mail: akkanrad@hotmail.com

Received 28 September 2024; revision requested 18 November 2024; accepted 04 December 2024.

Epub: 16.01.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2024.242926

You may cite this article as: Kantarcı M, Aydın S, Oğul H, Kızılgöz V. New imaging techniques and trends in radiology. *Diagn Interv Radiol.* 2025;31(5):505-517.

questionnaires to determine the appropriate course of treatment. However, diagnosis can be challenging, as many indications and symptoms are non-specific in nature. For example, the presence of erythema, which is characterized by redness of the skin, may indicate a variety of disorders. Therefore, distinct diagnostic methods are required to identify the etiology of diseases and determine appropriate treatment or preventive measures.[4,5]

Historically, ancient physicians made medical diagnoses based on visual and auditory observations, occasionally supplemented by the examination of human specimens. For example, techniques such as examining bodily fluids, including urine and saliva, were commonly practiced before 400 B.C. In ancient Egypt and Mesopotamia, physicians were capable of diagnosing conditions related to the gastrointestinal and cardiovascular systems, cardiac rhythm, spleen, liver, and menstrual disorders. However, medical interventions were primarily limited to affluent and noble individuals.[5]

Hippocrates, who lived around 300 B.C., advocated the use of the mind and senses as diagnostic tools, earning him the title of the "Father of Medicine." He promoted a diagnostic process that included urine testing, skin color observation, and the examination of the lungs and other external indicators. He also observed the correlation between illness and heredity. Abu al-Qasim al-Zahrawi, an Arabic physician of the Islamic era, documented the first recorded instance of a hereditary genetic disorder, now known as hemophilia. He provided a detailed account of a family in Andalusia in which the males were affected by this condition.[5]

Main points

- Computed tomography (CT) scans will continue to be an essential part of contemporary medical diagnostics thanks to developments in resolution, velocity, radiation dose reduction, artificial intelligence (AI) integration, and personalized treatment. The development of portable CT scanners and the use of functional and multimodal imaging will enhance this technology's potential.
- Advancements in magnetic resonance imaging (MRI) systems are meant to improve accessibility, shorten scan times, and produce better-quality images in areas where MRI has historically had difficulties.
- AI technology can produce results from ultrasound (US) exams that are more accurate and consistent. The application of machine learning to US imaging has great potential to improve image quality, produce more unique and useful images, and possibly introduce new US imaging methods.
- Across all modalities, AI technologies are increasingly being used and showing an increased trajectory in both image production and evaluation.

During the Middle Ages, physicians employed various methods to determine the origins of physical imbalances. Uroscopy, the predominant technique, involved collecting the patient's urine in a specialized container called a "matula" and analyzing its color, odor, density, and the presence of precipitates. Physicians also examined the consistency and color of blood to distinguish between chronic and acute conditions. Pulse rate, strength, and rhythm were evaluated through palpation. Additionally, medical practices during this period often incorporated the interpretation of zodiac signs.[6]

In the 19th century, the introduction of diagnostic equipment such as X-rays and microscopes brought about a significant transformation in the field of diagnosing and treating disorders. In the early part of the century, doctors predominantly diagnosed diseases by analyzing symptoms and indications. During the 1850s, the use of instruments such as ophthalmoscopes, stethoscopes, and laryngoscopes enhanced doctors' sensory capabilities, leading to the development of novel diagnostic methods and approaches. During this era, a variety of diagnostic techniques were developed, including chemical testing, bacteriological tests, microscopic examinations, X-rays, and several other medical tests.[5,6]

The development of X-rays marked substantial advancements in medical imaging procedures. Wilhelm Conrad Roentgen discovered X-rays in November 1895, a discovery that earned him the Nobel Prize in 1901. Initially, radiologists used the term "plane film" to describe X-rays, employing them to diagnose bone fractures and chest abnormalities. Fluoroscopy, with its enhanced X-ray beam, facilitated the detection of a wide range of patient issues. In the 1920s, radiologists began using these procedures to diagnose disorders such as esophageal cancer, ulcers, and stomach conditions. Fluoroscopy ultimately evolved into computed tomography (CT).[7]

Numerous advanced imaging techniques have been developed, each with its principles, applications in medical labs, and advancements over time. The following techniques are essential for understanding their benefits and uses in diagnosing, managing, and treating various diseases, including cardiovascular conditions, cancer, neurological disorders, and trauma.

Advancements across modalities

1. Computed tomography

CT uses X-rays to generate highly detailed cross-sectional images of the body. The high-resolution imaging of tissues and organs aids in diagnosing internal injuries, cancers, and other illnesses. Hounsfield developed the first iteration of a CT scanner in the 1960s. CT, commonly known as X-ray CT, was first implemented in 1971 at Atkinson Morley Hospital in Wimbledon (now part of St George's Hospital). Sir Godfrey Hounsfield performed this pioneering brain scan under the guidance of Jamie Ambrose, MD, an expert neuroradiologist. The objective of the scan was to investigate less painful alternatives to existing methods of brain examination. CT technology has undergone significant developments since its introduction in the 1970s. These advancements have revolutionized the field of diagnosis and treatment planning by using X-rays and advanced algorithms to produce highly detailed cross-sectional images. The scanner designs used for image formation in CT are called generations. New generations have emerged with different arrangements of components and mechanical movements required for data collection. The main differences between CT generations relate to the number and arrangement of detectors, the shape of the X-ray beam, and the rotation of the tube and detectors. Based on a recent analysis by Mordor Intelligence, the CT market is projected to experience significant growth, with its value expected to rise from $8.14 billion in 2023 to $10.95 billion by 2028. This growth is anticipated to occur at a compound annual growth rate of 6.12%.[8]

The introduction of dual-energy CT (DECT) technology marked a substantial departure from traditional methods and paved the way for contemporary advancements in CT technology. DECT is a well-established technology with a significant and extensive background. Sir Godfrey Hounsfield devised a technique in the 1970s to differentiate calcium from iodine by using two distinct energy spectra from X-ray photons. This method relies on understanding the specific atomic numbers and unique K-edge characteristics of various substances. These features are essential for discerning the differing impacts of Compton scattering and the photoelectric effect in X-ray attenuation.[9]

In the early 1980s, DECT technology was primarily used for bone densitometry, as demonstrated by devices such as the Somatom DR manufactured by Siemens Healthcare. This device employed rapid tube potential switching to acquire two-photon energy spectra. Because of the limited computing capabilities of the hardware available at the time, DECT was mainly used for densitometry purposes. However, the 21st century witnessed notable progress in the therapeutic use of DECT, driven by rapid advancements in processing capabilities. During this period, scanners with dual sources, such as the Somatom Definition DS in 2006 and the Somatom Definition Flash in 2009, were introduced. Additionally, multilayer detectors, such as the Brilliance-64 in 2015, were also introduced. In 2010, General Electric Healthcare improved the technique of rapid tube potential switching with models such as the Revolution GSI and Discovery 750 HD.[10]

DECT enables the capture and reconstruction of a wide range of images. The kilovolt peak (kVp) images generated by DECT closely resemble those obtained from single-energy CT, as they replicate the characteristics of a single-energy spectrum. These images can be acquired using dual-layer, rapid kVp-switching, and split-filter DECT techniques. Dual-source DECT generates images by using a pair of kVp values or kVp-equivalent images, which are calculated by combining data from two distinct peaks using a weighted average. As a result, the reconstructions resemble images obtained using a single, user-selected kVp value. Virtual monoenergetic imaging replicates scans using photons at a specific energy level, which is advantageous due to the increased iodine attenuation at lower photon energy levels. In addition, material decomposition techniques exploit the different effects of Compton scattering and the photoelectric effect on X-ray attenuation. This allows for the production of images with enhanced or reduced iodine visibility and the exclusion of urine or calcium.[9,11]

Current studies on the cost-effectiveness of DECT reveal partially conflicting results for different areas of use. Although its use for incidental renal lesions and in the emergency department reduce costs, it is noted that the costs of cardiovascular system imaging sometimes increase. From this perspective, detailed studies on more specific usage areas are needed to determine the cost-effectiveness of DECT.[12-14]

Table 1 provides a summary of the areas in which DECT is used substantially more frequently. Recent studies suggest that it may also be beneficial in the evaluation of pulmonary perfusion, myocarditis, and the diagnosis of alveolar echinococcosis. These applications are in addition to those mentioned above.[15-18]

After August 2022, the Food and Drug Administration (FDA) approved two biomedical imaging technologies, developed in collaboration with the National Institute of Biomedical Imaging and Bioengineering (NIBIB), to be used in clinical settings. Both methods offer improvements in CT. Dr. Cynthia McCollough, the project lead and director of Mayo Clinic's CT Clinical Innovation Center, and her team have made a significant advancement by developing the first photon-counting detector (PCD)-CT system. This new system outperforms current CT technology and was described as the first major imaging advancement cleared by the FDA for CT in a decade.

Photon-counting CT (PCCT) is an advanced technological development in the field of energy-resolving, direct-conversion X-ray detectors. After 15 years of thorough study and development, this technique has recently been integrated into clinical CT equipment. The fundamental concepts of PCCT differ greatly from those of traditional CT detectors. The detectors used in traditional CT are known as energy-integrating detectors. These provide signals that are directly proportional to the total energy of photons received within a specific measurement interval. PCCT, however, uses PCDs to directly convert the energy of individual photons into electrical impulses. The device exclusively emits electrical pulses with heights exceeding the thresholds indicative of noise.[5] Therefore, this technology enables a significant reduction in electrical noise levels and an increase in the signal-to-noise ratio (SNR). Furthermore, it can also be utilized in dual-energy imaging. The advent of PCCT has the potential to transform the clinical CT field by leveraging its multiple inherent advantages and overcoming several constraints present in existing cutting-edge CT systems (Figure 1).[9] DECT requires specialized equipment and is limited to two energy levels. However, with the introduction of this novel detector, additional "buckets" are available to categorize X-ray energies, enhancing the ability to accurately represent material differences.

Table 1. DECT applications

Region	Material categorization/virtual monoenergetic beam	Quantification of iodone
Brain	Used to differentiate between tumors and bleeding	Used to distinguish between bleeding and contrast
Cardiac	Using low virtual monoenergetic KeV contributes to imaging myocardial fibrosis	
Lung		COVID-19 shows high iodine density around pulmonary opacity and increased perfusion in the lung parenchyma Reduced perfusion in the lung parenchyma within the area of pulmonary infarct indicates possible hypoperfused lung or pulmonary embolism.
Abdomen	Differentiates a hypoperfused segment of the bowel wall from one that is normally perfused Distinguishes between different types of tumors Aids in analyzing the composition of distinct kidney or gallstones	Iodine map imaging helps to better visualize iodine accumulation in the bowel wall, thereby improving diagnostic certainty for intramural hemorrhage
Vasculer imaging	Reduces the impact of blooming artifacts from calcified plaques	
Bones	VNC images can be used to distinguish chronic fractures from acute and non-displaced CT occult fractures	
Metallic artifacts	A high monoenergetic beam can help minimize metallic artifacts	

DECT, dual-energy computed tomography; COVIF-19, coronavirus disease-2019; VNC, virtual non-contrast images; CT, computed tomography.

Published in the Journal Radiology, clinical investigations have demonstrated that the new PCCT devices can effectively reduce noise by up to 47%. In addition, the new technique reduces the amount of contrast agent required for CT imaging. Due to the enhanced signal provided by the PCCT system, participants in the trial were able to achieve the same image quality as conventional CT systems using 30% less contrast agent. The PCCT systems offer superior spatial resolution compared with conventional systems, delivering the highest reported resolution for a clinical CT system.[19,20] Siemens developed a prototype PCD-CT system, and with financial support from McCollough through NIBIB, the team began scanning patients with approval from the Institutional Review Board. A total of 1,100 patients were examined in these tests, initially using a traditional CT system and subsequently with the advanced PCCT scanner, showcasing the benefits of the new technology. This device is the first product of its category available on the market.[20]

Another CT-based method approved by the FDA is artificial intelligence (AI)-assisted CT perfusion (CTP) imaging. An AI software was developed to assist in image reconstruction to reduce the elevated radiation dose in CTP. This software employs the K-space weighted image average technique to reduce noise in CTP images, resulting in lower radiation exposure for patients without compromising image processing quality or speed. Research has demonstrated that the software effectively decreases the radiation exposure of CTP by 50%–75% compared with the conventional CTP approach. Additional benefits of using this approach include no interruptions to the regular clinical workflow and no requirement for upgrades or modifications to existing CT hardware. The software has received FDA 510(k) clearance and is eligible for integration into clinical practice.[21]

The prospects for CT technology in the future are highly encouraging for both healthcare providers and patients. Advancements in resolution, velocity, radiation dose reduction, AI integration, and personalized medicine will ensure that CT scans remain a crucial component of modern medical diagnostics. The use of functional and multimodal imaging, along with the development of portable CT scanners, will further enhance the capabilities of this technology. In the future, the continuous progress of CT technology will lead to greater accuracy in diagnosis, improved treatment outcomes, and enhanced patient care.

Nowadays, thanks to advances in CT technology, especially cone beam and dual-source CT, arthrography is used in the diagnosis of many musculoskeletal pathologies. Compared with conventional magnetic resonance imaging (MRI) and MR arthrography, CT arthrography is superior in depicting chondral/osteochondral damage, loose bodies, chondral variations, and subarticular bone fractures.[22-27] Moreover, CT arthrography has excellent spatial resolution with multiplanar imaging capability and shorter examination times. Other indications for CT arthrography include patients with non-MRI-safe implantable devices or cardiac pacemakers and individuals with claustrophobia.[24,28]

Cone beam or flat-panel detector CT technology uses a cone-shaped X-ray beam and applies software programs with sophisticated algorithms, including back projection. Because of its perfect high spatial resolution, cone beam CT arthrography allows the optimal evaluation of cartilage and subchondral bone microarchitecture in the articular surface.[29,30] Recent studies have reported that cone beam CT scans can obtain images with very high resolution (75–300 μm slice thickness) with low-dose applications.[30-32] Lower radiation doses in cone beam CT technology are achieved through the smaller field of view, the use of a high-quality flat-panel detector system, and pulsed X-ray beams.[33]

In DECT, two different datasets are acquired at different voltage peak levels to separate materials based on tissue composition (e.g., urate mineralization, calcification, and iodine). This technique allows for the detection of gout tophi and the demonstration of bone marrow edema in vertebral compression fractures.[34,35] Recently, DECT has also been used to distinguish intra-articular iodinated contrast media from adjacent bone in CT arthrographic applications.[36-38]

2. Advancements in magnetic resonance imaging techniques and features

MRI was first implemented as a clinical diagnostic instrument in the early 1980s. Significant technological developments have occurred since its introduction. Advancements in various technical components, including data acquisition, image reconstruction, and hardware systems, have greatly impacted and propelled growth in other areas of MRI technology. The advancements in each component have generated new opportunities for growth in the others. Moreover, the swift integration of cutting-edge technologies derived from fundamental sciences and technical disciplines such as computer science, data processing, and semiconductors has resulted in revolutionary advancements in MRI technology (Figure 2).

Initial developments in MRI techniques focused on optimizing data acquisition protocols to achieve adequate spatial and temporal resolution, contrast, and imaging efficiency. An example of this delicate balance is the implementation of line reductions in the basic spin-echo protocol. However, this method has a drawback-the absence of frequency data, leading to a reduction in either image quality or image dimensions. Scanning time was reduced in methods such as the fast spin echo, echo planar imaging, or multi-echo approach by recording multiple lines following the radio frequency (RF) pulse. Despite the improvement, a major limitation was the rapid decline in signal intensity caused by energy transfer during T2 capture. This limitation allowed only 3–4 lines per RF pulse and led to a noticeable degradation in image quality.[39]

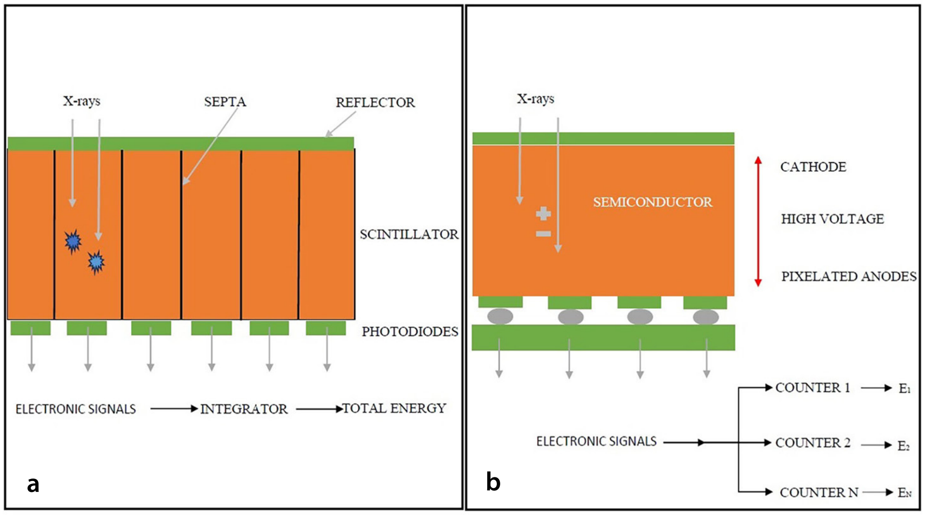


Figure 1. Illustrations of traditional CT **(a)** and photon-counting CT **(b)** detectors. CT, computed tomography.

2a. Acquisition procedures: advancements in parallel imaging techniques

Parallel imaging is currently employed in almost all clinical MRI scans to enable fast data capture for several reasons. Abdominal and cardiac scans often require patients to hold their breath to facilitate shorter scanning times. In certain situations, such as when multiple sequences occur after excitation pulses, blurring artifacts can arise, especially in imaging techniques such as turbo spin echo, due to significant T2 relaxation. These artifacts occur during the decomposition process while retrieving the lines. In other contexts, swift data collection is crucial to obtain extensive datasets efficiently.[40]

Parallel imaging reduces scanning time by using phased array coils to capture distinct perspectives of the tissue, thereby avoiding the need to scan a large portion of the region subjected to gradient encoding. However, the sensitivity of each coil element decreases rapidly with distance, which limits data collection to a specific tissue profile. A comprehensive image is generated by combining individual images from each coil. The maximum acceleration factor in parallel imaging is directly proportional to the number of coils. Typically, parallel imaging techniques employ coil arrays consisting of 4–8 coils. However, arrays with 32 or even 128 channels are available, particularly in cardiac imaging, resulting in a significant reduction in scanning time.

2b. Methods for reconstructing images for analysis

To prepare the data for meaningful information extraction, the initial steps of image capture, preprocessing, and segmentation are essential. During these processes, irrelevant or noise-based signals are eliminated. Patient movement is a common cause of noise. Sequential images are registered to correct motion artifacts, a process that can be achieved using algorithms specifically designed for medical imaging. The Insight ToolKit is currently considered the standard for MRI registration. It offers a range of algorithms for various operations, including transformations, similarity metrics, and contrast normalization.[41]

Machine learning applications have been increasingly used in recent trends in preprocessing and segmentation, such as denoising. Feature identification and classification have become important trends in machine learning techniques, primarily because these tasks require a large amount of manual effort. The abundance of imaging data obtained from MRI scans has increased the complexity of clinical diagnoses relying on MRIs, prompting the development of automated methods for data extraction and interpretation. Machine learning relies on algorithms generated from neural network architectures. These structures consist of nodes connected by weighted edges. Nodes receive inputs, multiply them by a set of parameters called weights, and then transport the resulting outputs through transfer functions such as sigmoid and hyperbolic tangent functions.[42]

Multi-information sourcing refers to the process of gathering and obtaining multiple sources of information. Models that use multiparametric techniques offer the substantial benefit of examining correlations between a large number of quantitative parameters, potentially leading to significantly improved accuracy. This contrasts with methods that analyze data using only one parameter. The time it takes for longitudinal (T1) and transverse (T2) relaxation, as well as the production of classical MR contrasts after the event, are all important metrics that can be obtained through MRI. However, this list is not exhaustive. Monitoring these metrics is done in conjunction with cutting-edge techniques for rapidly collecting data and performing computer analysis.[39]

Contemporary multiparametric analytical techniques use similar methods. These methods involve sampling both parameters and K-spaces simultaneously. These techniques require adjusting the collection parameters to capture data on the transient state, followed by undersampled K-space snapshots after each stimulation. Consequently, parametric maps are generated using a physical model based on the Bloch equations. Magnetic resonance fingerprinting (MRF) and quantitative transient-state imaging (QTI) are two examples of the various methods created as a direct outcome of this methodology.[43,44]

Modern multiparametric analytical techniques use similar methods by simultaneously sampling both parameters and K-spaces. These methods involve gathering transient-state data by adjusting the collection parameters and obtaining undersampled K-space snapshots after each stimulation. Parametric maps are generated using a physical model based on the Bloch equations. The methodologies of MRF and QTI have led to the development of these techniques.[45]

MRF is a technique that involves altering the settings of MRI sequences over time. This results in a series of MRI images with different weighting, and each type of tissue has a distinct MRI signal fingerprint. These fingerprints can be simulated using computational methods to generate a collection of tissue-specific fingerprints. During image reconstruction, the fingerprints obtained from the MRI data are compared with a dictionary. The fingerprint with the highest correlation is used to determine the MRI parameters for each voxel. After analyzing all the voxels, parametric maps are generated. The promising potential of MRF lies in its ability to accurately detect and identify specific structural characteristics, enabling the diagnosis of a

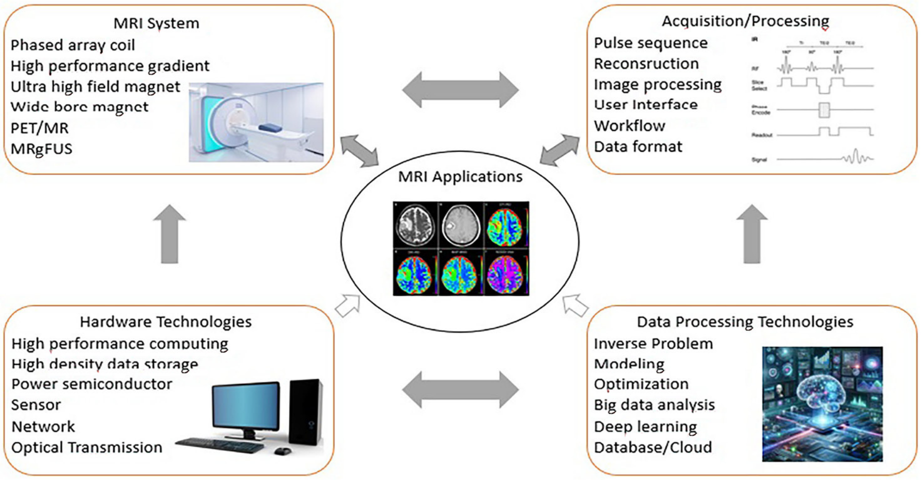


Figure 2. The advancements in MRI technology and the interconnections among its technical components. Advancements in fundamental sciences and engineering have a significant influence on MRI technology. Innovations in each component stimulate advancements in others. MRI, magnetic resonance imaging; MRgFUS, MR-guided focused ultrasound; PET, positron emission tomography.

wide range of clinical diseases. Novel methodologies, such as quantitative sequencing, enable the rapid and precise mapping of dynamic physiological processes. These approaches can evaluate blood flow in cardiac assessments by calculating scalar or vector velocities. A study was conducted to determine scalar velocities perpendicular to a vascular slice using multiparametric T1, T2, and proton density (PD) data. However, these methods are limited by their reliance on physical models to simulate the physiological events being mapped, which may result in potential data loss. Moreover, the complex nature of these models often requires considerable computational resources, thereby prolonging the time required for data collection.[39]

2c. Functional application through magnetic resonance imaging

Blood oxygenation-level dependent (BOLD) imaging, sometimes referred to as functional MRI (fMRI), was developed to indirectly assess neural activity in the brain by examining changes in blood oxygenation associated with brain activity. This type of imaging utilizes the neurovascular response of hyperemia, in which specific brain activity leads to increased blood oxygenation in the stimulated area (Figure 3). Conventional procedures for BOLD imaging typically employ T2 weighting and scan durations of less than 5 seconds to record the hemodynamic response function. Since its inception, the use of BOLD imaging has expanded significantly.[46]

The dependence of the BOLD signal on neurovascular mechanisms introduces specific constraints on fMRI, primarily because the hemodynamic response is slower than the underlying brain activity. This disparity means that the precise timing of neuronal spiking events is largely obscured. To isolate the signal activity associated with these events, mathematical processing techniques such as the general linear model or experimental block protocols are used. By employing these techniques, a temporal resolution of 100 milliseconds can be achieved, which is roughly one-tenth the speed of the brain activities being observed.[39]

An additional challenge encountered by fMRI is the constrained SNR, resulting from limitations in data acquisition and preprocessing. Researchers are actively exploring the use of strong magnetic fields to enhance the accuracy of anatomical imaging to address this challenge. Although most fMRI scans are conducted using three T fields, there is a growing trend toward employing seven T fields. Higher field strengths can reduce the need for spatial smoothing and improve the correlation coefficients of neuronal activity in resting-state networks (RSNs), indicating enhanced spatial resolution.[47]

An effective approach to overcome the time constraints of fMRI is to employ multimodal methods, which combine fMRI with techniques such as EEG or MEG. Both EEG and MEG provide quick temporal responses, capable of identifying brain events with millisecond precision. The reason for integrating these techniques with fMRI is their notably improved temporal resolution. Recent technical improvements enable the concurrent recording of EEG and fMRI signals, enhancing our comprehension of the spatial and temporal characteristics of physiological signals. Nevertheless, compared with fMRI alone, these integrated methodologies are less commonly employed. EEG has a poorer spatial resolution than fMRI, whereas MEG encounters difficulties in accurately determining the source of activity. Hence, to draw any experimental or clinical conclusions, it is imperative for experimental designs or clinical assessments utilizing these integrated methodologies to precisely ascertain the source of the signals.[48]

The persistent difficulties in understanding multimodal approaches have stimulated a longstanding desire to create alternative techniques that provide both precise spatial and temporal resolution. A novel method has been devised that combines the identification of extremely low magnetic fields generated by cerebral electrical activity with the detection of the hemodynamic response using fMRI. The technique, referred to as direct imaging of neuronal activity for fMRI, employs alternating K-space lines to capture the hemodynamic response while directly measuring the ultra-weak magnetic field using another K-space line. Thus far, this methodology has exclusively been utilized in animal models.[49]

A significant advancement in fMRI is the development of resting-state fMRI (RS-fMRI), which examines the inherent, involuntary oscillations in the BOLD signal with a frequency below 0.1 Hz without requiring any specific activities. The functional importance of these variances was first identified in 1995 a study where participants were instructed to abstain from engaging in any cognitive, verbal, or motor tasks. By analyzing the correlation between the BOLD signal time course in a particular brain region that is stimulated by bilateral finger tapping and the signals in other brain areas, the researchers discovered a strong association between changes in activity in the left somatosensory cortex and changes in activity in the corresponding region of the opposite hemisphere. This finding led to the deduction that these "resting networks" reflect the brain's functional connections. Following that, the analysis of spontaneous, synchronized fluctuations in activity across different regions of the brain has resulted in studies that have discovered a spectrum of 7–17 enduring networks, with 7 consistently recognized.[39,50]

RSNs in the human brain are mostly identified through the analysis of BOLD signals. This analysis is based on fMRI's capacity to detect neuronal activity. RS-fMRI relies on

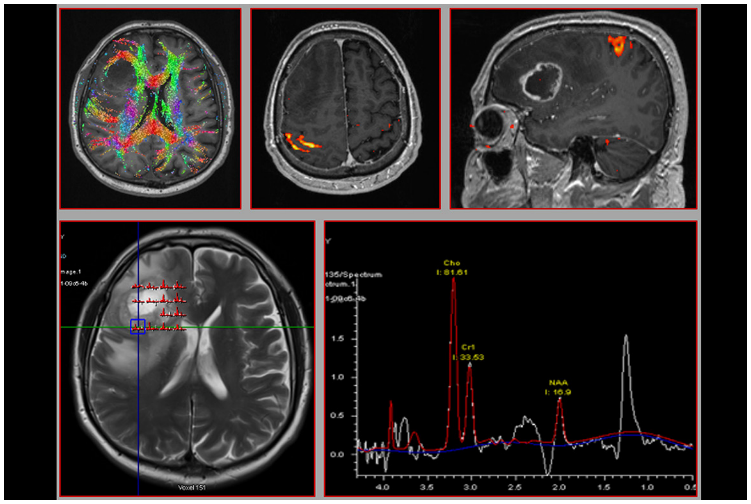


Figure 3. Multiparametric MR images (DTI, MRS, and BOLD fMRI) demonstrate functional activity in the right motor cortex. The images also evaluate the relationship between mass and the motor cortex. MR, magnetic resonance; DTI, diffusion tensor imaging; MRS, magnetic resonance spectroscopy; BOLD fMRI, blood oxygenation-level dependent functional magnetic resonance imaging.

the BOLD signal, which allows for the indirect monitoring of brain activity. This technique shares the advantages of fMRI, such as the ability to observe neural activity, but it has its intrinsic limitations. The primary constraint of fMRI is its temporal resolution, which is limited by the time it takes for the hemodynamic response. Therefore, one essential component of RS-fMRI use is the quantification of fluctuations in brain activity rather than directly recording instances of spiking.[39]

In the initial investigations of RSN functional connectivity, researchers chose specific areas of interest (ROIs) according to their own preferences chose specific regions of interest according to their preferences. Although the ROI technique is simple and easily understandable, its efficiency in discovering new networks is limited due to its dependence on user-defined regions. This is because it is restricted by specified criteria. As a result of this constraint, as well as progress in mathematical modeling and processing capacity, there has been a transition from imposing initial conditions on data to extracting patterns of brain activity directly from the unprocessed time series. An exemplary illustration of this novel methodology is independent component analysis (ICA), which posits that the time series signal arises from numerous spatiotemporal processes that are statistically independent of each other. Through the process of separating these autonomous signals, scientists are able to create chronological sequences for particular parts of the brain and organize them into maps that depict their spatial arrangement. RS-fMRI data can also be interpreted using graph theory, in which nodes represent activity sources and edges characterize the connectivity between these nodes. Unlike ICA, which primarily emphasizes the strength of correlations between distinct areas, graph theory specifically investigates the characteristics of network structure. The interconnections between nodes are characterized by graph metrics, including average path length, clustering coefficients, node degree, centrality measurements, and modularity levels. Graph theory is a potentially valuable tool for investigating how networks in the brain combine and separate. Modularity, a measure of the presence of functionally distinct components or modules within RSNs, is a key tool for characterizing functional changes in behavior, network disturbances, or diseases. This method has uncovered substantial modifications in situations such as stroke and psychiatric disorders.[39,51,52]

Theoretically, conclusions concerning causation based on directed functional connectivity can be expanded to include overall neural activity across the brain. Empirical investigations utilizing RS-fMRI have demonstrated that RSNs can be differentiated based on their metastability and synchronization. These observations have resulted in theories of brain function and behavior that propose that the human brain operates at maximal metastability when at rest, indicating an ideal state of network switching. Identifying the characteristics of RSNs, such as metastability, suggests that changes in directed connectivity could be used to evaluate the development of various brain states. This presents the methodological challenge of creating a descriptive methodology that links functional neuroimaging data to the overall dynamics of the entire brain. Recent efforts to address this challenge have pursued two primary methodologies.[53,54]

2d. Arthrographic applications in magnetic resonance imaging

Joint bone structures can be evaluated successfully using conventional radiographs and CT scans. However, these modalities do not enable the examination of soft tissue stabilizers. MRI, MR arthrography, and CT arthrography are the preferred imaging techniques for evaluating the labral, meniscal, fibrocartilaginous, capsular, and ligamentous structures of joints (Figures 4-6). Routine joint MR examination pulse sequences include fast spin-echo PD with fat suppression, T1- and T2-weighted fast spin-echo without fat saturation, and, occasionally, short tau inversion recovery (STIR). Conventional MRI sequences allow the non-invasive evaluation of tendon pathologies. However, labroligamentous and chondral lesions in these sequences are frequently overlooked. Direct MR arthrography with the intra-articular injection of diluted contrast media is a more sensitive imaging modality for evaluating stabilizers, such as the labrum, joint capsule, and ligaments.[55-60] In an imaging study that used arthroscopy as a reference standard, Gusmer et al.[61] found that conventional MRI has 86% sensitivity for detecting superior labral tears and 74% sensitivity for detecting posterior labral tears (Figure 7). However, despite improvements in image quality, routine MRIs may underestimate the exact extent of tears of the glenoid labrum.[60-62] Moreover, labrocapsular variant anomalies can be misdiagnosed as labral pathologies.[56,63]

Because of increased intra-articular fluid in patients with acute joint injuries, fluid-sensitive MR sequences such as PD, STIR, and T2-weighted imaging can reveal intra-articular damage, including labroligamentous, cartilaginous, and capsular injuries. However, in patients with chronic repetitive trauma, direct MR arthrography demonstrates clear diagnostic superiority over conventional MRI. Direct MR arthrography involves the intra-articular injection of diluted contrast media (gadolinium chelate). This technique allows for the optimal and separate evaluation of intra-articular structures with adequate capsular distension. Moreover, capsular distension in direct MR arthrography permits the leakage of contrast material into the labral substance or sublabral location in cases of labral tears or detachments (Figures 8 and 9). This makes it easier to identify pathologies of the glenoid or acetabular labrum.

Fluoroscopy-guided intra-articular injections for arthrography have been commonly employed since 1975.[64] However, many authors now advocate for performing injection procedures under ultrasonography guidance to avoid damaging anatomical structures along the injection pathway.[65-68] Real-time ultrasonographic guidance for arthrographic examination eliminates exposure to iodinated contrast material and ionizing radiation. In our routine practice, we use sonographic guidance for various approaches: the posterior approach for shoulder arthrography, the

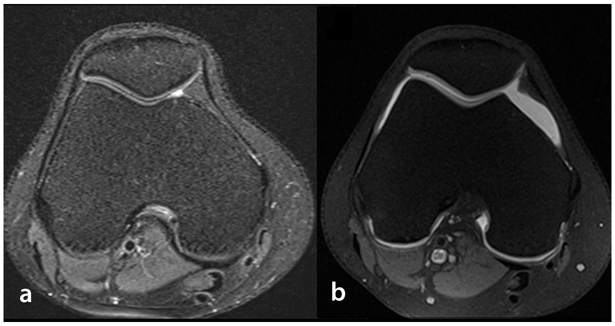


Figure 4. (b) Axial T1-weighted knee MR arthrogram obtained following intra-articular gadolinium injection shows the articular cartilage and capsule more clearly than pre-arthrographic axial PD MR imaging (**a**). MR, magnetic resonance; PD, proton density.

anteromedial approach for ankle arthrography, the anterolateral approach for hip arthrography, the dorsal–radial approach for wrist arthrography, the lateral approach for elbow arthrography, and the anterolateral approach for knee arthrography.

In arthrographic procedures, a sufficient volume of contrast solution is injected until the joint capsule is adequately dilated. The solution volume is determined based on the resistance encountered during injection and the patient's comfort level. The diluted gadolinium solution used for all joint arthrography procedures should have a concentration of 1:200. Table 1 shows the arthrographic solution volume and needle size for each joint.

A thin-section three-dimensional (3D) MR arthrography sequence, such as the fat-suppressed T1-weighted volumetric interpolated breath-hold examination (VIBE), allows for multiplanar reconstruction using submillimetric image slices. 3D VIBE MR arthrography not only provides excellent contrast for labroligamentous structures but also allows the optimal evaluation of the fibrocartilaginous complex and subchondral bone structure (Figure 10). In recent years, the 3D high-resolution T1-weighted VIBE MR arthrography sequence has been successfully employed for diagnosing glenoid bare spot, illustrating intra-articular small ligamentous structures, describing the aponeurotic expansion of the supraspinatus tendon, demonstrating glenoid cartilage defects accompanied by labral pathologies, and evaluating glenohumeral joint capacity for diagnosing primary adhesive capsulitis.[22-24,69-72] Lastly, MR arthrographic examinations with stress maneuvers have been successfully used to investigate capsular abnormalities of the shoulder joint.[73]

2e. Magnetic resonance spectroscopy and cerebrospinal fluid flowmetry

When placed in a strong magnetic field, hydrogen nuclei (protons) exhibit magnetic properties, serving as the source of measurable signals in MRI. The protons in water molecules are the primary source of the signal in MR examinations. However, protons in different molecules display slight magnetic variations, and this subtle difference enables the identification of small molecules in MR spectroscopy (MRS).[74] If the molecules are mobile and present in measurable quantities, MRS can depict these molecules within tissues on the MR spectrum (Figure 11).[75] The raw signal obtained by MRS is dominated by water,

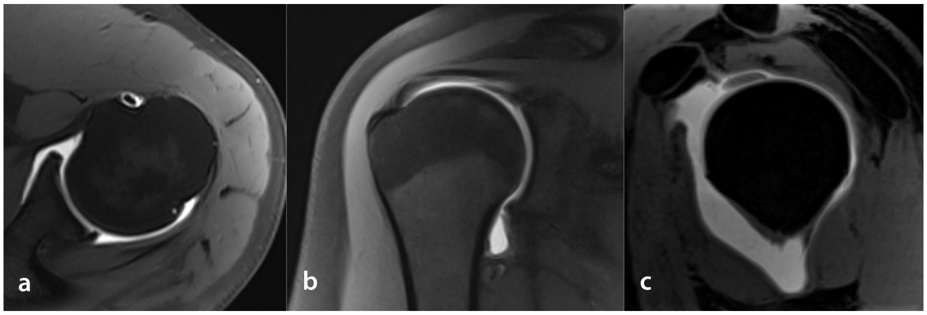


Figure 5. Axial **(a)**, coronal oblique **(b)**, and sagittal oblique **(c)** shoulder MR arthrograms optimally demonstrate the joint capsule, labroligamentous structures, and the underside of the rotator cuff tendons. MR, magnetic resonance.

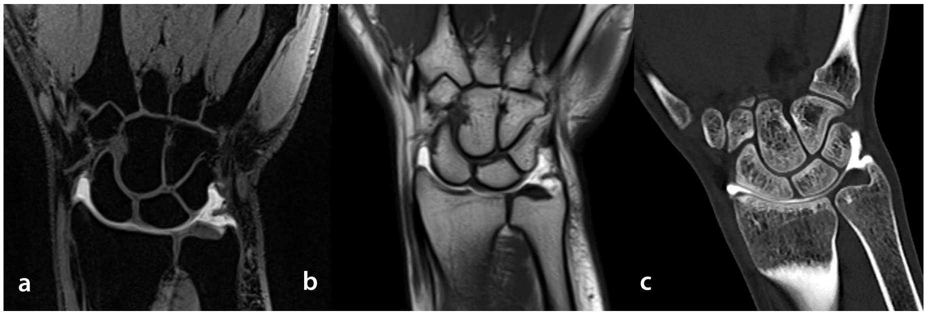


Figure 6. Coronal plane T1-weighted VIBE, TSE T1, and multi-detector CT arthrograms of the radiocarpal joint clearly reveal the cartilaginous surface, joint capsule, and triangular fibrocartilaginous complex. VIBE, volumetric interpolated breath-hold examination; CT, computed tomography.

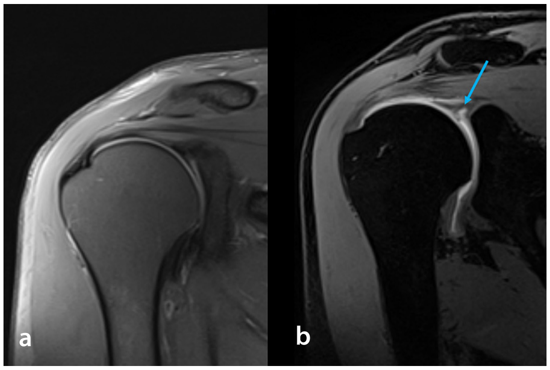


Figure 7. Coronal oblique plane PD MR imaging **(a)** of the right glenohumeral joint shows no pathology in the superior labrum; however, T1-weighted VIBE MR arthrography **(b)** reveals a type 2 SLAP lesion (blue arrow). PD, proton density; MR, magnetic resonance; VIBE, volumetric interpolated breath-hold examination.

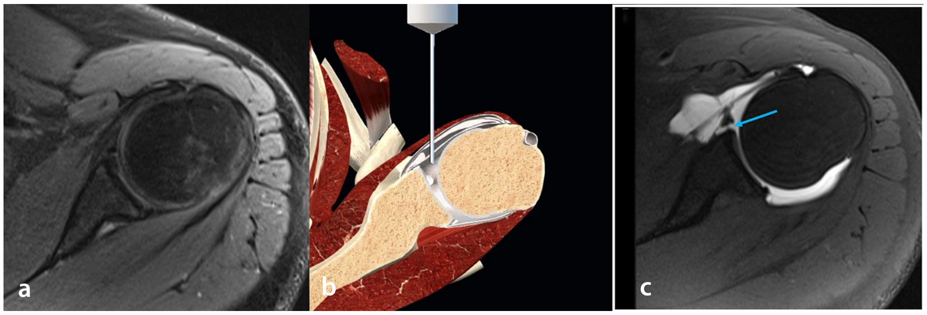


Figure 8. Transverse PD MR imaging **(a)** of the left glenohumeral joint shows no pathology in the anterior labrum; however, after Gd injection into the articular space **(b)**, SE T1-weighted MR arthrography **(c)** clearly reveals a fibrous Bankart lesion (blue arrow). [The illustration was created using Adobe Photoshop (Adobe Inc., 2021 Adobe Photoshop, https://www.adobe.com/products/photoshop.html) based on figures provided by the Complete Anatomy program (3D4 Medical, 2021. Complete Anatomy. Retrieved from https://3d4medical.com/)]. PD, proton density; MR, magnetic resonance; Gd, gadolinium; SE, spin-echo.

rendering signals from other metabolites invisible. To address this issue, water suppression techniques are employed, allowing for a clear and useful spectrum. MRS is not only used for differential diagnosis in neuroradiology, particularly in brain tissue, but also in other parts of the body,[76-78] focusing on specific metabolites of the targeted tissue.

Another advanced MRI technique, cerebrospinal fluid (CSF) flowmetry, is used to assess CSF through both qualitative and quantitative approaches (Figure 12). Time-resolved 2D phase-contrast MRI with velocity encoding is the most commonly employed method for this examination. The measured flow parameters in this technique reflect the pulsatile (to-and-fro) movement of CSF caused by vascular pulsations rather than the slow CSF transfer along the glymphatic pathway. This technique relies on the sequential, location-specific application of a pair of phase-encoding pulses in opposite directions. Stationary protons, which experience the same pulse in both instances, produce no signal. By contrast, moving protons, which encounter altered phase-encoding pulses, are rendered visible.[79] CSF flowmetry studies are particularly useful in evaluating clinical conditions such as normal pressure hydrocephalus, the patency of third ventriculostomy, aqueductal stenosis, and CSF flow at the cervicomedullary junction.[80]

2f. Artificial intelligence

AI has revolutionized MRI by introducing a wide range of applications that improve image acquisition, analysis, and therapeutic decision-making. The integration of AI into MRI has ushered in a new era in medical imaging, offering substantial benefits to both patients and healthcare providers. Table 2 summarizes how AI enhances various aspects of MRI, including improving image quality, facilitating disease diagnosis, and supporting treatment planning.[81]

3. Ultrasound

Ultrasound (US) is a versatile imaging technique widely used as an initial diagnostic method in various clinical scenarios worldwide. Continuous advancements in US technology provide new opportunities for medical diagnoses and therapies, solidifying its importance in medical imaging.

A 3D imaging method has been developed to overcome the limitations of traditional 2D US. This innovation allows the visualization of 3D anatomy, precise transducer adjustments for optimal disease monitoring, and accurate volume measurements. Several techniques have been developed for producing 3D US images; these include mechanical and free-hand scanning with linear arrays and the use of 2D arrays for real-time 3D imaging, also known as 4D US. Mechanical scanning utilizes a motorized mechanism to move a standard transducer and algorithms to construct 3D images from 2D scans. A motor/encoder ensures accurate information on the positions and orientations of the 2D US images, enabling the precise adjustment of the scanning geometry.[82]

Calibration is a crucial step in 3D reconstruction. It involves determining the position and angle of the position sensor relative to the US image. Various methods can achieve this. One successful approach enhances the spatial calibration of probes in 3D free-hand ultrasonic scanning. This method

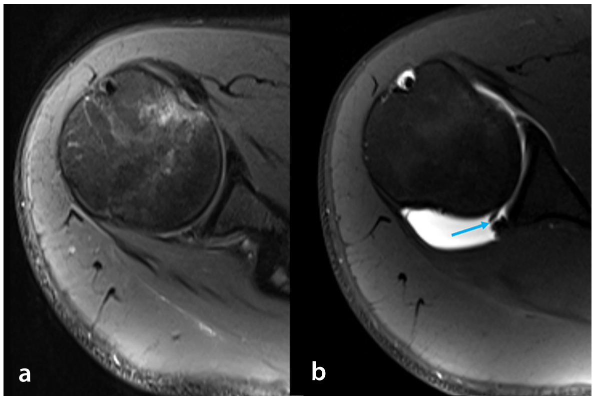


Figure 9. Transverse plane PD MR imaging **(a)** of the right glenohumeral joint shows no pathological findings in the posterior labrum; however, SE T1-weighted MR arthrography **(b)** clearly reveals a posterior labral defect (blue arrow). PD, proton density; MR, magnetic resonance; SE, spin-echo.

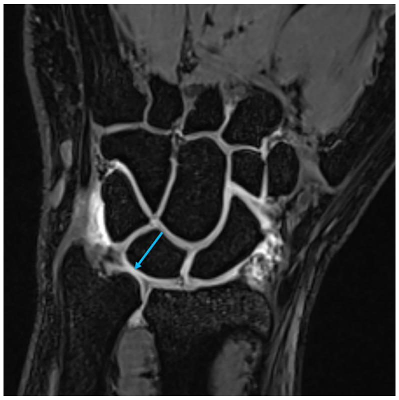


Figure 10. Coronal plane T1-weighted VIBE MR arthrography of the radiocarpal joint shows a central rupture (blue arrow) of the triangular fibrocartilaginous complex. VIBE, volumetric interpolated breath-hold examination; MR, magnetic resonance.

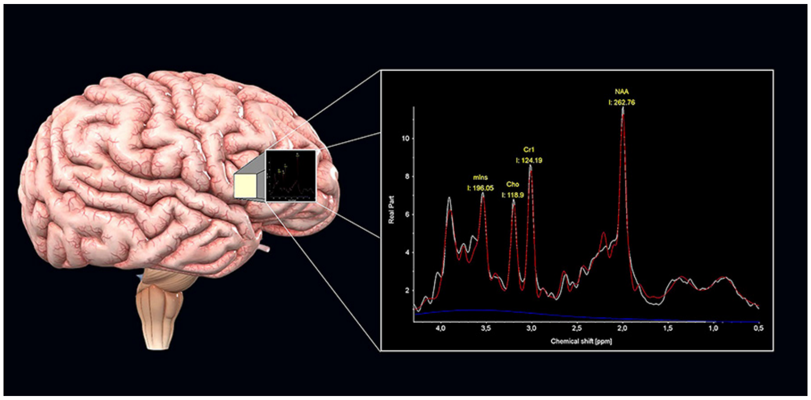


Figure 11. Measuring metabolites of the brain using MR spectroscopy [the illustration was created using Adobe Photoshop (Adobe Inc., 2021 Adobe Photoshop, https://www.adobe.com/products/photoshop.html) based on figures provided by the Complete Anatomy program (3D4 Medical, 2021. Complete Anatomy. Retrieved from https://3d4medical.com/)]. MR, magnetic resonance.

identifies similarity measures between two image sets-one from a 2D sweep and the other from a 3D reconstruction taken in a perpendicular sweep. However, a limitation of mechanical or free-hand scanning is the relatively slow volume capture rate, typically 2–3 volumes per second, which can hinder 3D imaging efficiency.[83]

The implementation of transducers equipped with 2D phased arrays for real-time 3D imaging has greatly enhanced the rate at which volume acquisition occurs. These transducers use electronic scanning to collect 3D data by producing a diverging beam in a pyramidal shape. The received echoes are then processed to create real-time 3D images. To further enhance high-volume imaging rates, a wideband 2D sparse array paired with multiline receiving has been proposed. This approach optimizes the use of a limited number of active components while maintaining a high level of accuracy and speed.[82,83]

A critical aspect of 3D imaging is the representation of the generated images, commonly achieved through multiplanar reformatting or volume rendering. However, the spatial resolution of 3D imaging is anisotropic and is typically inferior to that of 2D imaging. This limitation arises because the spacing between the collected 2D images increases with depth, resulting in reduced resolution at greater depths.[83]

Elastography is an advanced imaging technique that uses US to assess tissue stiffness, enhancing the diagnostic capabilities of B-mode US. Two primary methods of elastography are employed in evaluating breast lesions, shear wave elastography (SWE) and strain elastography. Although strain elastography requires operator expertise, SWE relies on focused radiation forces and eliminates the need for manual compression, making it operator independent.[84]

SWE is widely used in diagnosing tumoral and inflammatory pathologies in many organs, with research in these areas steadily growing. Moreover, recent studies suggest that SWE may also play a role in monitoring treatment efficacy.[84-86]

AI technology has the potential to create more accurate and repeatable outcomes in US examinations. AI and computer-assisted technologies can standardize medical processes, reduce training and examination durations, and improve the quality of US images across four main study areas. Leveraging machine learning in US imaging holds considerable promise for enhancing image quality, providing clearer and more practical visuals, and introducing novel US imaging techniques. Advancements in beamforming, super-resolution, and image enhancement often require hardware modifications, which are typically more complex than straightforward software upgrades. Despite these challenges, many recent research advancements outperform conventional reconstruction algorithms, which transform ultrasonic wave measurements into display visuals. The enhanced processing capabilities of medical devices now support the integration of increasingly sophisticated real-time solutions in a range of US imaging approaches. AI algorithms can aid healthcare professionals-including physicians, nurses, and technicians-in performing comprehensive US scans, thereby simplifying the learning pro-

Table 2. Concise overview of how AI improves several elements of MRI

Submission	Description
Image enhancement	Reduces noise and artifacts in MRI images. Enhances image resolution for finer anatomical details
Image reconstruction	Enables faster MRI scans. Reconstructs high-quality images from sparsely sampled data
Disease detection and diagnosis	Identifies and characterizes tumors in MRI scans. Aids in diagnosing conditions such as Alzheimer's using brain MRI. Assists in detecting heart diseases via cardiac MRI
Lesion segmentation	Accurately segments lesions in MRI scans, aiding in treatment planning
Functional MRI analysis	Maps brain regions activated during tasks or conditions, facilitating cognitive research
Diffusion MRI analysis	Reconstructs white matter tracts in the brain, which are valuable for neurosurgical planning
Quantitative imaging	Quantifies tissue properties (T1, T2, diffusion) for disease characterization. AI analyzes tissue perfusion in MRI, which is important for diagnosing conditions such as stroke
Automated reporting	Generates automated radiology reports by extracting findings from MRI scans
Treatment planning	Assists in radiotherapy planning by delineating target volumes on MRI
Monitoring disease progression	Tracks disease progression by analyzing changes in MRI scans over time
Predictive modeling	Predicts disease outcomes and treatment responses based on MRI data
Quality control	Performs quality checks on MRI scans, flagging artifacts and anomalies
Population studies	Analyzes large MRI datasets for trends, risk factors, and early disease indicators
Customization and personalization	Tailors MRI protocols to individual patients for optimized imaging

AI, artificial intelligence; MRI, magnetic resonance imaging.

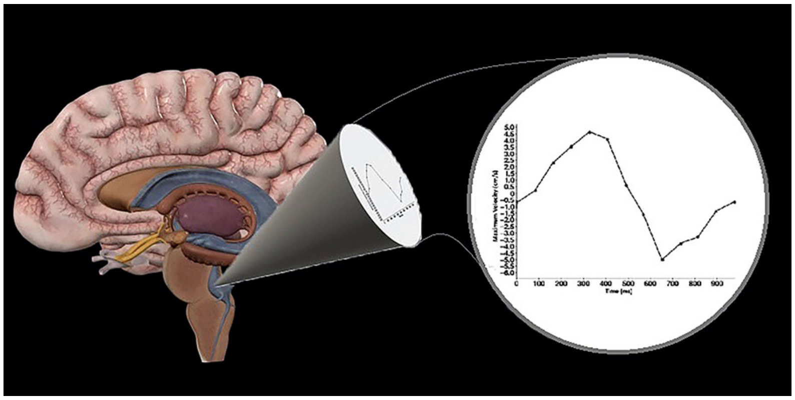


Figure 12. CSF flow analysis using CSF flowmetry [the illustration was created using Adobe Photoshop (Adobe Inc., 2021 Adobe Photoshop, https://www.adobe.com/products/photoshop.html) based on figures provided by the Complete Anatomy program (3D4 Medical, 2021. Complete Anatomy. Retrieved from https://3d4medical.com/)]. CSF, cerebrospinal fluid.

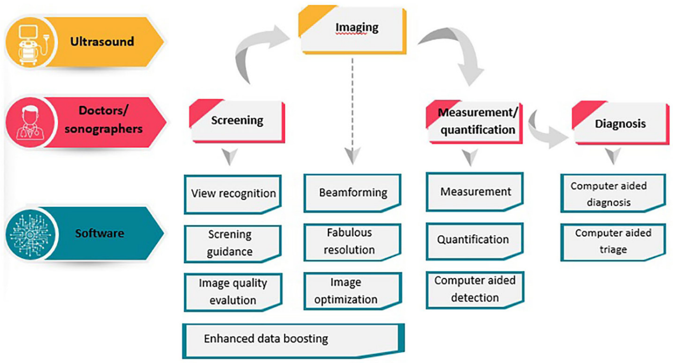


Figure 13. Effects of technological advancements on ultrasonographic imaging steps.

cess. Modern image processing algorithms used for measurement, quantification, and computer-aided detection have evolved beyond conventional feature engineering. The latest US imaging systems utilize advanced deep learning approaches. Computer-assisted diagnosis, triage, detection, and quantification are currently receiving considerable academic attention for their potential to reduce the workload of physicians (Figure 13).[87]

In conclusion, the field of radiography is being significantly influenced by technological advancements, potentially more so than other areas of medicine. Current developments in CT technology primarily focus on reducing the dosage of the ionizing radiation administered to patients. By contrast, progress in MRI systems is centered around improving accessibility, shortening scan durations, and generating high-quality images in regions where MRI has traditionally faced challenges. Furthermore, the development of portable devices for bedside use is becoming an increasingly important objective in both CT and MRI. Sonography innovations are advancing to enhance image quality and expand the applications of elastography. AI technology holds great potential for producing more accurate and repeatable results in US exams, enhancing image quality, generating clearer and more useful images, and even developing new US imaging techniques. Furthermore, AI technologies are being increasingly integrated into CT and MRI, with a growing focus on improving image production, enhancing image quality, and facilitating image evaluation.

Footnotes

Conflict of interest disclosure

Sonay Aydın, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

1. Hussain S, Mubeen I, Ullah N, et al. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *Biomed Res Int*. 2022;2022:5164970. [CrossRef]
2. Kasban H, El-Bendary M, Salama D. A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*. 2015;4(2):37-58. [CrossRef]
3. Roobottom CA, Mitchell G, Morgan-Hughes G. Radiation-reduction strategies in cardiac computed tomographic angiography. *Clin Radiol*. 2010;65(11):859-867. [CrossRef]
4. McPhee SJ, Papadakis MA, Rabow MW. Current medical diagnosis & treatment 2010: McGraw-Hill Medical New York: 2010. [CrossRef]
5. Mooney LR. A middle English verse compendium of astrological medicine. Med Hist. 1984;28(4):406-419. [CrossRef]
6. Berger D. A brief history of medical diagnosis and the birth of the clinical laboratory. Part 1--Ancient times through the 19th century. MLO Med Lab Obs. 1999;31(7):28-40. [CrossRef]
7. Bradley WG. History of medical imaging. *Proc Am Philos Soc*. 2008;152(3):349-361. [CrossRef]
8. Schulz RA, Stein JA, Pelc NJ. How CT happened: the early development of medical computed tomography. J Med Imaging. 2021;8(5):052110. [CrossRef]
9. Stański M, Michałowska I, Lemanowicz A, et al. Dual-energy and photon-counting computed tomography in vascular applications-technical background and post-processing techniques. *Diagnostics (Basel)*. 2024;14(12):1223. [CrossRef]
10. McCollough CH, Leng S, Yu L, Fletcher JG. Dual- and Multi-Energy CT: principles, technical approaches, and clinical applications. *Radiology*. 2015;276(3):637-653. [CrossRef]
11. Parakh A, Lennartz S, An C, et al. Dual-energy CT images: pearls and pitfalls. *Radiographics*. 2021;41(1):98-119. [CrossRef]
12. Toshav A. Economics of dual-energy CT: workflow, costs, and benefits. *Semin Ultrasound CT MR*. 2022;43(4):352-354. [CrossRef]
13. Wong WD, Mohammed MF, Nicolaou S, et al. Impact of dual-energy CT in the emergency department: increased radiologist confidence, reduced need for follow-up imaging, and projected cost benefit. *AJR Am J Roentgenol*. 2020;215(6):1528-1538. [CrossRef]
14. Burch RA, Siddiqui TA, Tou LC, Turner KB, Umair M. The cost effectiveness of coronary CT angiography and the effective utilization of CT-fractional flow reserve in the diagnosis of coronary artery disease. *J Cardiovasc Dev Dis*. 2023;10(1):25. [CrossRef]
15. Aydın S, Karavaş E, Ünver E, Şenbil DC, Kantarcı M. Long-term lung perfusion changes related to COVID-19: a dual energy computed tomography study. *Diagn Interv Radiol*. 2023;29(1):103-108. [CrossRef]
16. Aydin S, Kantarci M, Karavas E, Unver E, Yalcin S, Aydin F. Lung perfusion changes in COVID-19 pneumonia: a dual energy computerized tomography study. *Br J Radiol*. 2021;94(1125):20201380. [CrossRef]
17. Kantarcı M, Bayraktutan Ü, Akbulut A, et al. The value of dual-energy computed tomography in the evaluation of myocarditis. *Diagn Interv Radiol*. 2023;29(2):276-282. [CrossRef]
18. Cetin T, Kantarci M, Irgul B, et al. Quadruple-rule-out computed tomography angiography (QRO-CT): a novel dual-energy computed tomography technique for the diagnostic work-up of acute chest pain. *Diagnostics (Basel)*. 2023;13(17):2799. [CrossRef]
19. Rajendran K, Petersilka M, Henning A, et al. Full field-of-view, high-resolution, photon-counting detector CT: technical assessment and initial patient experience. *Phys Med Biol*. 2021;66(20):10. [CrossRef]
20. Rajendran K, Petersilka M, Henning A, et al. First clinical photon-counting detector CT system: technical evaluation. *Radiology*. 2022;303(1):130-138. [CrossRef]
21. Zhao C, Martin T, Shao X, Alger JR, Duddalwar V, Wang DJJ. Low dose CT perfusion with K-space weighted image average (KWIA). *IEEE Trans Med Imaging*. 2020;39(12):3879-3890. [CrossRef]
22. Gokce A, Guclu D, Unlu EN, Kazoglu I, Arican M, Ogul H. Comparison of conventional MR arthrography and 3D volumetric MR arthrography in detection of cartilage defects accompanying glenoid labrum pathologies. *Skeletal Radiol*. 2024;53(6):1081-1090. [CrossRef]
23. Ozel MA, Ogul H, Koksal A, et al. Detection of the glenoid bare spot by non-arthrographic MR imaging, conventional MR arthrography,

and 3D high-resolution T1-weighted VIBE MR arthrography: comparison with CT arthrography. *Eur Radiol*. 2023;33(5):3276-3285. [CrossRef]

24. Ogul H, Taydas O, Sakci Z, Altinsoy HB, Kantarci M. Posterior shoulder labrocapsular structures in all aspects; 3D volumetric MR arthrography study. *Br J Radiol*. 2021;94(1123):20201230. [CrossRef]
25. Polat G, Oğul H, Yalçin A, et al. Efficacy of the rotational traction method in the assessment of glenohumeral cartilage surface area in computed tomography arthrography. *J Comput Assist Tomogr*. 2019;43(2):345-349. [CrossRef]
26. Ogul H, Taydas O, Tuncer K, Polat G, Pirimoglu B, Kantarci M. MR arthrographic evaluation of the association between anterolateral soft tissue impingement and osteochondral lesion of the tibiotalar joint. *Radiol Med*. 2019;124(7):653-661. [CrossRef]
27. Ogul H, Cankaya B, Kantarci M. The distribution in joint recesses and adjacent synovial compartments of loose bodies determined on MR and CT arthrographies of ankle joint. *Br J Radiol*. 2022;95(1132):20201239. [CrossRef]
28. Zagarella A, Signorelli G, Muscogiuri G, et al. Overuse-related instability of the elbow: the role of CT-arthrography. *Insights Imaging*. 2022;12(1):140. [CrossRef]
29. Demehri S, Muhit A, Zbijewski W, et al. Assessment of image quality in soft tissue and bone visualization tasks for a dedicated extremity cone-beam CT system. *Eur Radiol*. 2015;25(6):1742-1751. [CrossRef]
30. Posadzy M, Desimpel J, Vanhoenacker F. Cone beam CT of the musculoskeletal system: clinical applications. *Insights Imaging*. 2018;9(1):35-45. [CrossRef]
31. Koskinen SK, Haapamäki VV, Salo J, et al. CT arthrography of the wrist using a novel, mobile, dedicated extremity cone-beam CT (CBCT). *Skeletal Radiol*. 2013;42(5):649-657. [CrossRef]
32. Koivisto J, Kiljunen T, Kadesjö N, Shi XQ, Wolff J. Effective radiation dose of a MSCT, two CBCT and one conventional radiography device in the ankle region. *J Foot Ankle Res*. 2015;8:8. [CrossRef]
33. Haridas H, Mohan A, Papisetti S, Ealla KK. Computed tomography: will the slices reveal the truth. *J Int Soc Prev Community Dent*. 2016;6(Suppl 2):85-92. [CrossRef]
34. Karaca L, Yuceler Z, Kantarci M, et al. The feasibility of dual-energy CT in differentiation of vertebral compression fractures. *Br J Radiol*. 2016;89(1057):20150300. [CrossRef]
35. Park EH, O'Donnell T, Fritz J. Dual-energy computed tomography applications in rheumatology. *Radiol Clin North Am*. 2024;62(5):849-863. [CrossRef]
36. Sandhu R, Aslan M, Obuchowski N, Primak A, Karim W, Subhas N. Dual-energy CT arthrography: a feasibility study. *Skeletal Radiol*. 2021;50(4):693-703. [CrossRef]
37. Foti G, Booz C, Buculo GM, et al. Dual-energy CT arthrography: advanced muscolo-skelatal applications in clinical practice. *Tomography*. 2023;9(4):1471-1484. [CrossRef]
38. Stern C, Graf DN, Bouaicha S, Wieser K, Rosskopf AB, Sutter R. Virtual non-contrast images calculated from dual-energy CT shoulder arthrography improve the detection of intraarticular loose bodies. *Skeletal Radiol*. 2022;51(8):1639-1647. [CrossRef]
39. Larrivee D. Introductory chapter: new advances in MRI clinical analysis. *New Advances in Magnetic Resonance Imaging: IntechOpen*. 2024. [CrossRef]
40. Gordon Y, Partovi S, Müller-Eschner M, et al. Dynamic contrast-enhanced magnetic resonance imaging: fundamentals and application to the evaluation of the peripheral perfusion. *Cardiovasc Diagn Ther*. 2014;4(2):147-164. [CrossRef]
41. Vadmal V, Junno G, Badve C, Huang W, Waite KA, Barnholtz-Sloan JS. MRI image analysis methods and applications: an algorithmic perspective using brain tumors as an exemplar. *Neurooncol Adv*. 2020;2(1):vdaa049. [CrossRef]
42. Khalil M, Ayad H, Adib A. Performance evaluation of feature extraction techniques in MR-brain image classification system. *Procedia Comput Sci*. 2018;127:218-225. [CrossRef]
43. Ma D, Gulani V, Seiberlich N, et al. Magnetic resonance fingerprinting. *Nature*. 2013;495(7440):187-192. [CrossRef]
44. Fayaz M, Torokeldiev N, Turdumamatov S, et al. An efficient methodology for brain MRI classification based on DWT and convolutional neural network. *Sensors (Basel)*. 2021;21(22):7480. [CrossRef]
45. Yacoub E, Van De Moortele PF, Shmuel A, Uğurbil K. Signal and noise characteristics of Hahn SE and GE BOLD fMRI at 7 T in humans. *Neuroimage*. 2005;24(3):738-750. [CrossRef]
46. Loued-Khenissi L, Döll O, Preuschoff K. An overview of functional magnetic resonance imaging techniques for organizational research. *Organ Res Methods*. 2019;22(1):17-45. [CrossRef]
47. Raimondo L, Oliveira ĹAF, Heij J, et al. Advances in resting state fMRI acquisitions for functional connectomics. *Neuroimage*. 2021;243:118503. [CrossRef]
48. Fleury M, Figueiredo P, Vourvopoulos A, Lécuyer A. Two is better? Combining EEG and fMRI for BCI and neurofeedback: a systematic review. *J Neural Eng*. 2023;20(5). [CrossRef]
49. Toi PT, Jang HJ, Min K, et al. In vivo direct imaging of neuronal activity at high temporospatial resolution. *Science*. 2022;378(6616):160-168. [CrossRef]
50. Damoiseaux JS, Rombouts SA, Barkhof F, et al. Consistent resting-state networks across healthy subjects. *Proc Natl Acad Sci U S A*. 2006;103(37):13848-13853. [CrossRef]
51. Corbetta M, Siegel JS, Shulman GL. On the low dimensionality of behavioral deficits and alterations of brain network connectivity after focal injury. *Cortex*. 2018;107:229-237. [CrossRef]
52. Biswal B, Yetkin FZ, Haughton VM, Hyde JS. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn Reson Med*. 1995;34(4):537-541. [CrossRef]
53. Beim Graben P, Jimenez-Marin A, Diez I, Cortes JM, Desroches M, Rodrigues S. Metastable resting state brain dynamics. *Front Comput Neurosci*. 2019;13:62. [CrossRef]
54. López-González A, Panda R, Ponce-Alvarez A, et al. Loss of consciousness reduces the stability of brain hubs and the heterogeneity of brain dynamics. *Commun Biol*. 2021;4(1):1037. [CrossRef]
55. Beltran J, Rosenberg ZS, Chandnani VP, Cuomo F, Beltran S, Rokito A. Glenohumeral instability: evaluation with MR arthrography. *Radiographics*. 1997;17(3):657-673. [CrossRef]
56. Ogul H, Tuncer K, Kose M, Pirimoglu B, Kantarci M. MR arthrographic characterization of posterior capsular folds in shoulder joints. *Br J Radiol*. 2019;92(1094):20180527. [CrossRef]
57. Shah N, Tung GA. Imaging signs of posterior glenohumeral instability. *AJR Am J Roentgenol*. 2009;192(3):730-735. [CrossRef]
58. Harish S, Nagar A, Moro J, Pugh D, Rebello R, O'Neill J. Imaging findings in posterior instability of the shoulder. *Skeletal Radiol*. 2008;37(8):693-707. [CrossRef]
59. Chiavaras MM, Harish S, Burr J. MR arthrographic assessment of suspected posteroinferior labral lesions using flexion, adduction, and internal rotation positioning of the arm: preliminary experience. *Skeletal Radiol*. 2010;39(5):481-488. [CrossRef]
60. Ogul H, Ayyildiz V, Pirimoglu B, et al. Magnetic resonance arthrographic demonstration of association of superior labrum anterior and posterior lesions with extended anterior labral tears. *J Comput Assist Tomogr*. 2019;43(1):51-60. [CrossRef]
61. Gusmer PB, Potter HG, Schatz JA, et al. Labral injuries: accuracy of detection with unenhanced MR imaging of the shoulder. *Radiology*. 1996;200(2):519-524. [CrossRef]
62. Lindauer KR, Major NM, Rougier-Chapman DP, Helms CA. MR imaging appearance of 180-360 degrees labral tears of the shoulder. *Skeletal Radiol*. 2005;34(2):74-79. [CrossRef]
63. Ogul H. Evaluation of posterosuperior labral tear with shoulder sonography after intra-articular injection. *Am J Phys Med Rehabil*. 2018;97(11):e110. [CrossRef]

64. Schneider R, Ghelman B, Kaye JJ. A simplified injection technique for shoulder arthrography. *Radiology*. 1975;114(3):738-739. [CrossRef]

65. Ogul H, Bayraktutan U, Yildirim OS, et al. Magnetic resonance arthrography of the glenohumeral joint: ultrasonography-guided technique using a posterior approach. *Eurasian J Med*. 2012;44(2):73-78. [CrossRef]

66. Ogul H, Bayraktutan U, Ozgokce M, et al. Ultrasound-guided shoulder MR arthrography: comparison of rotator interval and posterior approach. *Clin Imaging*. 2014;38(1):11-17. [CrossRef]

67. Zwar RB, Read JW, Noakes JB. Sonographically guided glenohumeral joint injection. *AJR Am J Roentgenol*. 2004;183:48-50. [CrossRef]

68. Souza PM, Aguiar RO, Marchiori E, Bardoe SA. Arthrography of the shoulder: a modified ultrasound guided technique of joint injection at the rotator interval. *Eur J Radiol*. 2010;74(3):29-32. [CrossRef]

69. Ogul H, Tas N, Tuncer K, et al. 3D volumetric MR arthrographic assessment of shoulder joint capacity in patients with primary adhesive capsulitis. *Br J Radiol*. 2019;92(1094):20180496. [CrossRef]

70. Guclu D, Ogul H, Unlu EN, et al. The 2D and 3D MR arthrographic description of aponeurotic expansion of supraspinatus tendon and biceps tendon anomaly in a large patient cohort. *Skeletal Radiol*. 2024;53(2):375. [CrossRef]

71. Ogul H, Karaca L, Can CE, et al. Anatomy, variants, and pathologies of the superior glenohumeral ligament: magnetic resonance imaging with three-dimensional volumetric interpolated breath-hold examination sequence and conventional magnetic resonance arthrography. *Korean J Radiol*. 2014;15(4):508-522. [CrossRef]

72. Cankaya B, Ogul H. An inconspicuous stabilizer of the subtalar joint: MR arthrographic anatomy of the posterior talocalcaneal ligament. *Skeletal Radiol*. 2021;50(4):705-710. [CrossRef]

73. Keles P, Ogul H, Tuncer K, Sakci Z, Ay M, Kantarci M. Magnetic resonance arthrography with positional manoeuvre for the diagnosis of synovial fold of posterior shoulder joint capsule. *Eur Radiol*. 2024. [CrossRef]

74. Sitter B, Sjøbakk TE, Larsson HBW, Kvistad KA. Clinical MR spectroscopy of the brain. *Tidsskr Nor Laegeforen*. 2019;139(6). [CrossRef]

75. Ross B, Bluml S. Magnetic resonance spectroscopy of the human brain. *Anat Rec*. 2001;265(2):54-84. [CrossRef]

76. Engelke K, Chaudry O, Gast L, et al. Magnetic resonance imaging techniques for the quantitative analysis of skeletal muscle: State of the art. *J Orthop Translat*. 2023;42:57-72. [CrossRef]

77. Fardanesh R, Marino MA, Avendano D, Leithner D, Pinker K, Thakur SB. Proton MR spectroscopy in the breast: Technical innovations and clinical applications. *J Magn Reson Imaging*. 2019;50(4):1033-1046. [CrossRef]

78. Verma S, Rajesh A, Fütterer JJ, et al. Prostate MRI and 3D MR spectroscopy: how we do it. *AJR Am J Roentgenol*. 2010;194(6):1414-1426. [CrossRef]

79. Battal B, Kocaoglu M, Bulakbasi N, Husmen G, Tuba Sanal H, Tayfun C. Cerebrospinal fluid flow imaging by using phase-contrast MR technique. *Br J Radiol*. 2011;84(1004):758-765. [CrossRef]

80. Mbonane S, Andronikou S. Interpretation and value of MR CSF flow studies for paediatric neurosurgery. *S Afr J Radiol*. 2013;17(1):26-29. [CrossRef]

81. Turkbey B, Haider MA. Deep learning-based artificial intelligence applications in prostate MRI: brief summary. *Br J Radiol*. 2022;95(1131):20210563. [CrossRef]

82. Golemati S, Cokkinos DD. Recent advances in vascular ultrasound imaging technology and their clinical implications. *Ultrasonics*. 2022;119:106599. [CrossRef]

83. Sciallero C, Trucco A. Wideband 2-D sparse array optimization combined with multiline reception for real-time 3-D medical ultrasound. *Ultrasonics*. 2021;111:106318. [CrossRef]

84. Ece B, Aydin S, Kantarci M. Shear wave elastography-correlated dose modifying: can we reduce corticosteroid doses in idiopathic granulomatous mastitis treatment? Preliminary results. *J Clin Med*. 2023;12(6):2265. [CrossRef]

85. Ece B, Aydin S. Can shear wave elastography help differentiate acute tonsillitis from normal tonsils in pediatric patients: a prospective preliminary study. *Children (Basel)*. 2023;10(4):704. [CrossRef]

86. Cetin T, Tokur O, Bozkurt HB, Aydin S, Memis KB, Kantarci M. Shear wave ultrasonographic elastography in pediatric spleens and its role in differential diagnosis. *Diagnostics (Basel)*. 2024;14(11):1142. [CrossRef]

87. Tenajas R, Miraut D, Illana CI, Alonso-Gonzalez R, Arias-Valcayo F, Herraiz JL. Recent advances in artificial intelligence-assisted ultrasound scanning. *Appl Sci*. 2023;13(6):3693. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253301

NEURORADIOLOGY

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Impact of a computed tomography-based artificial intelligence software on radiologists' workflow for detecting acute intracranial hemorrhage

Jimin Kim[1]
Jinhee Jang[2,3]
Se Won Oh[1]
Ha Young Lee[1]
Eun Jeong Min[4]
Jin Wook Choi[5]
Kook-Jin Ahn[2]

[1]The Catholic University of Korea, Eunpyeong St. Mary's Hospital College of Medicine, Department of Radiology, Seoul, Korea

[2]Seoul St. Mary's Hospital College of Medicine, The Catholic University of Korea, Department of Radiology, Seoul, Korea

[3]Applied Artificial Intelligence Research (A^{2}IR), Institute for Precision Health (IPH), University of California, Irvine, USA

[4]The Catholic University of Korea College of Medicine, Department of Medical Life Sciences, Seoul, Korea

[5]Ajou University School of Medicine, Ajou University Hospital, Department of Radiology, Suwon, Korea

Corresponding author: Jinhee Jang, Kook-Jin Ahn

E-mail: znee@catholic.ac.kr, ahn-kj@catholic.ac.kr

Received 14 February 2025; revision requested 05 April 2025; accepted 05 May 2025.

Epub: 07.07.2025

Publication date: 08.09.2025

DOI: 10.4274/dir.2025.253301

PURPOSE

To assess the impact of a commercially available computed tomography (CT)-based artificial intelligence (AI) software for detecting acute intracranial hemorrhage (AIH) on radiologists' diagnostic performance and workflow in a real-world clinical setting.

METHODS

This retrospective study included a total of 956 non-contrast brain CT scans obtained over a 70-day period, interpreted independently by 2 board-certified general radiologists. Of these, 541 scans were interpreted during the initial 35 days before the implementation of AI software, and the remaining 415 scans were interpreted during the subsequent 35 days, with reference to AIH probability scores generated by the software. To assess the software's impact on radiologists' performance in detecting AIH, performance before and after implementation was compared. Additionally, to evaluate the software's effect on radiologists' workflow, Kendall's Tau was used to assess the correlation between the daily chronological order of CT scans and the radiologists' reading order before and after implementation. The early diagnosis rate for AIH (defined as the proportion of AIH cases read within the first quartile by radiologists) and the median reading order of AIH cases were also compared before and after implementation.

RESULTS

A total of 956 initial CT scans from 956 patients [mean age: 63.14 ± 18.41 years; male patients: 447 (47%)] were included. There were no significant differences in accuracy [from 0.99 (95% confidence interval: 0.99–1.00) to 0.99 (0.98–1.00), $P = 0.343$], sensitivity [from 1.00 (0.99–1.00) to 1.00 (0.99–1.00), $P = 0.859$], or specificity [from 1.00 (0.99–1.00) to 0.99 (0.97–1.00), $P = 0.252$] following the implementation of the AI software. However, the daily correlation between the chronological order of CT scans and the radiologists' reading order significantly decreased [Kendall's Tau, from 0.61 (0.48–0.73) to 0.01 (0.00–0.26), $P < 0.001$]. Additionally, the early diagnosis rate significantly increased [from 0.49 (0.34–0.63) to 0.76 (0.60–0.93), $P = 0.013$], and the daily median reading order of AIH cases significantly decreased [from 7.25 (Q1–Q3: 3–10.75) to 1.5 (1–3), $P < 0.001$] after the implementation.

CONCLUSION

After the implementation of CT-based AI software for detecting AIH, the radiologists' daily reading order was considerably reprioritized to allow more rapid interpretation of AIH cases without compromising diagnostic performance in a real-world clinical setting.

CLINICAL SIGNIFICANCE

With the increasing number of CT scans and the growing burden on radiologists, optimizing the workflow for diagnosing AIH through CT-based AI software integration may enhance the prompt and efficient treatment of patients with AIH.

KEYWORDS

Acute intracranial hemorrhage, computed tomography, deep learning, artificial intelligence, radiologist, workflow, accuracy

You may cite this article as: Kim J, Jang J, Oh SW, et al. Impact of a computed tomography-based artificial intelligence software on radiologists' workflow for detecting acute intracranial hemorrhage. *Diagn Interv Radiol.* 2025;31(5):518-525.

Early and accurate detection of acute intracranial hemorrhage (AIH) on brain computed tomography (CT) is imperative due to the serious risks posed by this condition.[1-3] Timely diagnosis allows for immediate, life-saving intervention, whereas delayed detection can result in severe brain damage or death.[2-4] However, the rapidly increasing number of CT scans performed daily has placed a substantial burden on medical staff, including radiologists, potentially compromising the accuracy and timeliness of AIH diagnosis.[5,6]

In addition to the increasing workload, radiologists often face interruptions in their workflow due to various factors, such as urgent consultations, training of junior staff, and technical issues with imaging equipment.[7-9] These disruptions can lead to delays in image interpretation, increased cognitive load, and even diagnostic errors, particularly in high-stakes conditions such as AIH.[10,11] Such challenges underscore the importance of optimizing radiologists' workflow to ensure timely and accurate diagnoses.[12]

Recently, artificial intelligence (AI) has become a major focus in the field of neuroradiology, and numerous commercially available AI-based software programs have been developed for detecting acute cerebral findings.[13-18] Although previous studies have demonstrated the impressive standalone performance of these AI algorithms in diagnosing AIH and other stroke-related conditions on CT scans, their potential benefits for workflow optimization remain underexplored. Although early and prompt decision-making in AIH cases is critical for patient outcomes,[2-4] radiologists have traditionally relied on ambiguous prioritization systems such as stat, routine, or first-in, first-out (FIFO). This is largely because they are unable to assess the urgency of each exam in the worklist before opening it in the picture archiving and communication system (PACS).[19,20] To address this issue, some studies have shown that integrating AI algorithms into the PACS can greatly improve turnaround time (TAT) by prioritizing images based on urgency, thereby facilitating faster intervention and improved outcomes.[20-24] Therefore, evaluating the impact of AI software on radiologists' workflow in real-world settings is crucial for advancing its practical integration.

Main points

- A commercially available computed tomography-based artificial intelligence (AI) software was developed to ease the growing burden on radiologists to promptly diagnose acute intracranial hemorrhage (AIH).
- Evaluating AI software in a real-world clinical setting is essential for practical use.
- The implementation of this AI software considerably optimized radiologists' prioritization of reading order and enabled earlier reporting of AIH cases without compromising performance.
- The optimized workflow by the AI software integration is expected to improve the prompt and efficient treatment of patients with AIH.

This observational study aims to explore the impact of a commercially available CT-based AI software for detecting AIH on radiologists' diagnostic performance and their workflow in a real-world clinical setting.

Methods

The retrospective study was performed in line with the principles of the Declaration of Helsinki and approved by the Eunpyeong St. Mary's Hospital's Institutional Review Board (protocol number: PC24RASI0078, date: June 2024), and informed consent was waived according to the decision of the board committee.

Sample eligibility

A total of 1,375 non-contrast brain CT scans from patients with suspected AIH (including subdural, epidural, subarachnoid, intraparenchymal, and intraventricular hemorrhages) were potentially eligible over a 70-day period between December 1, 2023, and February 9, 2024. During this period, scans were included based on the following criteria: (1) the first CT scan performed during the patient's clinical course, (2) acceptable image quality for interpretation, and (3) availability of complete radiologist reports. All potentially eligible CT scans were reviewed by a board-certified neuroradiologist with 11 years of experience (J.K.) according to these criteria. After review, 273 follow-up scans, 140 scans with major metal artifacts caused by clips or coils, and 6 scans without radiologist interpretation were excluded. Ultimately, 956 non-contrast brain CT scans were included in this study.

To distinguish between study periods before and after AI software implementation, the boundary date was set as January 5, 2024, the date of implementation. Consequently, the pre-AI period was defined as the 35 days from December 1, 2023, to January 4, 2024, whereas the post-AI period covered the following 35 days from January 5 to February 9, 2024. Of the 956 brain CT scans, 541 were acquired during the pre-AI period, and the remaining 415 during the post-AI period (Figure 1).

Computed tomography scanning protocol

CT scans were performed using one of two CT machines at the institution. Machine A was a 128-slice single-source CT scanner (SOMATOM Edge, Siemens Healthineers, Forchheim, Germany) with a tube potential of 70–140 kVp and 20–800 mA; machine B was a dual-source CT scanner (SOMATOM Force, Siemens Healthineers, Germany) with

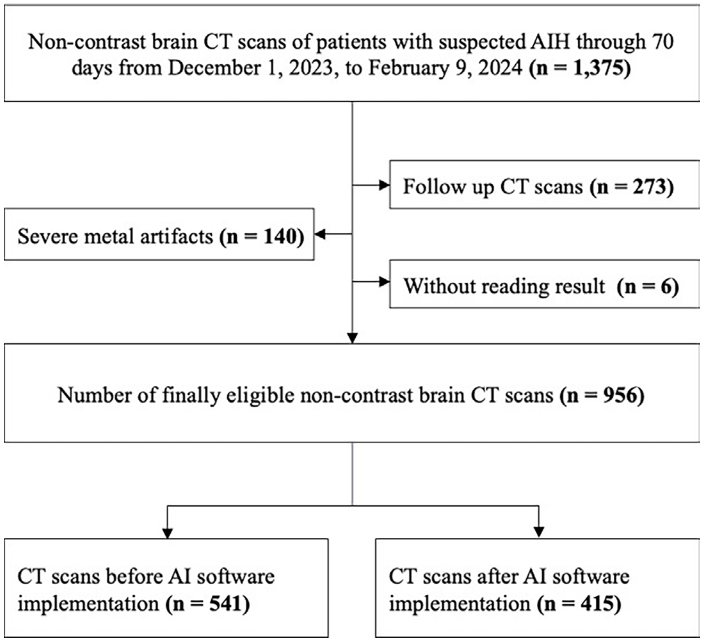

Non-contrast brain CT scans of patients with suspected AIH through 70 days from December 1, 2023, to February 9, 2024 (n = 1,375)
Follow up CT scans (n = 273)
Severe metal artifacts (n = 140)
Without reading result (n = 6)
Number of finally eligible non-contrast brain CT scans (n = 956)
CT scans before AI software implementation (n = 541)
CT scans after AI software implementation (n = 415)

Figure 1. Flowchart of study enrollment. AI, artificial intelligence; AIH, acute intracranial hemorrhage; CT, computed tomography.

a variable tube potential of 70–150 kVp and 20–1300 mA. The acquisition parameters were as follows: slice thickness, 4 mm without gap; rotation time, 1.0 s; pitch, 1; automatic tube voltage modulation (CARE kV, Siemens Healthineers, Germany) using a reference of 120 kV; automatic tube current selection (CAREDose 4D, Siemens Healthineers, Germany) using a reference of 250 mAs; and collimation of 128 × 0.6 for machine A and 192 × 0.6 for machine B.

Artificial intelligence software development

The commercially available CT-based AI software for detecting AIH (HyperInsight - ICH, version 2.0.1, Purple AI Inc., Korea) used in this study was developed using deep learning algorithms trained on 28,351 slices from 2,010 patients with AIH and 1,000 normal participants. The AIH detection process employed a joint convolutional and recurrent neural network-based sequence module that provided AIH probability scores (ranging from 0 to 100) on both a patient-wise and slice-wise basis. It also generated anomalies for patients with AIH by subtracting original CT images from restored images and post-processing them using unsupervised training on normal datasets. AI-assisted brain CT images showing AIH locations and scores were displayed to the radiologists on the PACS viewer alongside the original images.[18]

Ground truth for acute intracranial hemorrhage

To establish the ground truth for AIH, 2 board-certified neuroradiologists (S.W.O. and H.Y.L., with 17 and 19 years of experience in brain imaging, respectively) independently reviewed the same set of 956 non-contrast brain CT scans. The neuroradiologists diagnosed AIH based solely on CT findings and were blinded to patients' clinical information, previous reading results, and follow-up imaging. In cases of disagreement, the ground truth was determined by consensus, referring to other available imaging modalities.

Radiologists' computed tomography interpretation

Two board-certified general radiologists (H.B. and H.S., each with 10 years of experience in brain imaging without fellowship training in neuroradiology) routinely interpreted the enrolled non-contrast brain CT scans as part of clinical practice. These radiologists were blinded and unaware of the study's purpose and design throughout the entire study period. Therefore, they could freely refer to patients' clinical information and other available studies using the institution's PACS (ZeTTA PACS, version 1.0.0.42.10, TaeYoung Soft, Korea). Prior to AI software implementation, the two radiologists received brief training in using the software from a board-certified neuroradiologist (J.K.) for 1 day. The radiologists required minimal learning time with the AI software, as the probability scores were intuitively presented within the existing worklist interface. After implementation, the AIH probability scores generated by the software were integrated into the PACS worklist, allowing the radiologists to determine the reading order based on the scores. Figure 2 exemplifies the worklists before and after implementation. During the entire study period, CT scan completion time and the radiologists' final report time were automatically recorded on the PACS server of our institution.

Definition of the early diagnosis rate

Since early diagnosis of AIH is crucial for improving patient outcomes,[1-4] the early diagnosis rate for AIH cases was defined to assess the potential effectiveness of changes to the reading order. The first quartile of the radiologists' reading order was chosen as the threshold for defining early diagnosis, because the first quartile is commonly used to identify the highest-priority or most urgent cases in general medical practice.[25,26] By using the first quartile of reads, the aim was to assess the effectiveness of the prioritization by the AI software. The equation for the early diagnosis rate was defined as follows:

$$\begin{array}{l}\text{Early}\\ \text{Diagnosis}\\ \text{Rate}\end{array} = \frac{\begin{array}{c}\text{AIH cases read rapidly within the}\\ \text{first quartile by radiologists}\end{array}}{\text{Total AIH cases}} \qquad (1)$$

Statistical analysis

The sample size of the case group was calculated based on a significance level of 0.05, a statistical power of 0.8, a specificity of 0.90 from a previous meta-analysis, and a specificity of 0.984 from prior validation research, with a dropout rate of 10%.[13,17] The determined sample size for the study was 202 cases. Due to its explanatory nature, the sample size for the daily analysis was determined based on previous studies,[27] and a minimum of 1 month was selected for each period before and after AI software implementation. The stand-alone performance of the AI software after implementation was evaluated using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity.

First, a simple comparison of the radiologists' absolute TAT (the time gap between CT scan completion and the radiologists' final report) was conducted as a preliminary study. The TAT of cases with and without AIH between the pre-and post-AI periods was compared using an independent t-test, following the Shapiro–Wilk test for normality. This preliminary comparison aimed to explore the feasibility of conducting daily comparisons and to avoid bias arising from TAT comparisons.

Furthermore, to evaluate the impact of AI software on the radiologists' daily diagnostic performance for AIH, their accuracy, sensitivity, and specificity were calculated in both pre-and post-AI periods and compared between the two periods. Moreover, the impact of false negative and false positive cases generated by the AI software on radiologists' decisions was assessed in an additional sub-analysis.

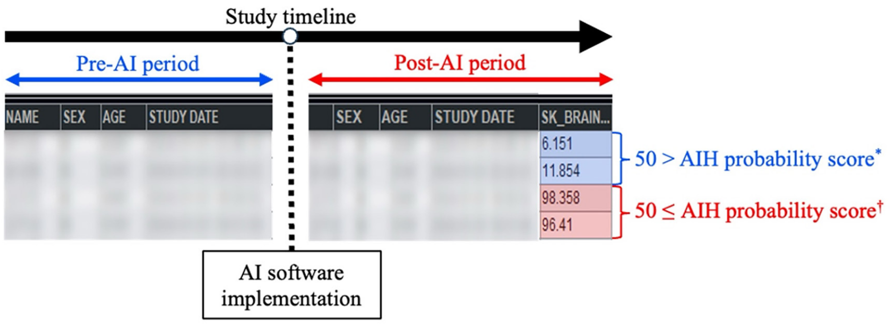


Figure 2. This figure provides examples of the worklists used in the study. The left worklist during the pre-AI period shows the radiologists' routine worklists before AI software implementation. By contrast, in the SK_BRAIN column, the AIH probability scores generated by the software were added to the worklist during the post-AI period. Consequently, the radiologists could use the score to predict that the cases with a blue background (*) were less likely to exhibit AIH, whereas the cases with a red background were more likely to exhibit AIH (†) after the implementation. AI, artificial intelligence; AIH, acute intracranial hemorrhage.

Lastly, the impact of the AI software on the radiologists' workflow was evaluated. The ordinal correlation between the chronological order of CT scans and the radiologists' reading order was measured using Kendall's Tau in both pre-and post-AI periods. These rank correlation coefficients were compared between the two periods. In addition, the modified reading order was evaluated to confirm whether it appropriately prioritized the rapid reading of AIH. For this evaluation, the daily early diagnosis rate for AIH cases and the median reading order of AIH were calculated in both pre- and post-AI periods and compared between the two periods.

Mean daily diagnostic performance; Kendall's Tau; early diagnosis rate for AIH cases; median reading order of AIH; and baseline characteristics including age, gender proportion, AIH incidence, Glasgow Coma Scale scores, and modified Rankin scale scores between the pre-and post-AI periods were compared using independent t-tests or Mann–Whitney U tests following the Shapiro–Wilk test. A visual summary of the comparison analyses is presented in Figure 3.

Continuous variables were described as means with 95% confidence intervals (CIs) using bootstrapping, and ordinal variables were described as medians with ranges from the 25th percentile (Q1) to the 75th percentile (Q3). The statistical software MedCalc (version 23.2.1, MedCalc Software Ltd, USA) was used for statistical analysis. A *P* value less than 0.05 was considered statistically significant.

Results

Patient characteristics

A total of 956 initial CT scans from 956 patients were included. Of these, 541 and 415 CT scans were acquired during the pre-and post-AI periods, respectively. The mean age of the total patient cohort was 63.14 years ± 18.41 (standard deviation), the proportion of male participants was 45%, and the incidence of AIH was 13%. There was no significant difference in median age [pre-AI period: 67 years (51–77); post-AI period: 67 (52–78); *P* = 0.558], number of male patients [pre-AI period: 246 (45%); post-AI period: 201 (48%); *P* = 0.363], AIH cases [pre-AI period: 72 (13%); post-AI period: 50 (12%); *P* = 0.681], median Glasgow Coma Scale score [pre-AI period: 15 (15–15); post-AI period: 15 (15–15); *P* = 0.831], and modified Rankin scale scores [pre-AI period: 0 (0–0); post-AI period: 0 (0–0); *P* = 0.295] before and after AI implementation. The number of daily CT scans [pre-AI period: 12 (7.25–17.75); post-AI period: 12 (10–19.75); *P* = 0.256] and daily AIH cases [pre-AI period: 1 (0–1.75); post-AI period: 2 (1–3); *P* = 0.063] were not significantly different. These results are summarized in Table 1.

Table 1. Baseline characteristics of patients

Baseline characteristics	Pre-AI period (n = 541)	Post-AI period (n = 415)	*P* value
Age (years)*	67 (51–77)	67 (52–78)	0.558
Number of male patients (%)	246 (45)	201 (48)	0.363
Number of AIH cases (%)	72 (13)	50 (12)	0.681
Glasgow Coma Scale score*	15 (15–15)	15 (15–15)	0.831
Modified Rankin Scale score*	0 (0–0)	0 (0–0)	0.295
Number of daily CT scans*	12 (7.25–17.75)	12 (10–19.75)	0.256
Number of daily AIH cases*	1 (0–1.75)	2 (1–3)	0.063

*The Mann–Whitney U test was used, and the values are presented as the median with the range between Q1 and Q3. AI, artificial intelligence; AIH, acute intracranial hemorrhage; CT, computed tomography.

Preliminary comparison of turnaround time

In the preliminary study, the mean TAT significantly decreased (from 1,610 min to 1,145 min, *P* < 0.001) after AI software implementation. When analyzed by cases with and without AIH, TAT significantly decreased in both cases with AIH (from 1,452 min to 870 min, *P* < 0.001) and without AIH (from 2,084 min to 1,184 min, *P* < 0.001) after AI software implementation. These preliminary results are illustrated in Figure 4.

Stand-alone performance of the artificial intelligence software

The prevalence of AIH in the post-AI period was 12%. After AI software implementation, the AUC for the standalone AI software was 0.99 (95% CI, 0.98–0.99) in detecting AIH. The accuracy, sensitivity, and specificity were 0.98 (95% CI, 0.97–0.99), 0.96 (95% CI, 0.86–0.99), and 0.99 (95% CI, 0.97–0.99), respectively, using a probability score cut-off of 50% for detecting AIH.

Diagnostic performance of radiologists

The radiologists' daily accuracy [from 0.99 (95% CI, 0.99–1.00) to 0.99 (95% CI, 0.98–1.00), *P* = 0.343], sensitivity [from 1.00 (95% CI, 0.99–1.00) to 1.00 (95% CI, 0.99–1.00), *P* = 0.859], and specificity [from 1.00 (95% CI, 0.99–1.00) to 0.99 (95% CI, 0.97–1.00), *P* = 0.252] for detecting AIH were not significant-

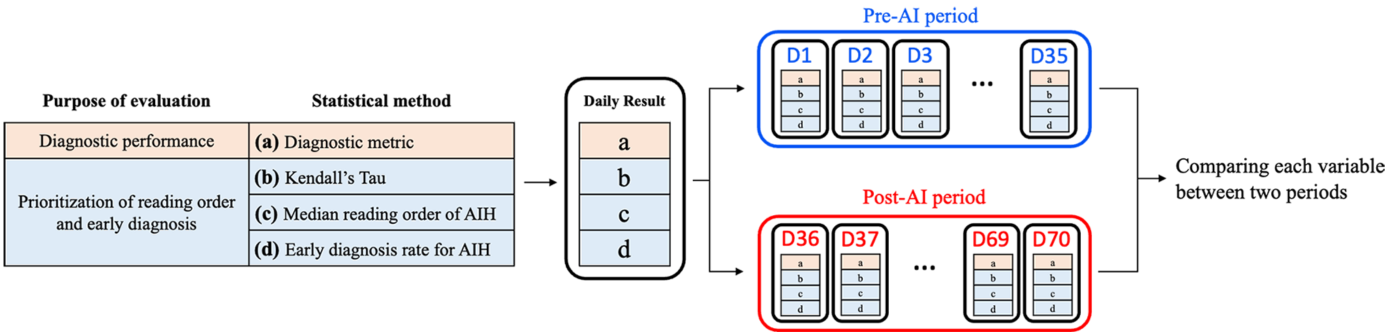


Figure 3. This figure presents a schematic representation of the statistical analyses used in this study. To assess the AI software's impact on the radiologists' diagnostic performance (a) in AIH detection, performance was compared before and after implementation. To assess the impact of the software on the radiologists' workflow, Kendall's Tau (b) was used to compare the correlation between the daily chronological order of the CT scan and the radiologists' reading order before and after the implementation. The median reading order of AIH (c) and the early diagnosis rate for AIH (d) (defined as the proportion of AIH cases read rapidly within the top quarter by radiologists) were compared before and after implementation. AI, artificial intelligence; AIH, acute intracranial hemorrhage; CT, computed tomography.

ly different after AI software implementation. These results are summarized in Table 2.

In an additional sub-analysis of false negative and false positive cases, there were two false negative and four false positive cases generated by the AI software. However, the radiologists' diagnoses and the ground truth for AIH were entirely identical even in these cases. Examples of cases with and without AIH integrated with the AI software are illustrated in Figure 5.

Prioritization of reading order and early diagnosis

The daily correlation between the chronological order of CT scans and the radiologists' reading order significantly decreased after AI software implementation [Kendall's Tau: from 0.61 (95% CI, 0.48–0.73) to 0.01 (95% CI, 0.00–0.26), $P < 0.001$]. The radiologists' daily early diagnosis rate of AIH significantly increased after AI software implementation [from 0.50 (0.23–1.00) to 1.00 (0.55–1.00), $P = 0.014$]. Furthermore, the radiologists' daily median reading order for AIH cases significantly decreased after AI software implementation [from 7.25 (3–11.75) to 1.5 (1–3), $P < 0.001$]. These results are summarized in Table 2 and illustrated in Figure 6.

Discussion

This study aimed to assess the impact of a commercially available CT-based AI software for AIH detection on radiologists' diagnostic performance and workflow. The software greatly optimized radiologists' reading prioritization and enabled them to read AIH cases more rapidly in daily practice. Furthermore, the AI software did not compromise the radiologists' diagnostic performance for detecting AIH, even in cases where the AI generated false positives or false negatives.

Regarding the radiologists' diagnostic performance for AIH, the impact of the AI software was negligible, and the radiologists were not influenced by the false negative or false positive results generated by the software. Several factors may explain this finding. First, the study design played a role. In this observational study, the readers had access to patient information and other examinations as part of routine clinical practice, unlike previous validation studies with controlled conditions where readers lacked clinical context.[17] Additionally, the diagnostic accuracy of board-certified radiologists for AIH is known to be particularly high in routine clinical settings.[1-3] Therefore, it is not surprising that the radiologists in this study–being board-certified and experienced in diagnosing AIH–maintained high performance. Notably, the minor changes in accuracy and specificity may indicate effective management of false positives by the AI software. In other words, potential false positives generated by the AI were either easily recognized or efficiently disregarded, thereby not compromising diagnostic outcomes. Consequently, our findings suggest that the AI software's impact on detection performance may be negligible–or at least not detrimental–when radiologists interpret images under routine conditions or already possess sufficient diagnostic expertise.[28,29]

To evaluate whether the AI software could influence the radiologists' actual reading order, we compared the correlation between the chronological order of CT scans and the radiologists' reading order before and after AI software implementation. Before the implementation, there was a high correlation between the two, suggesting that radiologists typically interpreted CT scans using a traditional stat or FIFO prioritization system. However, after implementation, a considerable dissociation between the two orders was observed, along with an increased early diagnosis rate of AIH. This suggests that the integrated AI software substantially altered the radiologists' reading order and facilitated prioritization of CT scans with AIH over those without. This shift in prioritization occurred because radiologists could estimate the ur-

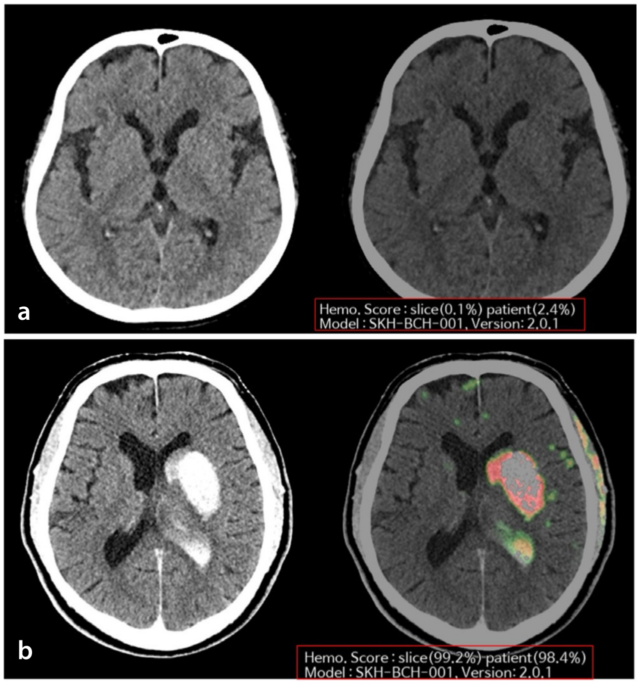

a
Hemo. Score : slice(0.1%) patient(2.4%)
Model : SKH-BCH-001, Version: 2.0.1
b
Hemo. Score : slice(99.2%) patient(98.4%)
Model : SKH-BCH-001, Version: 2.0.1

Figure 4. In each figure, the left image shows a non-contrast brain CT scan, whereas the right image shows an overlaid heatmap of AIH generated by the AI software. The red box below depicts the AIH probability score (hemo. score) for both the slice- and patient-wise levels, as well as the model and version information of the AI software. **(a)** AIH is not visible on the CT scan, resulting in an AIH probability score of less than 50%, with no heatmap. **(b)** There is AIH in the left basal ganglia, extending to the left lateral ventricle, resulting in a probability score of over 50% with a visible heatmap. AI, artificial intelligence; AIH, acute intracranial hemorrhage; CT, computed tomography.

Table 2. Radiologists' diagnostic performance, prioritization of reading order, and early diagnosis between the pre- and post-AI periods

Variables	Pre-AI period (35 days)	Post-AI period (35 days)	*P* value
Accuracy[†]	0.99 (0.99–1.00)	0.99 (0.98–1.00)	0.343
Sensitivity[†]	1.00 (0.99–1.00)	1.00 (0.99–1.00)	0.859
Specificity[†]	1.00 (0.99–1.00)	0.99 (0.97–1.00)	0.252
Kendall's Tau[†]	0.61 (0.48–0.73)	0.01 (0.00–0.26)	<0.001*
Early diagnosis rate[‡]	0.50 (0.23–1.00)	1.00 (0.55–1.00)	0.014*
Median reading order[‡]	7.25 (3.00–10.75)	1.50 (1.00–3.00)	<0.001*

*$P < 0.05$, statistical significance. [†] An independent t-test was used, and the values are presented as the mean with 95% CI. [‡] The Mann–Whitney U test was used, and the values are presented as the median with the range between Q1 and Q3. AI, artificial intelligence; CI, confidence interval.

gency of AIH cases by referring to the AIH probability score before opening a CT scan from their worklist. This predictability led to a remarkable increase in early diagnosis. After implementation, the median reading order of AIH cases considerably decreased, and the early diagnosis rate for AIH cases increased substantially. These changes signify that the radiologists' workflow was prioritized and optimized to allow for more rapid interpretation of AIH cases. Considering that non-contrast brain CT is the first-line approach for AIH, these improvements brought by the AI software may enhance not only the promptness but also the efficiency of clinical diagnosis and treatment for patients with suspected AIH.[1-4,6,24]

In terms of patient characteristics, the modified Rankin scale scores were not considerably different after AI software implementation. However, these findings should be interpreted with caution. Because the primary objective of this study was to evaluate the impact of AI integration on radiologists' workflow, the AI software was not utilized by physicians in clinical decision-making. Moreover, functional outcomes are influenced by a wide range of clinical variables, including age, neurological status, comorbidities, and treatment delays.[2-4,16] None of these factors were adjusted for in our analysis, as this was beyond the scope of the study. Therefore, the lack of observed improvement in functional outcomes does not imply that the AI software lacks clinical value. On the contrary, considering our findings demonstrating enhanced reading prioritization by AI and previous research indicating the greatest benefits of AI when used by clinicians,[17] it can be inferred that AI contributes to efficiency and potentially improves patient care in clinical environments. Consequently, this study remains important as it establishes a foundation for the broader adoption of AI in clinical practice.

In this study, we conducted an ordinal comparison on a daily basis rather than a simple TAT comparison between the pre-AI and post-AI periods, as the mean TAT for both cases with and without AIH had already decreased substantially in the preliminary study. Radiologists' TAT can be affected by numerous factors, including routine tasks, working days, or other unexpected circumstances,[7-11] and the radiologists in this study–who interpreted various imaging modalities across different body parts–may have been similarly influenced.[19,20] Therefore, our daily ordinal comparison of radiologists' reading order more accurately reflected their workflow in a routine real-world clinical setting than a simple TAT comparison. As a result, we mitigated potential bias and gained clearer insights into radiologists' workflow.

This study had several limitations. First, its retrospective observational design may have introduced uncontrolled bias that could have affected our results. Second, the findings were based on data from a single insti-

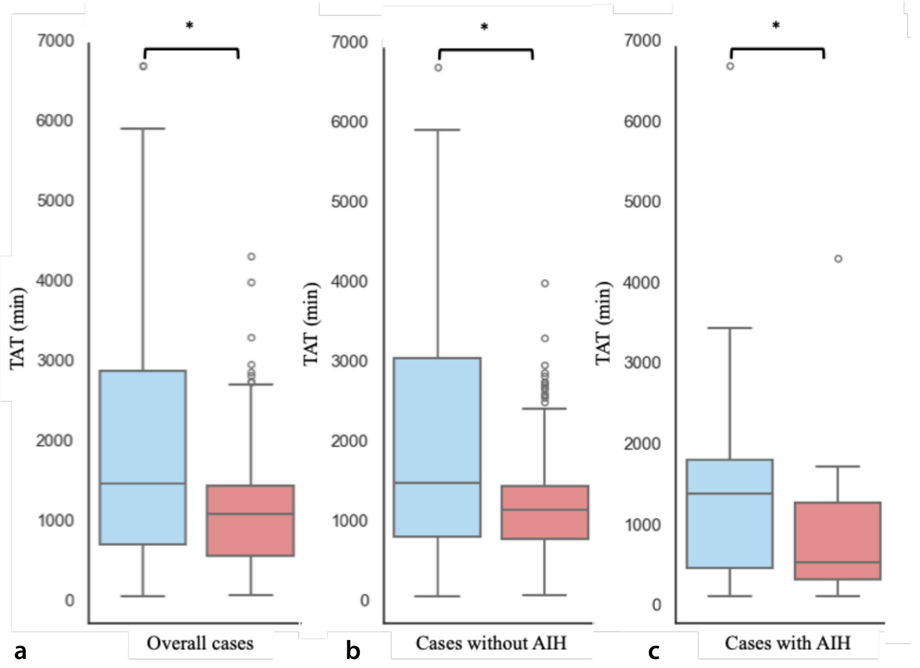


Figure 5. This figure illustrates box and whisker plot charts comparing TAT between the pre-AI (blue box) and post-AI (red box) periods. In the overall case **(a)**, the mean TAT considerably decreased after AI software implementation. The mean TATs of cases without **(b)** and with AIH **(c)** in the post-AI period were considerably shorter than those in the pre-AI period. An asterisk (*) indicates a statistically significant difference. AI, artificial intelligence; AIH, acute intracranial hemorrhage; TAT, turnaround time.

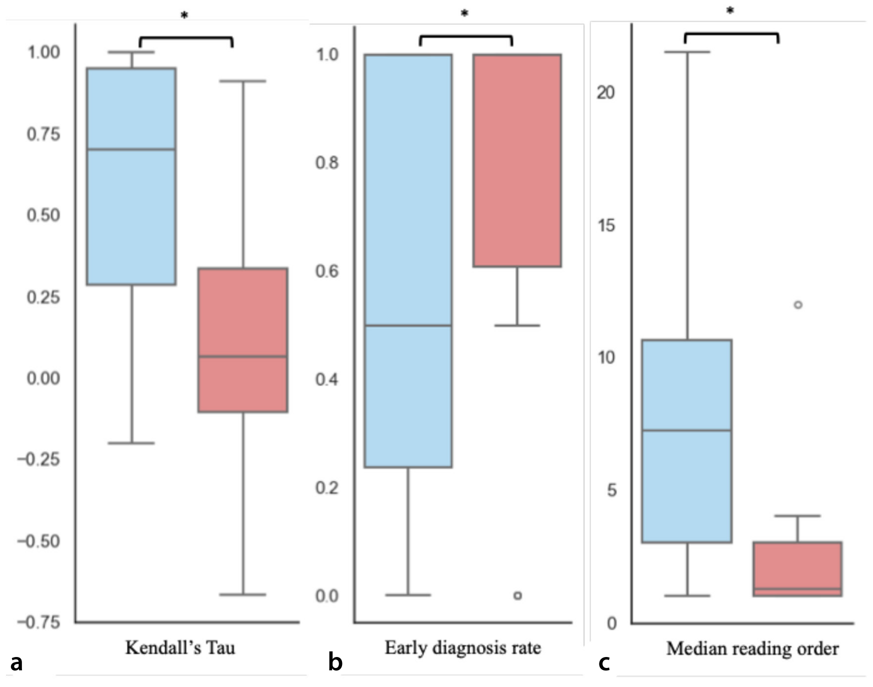


Figure 6. This figure illustrates box and whisker plot charts comparing variables between the pre-AI (blue box) and post-AI (red box) periods. **(a)** Kendall's Tau between the chronological order and **(b)** the radiologists' reading order considerably decreased, whereas the early diagnosis rate for AIH considerably increased after AI software implementation. **(c)** The median reading order of AIH decreased substantially after implementation. An asterisk (*) indicates a statistically significant difference. AI, artificial intelligence; AIH, acute intracranial hemorrhage.

tution using machines from a single vendor, which may limit the generalizability of the study. Additionally, radiologists' experience levels, institutional CT workflow protocols, and the availability of technical support may vary greatly across centers, potentially influencing both diagnostic performance and the impact of AI-driven prioritization.[30] Third, our statistical analysis of daily comparisons for radiologists' performance and workflow, while logically sound, may have unpredictably weakened statistical power by reducing the sample size from hundreds to dozens. To maintain statistical robustness without sacrificing temporal granularity, future research could employ rolling averages, time-series models that account for intraday variability, or extend the study period. Finally, this comparison study focused solely on the daily impact of AI software assistance on AIH detection within the radiologists' workflow and did not assess broader real-world challenges. For instance, integrating AI into clinical workflows requires substantial computational resources and careful implementation planning. Therefore, additional prospective multicenter trials involving multiple vendors, a larger reader cohort, and diverse clinical settings are needed to mitigate potential selection bias and improve generalizability.[30,31]

In conclusion, the integration of CT-based AI software for detecting AIH considerably enhanced the prioritization of radiologists' reading order and accelerated their interpretation of AIH cases while maintaining diagnostic performance by optimizing workflows in real-world clinical settings. Consequently, with the increasing number of CT scans and the growing demands placed on radiologists, AI software is expected to improve workflow efficiency and support the prompt and effective treatment of patients with AIH.

Acknowledgement

We sincerely thank two contributors - Dr. Hosuk Song and Dr. Hokyun Byun for their valuable contributions and support in the development of this manuscript.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

This work was supported by the National Research Foundation of Korea funded by the Korean government (MSIT) [RS-2023-00208409]; and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [HR22C173401].

References

1. Waqas M, Vakharia K, Munich SA, et al Initial emergency room triage of acute ischemic stroke. *Neurosurgery*. 2019;85(suppl_1):S38-S46. [Crossref]
2. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. *Lancet Neurol.* 2010;9(2):167-176. [Crossref]
3. Broderick JP, Brott TG, Duldner JE, Tomsick T, Huster G. Volume of intracerebral hemorrhage. A powerful and easy-to-use predictor of 30-day mortality. *Stroke*. 1993;24(7):987-993. [Crossref]
4. Forman R, Slota K, Ahmad F, et al. Intracerebral hemorrhage outcomes in the very elderly. *J Stroke Cerebrovasc Dis*. 2020;29(5):104695. [Crossref]
5. National Council on Radiation Protection and Measurements. Scientific Committee 6-2 on Radiation Exposure of the U.S. Population. *Ionizing Radiation Exposure of the Population of the United States*. Bethesda (MD): National Council on Radiation Protection and Measurements; 2009. (Report No.: 160). [Crossref]
6. Winder M, Owczarek AJ, Chudek J, Pilch-Kowalczyk J, Baron J. Are we overdoing it? Changes in diagnostic imaging workload during the years 2010-2020 including the impact of the SARS-CoV-2 pandemic. *Healthcare (Basel)*. 2021;9(11):1557. [Crossref]
7. Kansagra AP, Liu K, Yu JP. Disruption of radiologist workflow. *Curr Probl Diagn Radiol*. 2016;45(2):101-106. [Crossref]
8. Mamlouk MD, Saket RR, Hess CP, Dillon WP. Adding value in radiology: establishing a designated quality control radiologist in daily workflow. *J Am Coll Radiol*. 2015;12(8):838-841. [Crossref]
9. Kotter E, Ranschaert E. Challenges and solutions for introducing artificial intelligence (AI) in daily clinical workflow. *Eur Radiol*. 2021;31(1):5-7. [Crossref]
10. Balint BJ, Steenburg SD, Lin H, Shen C, Steele JL, Gunderman RB. Do telephone call interruptions have an impact on radiology resident diagnostic accuracy? *Acad Radiol*. 2014;21(12):1623-1628. [Crossref]
11. Yu JP, Kansagra AP, Mongan J. The radiologist's workflow environment: evaluation of disruptors and potential implications. *J Am Coll Radiol*. 2014;11(6):589-93. [Crossref]
12. Halsted MJ, Froehle CM. Design, implementation, and assessment of a radiology workflow management system. *AJR Am J Roentgenol*. 2008;191(2):321-7. [Crossref]
13. Agarwal S, Wood D, Grzeda M, et al. Systematic review of artificial intelligence for abnormality detection in high-volume neuroimaging and subgroup meta-analysis for intracranial hemorrhage detection. *Clin Neuroradiol*. 2023;33(4):943-956. [Crossref]
14. Mouridsen K, Thurner P, Zaharchuk G. Artificial intelligence applications in stroke. *Stroke*. 2020;51(8):2573-2579. [Crossref]
15. Segato A, Marzullo A, Calimeri F, De Momi E. Artificial intelligence for brain diseases: a systematic review. *APL Bioeng*. 2020;4(4):041503. [Crossref]
16. de Havenon A, Tirschwell DL, Heitsch L, et al. Variability of the modified Rankin scale score between day 90 and 1 year after ischemic stroke. *Neurol Clin Pract*. 2021;11(3):e239-e244. [Crossref]
17. Yun TJ, Choi JW, Han M, et al. Deep learning based automatic detection algorithm for acute intracranial haemorrhage: a pivotal randomized clinical trial. *NPJ Digit Med*. 2023;6(1):61. [Crossref]
18. Kim J, Oh SW, Lee HY, et al. Assessment of deep learning-based triage application for acute ischemic stroke on brain MRI in the ER. *Acad Radiol*. 2024;31(11):4621-4628. [Crossref]
19. Wesp W. Using STAT properly. *Radiol Manage*. 2006;28(1):26-30; quiz 31-33. [Crossref]
20. Gaskin CM, Patrie JT, Hanshew MD, Boatman DM, McWey RP. Impact of a reading priority scoring system on the prioritization of examination interpretations. *AJR Am J Roentgenol*. 2016;206(5):1031-1039. [Crossref]
21. Kotovich D, Twig G, Itsekson-Hayosh Z, et al. The impact on clinical outcomes after 1 year of implementation of an artificial intelligence solution for the detection of intracranial hemorrhage. *Int J Emerg Med*. 2023;16(1):50. [Crossref]
22. Zia A, Fletcher C, Bigwood S, et al. Retrospective analysis and prospective validation of an AI-based software for intracranial haemorrhage detection at a high-volume trauma centre. *Sci Rep*. 2022;12(1):19885. [Crossref]
23. Baltruschat I, Steinmeister L, Nickisch H, et al. Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *Eur Radiol*. 2021;31(6):3837-3845. [Crossref]
24. McWey RP, Hanshew MD, Patrie JT, Boatman DM, Gaskin CM. Impact of a four-point order-priority score on imaging examination performance times. *J Am Coll Radiol*. 2016;13(3):286-95.e5. [Crossref]
25. Maltby J, Williams G, McGarry J, Day L. Research methods for nursing and healthcare. Routledge; 2014. [Crossref]

26. Kim B, Romeijn S, van Buchem M, Mehrizi MHR, Grootjans W. A holistic approach to implementing artificial intelligence in radiology. *Insights Imaging*. 2024;15(1):22. [Crossref]
27. Savage CH, Tanwar M, Elkassem AA, et al. Prospective evaluation of artificial intelligence triage of intracranial hemorrhage on noncontrast head CT examinations. *AJR Am J Roentgenol*. 2024;223(5):e2431639. [Crossref]
28. Yang HK, Ko Y, Lee MH, et al. Initial performance of radiologists and radiology residents in interpreting low-dose (2-mSv) appendiceal CT. Erratum in: *AJR Am J Roentgenol*. 2016;206(4):901. [Crossref]
29. Labus S, Altmann MM, Huisman H, et al. A concurrent, deep learning-based computer-aided detection system for prostate multiparametric MRI: a performance study involving experienced and less-experienced radiologists. *Eur Radiol*. 2023;33(1):64-76. [Crossref]
30. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging (Bellingham)*. 2023;10(6):061104. [Crossref]
31. Panayides AS, Amini A, Filipovic ND, et al. AI in medical imaging informatics: current challenges and future directions. *IEEE J Biomed Health Inform*. 2020;24(7):1837-1857. [Crossref]

DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242720

PEDIATRIC RADIOLOGY

ORIGINAL ARTICLE

Copyright@Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Feasibility study of computed high b-value diffusion-weighted magnetic resonance imaging for pediatric posterior fossa tumors

Semra Delibalta[1]
Barış Genç[2]
Meltem Ceyhan Bilgici[2]
Kerim Aslan[2]

[1]Acıbadem Mehmet Ali Aydınlar University Faculty of Medicine, Atakent Hospital, Department of Radiology, İstanbul, Türkiye

[2]Ondokuz Mayıs University Faculty of Medicine, Department of Radiology, Samsun, Türkiye

PURPOSE

To evaluate the diagnostic efficacy of computed diffusion-weighted imaging (DWI) in pediatric posterior fossa tumors generated using high b-values.

METHODS

We retrospectively performed our study on 32 pediatric patients who had undergone brain magnetic resonance imaging for a posterior fossa tumor between January 2016 and January 2022. The DWIs were evaluated for each patient by two blinded radiologists. The computed DWI (cDWI) was mathematically derived using a mono-exponential model from images with b = 0 and 1,000 s/mm^2 and high b-values of 1,500, 2,000, 3,000, and 5,000 s/mm^2. The posterior fossa tumors were divided into two groups, low grade and high grade, and the tumor/thalamus signal intensity (SI) ratios were compared. The Mann–Whitney U test and receiver operating characteristic (ROC) curves were used to compare the diagnostic performance of the acquired DWI (DWI$_{1000}$), apparent diffusion coefficient (ADC)$_{1000}$ maps, and cDWI (cDWI$_{1500}$, cDWI$_{2000}$, cDWI$_{3000}$, and cDWI$_{5000}$).

RESULTS

The comparison of the two tumor groups revealed that the tumor/thalamus SI ratio on the DWI$_{1000}$ and cDWI (cDWI$_{1500}$, cDWI$_{2000}$, cDWI$_{3000}$, and cDWI$_{5000}$) was statistically significantly higher in high-grade tumors ($P < 0.001$). In the ROC curve analysis, higher sensitivity and specificity were detected in the cDWI$_{1500}$, cDWI$_{2000}$, cDWI$_{3000}$, and ADC$_{1000}$ maps (100%, 90.90%) compared with the DWI$_{1000}$ (80%, 81.80%). cDWI$_{3000}$ had the highest area under the curve (AUC) value compared with other parameters (AUC: 0.976).

CONCLUSION

cDWI generated using high b-values was successful in differentiating between low-grade and high-grade posterior fossa tumors without increasing imaging time.

CLINICAL SIGNIFICANCE

cDWI created using high b-values can provide additional information about tumor grade in pediatric posterior fossa tumors without requiring additional imaging time.

KEYWORDS

Computed diffusion-weighted imaging, high b-value, magnetic resonance imaging, pediatric posterior fossa tumors, synthetic diffusion-weighted imaging

Corresponding author: Semra Delibalta

E-mail: drsemradelibalta@gmail.com

Received 07 March 2024; revision requested 14 April 2024; last revision received 07 July 2024; accepted 22 July 2024.

Epub: 02.09.2024

Publication date: 08.09.2025

DOI: 10.4274/dir.2024.242720

Pediatric brain tumors are the most common childhood solid tumors and are frequently located in the posterior fossa.[1,2] The most common tumors in the posterior fossa in children are medulloblastoma (MB), pilocytic astrocytoma (PA), and ependymoma.[3,4]

Although conventional magnetic resonance imaging (MRI) is necessary for the diagnosis of brain tumors and the evaluation of their extent and location, it provides limited information on tumor type and grade.[5] Advanced MRI techniques such as diffusion-weighted imaging (DWI) contribute to the differential diagnosis of these tumors. Diffusion restriction and

You may cite this article as: Delibalta S, Genç B, Ceyhan Bilgici M, Aslan K. Feasibility study of computed high b-value diffusion-weighted magnetic resonance imaging for pediatric posterior fossa tumors. *Diagn Interv Radiol.* 2025;31(5):526-531.

low apparent diffusion coefficient (ADC) values are found more prominently in high-grade tumors with high cellularity than in low-grade tumors.[6] However, when using standard b-values (b = 1,000 s/mm^2), overlaps can be observed in the signal intensity (SI) of high-grade and low-grade tumors.[7,8] When DW images obtained using high b-values (b = 3,000 s/mm^2) and standard b-values in the differential diagnosis of high-grade and low-grade gliomas were compared, more successful results were obtained in examinations with high b-values.[9] However, at a field strength of 1.5T, higher b-values result in low image quality and a low signal-to-noise ratio (SNR).[10,11] Computed DWI (cDWI) is a synthetic DWI mathematically derived from an acquired DWI with two different b-values.[12] Synthetic DWI with high b-values exhibits stronger diffusion effects at a higher SNR than images obtained using existing b-values and can be generated without additional scanning time.[13,14] Studies have demonstrated that cDWI has improved lesion prominence compared with conventional DWI when examining the brain and other body regions.[13,15-19] To the best of our knowledge, no studies have investigated the diagnostic performance of calculated high b-values in pediatric posterior fossa tumors. In the present study, we aimed to evaluate the diagnostic performance of cDWI generated using high b-values in pediatric posterior fossa tumors.

Methods

This study was approved by the Ethics Committee of Ondokuz Mayıs University Faculty of Medicine and was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines (October 26, 2022, number: 2022/467). The requirement for informed consent was waived.

Main points

- Compared with images generated using existing b-values, synthetic diffusion-weighted imaging (DWI) with high b-values exhibits greater diffusion effects at a higher signal-to-noise ratio and may be produced without additional scanning time.
- The use of computed DWI (cDWI) with high b-values can help distinguish between low-grade and high-grade tumors without requiring more imaging time.
- For differentiating between low-grade and high-grade posterior fossa tumors, cDWI$_{1500}$, cDWI$_{2000}$, and cDWI$_{3000}$ perform better as diagnostic tools than the acquired DWI$_{1000}$ and apparent diffusion coefficient$_{1000}$ maps.

Patients

This study was conducted retrospectively in a single center after approval from the Ethics Committee, and the report was drafted in accordance with the Standards for Reporting of Diagnostic Accuracy Studies guidelines.[20] Between January 2016 and January 2022, 32 pediatric patients who had undergone preoperative brain MRI for posterior fossa tumors and who had not received treatment were included in the study. One patient without a histopathological diagnosis was excluded from the study, and three patients were excluded from the study because artifacts affected the evaluation of the DW images. Finally, 28 patients were included in the study (Figure 1).

Based on the World Health Organization 2021 classification, the patients were divided into two groups: low grade (grade 1 and 2 tumors) and high grade (grade 3 and 4 tumors).[21] The mean age of the low-grade tumor group was 7.5 ± 3.9 years (eight girls: 7 ± 3.4 years; five boys: 8.1 ± 4.0 years), and the mean age of the high-grade tumor group was 9.2 ± 4.3 years (six girls: 9.1 ± 5.1 years; nine boys: 9.2 ± 4.5 years).

Magnetic resonance imaging examination

All examinations were performed using 1.5T MRI (Achieva, Philips Healthcare, Best, Netherlands and Magnetom, SIEMENS AG, Erlangen, Germany) devices. All acquisitions were performed in the multiparametric MRI protocol, using T1WI, T2WI, fluid attenuated inversion recovery, dynamic contrast enhanced MRI, and DWI sequences. The acquisition parameters of the DWI are summarized in Table 1. cDWI was created based on images with b = 0 and 1,000 s/mm^2, with high b-values of 1,500, 2,000, 3,000, and 5,000 s/mm^2, using the mono-exponential model established in a study produced by our team.[14]

Image analysis

Images were evaluated by two radiologists, with evaluations and measurements performed independently of each other's assessment and without knowledge of the tumor pathology. Precontrast T2, precontrast T1, and postcontrast T1WIs were analyzed, and tumor boundaries were established while assessing the cystic, hemorrhagic, and necrotic components of the tumor. Using the volume of interest (VOI) approach and ITK-SNAP, measurements were taken from the solid portion of the tumor using DWI.[22] Similar measurements were calculated manually using ITK-SNAP software from the acquired DWI$_{1000}$, cDWI (b = 1,500, 2,000, 3,000, and 5000 s/mm^2), and ADC$_{1000}$ maps. In each patient, the tumor and thalamus SI ratio was calculated by measuring the right thalamus using the VOI method.

Statistical analysis

The IBM SPSS (version 22; IBM, Armonk, NY, USA) software program was used in all calculations. The Shapiro–Wilk test was used in all statistical studies to verify normal distribution. Descriptive statistics of the data are presented as n (%), and for normalized variables, mean ± standard deviation values are

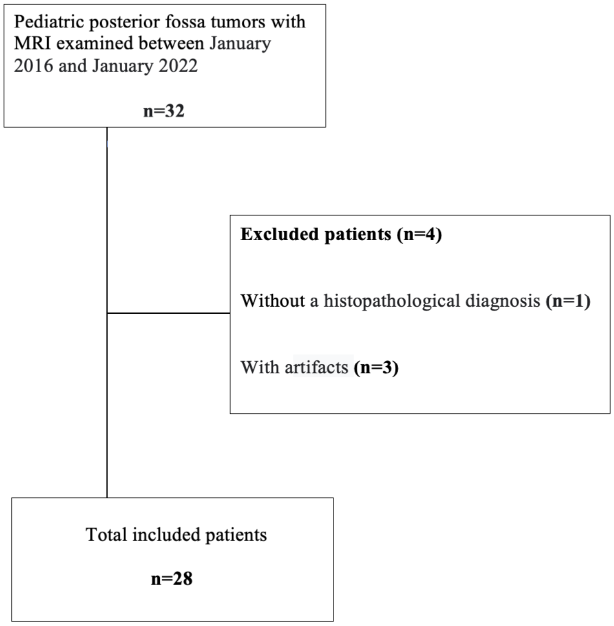

Pediatric posterior fossa tumors with MRI examined between January 2016 and January 2022
n=32
Excluded patients (n=4)
Without a histopathological diagnosis (n=1)
With artifacts (n=3)
Total included patients
n=28

Figure 1. Study flowchart. MRI, magnetic resonance imaging.

provided, whereas for non-normalized variables, the median (min–max) is provided. The Mann–Whitney U test was used for data with normal distribution, comparing the tumor/thalamus SI ratios of high-grade and low-grade tumors on DWI_{1000}, $cDWI_{1500}$, $cDWI_{2000}$, $cDWI_{3000}$, $cDWI_{5000}$, and ADC_{1000} maps. The receiver operating characteristic (ROC) curve was calculated for the diagnostic performance of DWI_{1000}, $cDWI_{1500}$, $cDWI_{2000}$, $cDWI_{3000}$, $cDWI_{5000}$, and ADC_{1000} maps in differentiating high–low grade tumors with the area under the curve (AUC). Youden's index was used to select the optimal predicted probability cut-off. The sensitivity and specificity of the DWI and ADC maps were calculated by determining the cut-off value using an ROC curve analysis. Interobserver correlation was evaluated using the intraclass correlation (ICC) coefficient, and κ values were interpreted as follows: κ = 0.00–0.20, slight agreement; κ = 0.21–0.40, fair agreement; κ = 0.41–0.60, moderate agreement; κ = 0.61–0.80, substantial agreement; and κ = 0.81–1.00, almost perfect agreement.[23] A *P* value <0.05 was considered statistically significant.

Results

In total, 13 low-grade [PA = 7 (54%), posterior fossa ependymoma (grade 2) = 3 (23%), low-grade tumor-diffuse astrocytoma = 3 (23%)] and 15 high-grade [MB = 13 (87%), posterior fossa ependymoma (grade 3) = 1 (1%), glioblastoma = 1 (1%)] tumors were included in our study. The tumor/thalamus SI ratios (median and min–max values) for DWI_{1000}, $cDWI_{1500}$, $cDWI_{2000}$, $cDWI_{3000}$, and $cDWI_{5000}$ in low-grade and high-grade tumors are reported in Table 2. The median (min–max) SI rates were higher in the high-grade tumors than in the low-grade tumors (P < 0.001). When the two tumor groups were compared, the tumor/thalamus SI ratio distributions were more clearly distinguished at higher b-values than at b = 1,000 s/mm^2 (Figure 2). In the ICC test, the kappa value was found to be greater than 0.75 for all parameters, with an almost perfect correlation between 0.82 and 0.95 (*P* < 0.001 for each comparison) (Table 2).

Table 1. Diffusion-weighted imaging sequence parameters

ssEPI DWI b1000		
Parameters	PHILIPS achieva	SIEMENS magnetom
Field of view (mm × mm)	240 × 240	229 × 229
Matrix	192 × 192	192 × 192
Slice thickness	3.50 mm	5 mm
Repetition time	4,200 ms	4,200 ms
Echo time	72 ms	105 ms
Flip angle	90°	90°
Calculated b-values	b1500, b2000, b3000, b5000	b1500, b2000, b3000, b5000

ssEPI, single-shot echo-planar imaging; DWI, diffusion-weighted imaging.

Table 2. Tumor/thalamus signal intensity ratios in diffusion-weighted imaging (DWI) and computed diffusion-weighted imaging at different b-values

Parameters	Low-grade tumors (n = 13) Median (min–max values)	High-grade tumors (n = 15) Median (min–max values)	*P*	ICC (κ values)	*P* (for ICC)
DWI_{1000}	1.09 (0.90–1.71)	1.62 (1.16–2.17)	<0.001	0.82	<0.001
$cDWI_{1500}$	1.00 (0.75–1.70)	1.75 (1.18–2.27)	<0.001	0.89	<0.001
$cDWI_{2000}$	0.82 (0.46–0.70)	1.89 (1.21–2.39)	<0.001	0.91	<0.001
$cDWI_{3000}$	0.59 (0.25–1.69)	1.99 (1.24–2.99)	<0.001	0.94	<0.001
$cDWI_{5000}$	0.38 (0.08–1.66)	2.81 (1.24–6.10)	<0.001	0.95	<0.001

The Mann–Whitney U test was used to compare the tumor/thalamus signal intensity ratios of high-grade and low-grade tumors. The intraclass correlation (ICC) was used to assess interobserver correlation. DWI, diffusion-weighted imaging; cDWI, computed diffusion-weighted imaging.

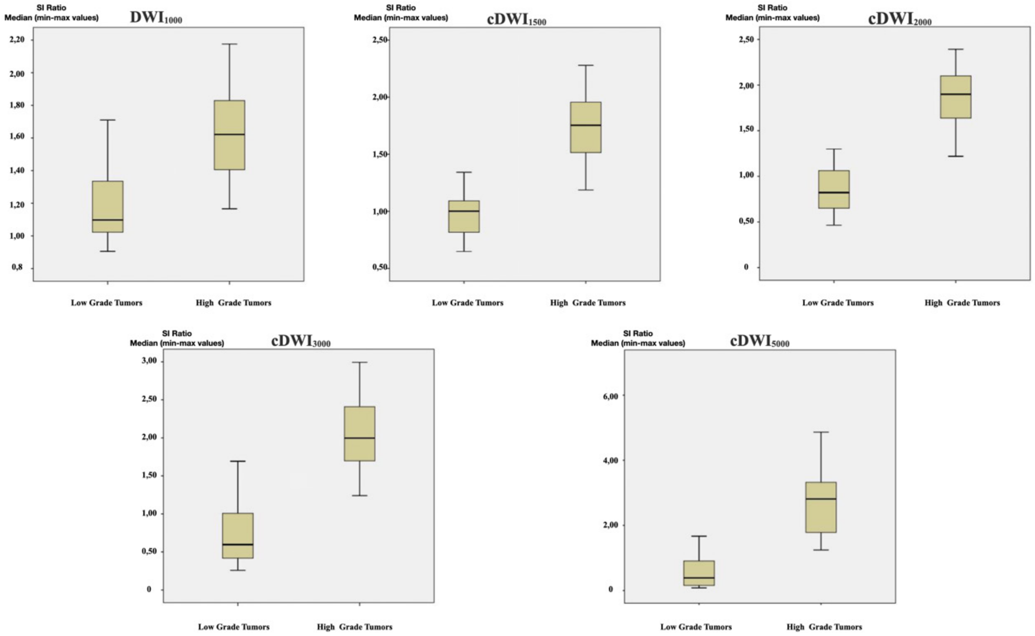


Figure 2. Box plot comparing tumor/thalamus signal intensity ratios in high-grade and low-grade tumors. Compared with diffusion-weighted imaging $(DWI)_{1000}$, the difference between the two groups is more pronounced in computed DWI ($cDWI_{1500}$, $cDWI_{2000}$, $cDWI_{3000}$, and $cDWI_{5000}$). SI, signal intensity.

In the ADC_{1000} maps, median (min–max) ADC values were found to be lower in the high-grade tumors than in the low-grade tumors [low-grade tumor: 1.1 (0.5–1.6) × 10^{-3} mm^2/s; high-grade tumor: 0.8 (0.6–1.0) × 10^{-3} mm^2/s, $P < 0.001$].

In the ROC curve analysis of the DWI_{1000}, cDWI, and ADC_{1000} maps, the AUC values (Figure 3) were found to be statistically significant in all parameters. The AUC value was higher in $cDWI_{3000}$ than in other parameters (AUC: 0.976, $P < 0.001$). In the ROC curve analysis, when optimal cut-off values were used, higher sensitivity and specificity were detected in cDWI (b = 1,500, 2,000, and 3,000 s/mm^2; 100%, 90.9%) than in DWI_{1000} (80%, 81.80%). The ADC_{1000} maps (100%, 90.90%) revealed higher sensitivity and specificity than DWI_{1000} (80%, 81.80%), whereas $cDWI_{5000}$ (93%, 81.80%) displayed higher sensitivity than DWI_{1000} but similar specificity (80%, 81.80%) (Table 3). The DWI_{1000}, cDWI, and ADC_{1000} maps of the two patients diagnosed with juvenile PA and MB are presented in Figures 4 and 5, respectively.

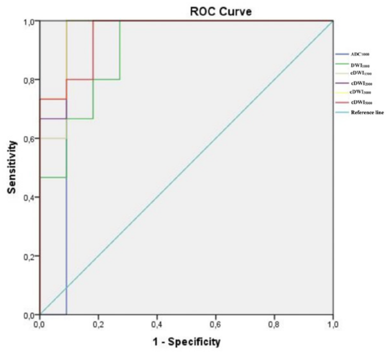


Figure 3. Receiver operating characteristic (ROC) curves at different b-values.

Table 3. Receiver operating characteristic curve analysis results on diffusion-weighted imaging (DWI), computed DWI (cDWI), and apparent diffusion coefficient (ADC)$_{1000}$ maps

Parameters	AUC	Cut-off	Sensitivity(%)	Specificity(%)	*P*
DWI_{1000}	0.903	1.36	80	81.80	<0.001
$cDWI_{1500}$	0.964	1.15	100	90.90	<0.001
$cDWI_{2000}$	0.97	1.15	100	90.90	<0.001
$cDWI_{3000}$	0.976	1.20	100	90.90	<0.001
$cDWI_{5000}$	0.958	1.22	93	81.80	<0.001
ADC_{1000}	0.909	0.00108	100	90.90	<0.001

AUC, area under the curve.

Discussion

In our study, we evaluated the benefits of cDWI created using high b-values for pediatric posterior fossa tumors compared with acquired DWI with standard b-values (b = 1,000 s/mm^2). We determined that the ADC_{1000} maps, DWI_{1000}, $cDWI_{1500}$, $cDWI_{2000}$, $cDWI_{3000}$, and $cDWI_{5000}$ were effective in distinguishing low–high grade tumors. Notably, our study determined that $cDWI_{3000}$ had a higher AUC value for diagnostic performance in the ROC curve analysis than other parameters. As demonstrated in Table 2, as b-values increased, the tumor/thalamus SI ratios decreased in low-grade tumors and increased in high-grade tumors. When compared with images using b = 1,000 s/mm^2, which are frequently used in standard examinations, we observed that the difference between the two groups increased with the increase in b-value. When compared with normal parenchyma areas, with the increase in b-values, a more significant signal reduction was observed in low-grade tumors (Figure 4) and a more pronounced signal in high-grade tumors (Figure 5). As a result, with the increase in b-values, a more significant contrast difference occurred between tumor and normal tissue. In a study using acquired DW images with b = 1,000 and b = 3,000 s/mm^2 on 3T MR to compare low-grade and high-grade differentiation in brain tumors, improved diagnostic performance (high sensitivity and specificity) was demonstrated with higher b-values.[9]

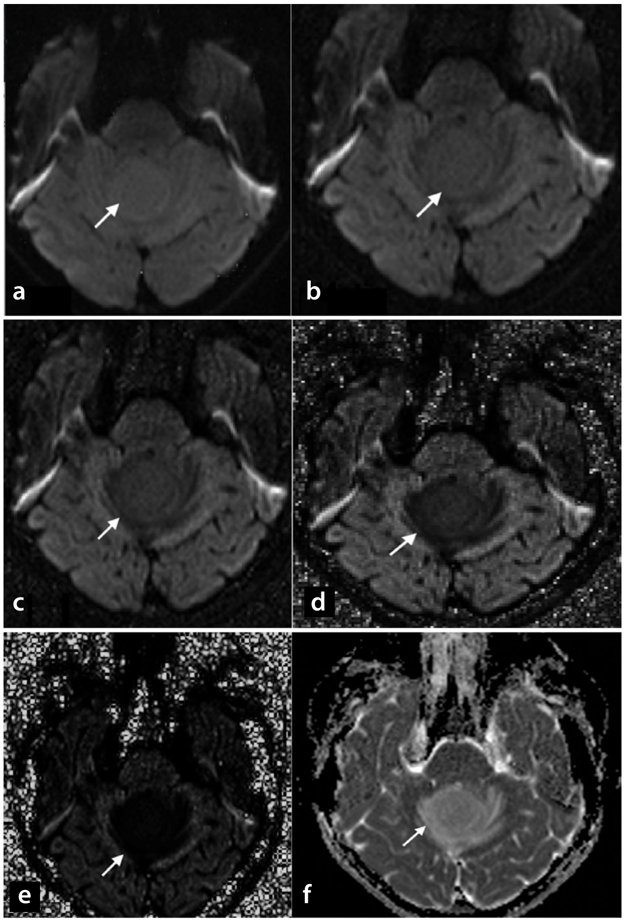


Figure 4. A 15-year-old female with juvenile pilocytic astrocytoma. **(a)** Diffusion-weighted imaging (DWI)$_{1000}$, **(b)** computed DWI (cDWI)$_{1500}$, **(c)** $cDWI_{2000}$, **(d)** $cDWI_{3000}$, **(e)** $cDWI_{5000}$, and **(f)** apparent diffusion coefficient (ADC)$_{1000}$ maps. In the mass located in the 4th ventricle, indicated by the arrow, on DWI with increased b-values, the signal loss of the tumor tissue is more prominent than that of the parenchyma. In addition, a high signal is observed in the parenchyma in the ADC_{1000} maps **(f)**.

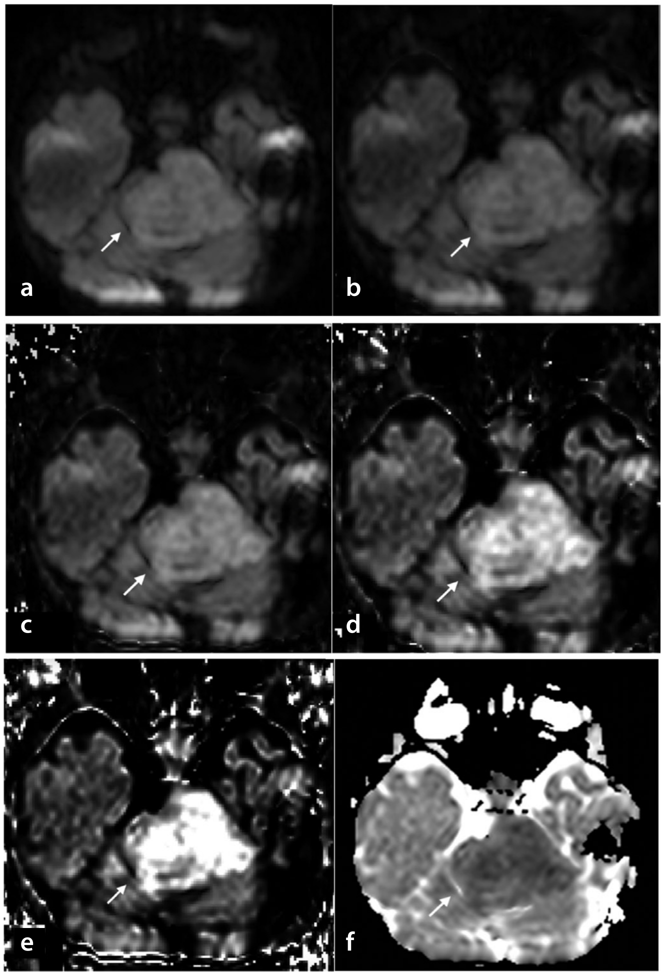


Figure 5. A 13-year-old female with medulloblastoma. **(a)** Diffusion-weighted imaging (DWI)$_{1000}$, **(b)** computed DWI (cDWI)$_{1500}$, **(c)** cDWI$_{2000}$, **(d)** cDWI$_{3000}$, **(e)** cDWI$_{5000}$, and **(f)** apparent diffusion coefficient (ADC)$_{1000}$ maps. In the mass located in the 4th ventricle, indicated by the arrow, as the b-values increase, the signal becomes evident in the tumor tissue, whereas a decrease in the signal is observed in the parenchyma. In addition, a low signal is observed in the parenchyma in the ADC$_{1000}$ maps **(f)**.

In the present study, higher sensitivity and specificity were identified in cDWI$_{1500}$, cDWI$_{2000}$, and cDWI$_{3000}$ compared with DWI$_{1000}$ without increasing acquisition time.

Previous studies on pediatric posterior fossa tumors and gliomas have revealed that high-grade tumors can be effectively distinguished from low-grade tumors with minimal ADC values.[8,24-26] In the present study, the median (min–max) ADC values were found to be lower in high-grade tumors than in low-grade tumors, which is consistent with the literature. In addition, in the present study, in the ROC curve analysis, we determined that cDWI$_{1500}$, cDWI$_{2000}$, and cDWI$_{3000}$ had higher AUC values than ADC$_{1000}$ maps, although cDWI$_{1500}$, cDWI$_{2000}$, and cDWI$_{3000}$ had similar sensitivity and specificity with ADC$_{1000}$ maps (Table 3).

In a study comparing cDWI and acquired DWI in patients with ischemic stroke, cDWI$_{1000}$ and cDWI$_{1500}$ had higher image quality and lesion prominence than acquired DWI$_{1000}$. However, in the present study, DWI$_{2000}$ and cDWI$_{2500}$ were not found to be an alternative to conventional DWI because of the low lesion detection rates.[15] Kamata et al.[16] reported that cDWI$_{3000}$ was more useful than DWI$_{1000}$ in diagnosing pediatric encephalitis/encephalopathy, and they obtained similar results for acquired DWI$_{3000}$. In a study investigating synthetic b-values in breast imaging, synthetic images for b1000 and b2000 were obtained and compared with acquired DWI$_{850}$. The results demonstrated that lesion prominence and image quality were optimal in cDWI$_{1200}$ and cDWI$_{1800}$. In breast imaging, improved lesion visibility and background suppression are theoretically expected with increasing b-values.[13] Similarly, in a study investigating diagnostic sensitivity in breast cancer, cDWI$_{1500}$ was found to be more sensitive than acquired DWI$_{1500}$.[27] In addition, Daimiel Naranjo et al.[28] revealed that cDWI$_{1200}$ increased the visibility of the tumor without increasing the scanning time, especially in dense breast tissue. In a study on prostate cancer, cDWI with a high b-value was compared with acquired DWI to detect SI differences between cancer and normal tissue, with cDWI identified as more effective. This study verified that cDWI had a better contrast ratio than real images with a high b-value.[29] This study has several limitations. First, it was retrospective and therefore cDWI at high b-values could not be compared with acquired DWI at high b-values. Second, in the literature, measurements have been calculated using region of interest and compared with ADCmin values. We used VOI in our study, which might produce some differences compared with the literature. Third, our study is the first on cDWI in brain tumors, and the results should be verified through further studies.

In conclusion, the present study demonstrated that the diagnostic performance of cDWI$_{1500}$, cDWI$_{2000}$, and cDWI$_{3000}$ is stronger in the differentiation of low-grade and high-grade posterior fossa tumors than that of acquired DWI$_{1000}$ and ADC$_{1000}$ maps. Moreover, the SI ratio between tumor and normal tissue became more pronounced with increasing b-values. Thus, cDWI created with high b-values can contribute to the differential diagnosis of low-grade and high-grade tumors without increasing the imaging time.

Conflict of interest disclosure

Meltem Ceyhan Bilgici, MD, is Section Editor in Diagnostic and Interventional Radiology. She had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

1. Moussalem C, Ftouni L, Abou Mrad Z, et al. Pediatric posterior fossa tumors outcomes: experience in a tertiary care center in the Middle East. *Clin Neurol Neurosurg*. 2020;197:106170. [CrossRef]
2. Paldino MJ, Faerber EN, Poussaint TY. Imaging tumors of the pediatric central nervous system. *Radiol Clin Nort Am*. 2011;49(4):589-616. [CrossRef]
3. Prasad KSV, Ravi D, Pallikonda V, Raman BVS. Clinicopathological study of pediatric posterior fossa tumors. *J Pediatr Neurosci*. 2017;12(3):245. [CrossRef]
4. Bidiwala S, Pittman T. Neural network classification of pediatric posterior fossa tumors using clinical and imaging data. *Pediatr Neurosurg*. 2004;40(1):8-15. [CrossRef]
5. Cha S. Update on brain tumor imaging: from anatomy to physiology. *AJNR Am J Neuroradiol*. 2006;27(3):475-487. [CrossRef]

6. Provenzale JM, Mukundan S, Barboriak DP. Diffusion-weighted and perfusion MR imaging for brain tumor characterization and assessment of treatment response. *Radiology*. 2006;239(3):632-649. [CrossRef]

7. Kono K, Inoue Y, Nakayama K, et al. The role of diffusion-weighted imaging in patients with brain tumors. *AJNR Am J Neuroradiol*. 2001;22(6):1081-1088. [CrossRef]

8. Jaremko JL, Jans LB, Coleman LT, Ditchfield MR. Value and limitations of diffusion-weighted imaging in grading and diagnosis of pediatric posterior fossa tumors. *AJNR Am J Neuroradiol*. 2010;31(9):1613-1616. Erratum in: *AJNR Am J Neuroradiol*. 201;31(10):E90. [CrossRef]

9. Seo HS, Chang KH, Na DG, Kwon BJ, Lee DH. High b-value diffusion (b = 3000 s/mm^2) MR imaging in cerebral gliomas at 3T: visual and quantitative comparisons with b = 1000 s/mm^2. *AJNR Am J Neuroradiol*. 2008;29(3):458-463. [CrossRef]

10. Kim HJ, Choi CG, Lee DH, Lee JH, Kim SJ, Suh DC. High-b-value diffusion-weighted MR imaging of hyperacute ischemic stroke at 1.5T. *AJNR Am J Neuroradiol*. 2005;26(2):208-215. [CrossRef]

11. Meyer JR, Gutierrez A, Mock B, et al. High-b-value diffusion-weighted MR imaging of suspected brain infarction. *AJNR Am J Neuroradiol*. 2000;21(10):1821-1829. [CrossRef]

12. Blackledge MD, Leach MO, Collins DJ, Koh DM. Computed diffusion-weighted MR imaging may improve tumor detection. *Radiology*. 2011;261(2):573-581. [CrossRef]

13. Bickel H, Polanec SH, Wengert G, et al. Diffusion-weighted MRI of breast cancer: improved lesion visibility and image quality using synthetic b-values. *J Mag Reson Imaging*. 2019;50(6):1754-1761. [CrossRef]

14. Higaki T, Nakamura Y, Tatsugami F, et al. Introduction to the technical aspects of computed diffusion-weighted imaging for radiologists. *Radiographics*. 2018;38(4):1131-1144. [CrossRef]

15. Sartoretti T, Sartoretti E, Wyss M, et al. Diffusion-weighted MRI of ischemic stroke at 3T: value of synthetic b-values. *Br J Radiol*. 2021;94(1121):20200869. [CrossRef]

16. Kamata Y, Shinohara Y, Kuya K, et al. Computed diffusion-weighted imaging for acute pediatric encephalitis/encephalopathy. *Acta radiol*. 2019;60(10):1341-1347. [CrossRef]

17. Fukukura Y, Kumagae Y, Hakamada H, et al. Computed diffusion-weighted MR imaging for visualization of pancreatic adenocarcinoma: comparison with acquired diffusion-weighted imaging. *Eur J Radiol*. 2017;95:39-45. [CrossRef]

18. Takeuchi M, Matsuzaki K, Harada M. Computed diffusion-weighted imaging for differentiating decidualized endometrioma from ovarian cancer. *Eur J Radiol*. 2016;85(5):1016-1019. [CrossRef]

19. Jendoubi S, Wagner M, Montagne S, et al. MRI for prostate cancer: can computed high b-value DWI replace native acquisitions? *Eur Radiol*. 2019;29(10):5197-5204. [CrossRef]

20. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. [CrossRef]

21. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol*. 2021;23(8):1231-1251. [CrossRef]

22. Yushkevich PA, Pashchinskiy A, Oguz I, et al. User-guided segmentation of multi-modality medical imaging datasets with ITK-SNAP. *Neuroinformatics*. 2019;17(1):83-102. [CrossRef]

23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174. [CrossRef]

24. Rumboldt Z, Camacho DL, Lake D, Welsh CT, Castillo M. Apparent diffusion coefficients for differentiation of cerebellar tumors in children. *AJNR Am J Neuroradiol.* 2006;27(6):1362-1369. [CrossRef]

25. Lee EJ, Lee SK, Agid R, Bae JM, Keller A, Terbrugge K. Preoperative grading of presumptive low-grade astrocytomas on MR imaging: diagnostic value of minimum apparent diffusion coefficient. *AJNR Am J Neuroradiol*. 2008;29(10):1872-1877. [CrossRef]

26. Chen Z, Ma L, Lou X, Zhou Z. Diagnostic value of minimum apparent diffusion coefficient values in prediction of neuroepithelial tumor grading. *J Magn Reson Imaging*. 2010;31(6):1331-1338. [CrossRef]

27. Park JH, Yun B, Jang M, et al. Comparison of the diagnostic performance of synthetic versus acquired high b-Value (1500 s/mm2) diffusion-weighted MRI in women with breast cancers. *J Magn Reson Imaging*. 2019;49(3):857-863. [CrossRef]

28. Daimiel Naranjo I, Lo Gullo R, Saccarelli C, et al. Diagnostic value of diffusion-weighted imaging with synthetic b-values in breast tumors: comparison with dynamic contrast-enhanced and multiparametric MRI. *Eur Radiol*. 2021;31(1):356-367. [CrossRef]

29. Ueno Y, Takahashi S, Ohno Y, et al. Computed diffusion-weighted MRI for prostate cancer detection: the influence of the combinations of b-values. *Br J Radiol*. 2015;88(1048):20140738. [CrossRef]