EISSN 1305-3612

DIR

Diagnostic and Interventional Radiology

TSR
1924
TURKISH SOCIETY
OF RADIOLOGY

dirjournal.org

VOLUME **32**
ISSUE **2**
March 2026

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, 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

Social Media Lead Editor

Tümay Bekçi, MD
Department of Radiology, Giresun University Faculty of Medicine, Giresun, Türkiye

Social Media Editors

Mahmut Bilal Doğan, MD
Department of Radiology, Marmara University Faculty of Medicine, İstanbul, Türkiye

Kendal Erincik, MD
Department of Radiology, İstanbul Medeniyet University Faculty of Medicine, İstanbul, Türkiye

Şule Gümürdü, MD
Department of Radiology, İstanbul Medeniyet University Faculty of Medicine, İstanbul, Türkiye

Gürsel Karacaoğlu, MD
Department of Radiology, İstanbul Medeniyet University Faculty of Medicine, İstanbul, Türkiye

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)*

**Aytekin Besim, MD, 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 Okan Akhan, MD, Editor in Chief.

Diagnostic and Interventional Radiology

Official Journal of the Turkish Society of Radiology

E-ISSN: 1305-3612

www.dirjournal.org

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)

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 (MEDLINE 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, Scopus, DOAJ, EBSCO, HINARI, EMBASE, CINAHL, ProQuest, Gale and CNKI..**

The journal is published online.

Owner: 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: March 2026

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.

DIR
Diagnostic and Interventional
Radiology

Volume 32 • Issue 2

March 2026

www.dirjournal.org

Contents

EDITORIAL

132 Editorial **A three-year summary of the *Diagnostic and Interventional Radiology Journal*.** *Mehmet Ruhi Onur*

ABDOMINAL IMAGING

133 Original Article **Multiparametric magnetic resonance imaging, diffusion-weighted magnetic resonance imaging, and magnetic resonance elastography: differentiating benign and malignant liver lesions.** *Avaz Jabiyev, Muşturay Karçaaltıncaba, Ali Devrim Karaosmanoğlu, Deniz Akata, Mustafa Nasuh Özmen, İlkay Sedakat İdilman*

139 Original Article **Role of ureteral wall thickness and computed tomography imaging in predicting spontaneous passage of ureteral stones.** *Özlem Kadırhan, Sonay Aydın, Ercüment Keskin, Mecit Kantarcı*

149 Original Article **Abbreviated liver magnetic resonance imaging with a second-shot arterial phase image to assess the viability of treated hepatocellular carcinoma after non-radiation locoregional therapy.** *Il Wan Son, Seung Baek Hong, Nam Kyung Lee, Suk Kim, Hyung Il Seo, Young Mok Park, Byeong Gwan Noh, Jong Hyun Lee*

156 Original Article **Texture analysis enhances diagnostic accuracy of lesions scored as 5 in the Prostate Imaging Reporting and Data System in magnetic resonance imaging.** *Yan Bai, Xin Ru Xie, Ying Hou, Yu Dong Zhang, Hai Bin Shi, Chen Jiang Wu*

ARTIFICIAL INTELLIGENCE AND INFORMATICS

164 Review **Systematic review of artificial intelligence competitions in radiology: a focus on design, evaluation, and trends.** *Muhammed Said Beşler, Ural Koç*

171 Invited Commentary **Reporting checklists: from a tool after publication to a tool before submission.** *Jingyu Zhong*

173 Original Article **Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties.** *Yasin Celal Güneş, Turay Cesur, Eren Çamur*

BREAST IMAGING

182 Original Article **Findings of suspicious calcifications on contrast-enhanced mammography and their pathological correlation.** *Dilşah Oral, İhsan Şebnem Örgüç, Hanife Seda Mavili, Teoman Coşkun*

191 Original Article **Magnetic resonance imaging-based artificial intelligence model predicts neoadjuvant therapy response in triple-negative breast cancer.** *Raşit Eren Büyüktoka, Zehra Hilal Adıbelli, Murat Sürücü, Özlem Özdemir, Yalçın İşler, Ali Murat Koç, Aslı Dilara Büyüktoka, Demet Kocatepe Çavdar, Özge Aslan, Serhat Değer, Ayşenur Oktay*

HEAD AND NECK IMAGING

200 Original Article **Magnetic resonance imaging findings in Ménière's disease: the impact of radiologist experience on hydrops imaging.** *Çağatay Cihan, Uğur Toprak, Emre Emekli, Armağan İncesulu, Hamit İpek*

INTERVENTIONAL RADIOLOGY

208 Original Article **Physiological indices for evaluating balloon angioplasty outcomes in below-the-knee artery lesions of patients with chronic limb-threatening ischemia.** *Murat Canyiğit, Muhammed Said Beşler, Turan Kaya*

214 Review **Short to mid-term outcomes of flow re-direction endoluminal device X (FRED™ X) in the management of intracranial aneurysms: a meta-analysis.** *Alperen Elek, Gülfem Nur Dindar, Sidar Karagöz, Semra Yücel, Eda Teomete, Celal Çınar, Mahmut Küsbeci, Egemen Öztürk, İsmail Oran*

221 Original Article **Efficacy and safety of ultrasound-guided bedside percutaneous cholecystostomy using the transhepatic approach and trocar technique in patients with acute cholecystitis.** *Ahmet Cem Demirşah, Berat Acu, Çiğdem Öztunalı*

228 Original Article **Pleural tail sign in computed tomography-guided lung biopsy: an imaging predictor of severe pneumothorax requiring chest tube placement.** *Jacob Jalil Hassan, Jakob Leonhardi, Timm Denecke, Anne Beeskow, Manuel Florian Struck, Anne-Kathrin Höhn, Sebastian Krämer, Armin Frille, Hans-Jonas Meye*r

PEDIATRIC RADIOLOGY

233 Original Article **Diagnostic accuracy of convolutional neural network algorithms to distinguish gastrointestinal obstruction on conventional radiographs in a pediatric population.** *Ercan Ayaz, Hasan Güçlü, Ayşe Betül Oktay*

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 2026; DOI: 10.4274/dir.2026.263962

EDITORIAL

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

A three-year summary of the *Diagnostic and Interventional Radiology Journal*.

Mehmet Ruhi Onur
Editor-in-Chief

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

More than three years ago, at the end of 2022, the executive committee of the Turkish Society of Radiology appointed me as the Chief Editor of *Diagnostic and Interventional Radiology* (*DIR*). I felt this was the most gratifying offer for my academic career. Then, starting in early 2023, I began my duties alongside my esteemed colleagues in the editorial board.

The performance of a scientific journal can generally be evaluated under four criteria: the number of articles submitted, the time taken to review articles, the publication time after acceptance, and the journal's citation status. When evaluating the number of submissions to our journal, we observed a decrease in 2023 compared to 2022, likely due to the introduction of the article processing charge (APC) at the end of 2022. However, we anticipate a gradual increase in submissions for 2024 and 2025. Regarding article review times, we achieved the fastest review time in 2025, compared to the previous three years.

As for the journal's citation factor, like many scientific journals post-pandemic, we experienced a decrease; in 2023, our citation factor was 1.4. As the editorial board, we assessed this situation and focused on publishing articles that contribute to the literature, possess high scientific quality, and address current topics. This strategy yielded results: the journal's impact factor for publications in 2025 was 1.7, and it is expected to reach 2.6 by June 2026. The rising immediacy index values in recent years (2024 - 2.1; 2025 - 2.5) indicate the potential for an increased impact factor in the coming years.

Beyond numerical data representing submission criteria, evaluation processes, publication periods, and impact factors, perhaps the most crucial determinant of a scientific journal's place in international literature is its adherence to scientific principles in article evaluation and publication. Over the past three years, *DIR* has consistently upheld its scientific publishing principles, maintaining a self-citation rate of between 0.9% and 2%, thereby preventing artificial inflation of its impact factor.

I owe a debt of gratitude to the Turkish Society of Radiology, the owner of *DIR*, for allowing the journal's editorial board to operate with complete scientific freedom. I would also like to thank our section editors, who meticulously reviewed the articles submitted during my tenure as Editor-in-Chief, as well as our reviewers, the authors, and our readers, whose feedback has contributed to the journal's development. My thanks extend to Galenos Publishing House, which successfully managed all publishing processes for our journal.

It is a great honor for me to hand over the editorship of *DIR* to Dr. Şükrü Mehmet Ertürk, who has successfully represented our country in the international radiology community and has published numerous scientific articles and book chapters while also serving as a book editor. I am confident that Dr. Ertürk, who previously served as our journal's publication coordinator, will lead the journal towards achieving its goals with his vision and leadership. I congratulate him and wish the new editorial board every success.

You may cite this article as: Onur MR. Summary of the last three years in *Diagnostic and Interventional Radiology*. *Diagn Interv Radiol*. 2026;32(2):132.

DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253324

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Multiparametric magnetic resonance imaging, diffusion-weighted magnetic resonance imaging, and magnetic resonance elastography: differentiating benign and malignant liver lesions

Avaz Jabiyev
Muşturay Karçaaltıncaba
Ali Devrim Karaosmanoğlu
Deniz Akata
Mustafa Nasuh Özmen
İlkay Sedakat İdilman

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

PURPOSE

This study investigates the accuracy of multiparametric magnetic resonance imaging (mpMRI), diffusion-weighted imaging (DWI), and magnetic resonance elastography (MRE) in differentiating benign and malignant liver lesions.

METHODS

This retrospective study included patients with focal liver lesions who underwent MRI and MRE between 2018 and 2022. Based on histopathologic analyses or follow-up imaging findings, 70 solid liver lesions were retrospectively evaluated as benign (n = 20) or malignant (n = 50).

RESULTS

There was no statistically significant difference between the benign and malignant liver lesions in pre-contrast T1 relaxation times ($P > 0.05$). Malignant liver lesions had a significantly lower T2 value, contrast-enhancement ratio (CER), T1 relaxation time reduction (T1D), T1D percentage [T1D (%)], and apparent diffusion coefficient (ADC), along with a significantly higher stiffness value ($P < 0.05$). In receiver operating characteristic analysis, the following cut-off values were determined for differentiating malignant from benign lesions: a CER of 1.99 [area under the curve (AUC): 0.828, sensitivity 78.6%, specificity 73.2%], a T1D of 749.5 ms (AUC: 0.817, sensitivity 71.4%, specificity 78%), a T1D (%) reduction of 49.71% (AUC: 0.831, sensitivity 78.6%, specificity 73.2%), a T2 relaxation time of 74 ms (AUC: 0.705, sensitivity 65%, specificity 76.6%), an ADC of 1.275×10^{-3} mm^2/s (AUC: 0.861, sensitivity 89.5%, specificity 81.2%), and a stiffness of 3.77 kPa (AUC: 0.848, sensitivity 85%, specificity 75%).

CONCLUSION

Combined mpMRI, DWI, and MRE provide high diagnostic accuracy, with ADC and MRE offering superior performance in differentiating malignant from benign liver lesions.

CLINICAL SIGNIFICANCE

This article highlights the accuracy of mpMRI, MRE, and DWI in distinguishing between malignant and benign liver lesions. These findings support the integration of mpMRI, DWI, and MRE into clinical practice for non-invasive liver lesion characterization.

KEYWORDS

Multiparametric magnetic resonance imaging, liver, diffusion-weighted imaging, magnetic resonance elastography, focal lesion

Corresponding author: İlkay Sedakat İdilman

E-mail: isidilman@hacettepe.edu.tr

Received 16 March 2025; revision requested 10 April 2025; accepted 28 April 2025.

Epub: 03.06.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253324

You may cite this article as: Jabiyev A, Karçaaltıncaba A, Karaosmanoğlu AD, Akata D, Özmen MN, İdilman İS. Multiparametric magnetic resonance imaging, diffusion-weighted magnetic resonance imaging, and magnetic resonance elastography: differentiating benign and malignant liver lesions. *Diagn Interv Radiol*. 2026;32(2):133-138.

The detection of focal liver lesions (FLLs) is one of the most commonly encountered findings in abdominal imaging in clinical practice. Although the majority of liver lesions in non-cirrhotic livers are benign, an FLL can sometimes represent the first indication of metastatic liver disease from an unknown primary malignancy. Since management strategies differ substantially based on the lesion's nature, it is crucial to differentiate malignant lesions from benign ones. Although specific imaging characteristics are associated with typical benign and malignant FLLs, atypical findings may complicate diagnoses and cause unnecessary anxiety for both patients and physicians.

Magnetic resonance imaging (MRI) is recognized as the most accurate radiological method for characterizing liver lesions.[1] MRI examinations routinely incorporate diffusion-weighted imaging (DWI) along with conventional sequences, using gadolinium-based extracellular or hepatospecific contrast agents in post-contrast multiphase studies to evaluate FLLs. Although hepatospecific contrast agents share properties with extracellular agents in dynamic imaging, the additional diagnostic information obtained during the hepatospecific phase enhances the differential diagnosis of liver lesions and improves the detection of small FLLs.[2] The combination of apparent diffusion coefficient (ADC) values, which decrease in malignancy, further improves the accuracy of MRI in characterizing FLLs.[3] Emerging MRI techniques in liver imaging, such as multiparametric MRI (mpMRI)–including T1, T2, and T2* mapping–and magnetic resonance elastography (MRE), have proven effective in imaging diffuse liver diseases.[4-6] However, the literature includes limited studies evaluating the characteristics of MRE and mpMRI and their roles in differentiating FLLs.[7-9] In this retrospective study, we aimed to demonstrate the mpMRI, MRE, and DWI characteristics of FLLs and evaluate the role of these quantitative measures in their characterization.

Main points

- Emerging magnetic resonance imaging techniques can effectively aid in distinguishing between malignant and benign liver lesions.
- Malignant liver lesions exhibit significantly lower T2, contrast-enhancement ratio, T1 relaxation time reduction (T1D), T1D percentage, and apparent diffusion coefficient (ADC) values while showing notably higher stiffness values.
- ADC values and lesion stiffness demonstrate slightly better performance in differentiating malignant from benign liver lesions.

Methods

This retrospective study received approval from the Hacettepe University Ethics Committee on July 26, 2022 (GO 22/380). A total of 50 patients with 70 lesions who underwent liver MRI, MRE, and mpMRI between January 1, 2018, and February 21, 2022, were included. Indications for MRI included suspicion of an FLL on ultrasound or computed tomography, follow-up imaging for chronic liver disease, and preoperative or follow-up imaging in patients with primary tumors outside the liver. Patients under 18 years of age; those who had undergone chemoembolization, radioembolization, or radiofrequency ablation; and those with lesions smaller than 1 cm (to avoid partial volume artifacts) were excluded from the study (Figure 1).

Magnetic resonance imaging examinations

All MRI examinations were performed on a 1.5-T system (Magnetom Aera, Siemens Healthcare, Erlangen, Germany). A 30-channel phased-array body coil was used, and patients were scanned in the supine position. All patients underwent MRI after a fasting period of 4–6 hours. A 3-plane localization gradient echo sequence was performed at the beginning of the examination. The standard liver MRI protocol included in-phase and out-of-phase sequences, coronal T2 HASTE, axial fat-suppressed T2, DWI, axial 3D dynamic T1, and axial and coronal hepatobiliary phase images obtained at the 20th minute after administration of Gd-EOB-DTPA (Primovist; Bayer-Schering Pharma AG, Berlin, Germany).

DWI and ADC mapping were performed at b-values of 50, 400, and 800 s/mm^2. The sequence parameters were as follows: a repetition time (TR) of 6200 ms, an echo time (TE) of 54 ms, a flip angle of 60°, a field of view (FOV) of 380 × 300 mm^2, a slice thickness of 8 mm, a matrix size of 192 × 144, a number of excitations (NEX) of 3, and a total acquisition time of 3 minutes.

Furthermore, T1 mapping was conducted using a B1 inhomogeneity-corrected method with variable flip angles. The sequence parameters were as follows: a TR of 4.4 ms, a TE of 2.1 ms, flip angles of 3° and 15°, a matrix size of 256 × 156, a FOV of 380 × 300 mm, a slice thickness of 4 mm, and an acquisition time of 1.5 minutes. For T2 mapping, various TEs were used with an SSFP-based true fast imaging with steady precession sequence and an exponential signal decay model. The parameters were as follows: a TR of 166 ms; TEs of 0, 25, and 55 ms; a flip angle of 70°; an FOV of 420 × 260 mm; a slice thickness of 10 mm; a matrix size of 192 × 192; a NEX of 1; and an acquisition time of 1.2 minutes. In addition, T2* mapping, used to evaluate hepatic iron load, was performed with the following parameters: a TR of 200 ms; TEs of 0.93, 2.1, 3.35, 4, 4.56, 5.77, 6.98, 8.19, 9.4, 10.61, 11.82, 13.03, and 14.24 ms; a flip angle of 20°; a slice thickness of 10 mm; an FOV of 400 × 300 mm; and a matrix size of 160 × 85.

MRE was performed using an active driver that generated mechanical waves at 60 Hz and a modified 2D gradient-recalled echo

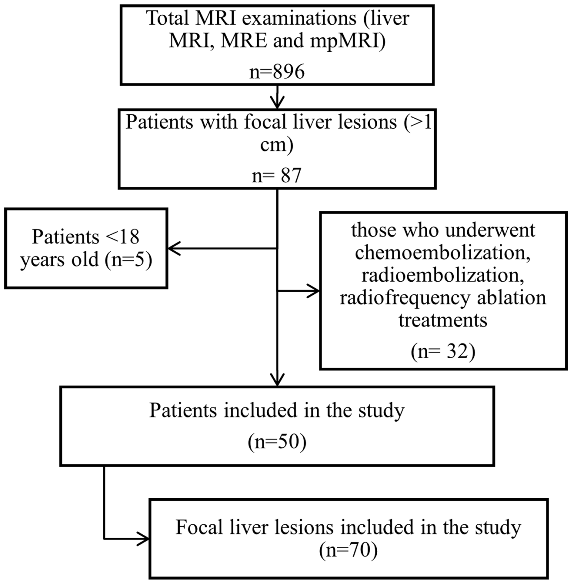

Total MRI examinations (liver MRI, MRE and mpMRI)
n=896
Patients with focal liver lesions (>1 cm)
n= 87
Patients <18 years old (n=5)
those who underwent chemoembolization, radioembolization, radiofrequency ablation treatments
(n= 32)
Patients included in the study
(n=50)
Focal liver lesions included in the study
(n=70)

Figure 1. Study flowchart. MRI, magnetic resonance imaging; MRE, magnetic resonance elastography; mpMRI, multiparametric magnetic resonance imaging.

sequence. The sequence parameters were as follows: a TR of 50 ms, a TE of 21 ms, a flip angle of 25°, a bandwidth of 31.25 kHz, a matrix size of 256 × 128, and an acquisition time of 2.5 minutes. Depending on liver size, four slices, each 10-mm thick, were obtained from the largest portion of the liver during a breath-hold. All MRI, mpMRI, MRE, and DWI sequences were performed during the same imaging session.

Imaging analysis

All data were transferred to a workstation (Syngo.via Siemens, Erlangen, Germany) for analysis. The MR images were reviewed by one radiologist (A.J., with 5 years of experience) under the supervision of a senior radiologist with 16 years of experience (I.S.I.). Lesion measurements were performed using a free-hand region of interest (ROI) that included a sufficiently large portion of the lesion while maintaining a thin margin outside the lesion's periphery to avoid partial volume artifacts. Free-hand ROIs were also drawn on the magnitude images to include FLLs and were copied onto the stiffness map, which provided liver stiffness values in kPa.

The average T1 relaxation time values before and after contrast–pre-contrast (pre-T1 value) and at 20 minutes post-contrast on hepatobiliary phase images (post-T1 value)–were used to calculate the contrast-enhancement ratio (CER), as previously described by Yoshimura et al.[10] Additionally, the decrease in T1 relaxation time [T1 relaxation time reduction (T1D)] and the percentage reduction in T1 relaxation time [T1D (%)] were calculated from these measurements, as outlined by Peng et al.[11] Subsequently, ADC values were calculated for the lesions using diffusion-weighted images. This measurement was performed using an ROI that included a sufficiently large portion of the lesion while preserving a thin margin outside the lesion's periphery on the ADC mapping images, in consensus with two experienced readers (A.J. and İ.S.İ.).

Statistical analysis

Statistical analysis was performed using SPSS version 23.0 (IBM Inc., Armonk, NY, USA) and Microsoft Excel (Microsoft Corporation, 2018). Categorical variables were summarized as counts and percentages, whereas continuous variables were expressed as means and standard deviations (minimum–maximum). The Student's t-test was used to compare normally distributed numerical variables between two independent groups, and the Mann–Whitney U test was used for non-normally distributed variables. Receiver operating characteristic (ROC) curves were used to assess the diagnostic performance of the MRI parameters, and the optimal threshold value was identified to maximize sensitivity and specificity in distinguishing malignant from benign lesions. The areas under the ROC (AUROC) curves were calculated, and the difference between two independent AUROC curves was evaluated using z statistics (http://vassarstats.net/roc_comp.html). For all tests, a two-tailed *P* value of less than 0.05 was considered statistically significant.

Results

A total of 70 solid lesions from 50 patients (mean age: 54.3 ± 13.7 years) were included in the evaluation. Of the patients, 16 (32%) were women and 34 (68%) were men. Seventeen patients had multiple lesions (14 had two lesions, and 3 had three lesions) assessed simultaneously. The lesions were classified as benign or malignant based on the histopathology (30%) or typical imaging characteristics and follow-up findings (70%). In total, 20 lesions (28.6%) were classified as benign and 50 lesions (71.4%) as malignant.

Among the benign lesions, 16 (80%) were hemangiomas, 3 (15%) were focal nodular hyperplasia, and 1 (5%) was a hepatocellular adenoma. Among the malignant lesions, 30 (60%) were metastases, 15 (30%) were hepatocellular carcinoma (HCC), 3 (6%) were lymphoma, and 2 (4%) were cholangiocarcinoma. All patients with HCC had chronic liver disease, with 14 diagnosed with cirrhosis.

The mean pre-T1 and post-T1 values were 1,338.8 ± 393.9 and 719.5 ± 260.1 ms, respectively. The mean T2 value was 70.5 ± 19.8 ms. The mean CER was 2.02 ± 0.7, the mean T1D was 634.45 ± 367.44 ms, and the mean percentage T1D was 44.77% ± 17.30%. The mean ADC value was 1.23 ± 0.46 × 10^{-3} mm^2/s, and the mean lesion stiffness was 4.6 ± 1.6 kPa. Malignant lesions had significantly lower T2 , CER, T1D, T1D (%), and ADC values and significantly higher stiffness values ($P < 0.05$). The characteristics of the patient population are summarized in Table 1 and Figure 2.

ROC analysis for the differentiation of malignant versus benign lesions demonstrated that the mean T2, CER, T1D, T1D (%), ADC, and lesion stiffness values all had an AUROC curve greater than 0.6 (Table 2, Figure 3). The mean ADC and lesion stiffness performed slightly better [area under the curve (AUC): 0.861 and 0.848, respectively] than CER, T1D, and T1D (%) (AUC: 0.828, 0.817, and 0.831, respectively) in differentiating malignant from benign lesions. There was no statistically significant difference between the ADC and MRE AUCs for differentiating malignant from benign lesions ($z = 0.12$, $P = 0.904$).

Discussion

In this study, we evaluated the characteristics of FLLs using mpMRI, DWI, and MRE. We highlighted the effectiveness of these

Table 1. Magnetic resonance imaging characteristics of liver lesions

Parameter	All lesions (n = 70)	Benign lesions (n = 20)	Malignant lesions (n = 50)	*P* value
Pre-T1 value (ms) (n = 70)	1,338.8 ± 393.9 (674–2484)	1,474.9 ± 465.4	1,284.3 ± 351.9	0.110
Post-T1 value (ms) (n = 55)	719.5 ± 260.1 (284–1611)	629.6 ± 230.9	750.2 ± 265.0	0.135
T2 value (ms) (n = 64)	70.5 ± 19.8 (42–122)	82.5 ± 23.3	65.0 ± 15.4	**0.005**
CER (n = 55)	2.02 ± 0.7 (1.04–4.32)	2.71 ± 0.86	1.79 ± 0.52	**0.002**
T1D (ms) (n = 55)	634.5 ± 367.4 (52–1586)	964.0 ± 379.6	521.9 ± 290.6	**<0.001**
T1D (%) (n = 55)	44.8 ± 17.3 (3.96–77.12)	59.4 ± 13.3	39.8 ± 15.7	**<0.001**
ADC (×10^{-3} mm^2/s) (n = 67)	1.23 ± 0.46 (0.290–2.271)	1.60 ± 0.29	1.10 ± 0.43	**<0.001**
MR elastography (kPa) (n = 25)	4.6 ± 1.6 (2.3–8.9)	3.6 ± 1.2	5.4 ± 1.6	**0.004**

Values are presented as mean ± standard deviation, with ranges in parentheses. CER, contrast-enhancement ratio; T1D, T1 relaxation time reduction; T1D (%), T1 relaxation time reduction percentage; ADC, apparent diffusion coefficient.

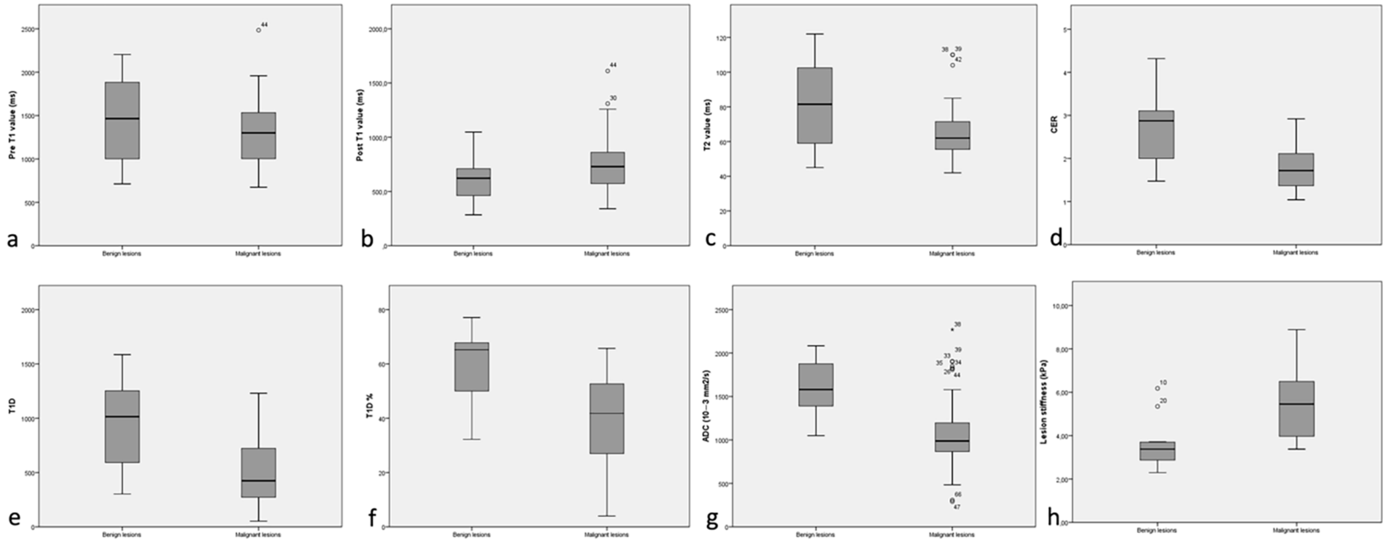


Figure 2. Bar graphs comparing mpMRI, ADC, and MRE measurements between benign and malignant lesions: **(a)** pre-T1 value, **(b)** post-T1 value, **(c)** T2 value, **(d)** CER, **(e)** T1D, (f) T1D (%), **(g)** ADC, **(h)** MRE. mpMRI, multiparametric magnetic resonance imaging, ADC, apparent diffusion coefficient; MRE, magnetic resonance elastography; CER, contrast-enhancement ratio; T1D, T1 relaxation time reduction; T1D (%), T1 relaxation time reduction percentage; ADC, apparent diffusion coefficient.

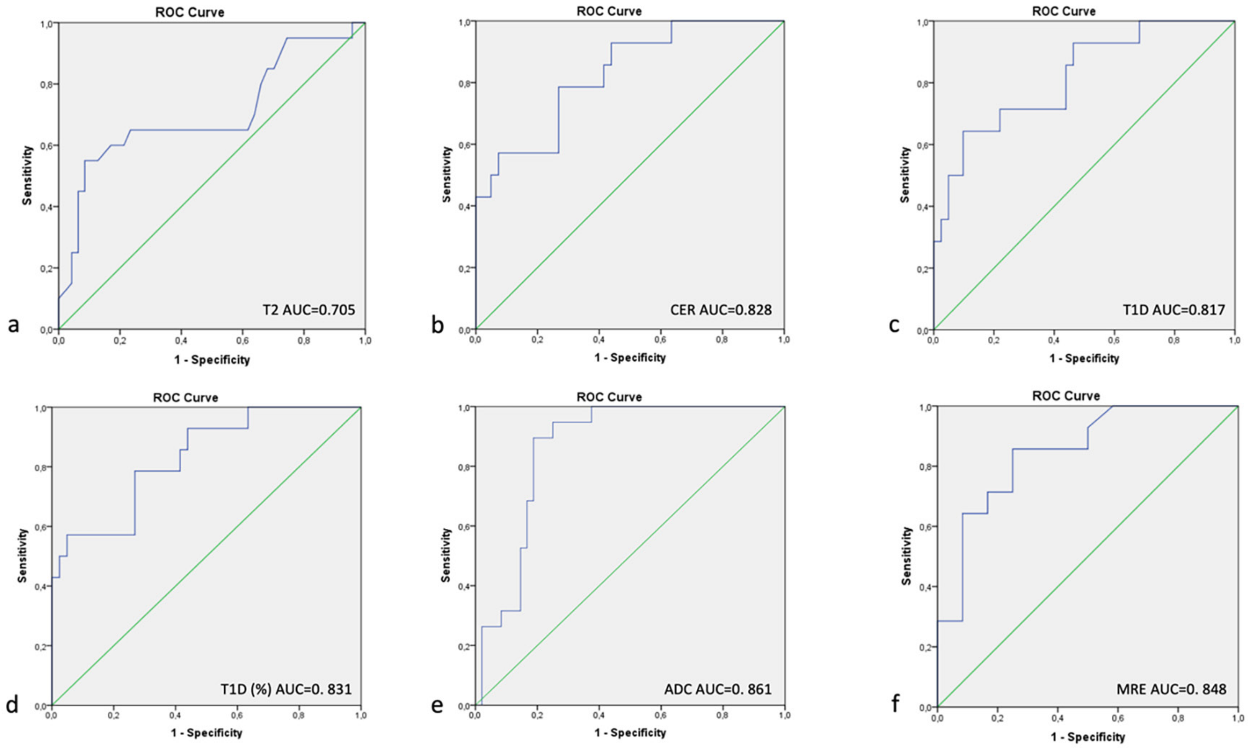


Figure 3. ROC curves for the differentiation of malignant versus benign focal liver lesions: **(a)** T2 relaxation time, **(b)** CER, **(c)** T1D, **(d)** T1D (%), **(e)** ADC, **(f)** MRE. ROC, receiver operating characteristic; CER, contrast-enhancement ratio; T1D, T1 relaxation time reduction; T1D (%), T1 relaxation time reduction percentage; ADC, apparent diffusion coefficient; MRE, magnetic resonance elastography.

imaging techniques in distinguishing between malignant and benign liver lesions. Our findings indicated that malignant lesions had significantly lower T2, CER, T1D, T1D (%), and ADC values while exhibiting significantly higher stiffness values ($P < 0.05$). Notably, the mean ADC and lesion stiffness performed slightly better, with AUC values of 0.861 and 0.848, respectively, compared with T2, CER, T1D, and T1D (%), which had AUC values of 0.705, 0.828, 0.817, and 0.831, respectively, in differentiating malignant from benign liver lesions.

Several studies have investigated the role of mpMRI in differentiating various FLLs. In a previous study, Mio et al.[12] demonstrated that T1 mapping using the phase-sensitive inversion recovery technique was useful for the differential diagnosis of hemangiomas, liver parenchymal cysts, HCC, and metastases. They found that T1D (%) values were high in hemangiomas, similar to our findings, whereas lower values were observed in HCC and metastases. A threshold value >50 was

Table 2. Cut-off values for the differentiation of malignant versus benign liver lesions

Parameter	Cut-off value	AUC (95% CI)	Sensitivity	Specificity
T2 relaxation time (ms)	≤74	0.705 (0.553–0.857)	0.650	0.766
CER	≤1.99	0.828 (0.705–0.950)	0.786	0.732
T1D (ms)	≤749	0.817 (0.687–0.947)	0.714	0.780
T1D (%)	≤49.71	0.831 (0.708–0.954)	0.786	0.732
ADC ($\times 10^{-3}$ mm^2/s)	≤1.275	0.861 (0.774–0.948)	0.895	0.812
MR elastography (kPa)	≥3.77	0.848 (0.697–0.999)	0.857	0.750

CER, contrast-enhancement ratio; T1D, T1 relaxation time reduction; T1D (%), T1 relaxation time reduction percentage; ADC, apparent diffusion coefficient; CI, confidence interval; MR, magnetic resonance.

established for differentiating hemangiomas from HCC, achieving 78.8% sensitivity and 100% specificity. A threshold >39 was found to distinguish hemangiomas from metastases, with 60% sensitivity and 97% specificity.

Yoshimura et al.[10] evaluated the role of CER in differentiating metastases from hemangiomas, reporting a threshold value of 1.6, with 100% sensitivity and 95% specificity. Peng et al.[11] assessed the role of T1D and T1D (%) using a dual flip angle VIBE 3D gradient echo sequence for differentiating HCC, focal nodular hyperplasia, and hemangiomas, finding that T1D and T1D (%) were significantly lower in HCC than in hemangiomas, consistent with our study. Wang et al.[13] compared the diagnostic value of T1 mapping and DWI for distinguishing benign and malignant FLLs. Significant differences were observed in native T1, enhanced T1, the percentage change in T1 relaxation time (ΔT1%), and ADC between benign and malignant FLLs. They also reported a similar ADC cut-off value (1.25 × 10^{-3} mm^2/s) for differentiating malignant lesions. Furthermore, they demonstrated that ADC was significantly positively correlated with T1 and ΔT1% and negatively correlated with enhanced T1.

Only one study has directly compared T2 values between malignant and benign FLLs, reporting significantly lower T2 and ADC values in malignant lesions. The researchers found a higher AUC for T2 (0.932), with a cut-off value of 107 ms, compared with an AUC of 0.874 for ADC with a cut-off of 1.25 × 10^{-3} mm^2/s.[14] In our study, however, the AUC for ADC was higher than in that report.

In our study, the threshold stiffness value for differentiating benign and malignant lesions was found to be 3.77 kPa or higher, with a sensitivity of 85% and specificity of 75%. Previous studies have reported similar findings regarding the differentiation of malignant and benign FLLs using MRE. Venkatesh et al.[7] conducted a preliminary study, revealing that malignant lesions had higher stiffness values, with a threshold of 5 kPa and 100% accuracy. Dominguez et al.[15] reported that benign and malignant liver lesions could be distinguished using a threshold value of 5.78–5.82 kPa, achieving 75%–85% accuracy, 64.7%–82.8% sensitivity, and 88% specificity. Another study by Hennedige et al.[8] evaluated 124 FLLs with MRE and DWI and observed significantly higher accuracy for MRE than for DWI (0.986 vs. 0.82, $P = 0.0016$). Abdelgawad et al.[16] also evaluated 124 FLLs using MRE and DWI. They found a strong negative correlation between the ADC of FLLs and MRE stiffness and reported a cut-off value of 4.23 kPa, with an AUC of 0.991 for MRE and a cut-off value of 1.43 × 10^{-3} mm^2/s with an AUC of 0.894 for DWI. Our study, based on a limited number of patients, observed similar AUCs for ADC and MRE, with no statistically significant difference.

The current study has several strengths and limitations. This is the first study to evaluate mpMRI, DWI, and MRE for differentiating malignant versus benign liver lesions, demonstrating various AUCs and enabling comparisons within the same population. The number of included lesions was limited, with most diagnosed based on follow-up rather than histopathologic evaluation. The retrospective nature of the study also led to restricted availability of MRE data for some lesions. However, the statistically significant findings, consistent with previous studies, underscore the importance of MRE in the differential diagnosis of benign and malignant FLLs. Additionally, subgroup analyses were not performed due to the limited sample size in each group. This highlights the need for more comprehensive prospective studies with a homogeneous distribution of lesion types to further clarify the role of mpMRI, DWI, and MRE in the characterization of FLLs.

In conclusion, mpMRI, DWI, and MRE can be used for the differentiation of solid liver lesions, with ADC and lesion stiffness performing slightly better than CER, T1D, and T1D (%). Comprehensive future studies involving a larger number of patients and lesions will enable the comparison of different techniques and demonstrate the impact of their combined application on diagnostic accuracy.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Pang EH, Harris AC, Chang SD. Approach to the solitary liver lesion: imaging and when to biopsy. *Can Assoc Radiol J*. 2016;67(2):130-148. [Crossref]
2. Hodler J, Kubik-Huch RA, von Schulthess GK, editors. Diseases of the Abdomen and Pelvis 2018-2021: Diagnostic Imaging - IDKD Book [Internet]. Cham (CH): Springer; 2018. [Crossref]
3. Taouli B, Koh DM. Diffusion-weighted MR imaging of the liver. *Radiology*. 2010;254(1):47-66. [Crossref]
4. Idilman IS, Li J, Yin M, Venkatesh SK. MR elastography of liver: current status and future perspectives. *Abdom Radiol (NY)*. 2020;45(11):3444-3462. [Crossref]
5. Banerjee R, Pavlides M, Tunnicliffe EM, et al. Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease. *J Hepatol*. 2014;60(1):69-77. [Crossref]
6. Thomaides-Brears HB, Lepe R, Banerjee R, Duncker C. Multiparametric MR mapping in clinical decision-making for diffuse liver disease. *Abdom Radiol (NY)*. 2020;45(11):3507-3522. [Crossref]
7. Venkatesh SK, Yin M, Glockner JF, Takahashi N, Araoz PA, Talwalkar JA, Ehman RL. MR elastography of liver tumors: preliminary results. *AJR Am J Roentgenol*. 2008;190(6):1534-1540. [Crossref]
8. Hennedige TP, Hallinan JT, Leung FP, et al. Comparison of magnetic resonance elastography and diffusion-weighted imaging for differentiating benign and malignant liver lesions. *Eur Radiol*. 2016;26(2):398-406. [Crossref]
9. Hectors SJ, Wagner M, Bane O, et al. Quantification of hepatocellular carcinoma heterogeneity with multiparametric magnetic resonance imaging. *Sci Rep*. 2017;7(1):2452. [Crossref]

10. Yoshimura N, Saito K, Saguchi T, et al. Distinguishing hepatic hemangiomas from metastatic tumors using T1 mapping on gadoxetic-acid-enhanced MRI. *Magn Reson Imaging*. 2013 Jan;31(1):23-27. [Crossref]

11. Peng Z, Li C, Chan T, et al. Quantitative evaluation of Gd-EOB-DTPA uptake in focal liver lesions by using T1 mapping: differences between hepatocellular carcinoma, hepatic focal nodular hyperplasia and cavernous hemangioma. *Oncotarget*. 2017;8(39):65435-65444. [Crossref]

12. Mio M, Fujiwara Y, Tani K, Toyofuku T, Maeda T, Inoue T. Quantitative evaluation of focal liver lesions with T1 mapping using a phase-sensitive inversion recovery sequence on gadoxetic acid-enhanced *MRI*. *Eur J Radiol Open*. 2020;8:100312. [Crossref]

13. Wang F, Yang Q, Zhang Y, Liu J, Liu M, Zhu J. 3D variable flip angle T1 mapping for differentiating benign and malignant liver lesions at 3T: comparison with diffusion weighted imaging. *BMC Med Imaging*. 2022;22(1):146. [Crossref]

14. Cieszanowski A, Anysz-Grodzicka A, Szeszkowski W, et al. Characterization of focal liver lesions using quantitative techniques: comparison of apparent diffusion coefficient values and T2 relaxation times. *Eur Radiol*. 2012;22(11):2514-24. [Crossref]

15. Dominguez A, Fino D, Spina JC, et al. Assessment of SE-MRE-derived shear stiffness at 3.0 Tesla for solid liver tumors characterization. *Abdom Radiol (NY)*. 2021;46(5):1904-1911. [Crossref]

16. Abdelgawad MS, Elseady BA, ELabd OL, Kohla MS, Samea MESA. Comparison of magnetic resonance elastography and diffusion-weighted imaging for differentiating benign and malignant liver lesions. *Egypt J Radiol Nucl Med*. 2024;55. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253304

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Role of ureteral wall thickness and computed tomography imaging in predicting spontaneous passage of ureteral stones

Özlem Kadırhan[1]
Sonay Aydın[1]
Ercüment Keskin[2]
Mecit Kantarcı[1,3]

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

[2] Erzincan Binali Yıldırım University Faculty of Medicine, Department of Urology, Erzincan, Türkiye

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

PURPOSE

Urolithiasis is a common health problem with a high recurrence rate, and effectively balancing follow-up with intervention is important for patient safety. In this context, our study aims to identify criteria that can predict the likelihood of spontaneous passage (SP) of ureteral stones.

METHODS

A retrospective analysis was performed on 2,773 patients who presented to our hospital with renal colic over a 4-year period. The study included 897 patients with unilateral ureteral stones measuring ≤10 mm, identified using non-contrast computed tomography, and inflammatory serum markers assessed through biochemical testing. Variables analyzed to predict the likelihood of SP included stone size, lateralization and location, ureteral wall thickness (UWT) at the stone site, stone density, degree of hydronephrosis (HN), ureteral length, parenchymal thickness and density, and various biochemical parameters.

RESULTS

It was determined that the SP of ureteral stones was considerably affected by larger stone size (right >6.5 mm, left >6 mm), higher stone density (>957 Hounsfield units), increased UWT (>1.7 mm), presence of high-grade HN (grade ≥2), and elevated neutrophil-lymphocyte ratio (NLR) (>2.15) and platelet-lymphocyte ratio (PLR) (>10.28) values in blood. No statistically significant relationship was observed between SP and ureteral length, renal parenchymal thickness, or renal parenchymal density. It was found that when the UWT at the level of the ureteral stone exceeded 1.7 mm, the risk of the stone not passing spontaneously increased by 706.5 times in univariate logistic regression (LR) analysis and by 337.9 times in multivariate LR analysis compared with individuals with a wall thickness below this threshold.

CONCLUSION

Our study demonstrated that, in addition to stone size and location, increased UWT at the stone level, higher stone density, the presence of concomitant high-grade HN, and elevated NLR and PLR values in the blood could be used as criteria to determine the likelihood of SP of ureteral stones. According to our results, UWT was shown to be a stronger risk factor for failure of SP than stone size.

CLINICAL SIGNIFICANCE

The findings indicate that wall thickness around ureteral stones is a risk factor with a higher negative predictive value for SP than the stone size and location.

KEYWORDS

Ureteral stone, spontaneous passage, ureteral wall thickness, stone density, neutrophil-lymphocyte ratio, platelet-lymphocyte ratio, computed tomography, hydronephrosis

Corresponding author: Özlem Kadırhan

E-mail: ozlemkkadirhan@gmail.com

Received 21 February 2025; revision requested 26 March 2025; accepted 02 June 2025.

Epub: 08.07.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253304

You may cite this article as: Kadırhan Ö, Aydın S, Keskin E, Kantarcı M. Role of ureteral wall thickness and computed tomography imaging in predicting spontaneous passage of ureteral stones. *Diagn Interv Radiol*. 2026;32(2):139-148.

Ureteral stones account for approximately 20% of all urinary tract stones. If stones formed in the kidney pass into the ureter, they may cause severe, sudden pain known as renal colic.[1] It has been reported that 5%–12% of individuals in developed countries present to the emergency department with renal colic at least once in their lifetime.[2-4] Urinary tract infection and sudden deterioration in renal function can also occur in the setting of urolithiasis.[1] In such cases, prompt identification of an appropriate treatment regimen based on the likelihood of spontaneous passage (SP) is crucial to prevent the potential development of urosepsis.

According to current guidelines, for patients with uncomplicated distal ureteral stones ≤10 mm, a period of observation or medical expulsive therapy (MET) is recommended. However, in some cases of urolithiasis, individualized treatment may be necessary.[5] This is because long-term observation, based on parameters such as stone size and location, may not be sufficient; if the stone does not pass spontaneously, patients may continue to experience severe colic pain, urosepsis, impaired renal function, or reduced quality of life due to the obstructing stone. To reduce the risk of non-passage and associated complications, additional predictive parameters and multicenter studies are needed to better assess the likelihood of SP.[1,6,7] For example, impacted ureteral stones–regardless of their size or location–may lead to increased ureteral wall thickness (UWT) at the site due to local inflammation, hypertrophy, and edema, thereby preventing SP. Moreover, these changes can increase the risk of acute complications during minimally invasive procedures, such as intraoperative bleeding and ureteral perforation, and may also prolong operative time.[8,9] In this context, informing the operating surgeon of these findings may reduce treatment failure, help better prepare for potential intraoperative complications, and support consideration of alternative treatment protocols. Although there are limited studies investigating the relationship between UWT around the stone, SP, and treatment success using non-contrast computed tomography (CT), the sample sizes in these studies are relatively small, indicating the need for further research in this area.[10-16] Additionally, although some studies examine serum inflammation markers in SP of ureteral calculi, exclusion criteria have not been standardized, particularly regarding the effects of concurrent diseases. In cases where the effects of other variables are assessed simultaneously with MET, the impact of the MET agent on markers used to predict SP remains unclear.

Main points

- Urolithiasis is a common cause of emergency room visits due to renal colic. In the absence of spontaneous passage (SP), urosepsis and sudden loss of renal function may occur. Therefore, appropriate management of the balance between follow-up and intervention is crucial for patient safety.
- The predictive value of the criteria regarding stone size and location, which are frequently used in SP of ureteral stones, is limited in some cases.
- Our study shows that the wall thickness around the ureteral stone is a risk factor with a higher negative predictive rate for the absence of SP than the stone size and location.
- Stone density, increasing degree of hydronephrosis, and elevated platelet-lymphocyte ratio and neutrophil-lymphocyte ratio values can also serve as additional parameters to enhance the predictive accuracy for SP of stones.

Detecting the probability of SP of a ureteral stone allows for selecting the most appropriate treatment method more quickly, limiting delays in patient management and reducing the risk of complications. Numerous studies in the literature focus on stone size and location to predict the likelihood of SP. In addition to these established parameters, only a few studies have attempted to predict SP based on biochemical indicators and other urinary system factors. However, these studies are insufficient to establish standardized values. Therefore, we aim to identify useful imaging and laboratory parameters that could enhance the predictive accuracy for SP of ureteral stones.

Methods

Ethics committee approval

The study was carried out with the permission of Erzincan Binali Yıldırım University Clinical Ethics Committee (decision number: 2023-15/8, date: 07/09/2023). This study was conducted in accordance with the Declaration of Helsinki. Due to its retrospective design, informed consent forms were not acquired as the data collected from patients did not contain any identifiable information.

Scope of the study

Between January 1, 2019, and December 31, 2023, patients with unilateral ureteral stones ≤10 mm who were admitted to our hospital's emergency department and/or urology clinic after their first renal colic episode, underwent non-contrast CT with a stone protocol, and had inflammatory serum markers assessed by biochemical tests were retrospectively screened without age restriction and included in the study. The exclusion criteria are presented in the diagram illustrating the study population (Figure 1). Data on patients' gender, age, treatment protocols, and biochemical test results were obtained from medical records.

Radiological assessment

All participants underwent CT scans using a 128-slice multi-detector CT scanner (Siemens Somatom, Siemens Healthcare, Forchheim, Germany) following a non-contrast stone protocol (kV: 120, mAs: automated current modulation; slice thickness: 1.5 mm). Image assessments were performed using the picture archiving and communication system (PACS) archive with Syngo.via software (Siemens Somatom, Siemens Healthcare, Forchheim, Germany). Soft tissue window settings (width 300; level 40) were applied to axial, coronal, and sagittal plane images.

The ureteral stone's lateralization (right/left), location, maximum axial diameter (mm), average density [Hounsfield units (HU)], UWT (mm), degree of hydronephrosis (HN) (grade 0–4), average renal parenchymal density (HU), and ureter length were assessed by a radiologist with 4 years of experience based on the non-contrast CT scans performed at the time of patients' admission.

The diameter and density of the ureteral stone were measured at the level of its greatest transverse dimension using the freehand region of interest (ROI) method (Figure 2a, b). The location of the ureteral stone was categorized as proximal, mid, or distal by dividing the ureteral length into three equal segments. UWT was determined by measuring the soft tissue density, including the ureteral wall and periureteral edema, at the level where this density was most prominent (Figure 2c). Renal parenchymal thickness was measured at the upper, mid, and lower poles on sagittal plane scans, avoiding areas with space-occupying lesions. The thickest area in each section was measured, and the average of these three measurements was used to determine the parenchymal thickness (Figure 3a). Renal parenchymal density was assessed by obtaining three measurements from the most homogeneous and thickest areas without space-occupying le-

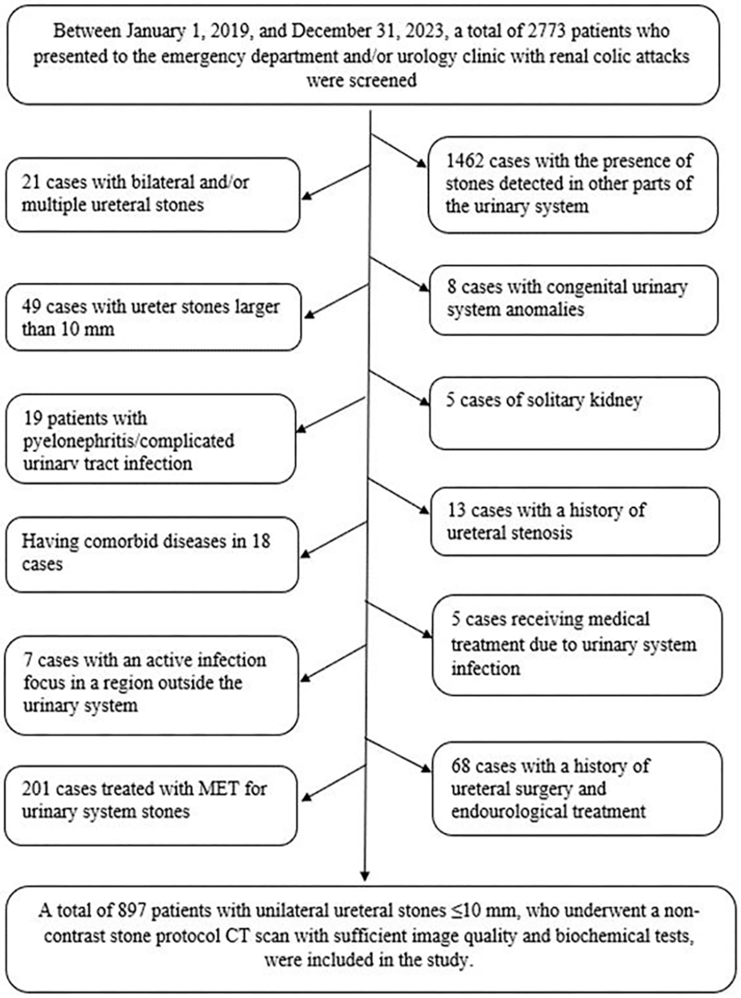

Between January 1, 2019, and December 31, 2023, a total of 2773 patients who presented to the emergency department and/or urology clinic with renal colic attacks were screened
21 cases with bilateral and/or multiple ureteral stones
1462 cases with the presence of stones detected in other parts of the urinary system
49 cases with ureter stones larger than 10 mm
8 cases with congenital urinary system anomalies
19 patients with pyelonephritis/complicated urinary tract infection
5 cases of solitary kidney
Having comorbid diseases in 18 cases
13 cases with a history of ureteral stenosis
7 cases with an active infection focus in a region outside the urinary system
5 cases receiving medical treatment due to urinary system infection
201 cases treated with MET for urinary system stones
68 cases with a history of ureteral surgery and endourological treatment
A total of 897 patients with unilateral ureteral stones ≤10 mm, who underwent a non-contrast stone protocol CT scan with sufficient image quality and biochemical tests, were included in the study.

Figure 1. Scheme showing the study population. MET, medical expulsive therapy; CT, computed tomography.

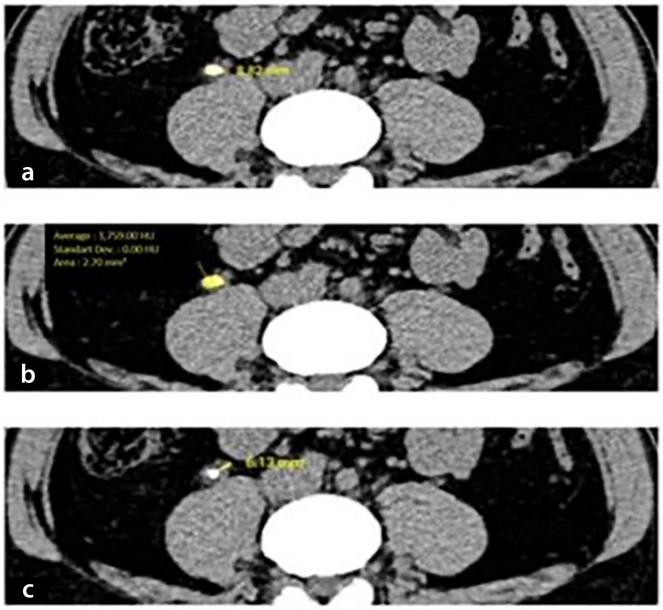


Figure 2. Non-contrast axial plane CT images: the diameter **(a)** and density **(b)** of the ureteral stone are measured at a single level where the stone's diameter is widest. Ureteral wall thickness **(c)** is measured at the level where the soft tissue density, consisting of the ureter wall and periureteral edema surrounding the stone, is highest, using the freehand ROI option. CT, computed tomography; ROI, region of interest.

sions at the upper, mid, and lower pole levels using the freehand ROI method. The average of these values was recorded (Figure 3b). Ureteral length was measured using reformatted CT images, with the ureteropelvic junction and ureterovesical junction as the starting and ending points, respectively. The measurement was based on the number of transverse slices, each represented by a single axial line, multiplied by the slice thickness parameter (Figure 4). The presence and grading of HN in the collecting system, secondary to the ureteral stone, were evaluated based on a commonly used CT grading classification system (Figure 5).

Biochemical assessment

Based on the biochemical tests conducted at the time of the patient's initial presentation, the neutrophil-lymphocyte ratio (NLR) and platelet-lymphocyte ratio (PLR) values were determined.

After applying the exclusion criteria, the SP status of ureteral stones was evaluated using non-contrast CT images accessed via the PACS system, and the mean follow-up durations were recorded based on patient follow-up records. Following the assessment of SP status, the effects of patients' demographic characteristics, initial findings on non-contrast CT, and biochemical test results on SP were analyzed.

Statistical Analysis

The conformity of the data to a normal distribution was verified using the Shapiro–Wilk test and Q–Q plot. Accordingly, parametric tests were used for inferential statistics. The Student's t-test was applied to compare parameters between two independent groups. The Mann–Whitney U test was used to compare median values of parameters that did not follow a normal distribution between two groups, and the chi-square test was employed to evaluate categorical variables expressed as percentages. Descriptive statistics were presented as mean ± standard deviation for normally distributed numerical variables and as number and percentage for categorical variables. In the study, the effects of age, gender, and selected clinical and laboratory characteristics on the risk of spontaneous stone passage were first analyzed using univariate logistic regression (LR). Variables found to be significant were then analyzed using stepwise multivariate LR (enter method). Optimum cut-off values were determined by receiver operating characteristic analysis. Statistical analyses were performed using SPSS version 25.0 (IBM Corporation, Armonk, NY, United States), and *P* values less than 0.05 were considered statistically significant.

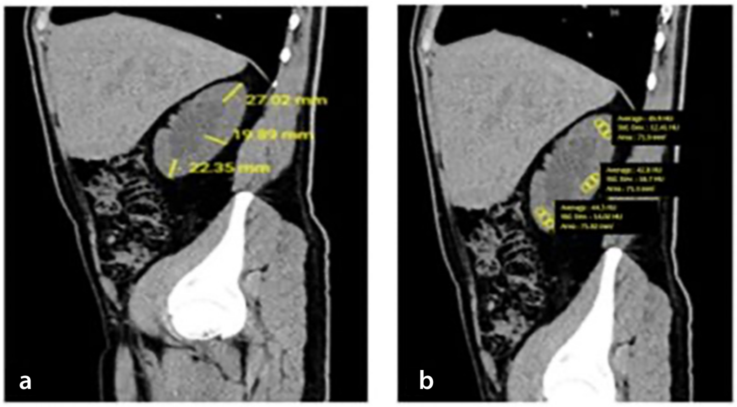


Figure 3. Sagittal non-contrast CT scan: **(a)** parenchymal thickness (mm) and **(b)** parenchymal density (HU) were measured using the freehand ROI option at the three thickest and most homogeneous levels without space-occupying lesions at the upper, middle, and lower poles of the kidney. The mean values of these measurements were calculated. CT, computed tomography; HU, Hounsfield units.

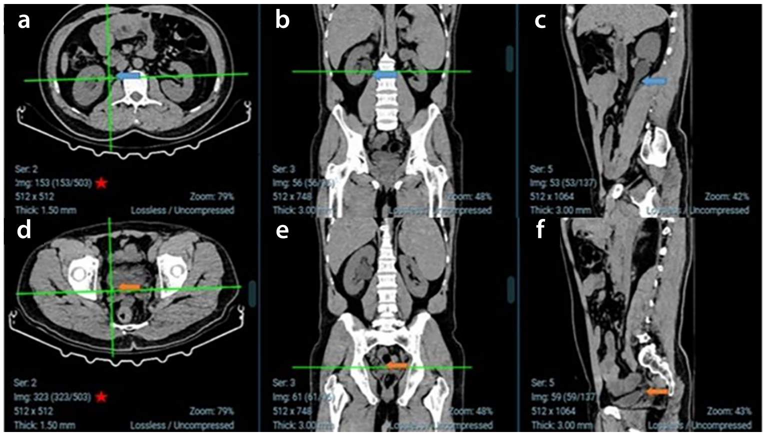


Figure 4. Measurement of ureteral length: The ureteropelvic junction (**a**, transverse; **b**, coronal; **c**, sagittal; blue arrow) and ureterovesical junction (**d**, transverse; **e**, coronal; **f**, sagittal; orange arrow) levels were identified on reformatted CT images. Ureteral length was determined by counting the number of transverse lines (red stars) between the two levels based on the section thickness parameter. CT, computed tomography.

Results

A total of 897 patients who presented to our hospital's emergency department or urology outpatient clinic with renal colic attacks over a 4-year period and met the inclusion criteria were included in the study without any age restriction. The median follow-up duration for SP was 4 weeks (±2 weeks) according to patient medical records.

The study population consisted of 72.9% (n = 654) male and 27.1% (n = 243) female participants, with a mean age of 46.05 ± 13.84 years. Among the 897 cases included, 384 (42.8%) showed SP of the ureteral stone during follow-up, whereas 513 (57.2%) did not experience SP (Table 1). Of the 513 patients without SP of ureteral stones, 88 were managed using extracorporeal shock wave lithotripsy (ESWL), whereas the remaining 425 cases were treated by ureterorenoscopy and ureteroscopic laser lithotripsy.

The clinical parameters of the cases were compared based on SP status. It was found that the average UWT at the stone level in cases without SP (2.41 ± 0.48 mm) was substantially higher than in cases with SP (1.41 ± 0.27 mm). In addition, the mean ureteral stone sizes and stone densities in cases without SP were considerably greater than those in cases with SP. Furthermore, the incidence of SP was substantially higher in distal ureteral stones (n = 195, 50.78%) than in proximal stones (n = 66, 17.18%) based on stone location. The study also observed that PLR and NLR values were considerably higher in cases without SP than in those with SP. Among the cases studied, the frequency of stone SP was substantially higher in patients without HN or with grade 1 HN (14.6% and 57.3%, respectively), and the likelihood of SP decreased considerably as the degree of HN increased (Table 2).

It was also found that mean values of ureter length, renal parenchymal density, and parenchymal thickness were considerably higher in male patients than in female patients. No significant differences between genders were observed for other parameters, including SP (Tables 3, 4).

According to the LR analysis of factors associated with the absence of SP in ureteral stones, UWT, left/right ureteral stone sizes, stone densities, renal parenchymal densities, and the presence of HN were identified as statistically significant risk factors. Patients with UWT values greater than 1.7 mm at the stone level had a 706.5-fold higher risk of absence of SP than those with UWT values less than 1.7 mm (Table 5).

According to the multivariate LR analysis of variables found significant in the univariate analysis, UWT at the stone level, left/right ureter stone size, stone density, PLR, and NLR values were identified as statistically significant risk factors for the absence of stone SP. When all these factors were present simultaneously, patients with UWT values greater than 1.7 mm at the stone level had a 337.98-fold higher risk of not passing stones spontaneously than those with UWT less than 1.7 mm (Table 6).

Discussion

The prevalence of ureteral stones and the frequency of related hospital admissions are increasing. Therefore, to avoid adding to the healthcare burden, conservative treatment should not be overlooked in cases with a likelihood of SP. However, early planning of invasive treatment is crucial in cases without SP probability, as delayed intervention may lead to acute renal failure. In such cases, ureteral stones can be managed using minimally invasive methods, thanks to advances in ESWL and endourological techniques.[17,18] Depending on the location and size of the stone, treatment success rates of 68%–90% have been reported for ESWL and 80%–97% for endourological methods.[5,19]

Despite these high success and stone-free rates, minimally invasive treatments are costly and carry potential risks, including hematoma formation, urinary tract infections, and urinary extravasation. Therefore, accurately predicting the likelihood of SP and avoiding overtreatment remains critical.[20-22] In this context, we aimed to identify certain indicators–and their standardized values–that have not been sufficiently investigated in the literature but may help predict SP of ureteral stones and guide clinical decision-making.

In our study, we demonstrated a statistically significant relationship between SP and several variables: UWT at the stone level, stone size, stone density, stone location, HN grade, PLR, and NLR. Stones located in the upper ureter were less likely to pass spontaneously than those in lower positions. Similarly, high-density stones, low-density stones, and stones accompanied by high-grade HN exhibited a lower rate of SP than stones in patients without HN or with low-grade HN. The likelihood of SP decreased as UWT, stone size, PLR, and NLR increased. No statistically significant relationship was found between SP and age or gender.

Statistically significant predictive values were identified through univariate and multivariate LR analyses for the following variables: UWT >1.7 mm at the stone level, ureteral stone dimensions (right >6.5 mm, left >6 mm), stone density >957 HU, PLR >10.28, NLR >2.15, and high-grade HN (≥ grade 2). These results indicate that the predictive value of UWT at the ureteral stone level may offer a stronger prediction of SP than stone size alone, which is a key factor in the formation of various clinical guidelines.

Stone size and location

Various studies in the literature have examined the effect of ureteral stone size and location on SP.[7,23,24] These studies generally show a positive correlation between smaller, distally located stones and higher SP rates.[25-28] Reported SP rates based on ureteral stone location range from 45%–79% for the lower ureter to 22%–60% for the middle and 12–48% for the upper level.[29,30] In our study group, SP was more likely in lower ureteral calculi and decreased progressively at higher levels (distal: 50.78%, middle: 32.03%, upper: 17.18%), consistent with findings in the literature. Although our SP percentages fall within the reported ranges, they are lower than the average values. This may be attributed to the relatively larger average stone sizes in our cohort (right: 6.29 mm; left: 6.4 mm). Stone size is another key factor often used to predict the SP of ureteral stones. The literature indicates that SP occurs in 68%–98% of ureteral stones ≤4 mm and in 25%–67% of 5–10 mm stones. With MET, SP rates for 5–10 mm stones can reach up to 83%.[30-37] Consistent with these findings, our study showed that the probability of SP decreased as stone size increased. Specifically, stone sizes of 6–6.5 mm in the ureter were identified as statistically significant risk factors for spontaneous non-passage in both univariate and multivariate regression analyses. However, the literature lacks standardization regarding the imaging plane used to measure stone size. One study reported that axial plane measurements–commonly used–can underestimate the actual stone burden by up to 20%.[38,39]

Yoshida et al.[1] and Lee et al.[6] have reported that measuring stone size in the longitudinal plane is more valuable, as it reflects a larger contact surface with the ureteral mucosa. A greater contact area is associated with increased mucosal inflammation and edema, thereby reducing the likelihood of secondary SP. In our study, stone size was defined as the larger of the measurements obtained in the transverse and longitudinal planes; however, a direct comparison between measurements from different planes was not performed. Therefore, although our study incorporates dimensional data in line with previous work, it may be considered relatively limited due to the lack of direct comparison between imaging planes.

Ureteral wall thickness around the ureteral stone

When a ureteral stone is impacted, an increase in UWT develops due to inflammation, periureteral edema, hypertrophy, and fibrosis resulting from stone irritation at the site of impaction.[1,10,16,40] Studies have also shown that increased UWT is associated with higher intraoperative complication rates and lower stone-free rates during ureteroscopic procedures.[11,41]

The effect of UWT at the level of the ureteral stone on SP is controversial, and few studies have investigated this issue. According to the study by Yoshida et al.[1], the probability of a 4-week SP in patients with low UWT at the ureteral stone level (76.4%) was considerably higher than in those with high UWT (14.7%). This research identified a threshold value of 2.71 mm for predicting SP and showed that when UWT is evaluated alongside established parameters such as stone size and location, the accuracy of SP prediction approaches 90%.[1] In our study, the mean UWT at the stone level in cases without SP was found to be 2.41 mm, which was similar to the mean thickness of 2.4 mm reported by Coşkun and Can[42] and lower than the 2.78 mm reported by Selvi et al.[43] This difference may be related to variations in the exclusion criteria and the inclusion of patients with metabolic syndrome who exhibited more heterogeneous characteristics in the study by Selvi et al.[43] However, the common finding across all these studies is that lower UWT values are associated with a higher likelihood of SP at the ureteral stone level.

Table 1. Distribution of socio-demographic and clinical characteristics of all cases

Variables	Mean ± SD	M (min-max)
Age	46.05 ± 13.84	45 (11–81)
Men/women n (%)	654 (72.9%)/243 (27.1%)	
Ureteral wall thickness at stone level (mm)	1.84 ± 0.62	1.7 (0.51–5.12)
Left ureteral stone size (mm)	6.4 ± 1.7	6.5 (3–10)
Right ureteral stone size (mm)	6.29 ± 1.7	6 (3–10)
Stone density (HU)	1,023.04 ± 471.64	957 (235–2068)
Ureteral length (mm)	223.81 ± 15.21	225 (172–264)
Kidney parenchymal density (HU)	38.08 ± 3.81	39 (25–47)
Kidney parenchymal thickness (mm)	18.21 ± 2.75	18 (6–28.77)
PLR	10.09 ± 3.8	
NLR	2.05 ± 0.5	
Accompanying hydronephrosis		
Grade 1	348 (38.8%)	
Grade 2	321 (35.8%)	
Grade 3	138 (15.4%)	
Grade 4	9 (1%)	
No	81 (9%)	
Spontaneous passage		
No	513 (57.2%)	
Yes	384 (42.8%)	

SD, standard deviation; M, median; HU, Hounsfield units; PLR, platelet-lymphocyte ratio; NLR, neutrophil-lymphocyte ratio; min-max, minimum-maximum.

Table 2. Comparison of the quantitative clinical characteristics of cases according to spontaneous stone passage status

Variables	Spontaneous stone passage No (n = 513) Mean ± SD	 Yes (n = 384) Mean ± SD	*P*
Age	46.51 ± 13,53	45.71 ± 14.09	0.620*
Ureteral wall thickness at stone level (mm)	2.41 ± 0.48	1.41 ± 0.27	**<0.001***
Left ureteral stone size (mm)	7.64 ± 1.29	5.44 ± 1.32	**<0.001***
Right ureteral stone size (mm)	7.66 ± 1.39	5.3 ± 1.12	**<0.001***
Stone density (HU)	1,355.29 ± 418.91	774.34 ± 336.89	**<0.001***
Ureteral length (mm)	224.44 ± 17.98	223.35 ± 12.79	0.541*
Kidney parenchymal density (HU)	37.63 ± 4.14	38.42 ± 3.52	0.083*
Kidney parenchymal thickness (mm)	18.14 ± 2.95	18.25 ± 2.60	0.730*
Stone location (%)			**<0.001***
Upper	255 (49.70%)	66 (17.18%)	
Middle	140 (27.29%)	123 (32.03%)	
Lower	118 (23%)	195 (50.78%)	
Accompanying hydronephrosis			
No	6 (1.6%)	75 (14.6%)	**<0.001+**
Grade 1	48 (14.1%)	294 (57.3%)	
Grade 2	195 (50.8%)	126 (24.6%)	
Grade 3 or above	129 (33.6%)	18 (3.5%)	
PLR	11.51 ± 4.79	9.06 ± 2.29	**<0.001****
NLR	2.35 ± 0.53	1.65 ± 0.47	**<0.001****

SD, standard deviation; **P* value obtained from the Student's t-test; $^{+}$*P* value was obtained from the chi-square test; **Mann–Whitney U test; HU, Hounsfield units; PLR, platelet-lymphocyte ratio; NLR, neutrophil-lymphocyte ratio.

Table 3. Comparison of the quantitative clinical characteristics between female and male patients

Variables	Men (n = 654) Mean ± SD	Women (n = 243) Mean ± SD	*P*
Age	45.5 ± 13.27	47.53 ± 15.24	0.260
Ureteral wall thickness at stone level (mm)	1.83 ± 0.6	1.88 ± 0.67	0.507
Left ureteral stone size (mm)	6.29 ± 1.7	6.67 ± 1.69	0.224
Right ureteral stone size (mm)	6.35 ± 1.74	6.11 ± 1.57	0.466
Stone density (HU)	1,043.85 ± 485.88	967.02 ± 428.82	0.211
Ureteral length (mm)	225.38 ± 14.88	219.59 ± 15.38	**0.003**
Kidney parenchymal density (HU)	38.46 ± 3.64	37.07 ± 4.09	**0.005**
Kidney parenchymal thickness (mm)	18.48 ± 2.73	17.46 ± 2.68	**0.004**

SD: standard deviation; *P* value obtained from the Student's t-test; HU, Hounsfield units.

Table 4. Comparison of the presence of hydronephrosis and spontaneous passage between female and male patients

	Men (n = 654)	Women (n = 243)	*P*
Accompanying hydronephrosis			
No	44 (10.1%)	10 (6.2%)	0.772
Grade 1	164 (37.6%)	68 (42%)	
Grade 2	160 (36.7%)	54 (33.3%)	
Grade 3 or above	66 (15.6%)	30 (18.5%)	
Spontaneous passage			
No	372 (56.9%)	141 (58%)	0.859
Yes	282 (43.1%)	102 (42%)	

The *P* value was obtained from the chi-square test.

Table 5. Examination of factors associated with the absence of spontaneous stone passage using univariate logistic regression analysis

Variables	Odds (95% CI)	*P*
Age > 45	1,193 (0.754–1,888)	0.452
Women	0.954 (0.569–1,599)	0.859
Ureteral wall thickness at stone level (mm) > 1.7	706.500 (157.638–3166.395)	**<0.001**
Left ureteral stone size (mm) > 6.5	6.061 (3,386–10,850)	**<0.001**
Right ureteral stone size (mm) > 6	7,046 (3,826–12,975)	**<0.001**
Stone density (HU) > 957	13,907 (7,841–24,667)	**<0.001**
Ureteral length (mm) > 225	1,375 (0.867–2,179)	0.176
Kidney parenchymal density (HU) < 39	1,646 (1,037–2,661)	0.034
Kidney parenchymal thickness (mm) > 18	1,197 (0.753–1,904)	0.447
Accompanying hydronephrosis		
Grade 1	2,296 (0.499–10,555)	0.286
Grade 2	19,345 (4,353–85,976)	**<0.001**
Grade 3 or above	89,583 (16,788–478.030)	**<0.001**
PLR > 10.28	3.99 (1,317–10,358)	**<0.001**
NLR > 2.15	3,746 (2,473–5,642)	**<0.001**

CI, confidence interval; Odds, odds ratio; PLR, platelet-lymphocyte ratio; NLR, neutrophil-lymphocyte ratio; HU, Hounsfield units.

Table 6. Examination of factors associated with the absence of spontaneous stone passage using multivariate logistic regression analysis

Variables	Odds (95% CI)	*P*
Ureteral wall thickness at stone level (mm) > 1.7	337.977 (58.270–1960.337)	**<0.001**
Left ureteral stone size (mm) > 6.5	5,429 (1,319–22,343)	**0.019**
Right ureteral stone size (mm) > 6	20,657 (3,170–134.609)	**0.002**
Stone density (HU) > 957	4,349 (1,170–16,165)	**0.028**
Kidney parenchymal density (HU) < 39	1,603 (0.452–5,688)	0.465
Accompanying hydronephrosis ref: none	1	0.399
Grade 1	0.536 (0.008–36,147)	0.772
Grade 2	1,858 (0.027–127.253)	0.774
Grade 3 or above	1,536 (0.021–110.372)	0.844
PLR > 10.28	7.49 (4,192–11,983)	**0.004**
NLR > 2.15	2,072 (1,127–3,219)	**<0.001**

CI, confidence interval; Odds, odds ratio; PLR, platelet-lymphocyte ratio; NLR, neutrophil-lymphocyte ratio; HU, Hounsfield units.

In addition, our study demonstrated that a UWT greater than 1.7 mm was a risk factor for non-SP of ureteral stones, with an odds ratio of 706.5 in univariate LR analysis and 337.9 in multivariate LR analysis when combined with other parameters (stone size, density, NLR, and PLR). In light of these findings, UWT appears to offer considerable superiority as a risk factor compared with stone size, which remains an important parameter in current clinical practice. In the study by Yoshida et al.[1], although UWT was an important predictor of non-SP in LR analyses, its risk ratio was low compared with stone size and stone location.[1] In the study by Cumpanas et al.[44], UWT was a considerable risk factor in univariate LR analysis but lost its importance in multivariate analysis, where linear stone dimensions emerged as the strongest predictor of non-SP. Conversely, in studies by Coşkun and Can[42] and Selvi et al.[43], after standardizing stone sizes between SP and non-SP groups, UWT was shown to be the most important predictor of non-SP in both univariate and multivariate LR analyses, outperforming other parameters.[42,43] A review of the current literature reveals that studies exploring the role of UWT in predicting the SP of ureteral stones remain limited, and no standardized criteria have yet been established for patient management. Therefore, further research is warranted to validate our findings and support the development of standardized predictive tools.

Degree of hydronephrosis

When examining the relationship between HN, which can develop due to obstructive ureteral stones, and the SP of the stone, studies have reported that the likelihood of SP decreases as the degree of HN increases.[43] Consistent with the literature, our study observed that the SP rates of ureteral stones decreased in proportion to the degree of HN.

Stone density

In the literature, various studies have investigated the effect of stone density (HU) on SP and the success of ESWL using non-contrast CT.[45] In a study by Coşkun and Can[42], the probability of SP was reported to be high in cases with stones of lower density. In our study group, the mean stone density in cases without SP (1,355.29 ± 418.91 HU) was statistically significantly higher than in those with SP (774.34 ± 336.89 HU). Furthermore, our results showed that a stone density above 957 HU was a statistically significant risk factor for non-SP in both univariate and multivariate

LR analyses. In contrast, Balci et al.[46] reported no statistically significant difference in stone density between cases with and without SP.

Ureter length

To our knowledge, only one study in the literature has investigated the effect of ureteral stone presence on ureter length in relation to SP. In a randomized study by Coşkun and Can[42], which compared groups with and without SP in 50 ureteral stones, the average ureter length in the non-SP group was 199 mm, and the presence of the stone was found to have no statistically significant effect on SP. Our study is the second in the literature to examine the impact of ureter length on the SP of ureteral stones and features a larger sample size than the previously reported study. In our cohort of 897 cases–513 of which did not experience SP–there was no statistically significant difference in ureter length between the SP and non-SP groups. The average ureter length in the non-SP group was 225 mm. In both studies, no statistically significant association between ureter length and SP was found in either univariate or multivariate regression analyses.[42]

Renal parenchymal thickness and density

When evaluating the potential for predicting SP based on renal parenchymal thickness and density, which may be affected by HN secondary to ureteral stones, no statistically significant differences were observed. In cases without SP, the mean parenchymal thickness was 18.14 mm and the mean parenchymal density was 37.63 HU, whereas in cases with SP, these values were 18.25 mm and 38.42 HU, respectively. Our study is the second in the literature to examine the effect of parenchymal thickness and density on the likelihood of SP in ureteral stones and includes the largest sample size to date. In the first published study on this topic, the mean parenchymal thickness in cases without SP was reported as 21.6 mm and the parenchymal density as 33.9 HU.[42] That study included 100 patients with equal gender distribution and SP status. Although it had a smaller sample size than our study, the results were also not statistically significant, consistent with our findings.

Inflammatory serum markers

In the literature, it has been reported that impacted ureteral stones cause a systemic inflammatory response due to obstruction and ureteral trauma, leading to elevated levels of certain blood markers such as white blood cell (WBC) count, neutrophil count, C-reactive protein (CRP), and procalcitonin. Conversely, some studies on similar parameters have shown a statistically significant relationship between these markers and a decreased probability of SP.[7-13,47,48] However, the study by Sfoungaristos suggested that increased WBC and neutrophil levels may stimulate ureteral peristalsis, thereby facilitating SP.[15] In contrast, a study by Cilesiz et al.[49] reported no statistically significant difference between SP and WBC or CRP values.

In contrast to the previously discussed inflammatory serum indicators, the association between NLR and PLR markers and the SP of ureteral stones has been examined in only a limited number of studies.[39,50] In the study by Abou Heidar et al.[48], it was shown that increased NLR (>2.87) and PLR (>10.42) values were associated with decreased SP rates in both univariate and multivariate analyses. In a recent study by Aghaways et al.[51], the NLR and PLR values were measured as 2.63 ± 1.35 and 11.47 ± 4.86, respectively, in patients without SP, and a statistically significant relationship was found between elevated values and a lower probability of SP. In our study, the NLR (2.35 ± 0.53) and PLR (11.51 ± 4.79) values in cases without SP were statistically significantly higher than in those with SP, with cutoff values of 2.15 for NLR and 10.28 for PLR. These results indicate a relationship between high NLR and PLR values and unsuccessful SP of ureteral stones. However, studies by Coşkun and Can[42], Ahmed et al.[52], and Senel et al.[53] reported no statistically significant relationship between SP of ureteral stones and NLR or PLR values.

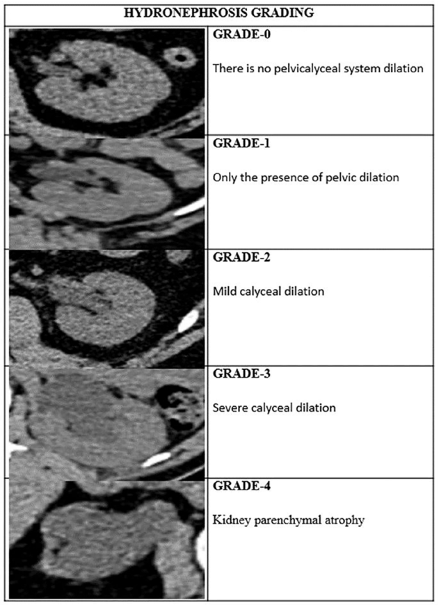

HYDRONEPHROSIS GRADING
GRADE-0
There is no pelvicalyceal system dilation
GRADE-1
Only the presence of pelvic dilation
GRADE-2
Mild calyceal dilation
GRADE-3
Severe calyceal dilation
GRADE-4
Kidney parenchymal atrophy

Figure 5. Hydronephrosis grading system.[54]

The incidence of ureteral stones has been reported to be approximately 12% in adult men and 6% in adult women.[33] In our study, a male predominance (n = 654, 72.9%) was observed among patients with ureteral stones, consistent with the literature. However, no statistically significant association was found between the SP of ureteral stones and gender.

The limitations of this study include its retrospective design and single-center setting. Another limitation is that all imaging measurements were performed by a single radiologist; therefore, interobserver variability was not assessed. In addition, stone composition was not determined in this study. Patients who received MET, known to facilitate SP, and those who received recent anti-inflammatory treatment, which could affect biochemical results, were excluded from the study. However, more useful results could be obtained from randomized studies comparing the data we obtained regarding SP of ureteral stones with cases who received MET or anti-inflammatory treatment. Among the parameters analyzed in this study, none can be considered entirely novel compared with the existing literature, which may be regarded as a limitation. Nevertheless, the comprehensive evaluation of these parameters within a relatively broad population contributes to the literature by providing a more holistic perspective. Furthermore, by confirming the diagnostic value of UWT, the study offers a distinctive and noteworthy finding.

In conclusion, the accurate prediction of the probability of SP remains debated, and additional criteria are needed for personalized patient-specific follow-up and treatment management. The results of our study indicate that, alongside large stone size and proximal stone location, high stone density, increased UWT, considerable HN at the stone's proximal site, and elevated NLR and PLR values in the blood are statistically significantly and negatively associated with SP.

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. Yoshida T, Inoue T, Taguchi M, Omura N, Kinoshita H, Matsuda T. Ureteral wall thickness as a significant factor in predicting spontaneous passage of ureteral stones of ≤10 mm: a preliminary report. *World J Urol*. 2019;37(5):913-919. [Crossref]
2. Fukuhara H, Ichiyanagi O, Kakizaki H, Naito S, Tsuchiya N. Clinical relevance of seasonal changes in the prevalence of ureterolithiasis in the diagnosis of renal colic. *Urolithiasis*. 2016;44(6):529-537. [Crossref]
3. Hong DY, Kim JW, Lee KR, Park SO, Baek KJ. Epidemiologic and clinical characteristics of patients presenting with renal colic in Korea. *Urol J*. 2015;12(3):2148-2153. [Crossref]
4. Cepeda Delgado M, López Izquierdo R, Amón Sesmero JH, Del Pozo Vegas C, Álvarez Manzanares J. Epidemiological characteristics of renal colic and climate-related causes in a continental area in Spain. *Urol Int*. 2015;95(3):309-313. [Crossref]
5. Türk C, Petrík A, Sarica K, et al. EAU Guidelines on diagnosis and conservative management of urolithiasis. *Eur Urol*. 2016;69(3):468-474. [Crossref]
6. Lee SR, Jeon HG, Park DS, Choi YD. Longitudinal stone diameter on coronal reconstruction of computed tomography as a predictor of ureteral stone expulsion in medical expulsive therapy. *Urology*. 2012;80(4):784-789. [Crossref]
7. Coelho-Souza SA, Jenkins SR, Casarin A, et al. *The* effect of light on bacterial activity in a seaweed holobiont. *Microb Ecol*. 2017;74(4):868-876. [Crossref]
8. Mugiya S, Ito T, Maruyama S, Hadano S, Nagae H. Endoscopic features of impacted ureteral stones. *J Urol*. 2004;171(1):89-91. [Crossref]
9. Deliveliotis C, Chrisofos M, Albanis S, Serafetinides E, Varkarakis J, Protogerou V. Management and follow-up of impacted ureteral stones. *Urol Int*. 2003;70(4):269-272. [Crossref]
10. Sarica K, Kafkasli A, Yazici Ö, et al. Ureteral wall thickness at the impacted ureteral stone site: a critical predictor for success rates after SWL. *Urolithiasis*. 2015;43(1):83-88. [Crossref]
11. Yoshida T, Inoue T, Omura N, et al. Ureteral wall thickness as a preoperative indicator of impacted stones in patients with ureteral stones undergoing ureteroscopic lithotripsy. *Urology*. 2017;106:45-49. [Crossref]
12. Aldaqadossi HA. Stone expulsion rate of small distal ureteric calculi could be predicted with plasma C-reactive protein. *Urolithiasis*. 2013;41(3):235-239. [Crossref]
13. Lee KS, Ha JS, Koo KC. Significance of neutrophil-to-lymphocyte ratio as a novel indicator of spontaneous ureter stone passage. *Yonsei Med J*. 2017;58(5):988-993. [Crossref]
14. Özcan C, Aydoğdu O, Senocak C, et al. Predictive factors for spontaneous stone passage and the potential role of serum C-reactive protein in patients with 4 to 10 mm distal ureteral stones: a prospective clinical study. *J Urol*. 2015;194(4):1009-1013. [Crossref]
15. Sfoungaristos S, Kavouras A, Katafigiotis I, Perimenis P. Role of white blood cell and neutrophil counts in predicting spontaneous stone passage in patients with renal colic. *BJU Int*. 2012;110(8 Pt B):E339-345. [Crossref]
16. Elibol O, Safak KY, Buz A, Eryildirim B, Erdem K, Sarica K. Radiological noninvasive assessment of ureteral stone impaction into the ureteric wall: a critical evaluation with objective radiological parameters. *Investig Clin Urol*. 2017;58(5):339-345. [Crossref]
17. Miller OF, Kane CJ. Time to stone passage for observed ureteral calculi: a guide for patient education. *J Urol*. 1999;162(3 Pt 1):688-690; discussion 690-691. [Crossref]
18. Hermanns T, Sauermann P, Rufibach K, Frauenfelder T, Sulser T, Strebel RT. Is there a role for tamsulosin in the treatment of distal ureteral stones of 7 mm or less? Results of a randomised, double-blind, placebo-controlled trial. *Eur Urol*. 2009;56(3):407-412. [Crossref]
19. Dasgupta R, Cameron S, Aucott L, et al. Shockwave lithotripsy versus ureteroscopic treatment as Therapeutic Interventions for Stones of the Ureter (TISU): a multicentre randomised controlled non-inferiority trial. *Eur Urol*. 2021;80(1):46-54. [Crossref]
20. Skolarikos A, Alivizatos G, de la Rosette J. Extracorporeal shock wave lithotripsy 25 years later: complications and their prevention. *Eur Urol*. 2006;50(5):981-990; discussion 990. [Crossref]
21. Perez Castro E, Osther PJ, Jinga V, et al. Differences in ureteroscopic stone treatment and outcomes for distal, mid-, proximal, or multiple ureteral locations: the Clinical Research Office of the Endourological Society ureteroscopy global study. *Eur Urol*. 2014;66(1):102-109. [Crossref]
22. Drake T, Grivas N, Dabestani S, et al. What are the benefits and harms of ureteroscopy compared with shock-wave lithotripsy in the treatment of upper ureteral stones? A systematic review. *Eur Urol*. 2017;72(5):772-786. [Crossref]
23. Preminger GM, Tiselius HG, Assimos DG, et al. 2007 Guideline for the management of ureteral calculi. *Eur Urol*. 2007;52(6):1610-1631. [Crossref]
24. Keller EX, De Coninck V, Audouin M, Doizi S, Daudon M, Traxer O. Stone composition independently predicts stone size in 18,029 spontaneously passed stones. *World J Urol*. 2019;37(11):2493-2499. [Crossref]
25. Jendeberg J, Geijer H, Alshamari M, Cierzniak B, Lidén M. Size matters: the width and location of a ureteral stone accurately predict the chance of spontaneous passage. *Eur Radiol*. 2017;27(11):4775-4785. [Crossref]
26. Türk C, Knoll T, Seitz C, Skolarikos A, Chapple C, McClinton S; European Association of Urology. Medical expulsive therapy for ureterolithiasis: the EAU recommendations in 2016. *Eur Urol*. 2017;71(4):504-507. [Crossref]
27. Brubaker WD, Dallas KB, Elliott CS, et al. Payer type, race/ethnicity, and the timing of surgical management of urinary stone disease. *J Endourol*. 2019;33(2):152-158. [Crossref]
28. Matlaga BR, Jansen JP, Meckley LM, Byrne TW, Lingeman JE. Treatment of ureteral and renal stones: a systematic review and meta-analysis of randomized, controlled trials. *J Urol*. 2012;188(1):130-137. [Crossref]
29. Skolarikos A, Laguna MP, Alivizatos G, Kural AR, de la Rosette JJ. The role for active monitoring in urinary stones: a systematic review. *J Endourol*. 2010;24(6):923-930. [Crossref]
30. Ramasamy V, Aarthy P, Sharma V, Thakur APS. Role of inflammatory markers and their trends in predicting the outcome of medical expulsive therapy for distal ureteric calculus. *Urol Ann*. 2022;14(1):8-14. [Crossref]
31. Assimos D, Krambeck A, Miller NL, et al. Surgical management of stones: American Urological Association/Endourological Society Guideline, PART I. *J Urol*. 2016;196(4):1153-1160. [Crossref]
32. Preminger GM, Tiselius HG, Assimos DG, et al. 2007 guideline for the management of ureteral calculi. *J Urol*. 2007;178(6):2418-2434. [Crossref]
33. Furyk JS, Chu K, Banks C, et al. Distal ureteric stones and tamsulosin: a double-blind, placebo-controlled, randomized, multicenter trial. *Ann Emerg Med*. 2016;67(1):86-95.e2. [Crossref]
34. Mokhless I, Zahran AR, Youssif M, Fahmy A. Tamsulosin for the management of distal ureteral stones in children: a prospective randomized study. *J Pediatr Urol*. 2012;8(5):544-548. [Crossref]
35. Aydogdu O, Burgu B, Gucuk A, Suer E, Soygur T. Effectiveness of doxazosin in treatment of distal ureteral stones in children. *J Urol*. 2009;182(6):2880-2884. [Crossref]
36. Demehri S, Steigner ML, Sodickson AD, Houseman EA, Rybicki FJ, Silverman SG. CT-based determination of maximum ureteral stone area: a predictor of spontaneous passage. *AJR Am J Roentgenol*. 2012;198(3):603-608. [Crossref]
37. Chau LH, Tai DC, Fung BT, Li JC, Fan CW, Li MK. Medical expulsive therapy using alfuzosin for patient presenting with ureteral stone less than 10mm: a prospective randomized controlled trial. *Int J Urol*. 2011;18(7):510-514. [Crossref]
38. Metser U, Ghai S, Ong YY, Lockwood G, Radomski SB. Assessment of urinary tract calculi with 64-MDCT: the axial versus coronal plane. *AJR Am J Roentgenol*. 2009;192(6):1509-1513. [Crossref]

39. Kadihasanoglu M, Marien T, Miller NL. Ureteral stone diameter on computerized tomography coronal reconstructions is clinically important and under-reported. *Urology*. 2017;102:54-60. [Crossref]

40. Özbir S, Can O, Atalay HA, Canat HL, Çakır SS, Ötünçtemur A. Formula for predicting the impaction of ureteral stones. *Urolithiasis*. 2020;48(4):353-360. [Crossref]

41. Legemate JD, Wijnstok NJ, Matsuda T, et al. Characteristics and outcomes of ureteroscopic treatment in 2650 patients with impacted ureteral stones. *World J Urol*. 2017;35(10):1497-1506. [Crossref]

42. Coşkun A, Can U. Is it possible to predict spontaneous passage of a ureteral stone? An up-to-date comment on the current problem with new concepts concerning the patient and the stone. *Cent European J Urol*. 2022;75(3):305-310. [Crossref]

43. Selvi I, Baydilli N, Tokmak TT, Akinsal EC, Basar H. CT-related parameters and Framingham score as predictors of spontaneous passage of ureteral stones ≤ 10 mm: results from a prospective, observational, multicenter study. *Urolithiasis*. 2021;49(3):227-237. [Crossref]

44. Cumpanas AD, Camp B, Tran CM, et al. Prospective evaluation of ureteral wall thickness as a means to predict spontaneous stone passage: is it beneficial? *J Endourol*. 2024. [Crossref]

45. Nakasato T, Morita J, Ogawa Y. Evaluation of Hounsfield units as a predictive factor for the outcome of extracorporeal shock wave lithotripsy and stone composition. *Urolithiasis*. 2015;43(1):69-75. [Crossref]

46. Balci M, Tuncel A, Aydin O, et al. Tamsulosin versus nifedipin in medical expulsive therapy for distal ureteral stones and the predictive value of Hounsfield unit in stone expulsion. *Ren Fail*. 2014;36(10):1541-1544. [Crossref]

47. Park CH, Ha JY, Park CH, Kim CI, Kim KS, Kim BH. Relationship between spontaneous passage rates of ureteral stones less than 8 mm and serum C-reactive protein levels and neutrophil percentages. *Korean J Urol*. 2013;54(9):615-618. [Crossref]

48. Abou Heidar N, Labban M, Bustros G, Nasr R. Inflammatory serum markers predicting spontaneous ureteral stone passage. *Clin Exp Nephrol*. 2020;24(3):277-283. [Crossref]

49. Cilesiz NC, Ozkan A, Kalkanli A, et al. Can serum procalcitonin levels be useful in predicting spontaneous ureteral stone passage? *BMC Urol*. 2020;20:1-6. [Crossref]

50. Proctor MJ, Morrison DS, Talwar D, et al. A comparison of inflammation-based prognostic scores in patients with cancer. A Glasgow Inflammation Outcome Study. *Eur J Cancer*. 2011;47(17):2633-2641. [Crossref]

51. Aghaways I, Ibrahim R, Bapir R, Salih RQ, Salih KM, Abdulla BA. The role of inflammatory serum markers and ureteral wall thickness on spontaneous passage of ureteral stone < 10 mm: a prospective cohort study. *Ann Med Surg (Lond)*. 2022;80:104198. [Crossref]

52. Ahmed AF, Gabr AH, Emara AA, Ali M, Abdel-Aziz AS, Alshahrani S. Factors predicting the spontaneous passage of a ureteric calculus of ≤10 mm. *Arab J Urol*. 2015;13(2):84-90. [Crossref]

53. Senel C, Aykanat IC, Asfuroglu A, Keten T, Balci M, Aslan Y, Tuncel A. What is the role of inflammatory markers in predicting spontaneous ureteral stone passage? *Aktuelle Urol*. 2022;53(5):448-453. English. [Crossref]

54. Onen A. Grading of hydronephrosis: an ongoing challenge. *Front Pediatr*. 2020;27;8:458. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253482

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Abbreviated liver magnetic resonance imaging with a second-shot arterial phase image to assess the viability of treated hepatocellular carcinoma after non-radiation locoregional therapy

Il Wan Son[1]
Seung Baek Hong[2]
Nam Kyung Lee[2]
Suk Kim[2]
Hyung Il Seo[3]
Young Mok Park[3]
Byeong Gwan Noh [3]
Jong Hyun Lee[4]

[1]Busan Centum Hospital, Clinic of Radiology, Busan, Korea

[2]Department of Radiology, Biomedical Research Institute, Pusan National University Hospital, Pusan National University School of Medicine, Busan, Korea

[3]Department of Surgery, Biomedical Research Institute, Pusan National University Hospital, Pusan National University School of Medicine, Busan, Korea

[4]Department of Internal Medicine, Biomedical Research Institute, Pusan National University Hospital, Pusan National University School of Medicine, Busan, Korea

PURPOSE

To evaluate the feasibility of abbreviated liver magnetic resonance imaging (AMRI) with a second-shot arterial phase (SSAP) image for the viability of treated hepatocellular carcinoma (HCC) after non-radiation locoregional therapy (LRT).

METHODS

We retrospectively enrolled patients with non-radiation LRT for HCC who underwent the modified gadoxetic acid-enhanced liver MRI protocol, which includes routine dynamic and SSAP imaging after the first and second injection of gadoxetic acid, respectively (6 mL and 4 mL, respectively), and an available reference standard for tumor viability in the treated HCC between March 2021 and February 2022. Two radiologists independently reviewed the full-protocol MRI (FP-MRI) and AMRI with SSAP. For the FP-MRI, observations were assigned using the Liver Imaging Reporting and Data System treatment response (LR-TR) algorithm v.2024. In the AMRI with SSAP, the observations were assigned using the abbreviated LR-TR category according to the arterial mass-like enhancement in SSAP. Ancillary features, such as diffusion restriction and T2-weighted mild-to-moderate hyperintensity, were also optionally used.

RESULTS

Of the 95 patients (70 men and 25 women; mean age, 68.7 years), 42 (44.2%) had viable lesions and 53 (55.8%) had non-viable lesions. The scan time of the simulated AMRI was significantly shorter than the FP-MRI (7.6±0.49 and 23.6±0.50 min, respectively; $p<0.001$). For evaluating the viability of treated HCC, there were no significant differences in the sensitivity and specificity between the FP-MRI and AMRI with SSAP (sensitivity, 85.7% vs. 80.1%, $P = 0.500$; specificity, 96.2% vs. 96.2%, $P = 1.000$).

CONCLUSION

The abbreviated LR-TR score in AMRI with SSAP showed non-inferior diagnostic performance to FP-MRI in terms of evaluating the viability for the treated HCC, which may be helpful in clinical practice alongside a decreased scan time.

CLINICAL SIGNIFICANCE

Abbreviated liver MRI with SSAP may be helpful for evaluating the viability of treated HCC in practice, while also providing a decreased scan time.

KEYWORDS

Hepatocellular carcinoma, Liver Imaging Reporting and Data System, treatment response, magnetic resonance imaging, gadoxetic acid

Corresponding author: Seung Baek Hong

E-mail: cinematiclife7@hanmail.net

Received 02 June 2025; revision requested 29 June 2025; accepted 23 July 2025.

Epub: 11.09.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253482

You may cite this article as: Son IW, Hong SB, Lee NK, et al. Abbreviated liver magnetic resonance imaging with a second-shot arterial phase image to assess the viability of treated hepatocellular carcinoma after non-radiation locoregional therapy. *Diagn Interv Radiol*. 2026;32(2):149-155.

Hepatocellular carcinoma (HCC) is the fourth leading cause of cancer-related deaths and the sixth most common cancer in the world.[1] Dynamic contrast-enhanced computed tomography (CT) and magnetic resonance imaging (MRI) are imaging modalities broadly used to assess the response of HCC to locoregional therapy (LRT). Because of the significant correlation between treatment response and patient prognosis, the precise and reliable evaluation of treatment response using imaging tests is crucial.[2]

The Liver Imaging Reporting and Data System (LI-RADS) introduced a treatment response algorithm, wherein, after LRT, the standardized approach can be applied to evaluate the treatment response using contrast-enhanced CT or MRI.[3] Unlike the modified response evaluation criteria for solid tumors, per-lesion treatment response is assessed using the LI-RADS treatment response (LR-TR) algorithm. Treated lesions can be categorized into three LR-TR categories, namely viable, non-viable, and equivocal.[4] In 2024, a revised version of the LR-TR algorithm, divided into categories for post-radiation therapy and non-radiation LRT groups, was released. The LI-RADS non-radiation TR algorithm v.2024 adopts a single major feature, "mass-like enhancement (any degree, any phase)" for assessing viability of treated HCC on CT or MRI. Additionally, for a treated lesion with uncertain mass-like enhancement, two MRI-based ancillary features, such as "diffusion restriction (any degree)" or "mild to moderate T2 hyperintensity", can optionally be used to upgrade from LR-TR equivocal to LR-TR viable.[5]

Gadoxetic acid (Primovist; Bayer Pharma, Berlin, Germany) is a hepatocyte-specific contrast agent in liver MRI, used to identify and characterize various hepatic lesions because it provides the additional benefit of delayed hepatobiliary phase (HBP) imaging.[6-8] Although gadoxetic acid liver MRI provides the aforementioned advantage, arterial phase (AP) images are more frequently degraded than those of other gadolinium-based contrast agents because of contrast-related transient severe motion (TSM).[9] To date, various strategies, such as advanced motion-insensitive MRI sequences, modifications to the injection protocol, and multiple APs, have been reported.[10-17] Several studies reported the usefulness of second-shot arterial phase (SSAP) images in gadoxetic acid-enhanced liver MRI.[17-19] Park et al.[17] reported that SSAP images showed significantly fewer motion artifacts compared with the original AP images.

Main points

- The abbreviated Liver Imaging Reporting and Data System treatment response score in abbreviated liver magnetic resonance imaging (AMRI) with second-shot arterial phase (SSAP) showed non-inferior diagnostic performance compared with full-protocol MRI (FP-MRI) in evaluating the viability for the treatment of hepatocellular carcinoma.
- The scan time of the AMRI can be significantly shorter compared with FP-MRI.
- The mean motion score of the original AP was significantly higher than that of the three SSAPs.

Several studies on SSAP have demonstrated the potential of abbreviated liver MRI (AMRI) with preserved diagnostic performance, compared with full-protocol MRI (FP-MRI). For evaluating hepatic metastasis, AMRI had a significantly shorter acquisition time compared to FP-MRI, while maintaining image quality, diagnostic performance, and visual vascularity.[18] To assess HCC using AMRI with SSAP, the modified LI-RADS category incorporating HBP hypointensity as a major feature demonstrated a high concordance rate (97.4%) with the standard LI-RADS category. In addition, the recall rate was reduced not only in the surveillance but also diagnosis of HCC with AMRI using the SSAP protocol.[19]

Although several studies have investigated SSAP,[17-19] further validation is required to establish the usefulness of the SSAP protocol in evaluating the viability of HCC after LRT. Therefore, we investigated the efficacy of AMRI with SSAP in determining the LR-TR category in patients with HCC following non-radiation LRT.

Methods

This retrospective study was approved by the Ethics Committee: Institutional Review Board of Pusan National University Hospital (approval number: 22504-018-150; approval date: May 9, 2025). Due to the retrospective nature of the study, the requirement for informed consent was waived.

Patients

At our institution, liver MRI was performed between March 2021 and February 2022 using a modified injection protocol that included routine dynamic and SSAP imaging after the first (6 mL) and second (4 mL) injections, respectively. The excellence of the modified injection protocol has been reported in previous studies.[17-19]

We used our institute's electronic database to identify eligible patients. The inclusion criteria were as follows: 1) patients with non-radiation LRT for HCC who underwent the modified liver MRI protocol, and 2) an available reference standard for tumor viability in the treated observation. Before LRT, the diagnosis of HCC was made using the reference standard defined by LI-RADS v.2018[3] or pathologic results. The exclusion criteria included: 1) misregistration of subtraction images and 2) diffuse infiltrative HCC. For patients with multiple lesions, the largest targeted lesion per patient was selected for analysis. A flow diagram of our study is provided in Figure 1.

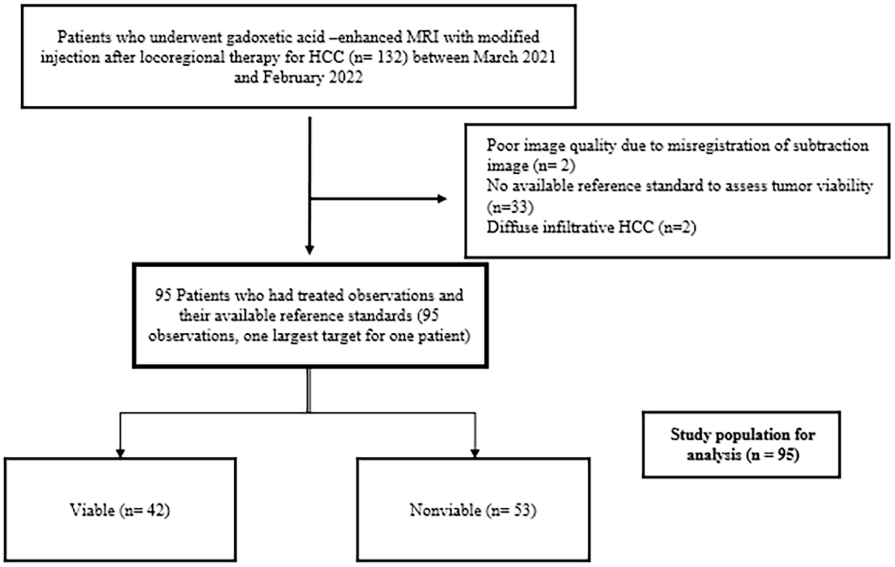

Patients who underwent gadoxetic acid –enhanced MRI with modified injection after locoregional therapy for HCC (n= 132) between March 2021 and February 2022
Poor image quality due to misregistration of subtraction image (n= 2)
No available reference standard to assess tumor viability (n=33)
Diffuse infiltrative HCC (n=2)
95 Patients who had treated observations and their available reference standards (95 observations, one largest target for one patient)
Study population for analysis (n = 95)
Viable (n= 42)
Nonviable (n= 53)

Figure 1. Flow diagram of the study. HCC, hepatocellular carcinoma.

Reference standard

Reference standards for viable tumors in the treated observations were as follows: 1) pathological confirmation (interval between MRI and operation <4 weeks) or 2) tumor staining in transcatheter arterial chemoembolization. Digital subtraction angiography (DSA) images were reviewed to evaluate the tumor staining. Non-viable tumors were considered based on the following: 1) pathologic confirmation (total necrosis) or 2) stability or a decrease in the size of the targeted lesion on follow-up images (at least a 6-month interval from MRI) with no evidence of treatment.

Magnetic resonance imaging techniques

All eligible patients had MRI examinations using a 3.0 T MR scanner (Magnetom Skyra; Siemens Healthineers, Erlagen, Germany) with a 32-element spine matrix coil and a 30-element body matrix coil. Non-contrast-enhanced sequences, such as T1-weighted dual-gradient echo in/out-of-phase sequences, T2-weighted breath-hold half-Fourier acquisition single shot turbo spin echo images, T2-weighted respiratory-triggered single shot images, and diffusion-weighted echo planar images with three b-values (0, 500, and 1,000 s/mm^2). For liver MRI, all patients received 10 mL of gadoxetic acid (Primovist; Bayer Schering Pharma) at a rate of 1 mL/s. Routine AP images (15–20 seconds after gadoxetic acid injection), portal venous (60–90 seconds after gadoxetic acid injection), transitional (180 seconds after gadoxetic acid injection), and HBP (20 minutes after gadoxetic acid injection) images were acquired. After the end of the routine MRI, 4 mL of gadoxetic acid was administered. SSAP images were subsequently acquired in the same manner (Figure 2).

Routine AP and SSAP images were acquired in one and three-phased, respectively. Subtracted images were acquired from the SSAP images. Details of the MRI sequence parameters are provided in Table 1.

Image analysis

One of the radiologists with 22 years of experience in abdominal radiology collected the MRI images and information on the size and location of the target lesions before the review process.

Two radiologists, with 13 and 11 years of experience in liver MRI, respectively, independently and randomly reviewed the FP-MRI and simulated AMRI sets, while being unaware of the clinical data and reference standard. During the review, the reviewers analysed two separate MRI sessions with a 4-week interval to reduce recall bias.

First review session

For FP-MRI, both reviewers classified observations based on the LR-TR algorithms (LR-TR viable, LR-TR equivocal, or LR-TR non-viable). Considering both routine AP and SSAP, the "viable" category was considered for mass-like enhancement (any degree, any phase) within or along the targeted lesion. In addition, ancillary features, such as diffusion restriction and T2-weighted mild-to-moderate hyperintensity, were optionally used.[5]

During the first session, the degree of motion artifact in routine AP, and first, second, and third-phase SSAP was recorded using the following scoring scale:[9] 1) no motion artifact; 2) minimal degree; 3) moderate degree, not significantly affecting diagnosis; 4) severe degree, degraded but interpretable images; and 5) extensive degree, non-diagnostic images.

Second review session

The simulated AMRI set comprised diffusion-weighted images, T2-weighted images, HBP images, and SSAP images (Figure 2). Both reviewers analyzed the arterial hyperenhancement using the SSAP image and its subtraction image. Accordingly, a modified version of the LR-TR algorithm using the abbreviated LR-TR (abbLR-TR) categories (abbLR-TR viable, abbLR-TR equivocal, or abbLR-TR non-viable) was devised. Both reviewers assigned the abbLR-TR classification according to the mass-like enhancement us-

Table 1. Details of the MRI sequence parameters

	Sequence	
	T1 VIBE (routine)*	T1 VIBE (SSAP)
Repetition time, *ms*	4.0	4.0
Echo time, *ms*	1.9	1.9
Flip angle, °	13	13
Field of view	380 x 285	380 x 285
Matrix	384 x 202	320 x 144
Section thickness, *mm*	3.5	3.5
Acquisition time (sec)	15	15
No. of phases acquired	1	3
Parallel acceleration factor†	2 x 2	2 x 2

*Data for the pre-contrast, arterial, portal venous, transitional, and hepatobiliary phases.
†Data are shown as phase direction acceleration factor × partition direction acceleration factor.
MRI, magnetic resonance imaging; SSAP, second-shot arterial phase.

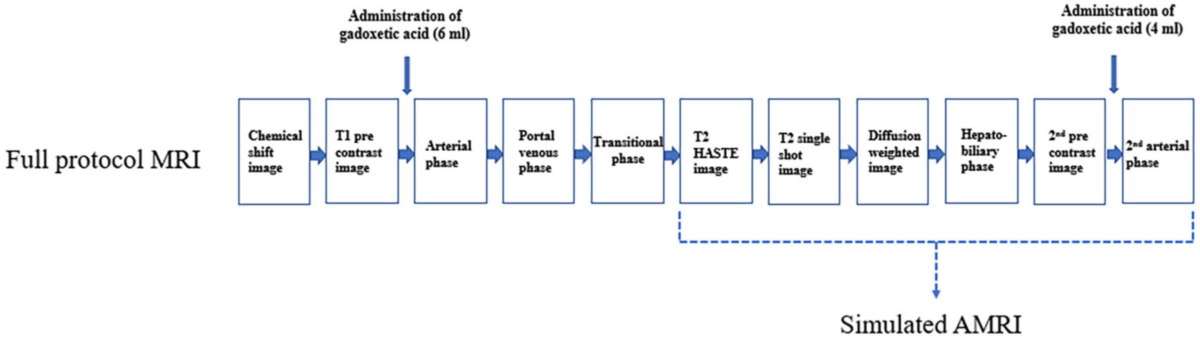


Figure 2. Liver magnetic resonance imaging with modified injection protocol. MRI, magnetic resonance imaging; AMRI, abbreviated liver magnetic resonance imaging.

ing a modified version of the LI-RADS treatment response algorithm. Ancillary features, such as diffusion restriction and T2-weighted mild to moderate hyperintensity, were also optionally used.

Disagreements during the review sessions regarding the motion artifact score and categorizations using LR-TR and abbLR-TR in the targeted lesions were resolved by consensus.

Scan time

For the patients included in the study, one board-certified radiologist compared the scan time between the FP-MRI and simulated AMRI.

Statistical analysis

The scan time between the two protocols (FP-MRI and simulated AMRI) was compared using the Student's t-test. Per-lesion sensitivity and specificity were compared between the two imaging sessions using McNemar's test. The Wilcoxon rank-sum test was used to compare the motion scores between routine AP and each phase of the SSAP in the FP-MRI. We evaluated the inter-observer agreement for the viability evaluation of treated lesions and motion artifact scores using Cohen's kappa. Agreement was defined as poor (kappa=0–0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), or excellent (0.81–1.00). Statistical significance was set at $P < 0.05$. All statistical analyses were conducted using the SPSS software for Windows (v.27.0; IBM Corp., Armonk, NY, USA).

Result

Patient demographics

A total of 132 patients underwent gadoxetic acid-enhanced MRI with modified injection after LRT for HCC. After exclusion, 95 patients with treated observations and their reference standards (95 observations of the largest target for each patient) were included in the study (Figure 1). The causes of chronic liver disease in the patients were chronic hepatitis B (n = 59), chronic hepatitis C (n = 22), alcohol (n = 13), and others (n = 1). Of the 95 patients, 42 (44.2%) had viable lesions and 53 (55.8%) had non-viable lesions. Tumor viability was assessed using histopathology and DSA. Only three viable lesions and one non-viable lesion were confirmed surgically. Other lesions were confirmed using DSA (Table 2).

Scan time

The scan time of the simulated AMRI was significantly shorter compared with the FP-MRI (7.6 ± 0.49 and 23.6 ± 0.50 minutes, respectively; $P < 0.001$).

Diagnostic performance for evaluating viability of treated HCC in the abbreviated liver and full-protocol magnetic resonance imaging

Using the FP-MRI, 36, 5, and 1 lesions were assigned to the LR-TR viable, LR-TR equivocal, and LR-TR non-viable categories, respectively, among 42 viable lesions. Using the AMRI with SSAP, 34, 6, and 2 lesions were assigned to the abbLR-TR viable, abbLR-TR equivocal, and abbLR-TR non-viable categories, respectively, among 42 viable lesions. For evaluating the viability of treated HCC, no significant differences were observed in the sensitivity and specificity between the FP-MRI and AMRI with SSAP [sensitivity, 85.7% (36/42) vs. 80.1% (34/42), respectively, $P = 0.500$; specificity, 96.2% (51/53) vs. 96.2% (51/53), respectively, $P = 1.000$] (Table 3, Figures 3, and 4).

Comparison of the respiratory motion artifacts between the original arterial phase and the second-shot arterial phase

The mean motion score of the original AP was significantly higher compared with the three SSAPs (1.25 vs. 1.04, 1.02, and 1.01; all $P < 0.001$) (Table 4). TSM was not observed in the original AP and SSAP.

Inter-observer agreement

Inter-observer agreement was good for the viability of the target lesion using the LR-TR algorithm in the FP-MRI and the abbLR-TR category in the AMRI with SSAP (k = 0.79 and k = 0.79, respectively). The inter-observer agreements were good for the motion scores on the original AP and the three SSAPs (k = 0.76 and k = 0.65, 0.66, and 0.66, respectively).

Table 2. Patient demographics

Characteristics	Patients
Age	68.7 years old
Male: female	70: 25
Etiology	
Chronic hepatitis B	59
Chronic hepatitis C	22
Alcohol	13
Others	1
Child-Pugh classification	
Previous locoregional treatment	
RFA	22
TACE	73
Reference standards for tumor viability	
Viable	42
Pathologically confirmed	3
Confirmed with DSA	39
Non-viable	53
Pathologically confirmed	1
Clinically confirmed	52
Previous gadoxetic acid-enhanced MRI	88

RFA, radiofrequency ablation; TACE, transcatheter arterial chemoembolization; MRI, magnetic resonance imaging; DSA, digital subtraction angiography.

Table 3. Diagnostic performance for the full-protocol MRI and AMRI with SSAP

	Full protocol MRI	AMRI with SSAP	*P* value
Sensitivity	36/42	34/42	0.500
Specificity	51/53	51/53	1.000

MRI, magnetic resonance imaging; AMRI, abbreviated liver magnetic resonance imaging; SSAP, second-shot arterial phase.

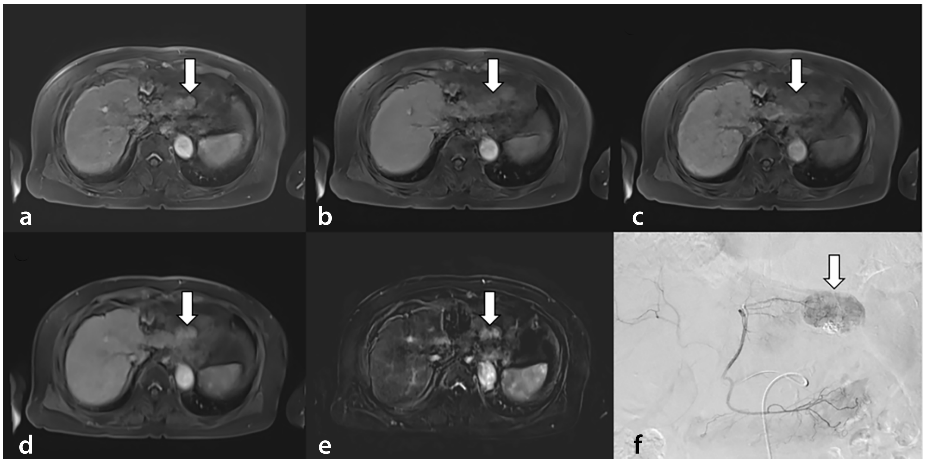


Figure 3. Hepatocellular carcinoma treated with transcatheter arterial chemoembolization (TACE) in a 60-year-old male. On gadoxetic acid-enhanced liver magnetic resonance imaging, a TACE-treated lesion (arrow) in segment II shows mass-like arterial enhancement and washout in the original arterial phase image **(a)** and portal venous phase **(b)**, respectively. In the hepatobiliary phase, it shows hypointensity (arrow) **(c)**. The second-shot arterial phase images, with or without subtraction, demonstrate mass-like arterial enhancement on the treated lesion **(d and e)**. This treated lesion is assigned as Liver Imaging Reporting and Data System treatment response (LR-TR)-viable. When applying the abbreviated LR-TR (abbLR-TR) category, it is assigned as abbLR-TR-viable. The digital subtraction angiography examination demonstrates the presence of the tumor stain (arrow) in the subsequent TACE session **(f)**.

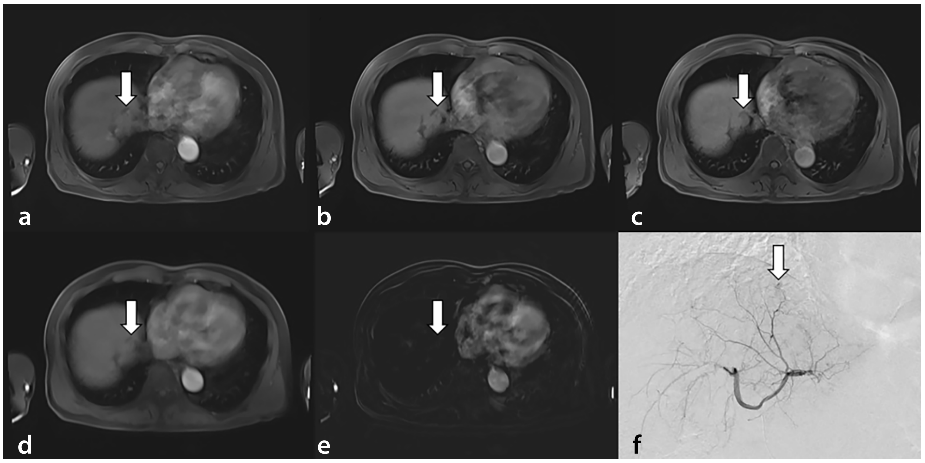


Figure 4. Hepatocellular carcinoma treated with transcatheter arterial chemoembolization (TACE) in a 63-year-old male. On gadoxetic acid-enhanced liver magnetic resonance imaging, a TACE-treated lesion (arrow) in the liver dome shows mass-like arterial enhancement and no washout in the original arterial phase image **(a)** and portal venous phase **(b)**, respectively. In the hepatobiliary phase, it shows hypointensity (arrow) **(c)**. The second- shot arterial phase images, with or without subtraction, demonstrate no mass-like arterial enhancement on the treated lesion **(d and e)**. This treated lesion is assigned as Liver Imaging Reporting and Data System treatment response (LR-TR)-viable. When applying the abbreviated LR-TR (abbLR-TR) category, it is assigned as abbLR-TR non-viable. The digital subtraction angiography examination demonstrates the presence of the tumor stain (arrow) in the subsequent TACE session **(f)**.

Table 4. Mean motion scores

	Mean motion score	*P* value (vs. original AP)
Original AP	1.25	
SSAP1	1.04	<0.001
SSAP2	1.02	<0.001
SSAP3	1.01	<0.001

AP, arterial phase; SSAP, second-shot arterial phase.

Discussion

In this study, the scan time of the simulated AMRI was significantly shorter compared with the FP-MRI. Contrastingly, the abbLR-TR category in the AMRI with SSAP showed diagnostic performance comparable to that of the LR-TR algorithm in the FP-MRI. In addition, the mean motion score of the original AP was significantly higher than that of SSAP.

In previous studies on SSAP, the image acquisition times for AMRI were significantly shorter; they also demonstrated no significant difference between the FP-MRI and AMRI with SSAP in the diagnosis of HCC or hepatic metastasis,[18,19] which are consistent with our results. However, our study focused on the viability of HCC after LRT.

In the current study, the reference standards for tumor viability included only a few pathologically confirmed results [viable, 7.1% (3/42); non-viable, 1.9% (1/53)]. In a previous meta-analysis for the LR-TR algorithm, the pooled specificity and sensitivity of the LR-TR viable category were 96% [95% confidence interval (CI): 91%–99%] and 63% (95% CI: 39%–81%), respectively.[20] The same study also conducted a meta-regression study for the reference standard. The researchers reported that studies using reference standard to imaging findings (imaging follow-up or imaging/pathologic results) demonstrated significantly higher sensitivity compared to those using pathology alone (81% vs. 48%, respectively). However, LR-TR v2024 was not adopted in the studies included in the meta-analysis. Recently, Zhou et al.[21] reported that LI-RADS non-radiation TRA v.2024 improved the sensitivity (85.5% and 87.2%) of assessing the viability of treated HCC, using ancillary features with reference to the only pathologic results. The sensitivity of our study for treated lesions with FP-MRI and AMRI with SSAP was 85.7% and 80.9%, respectively. In addition, the specificities were 96.2% for both protocols. Accordingly, our results showed a diagnostic performance comparable to those of previous studies. However, further studies evaluating the diagnostic performance of LR-TR A v.2024 are required.

Hepatobiliary-specific MR contrast agents (e.g., gadoxetic acid, gadobenate dimeglumine, mangafodipir trisodium) have previously been used in the diagnosis of HCC. These contrast agents, taken up by hepatocytes, can provide the T1 shortening effect of normal liver parenchyma, resulting in high signal intensity on HBP. For gadoxetic acid, 50% of the injected contrast media is transported to hepatocytes. Gadoxetic acid acts as an extracellular contrast agent in the arterial and portal venous phases. In addition, an HBP image can be obtained with a relatively short delay of 10–20 minutes after injection.[22] Therefore, gadoxetic acid is useful for the detection of HCC.[23,24] In evaluating the viability of treated HCC using gadoxetic acid-enhanced liver MRI, the HBP image can aid in the detection of a lesion treated with LRT. However, according to LR-TR v.2024, HBP hypointensity is not accepted as an ancillary feature for assessing the viability of treated HCC. Accordingly, it was not applied in either the FP-MRI or AMRI in this study. Although the presence of HBP hypointensity is not a major or ancillary feature in the LR-TR v.2024, the inclusion of an HBP image may be necessary for the AMRI protocol for assessing the viability of treated HCC after LRT. This is because the detection of newly developed HCC is also clinically important for patients with a history of LRT. For treatment-naïve lesions, HBP hypointensity was reported as a useful feature for the detection of HCC. Kim et al.[23] reported that the extension of washout to the transitional phase or HBP allowed for higher sensitivity without a reduction in specificity, rather than restricting it to the PVP after excluding typical hemangiomas and nodules with a targetoid appearance. In another study by Joo and colleagues, the diagnostic criteria extending washout to the HBP demonstrated higher sensitivity than those limiting washout to the PVP, with little loss of specificity.[24]

Recently, LR-TR v.2024 adopted a single major feature, "mass-like enhancement (any degree, any phase)" for assessing the viability of treated HCC on CT or MRI.[5] After LRT of the hypervascular HCC, the representative imaging finding suggesting viable HCC is "mass-like enhancement (any degree, any phase)". Zhou et al.[21] reported that LI-RADS non-radiation TRA v.2024 without ancillary features, using a single major feature, "mass-like enhancement (any degree, any phase)" for assessing the viability of treated HCC, demonstrated higher sensitivity than LI-RADS TRA v.2017 (80.3% and 81.1% vs. 79.1% and 79.9%, respectively). This result supports the diagnostic value of "mass-like enhancement (any degree, any phase)" as representative imaging findings, suggesting viable HCC. Zhou et al.[21] also reported that LI-RADS non-radiation TRA v.2024 with ancillary features provided significantly higher sensitivity than LI-RADS TRA v.2017 (85.5% and 87.2% vs. 79.9% and 79.1%; all $P < 0.001$). This result emphasized the importance of the ancillary features, such as diffusion restriction and mild-to-moderate T2 hyperintensity, in evaluating the viability of treated HCC after radiation-free LRT. In addition, this result supports the inclusion of diffusion-weighted imaging and T2-weighted imaging in the AMRI protocol for assessing the viability of treated HCC after LRT.

The mean motion score of the original AP was significantly higher than that of the three SSAPs. These results are consistent with those of Park et al.[17] However, TSM was not observed in our group. A previous study including a large number of patients demonstrated that the presence of hepatitis B and previous experience with gadoxetic acid-enhanced MRI were negative risk factors for TSM.[25] Our study also included a large portion of patients with previous experience with gadoxetic acid-enhanced MRI (88/95) and chronic hepatitis B (59/95). These factors are believed to be the reasons for the lack of TSM. In addition to the relative motion insensitivity of SSAP, the multiple APs (three-phase, in our study) of SSAP may also be one of the reasons for the non-inferior sensitivity to FP-MRI in our study. In a previous study conducted by Hong et al.[15], multiple APs had a lower incidence of TSM than single AP and significantly improved sensitivity for diagnosing HCC (≤3 cm), without a significant decrease in specificity.

This study has some limitations. First, it was a retrospective study that included a relatively small number of patients; thus, there may have been selection bias. However, the data collection was performed consecutively. A prospective study is required to evaluate the diagnostic performance of AMRI using SSAP. Second, reference standards for tumor viability included few pathologically confirmed results [viable, 7.1% (3/42); non-viable, 1.9% (1/53)], and our study included patients treated with only non-radiation LRT. However, the diagnostic performance of the abbLR-TR category for AMRI with SSAP was not significantly different from that reported in previous studies.[20,21]

In conclusion, the abbLR-TR category in the AMRI with SSAP showed non-inferior diagnostic performance compared to FP-MRI in evaluating the viability of the treated HCC after non-radiation LRT. Therefore, this abbreviated protocol may serve as a faster and more convenient alternative for post-treatment surveillance following non-radiation LRT.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Allard MA, Sebagh M, Ruiz A, et al. Does pathological response after transarterial chemoembolization for hepatocellular carcinoma in cirrhotic patients with cirrhosis predict outcome after liver resection or transplantation? *J Hepatol*. 2015;63(1):83-92. [CrossRef]
2. Ho MH, Yu CY, Chung KP, et al. Locoregional therapy-induced tumor necrosis as a predictor of recurrence after liver transplant in patients with hepatocellular carcinoma. *Ann Surg Oncol*. 2011;18(13):3632-3639. [CrossRef]
3. Chernyak V, Fowler KJ, Kamaya A, et al. Liver Imaging Reporting and Data System (LI-RADS) version 2018: imaging of hepatocellular carcinoma in at-risk patients. *Radiology*. 2018;289(3):816-830. [CrossRef]
4. Kielar A, Fowler KJ, Lewis S, et al. Locoregional therapies for hepatocellular carcinoma and the new LI-RADS treatment response algorithm. *Abdom Radiol (NY)*. 2018;43(1):218-230. [CrossRef]
5. American College of Radiology website. LI-RADS Ct/Mri Nonradiation Tra V2024 Core 2024 [cited 2024 Apr 18]. [CrossRef]
6. Sun HY, Lee JM, Shin CI, et al. Gadoxetic acid-enhanced magnetic resonance imaging for differentiating small hepatocellular carcinomas (< or =2 cm in diameter) from arterial enhancing pseudolesions: special emphasis on hepatobiliary phase imaging. *Invest Radiol*. 2010;45(2):96-103. [CrossRef]
7. Kim YK, Lee MW, Lee WJ, et al. Diagnostic accuracy and sensitivity of diffusion-weighted and of gadoxetic acid-enhanced 3-T MR imaging alone or in combination in the detection of small liver metastasis (≤ 1.5 cm in diameter). *Invest Radiol*. 2012;47(3):159-166. [CrossRef]
8. Park YS, Lee CH, Kim JW, Shin S, Park CM. Differentiation of hepatocellular carcinoma from its various mimickers in liver magnetic resonance imaging: what are the tips when using hepatocyte-specific agents? *World J Gastroenterol*. 2016;22(1):284-299. [CrossRef]
9. Davenport MS, Viglianti BL, Al-Hawary MM, et al. Comparison of acute transient dyspnea after intravenous administration of gadoxetate disodium and gadobenate dimeglumine: effect on arterial phase image quality. *Radiology*. 2013;266(2):452-461. [CrossRef]
10. Hong SB, Lee NK, Kim S, et al. Modified CAIPIRINHA-VIBE without view-sharing on

gadoxetic acid-enhanced multi-arterial phase MR imaging for diagnosing hepatocellular carcinoma: comparison with the CAIPIRINHA-Dixon-TWIST-VIBE. *Eur Radiol*. 2019;29(7):3574-3583. [CrossRef]

11. Pietryga JA, Burke LM, Marin D, Jaffe TA, Bashir MR. Respiratory motion artifact affecting hepatic arterial phase imaging with gadoxetate disodium: examination recovery with a multiple arterial phase acquisition. *Radiology*. 2014;271(2):426-434. [CrossRef]
12. Park YS, Lee CH, Kim IS, et al. Usefulness of controlled aliasing in parallel imaging results in higher acceleration in gadoxetic acid-enhanced liver magnetic resonance imaging to clarify the hepatic arterial phase. *Invest Radiol*. 2014;49(3):183-188. [CrossRef]
13. Yoo JL, Lee CH, Park YS, et al. The short breath-hold technique, controlled aliasing in parallel imaging results in higher acceleration, can be the first step to overcoming a degraded hepatic arterial phase in liver magnetic resonance imaging: a prospective randomized control study. *Invest Radiol*. 2016;51(7):440-446. [CrossRef]
14. Yoon JH, Lee JM, Yu MH, et al. Evaluation of transient motion during gadoxetic acid-enhanced multiphasic liver magnetic resonance imaging using free-breathing golden-angle radial sparse parallel magnetic resonance imaging. *Invest Radiol*. 2018;53(1):52-61. [CrossRef]
15. Hong SB, Hong S, Choi SH, et al. Multiple arterial-phase mri with gadoxetic acid improves diagnosis of hepatocellular carcinoma </=3.0 cm. *Liver Int*. 2023;43(2):462-470. [CrossRef]
16. Polanec SH, Bickel H, Baltzer PAT, et al. Respiratory motion artifacts during arterial phase imaging with gadoxetic acid: can the injection protocol minimize this drawback? *J Magn Reson Imaging*. 2017;46(4):1107-1114. [CrossRef]
17. Park YS, Lee J, Kim JW, Park CM, Lee CH. Second shot arterial phase to overcome degraded hepatic arterial phase in liver MR imaging. *Eur Radiol*. 2019;29(6):2821-2829. [CrossRef]
18. Kim JW, Lee CH, Park YS, Lee J, Kim KA. Abbreviated gadoxetic acid-enhanced MRI with second-shot arterial phase imaging for liver metastasis evaluation. *Radiol Imaging Cancer*. 2019;1(1):e190006. [CrossRef]
19. Kim JW, Lee CH, Kim KA, Lee J, Park YS. Abbreviated MRI with second shot arterial phase for HCC evaluation: modified version of LI-RADS and recall reduction strategy. *Eur Radiol*. 2023;33(6):4401-4411. [CrossRef]
20. Youn SY, Kim DH, Choi SH, et al. Diagnostic performance of Liver Imaging Reporting and Data System treatment response algorithm: a systematic review and meta-analysis. *Eur Radiol*. 2021;31(7):4785-4793. [CrossRef]
21. Zhou S, Zhou G, Shen Y, et al. LI-RADS nonradiation treatment response algorithm version 2024: diagnostic performance and impact of ancillary features. *AJR Am J Roentgenol*. 2025;224(2):e2432035. [CrossRef]
22. Seale MK, Catalano OA, Saini S, Hahn PF, Sahani DV. Hepatobiliary-specific MR contrast agents: role in imaging the liver and biliary tree. *Radiographics*. 2009;29(6):1725-1748. [CrossRef]
23. Kim DH, Choi SH, Kim SY, Kim MJ, Lee SS, Byun JH. Gadoxetic acid-enhanced MRI of hepatocellular carcinoma: value of washout in transitional and hepatobiliary phases. *Radiology*. 2019;291(3):651-657. [CrossRef]
24. Joo I, Lee JM, Lee DH, Jeon JH, Han JK. Retrospective validation of a new diagnostic criterion for hepatocellular carcinoma on gadoxetic acid-enhanced MRI: can hypointensity on the hepatobiliary phase be used as an alternative to washout with the aid of ancillary features? *Eur Radiol*. 2019;29(4):1724-1732. [CrossRef]
25. Jang EB, Kim DW, Choi SH, et al. Transient severe motion artifacts on gadoxetic acid-enhanced mri: risk factor analysis in 2230 patients. *Eur Radiol*. 2022;32(12):8629-8638. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253640

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Texture analysis enhances diagnostic accuracy of lesions scored as 5 in the Prostate Imaging Reporting and Data System in magnetic resonance imaging

Yan Bai[1-3]
Xin Ru Xie[3]
Ying Hou[1,3]
Yu Dong Zhang[1,3]
Hai Bin Shi[1,3]*
Chen Jiang Wu[1,3]*

[1]The First Affiliated Hospital with Nanjing Medical University, Department of Radiology, Nanjing, China

[2]The First Affiliated Hospital of Baotou Medical College, Inner Mongolia University of Science and Technology, Baotou, China

[3]School of Medical Imaging, Nanjing Medical University, Nanjing, China

* There authors contributed equally to this work

Corresponding author: Chen Jiang Wu

E-mail: njmu_wcj@163.com

Received 08 September 2025; revision requested 05 October 2025; accepted 08 November 2025.

Epub: 01.12.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253640

PURPOSE

Prostatitis is frequently observed in false-positive lesions scored as 5 in the Prostate Imaging Reporting and Data System (PI-RADS), necessitating improved diagnostic tools. This study investigated the potential of magnetic resonance imaging (MRI) texture analysis of apparent diffusion coefficient (ADC) sequences to enhance the differentiation of prostatitis from prostate cancer (PCa) in PI-RADS 5 lesions.

METHODS

This retrospective study enrolled patients undergoing 3.0-T MRI with lesions scored as PI-RADS 5. Lesions were manually delineated on ADC maps, and texture features were extracted using FireVoxel. Clinical data and ADC texture parameters were collected. The diagnostic performance [area under the curve (AUC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV)] of the clinical data, ADC texture, and a combined model were calculated and compared using the DeLong test.

RESULTS

The final cohort included 189 patients with 189 PI-RADS 5 lesions (164 PCa, 25 prostatitis). The combined model, incorporating clinical indicators (age, prostate-specific antigen density) and ADC texture parameters (signal coefficient of variation, ADC percentile), revealed the optimal diagnostic performance: SEN 98.7%, SPE 60.0%, PPV 97.9%, NPV 71.6%, and AUC 93.1%. Bootstrap resampling verified the robustness of the model. Decision curve analysis indicated an improved net benefit with the combined model for guiding biopsy decisions.

CONCLUSION

ADC imaging texture parameters are valuable for the differential diagnosis of prostatitis from lesions scored as PI-RADS 5. Their combination with clinical indicators substantially improves diagnostic performance, providing valuable information to facilitate surgical decision-making and potentially reduce unnecessary biopsies.

CLINICAL SIGNIFICANCE

This study addresses a critical limitation of the current PI-RADS system, which exhibits a notable rate of false positives in high-risk PI-RADS 5 lesions. By demonstrating the added value of quantitative ADC texture analysis in this specific diagnostic challenge, this research offers a practical and potentially translatable approach to reducing the number of unnecessary biopsies for PI-RADS 5 lesions.

KEYWORDS

Prostate cancer, prostatitis, PI-RADS 5, magnetic resonance imaging, diffusion-weighted imaging, apparent diffusion coefficient, texture analysis, differential diagnosis

You may cite this article as: Bai Y, Xie XR, Hou Y, Zhang YD, Shi HB, Wu CJ. Texture analysis enhances diagnostic accuracy of lesions scored as 5 in the Prostate Imaging Reporting and Data System in magnetic resonance imaging. *Diagn Interv Radiol*. 2026;32(2):156-163.

Advances in magnetic resonance imaging (MRI) have substantially enhanced the detection and diagnosis of prostate cancer (PCa).[1-4] The introduction of the Prostate Imaging Reporting and Data System (PI-RADS)[5] has established a crucial communication bridge between radiologists and clinicians. However, a notable issue with PI-RADS scores of 4 and 5 (considered high-risk lesions) is the high rate of false positives.[6,7] Notably, 15%–35% of these high-risk lesions are histologically benign,[8] and false-positive results lead to unnecessary biopsies,[9,10] which are invasive procedures that carry risks of complications and can erode the trust between clinicians and radiologists. The PI-RADS scoring system,[11,12] mainly based on signal intensity, may not fully capture the pathological changes in lesions.

Previous studies have revealed that among lesions with a PI-RADS score of 4 or 5 in the prostate, 14.8% (27/182) are benign, with 81.5% (22/27) of these benign cases diagnosed as prostatitis.[13] Studies have identified that apparent diffusion coefficient (ADC) values can help reduce false positives in PI-RADS 4 and 5 lesions.[14] However, research has not fully explored the imaging data, and the effectiveness of reducing false positives is limited. The development of radiomics has brought new advancements to the diagnosis of PCa.[15] Studies have demonstrated that radiomics can be used to differentiate the malignancy of PCa and distinguish clinically significant PCa with PI-RADS scores of 4 and 5.[16] However, radiomics faces several challenges, such as poor model generalizability, lack of biological interpretability, and high computational costs. Texture analysis is an essential component of radiomics and plays a crucial role in medical image analysis. By quantifying grayscale patterns and intensity variations in images, texture analysis provides deep insights into tissue heterogeneity and pathological features. This method has demonstrated substantial diagnostic efficacy in various medical applications, particularly in tumor detection and grading.[17] Research has indicated that texture analysis can be used to distinguish between benign and malignant diseases and to assess the malignancy of PCa.[18] However, this approach has not been applied to detecting lesions with PI-RADS scores of 5. This study aims to use quantitative imaging biomarkers to differentiate between cancerous and non-cancerous lesions with PI-RADS 5 scores, exploring their diagnostic accuracy.

Main points

- Texture analysis of apparent diffusion coefficient (ADC) maps enhances differentiation of prostate cancer (PCa) from benign prostatitis in lesions scored as 5 in the Prostate Imaging Reporting and Data System (PI-RADS), reducing false positives and unnecessary biopsies.
- Age, prostate-specific antigen density, and ADC texture parameters (signal coefficient of variation, ADC percentile) are independent predictors for distinguishing prostatitis from PCa in PI-RADS 5 lesions.
- A combined model integrating clinical and ADC texture parameters achieves superior diagnostic accuracy (area under the curve 93.1%) and clinical utility for improved patient management.

Methods

Patient selection

This study was approved by the Ethics Committee of The First Affiliated Hospital of Nanjing Medical University (protocol number: 2023-SRFA-467, date: 03.14.2023), with informed consent waived due to the retrospective nature of the study. We identified consecutive patients who underwent standard pelvic MRI examination before treatment by reviewing the department database for records from the period of January 2021 to July 2024. Clinical characteristics were obtained from the patient records in our hospital. Histopathological results were verified through cognitive fusion biopsies, transurethral resection of the prostate, and radical prostatectomy. Patients with the following criteria were included: (a) no prior hormonal or radiation treatment; (b) diffusion-weighted imaging (DWI) performed on a 3.0-T MRI scanner and with unified sequence parameters; and (c) lesions with pathologically confirmed PCa or prostatitis, with MRI demonstrating at least one lesion with a diameter ≥1.5 cm.

Clinical and laboratory data

Demographic data included age, and clinical data included the prostate-specific antigen (PSA) level and serum white blood cell count. The prostate volume was calculated using the following formula: (maximum anteroposterior diameter) × (maximum transverse diameter) × (maximum longitudinal diameter) × 0.52. PSA density (PSAD) was calculated by dividing the PSA level by the prostate volume.

Magnetic resonance imaging acquisition

All patients underwent a standardized prostate examination using 3.0-T MRI (Verio/Skyra, Siemens, Erlangen, Germany; u770, United Imaging, Shanghai, China) related to the probability of subsequent PSA progression (26), which complied with the PI-RADS guidelines. The multi-parametric MRI protocol included the following: (1) T2-weighted images on axial, sagittal, and coronal planes; (2) DWI on the axial plane with automatically generated ADC maps; and (3) T1-weighted sequences on axial planes with and without fat saturation. The MRI protocol technical details are listed in Table 1.

Imaging and histological correlation

The pathological results were taken as the "gold standard." The biopsy method used was transperineal prostate biopsy guided by the fusion of ultrasound and MR, and the biopsy was completed by experienced urologists. The surgical methods included transurethral prostatectomy and radical prostatectomy. The prostate samples obtained through the above methods were uniformly processed and sent to the pathology department for diagnosis. A genitourinary pathologist with 10 years of experience in genitourinary histopathology reviewed all the sample sections. Pathological diagnosis was divided into benign and malignant. Benign lesions included benign prostatic hyperplasia, prostatitis, abscess, and normal prostate tissue. Cases of PCa were graded according to the 2005 International Society of Urological Pathology Modified Gleason Grading System. Malignant lesions were further classified into Gleason score (GS) pathological grades: GS = 3 + 3, 3 + 4, and 4 + 3 and GS ≥ 4 + 4.

Texture analysis

All data were transferred in Digital Imaging and Communications in Medicine format. The region of interest (ROI) was manually delineated slice by slice along the boundaries of the tumor by one radiologist (8 years of clinical experience in prostate imaging), and the segmentations were cross-checked by another two experienced genitourinary specialist radiologists. FireVoxel (CAI2R, New York, NY, USA),[19] was used for the three-dimensional segmentation of the prostate lesions. Since PI-RADS 5 lesions all have a diameter >1.5 cm, we only selected the largest main lesion for each patient. First order and geometrical features were automatically extracted by FireVoxel. Based on histogram analysis, the following parameters were derived from the ADC

Table 1. Prostatic magnetic resonance imaging parameters of the 3.0-T scanner

Parameters	Imaging sequence		
	T2WI	T1WI	DWI
Repetition time (msec)	6.000	600	6.000
Echo time	105	24	82
Field of view (cm^2)	22	22	22
Matrix	384 × 384	384 × 384	128 × 128
Section thickness (mm)	3.5	3.5	3.5
Flip angle (degree)	110	90	90
b value (sec/mm^2)	…	…	0, 500, 1,000, and 1,500
Number of slices	25	23	23
Number of averages	2	2	2
Bandwidth/pixel	180	180	2.060
Parallel factors	2	2	2
Acquisition time (min)	3:46	3:30	3:48

T2WI, T2-weighted imaging; T1WI, T1-weighted imaging; DWI, diffusion-weighted imaging.

map: (a) minimum; (b) maximum; (c) mean; (d) kurtosis, which is the degree of peakedness of a distribution; (e) skewness, which is a measure of the degree of asymmetry of a distribution; (f) entropy, which quantifies the randomness of the gray-level distribution in an image, with higher values indicating a more dispersed and complex distribution of gray-level values; (g) coefficient of variation (COV), which quantifies the relative variability of pixel values, normalized by the mean intensity, making it useful for comparing heterogeneity across different images or ROIs; and (h) variance, which quantifies the dispersion of pixel values around the mean in the ADC, with greater variance reflecting increased tissue heterogeneity. For the cumulative histogram, the 25th, 50th, and 75th percentiles of the tumor ADC were derived (the nth percentile is the point at which n% of the voxel values that form the histogram are found to the left).

Statistical analysis

MedCalc Statistical Software, version 23.0.8 (MedCalc Software Ltd, Ostend, Belgium) was used for statistical analysis, with $P < 0.05$ considered statistically significant. The PI-RADS 5 lesions were divided into a PCa group and an inflammation group based on the pathological results. The Shapiro–Wilk and Kolmogorov–Smirnov tests were used to test the normal distribution of measurement data, and the Levene test was used to test the homogeneity of variance of measurement data. According to the test results, the normal distribution data were expressed as mean ± standard deviation (SD), the skewed distribution data were expressed as the median [upper and lower quartile (M (Q1, Q3))], and the measurement data were expressed as n (%).

Univariate and multivariate analyses were used to screen the independent risk factors for identifying inflammation and PCa in patients with PI-RADS 5 scores. If two parameters were highly correlated (e.g., PSA and PSAD, $|r| > 0.7$), the variable with greater clinical significance or a smaller *P* value in the univariate analysis was retained. The independent risk factors were combined to establish clinical ADC texture and combined models; the receiver operating characteristic (ROC) curve was drawn for the different screened models. The efficacy of different factors and models in differentiating PCa from inflammation in PI-RADS 5 lesions was evaluated. The area under the ROC curve (AUC) was used for quantification. Diagnostic sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV) were calculated at the cut-off point with the largest Youden index. To perform decision curve analysis, R version 3.5.1 was used; this evaluates the clinical utility of the diagnostic models by assessing the net benefit of using the models to guide clinical decisions.

To evaluate the internal validity and predictive performance of the combined model, bootstrap resampling was used to generate 1,000 random samples with replacements (Python 3.13, with the sklearn.metrics and matplotlib.pyplot libraries). The ROC curve and its confidence interval (CI) were generated by calculating the mean and 95% CI of the bootstrap sample area under the curve (AUC) as well as the mean and 95% CI (SD multiplied by 1.96) of the interpolated true positive rate. This method provides a robust estimate of model performance and quantifies the uncertainty from the limited sample.

Results

Patient and lesion characteristics

The process of patient exclusion and inclusion is shown in Figure 1. A total of 189 patients were finally included, with a total of 189 tumor foci with a diameter > 1.5 cm detected in histological findings. Of the included lesions, 108 (57.1%) originated in the peripheral zone (PZ), 54 (28.5%) in the transition zone (TZ), and the remaining 27 (14.3%)in both the PZ and TZ. The median prostate volume was 27.5 mL (33.5–49.2 mL). The clinical characteristics of the patients and the tumor foci ROIs are summarized in Table 2. Representative MRI images and ADC histograms of the PCa and prostatitis cases are shown in Figures 2 and 3.

Clinical apparent diffusion coefficient texture parameters for predicting prostatitis from PI-RADS 5 lesions

Univariate and multivariate analyses were conducted to identify independent risk factors for differentiating inflammation in PI-RADS 5 lesions. Based on the clinical data, the results showed that age [odds ratio (OR): 1.081 (95% CI: 1.017–1.149)] and PSAD [OR: 35.540 (95% CI: 3.534–357.449)] are independent risk factors for diagnosing prostatitis. Based on texture feature data from whole-tumor ADC image analysis, ADC percentage values are independent risk factors for predicting prostatitis in PI-RADS 5 lesions (OR: 0.983–0.998), and signal COV has the highest OR value [OR: 1.587.241.411 (95% CI: 3.431–7.342 E + 11)] (Table 3). In these predictive parameters, age, PSAD, and signal COV are positively correlated with the prediction outcome, whereas the other variables show a negative correlation.

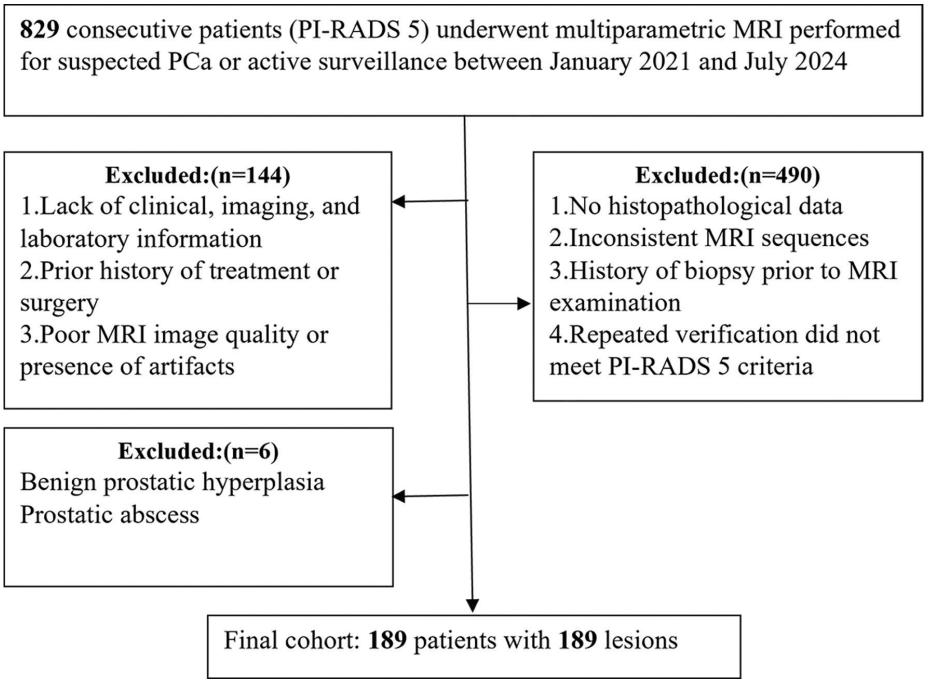

829 consecutive patients (PI-RADS 5) underwent multiparametric MRI performed for suspected PCa or active surveillance between January 2021 and July 2024
Excluded:(n=144)
1.Lack of clinical, imaging, and laboratory information
2.Prior history of treatment or surgery
3.Poor MRI image quality or presence of artifacts
Excluded:(n=490)
1.No histopathological data
2.Inconsistent MRI sequences
3.History of biopsy prior to MRI examination
4.Repeated verification did not meet PI-RADS 5 criteria
Excluded:(n=6)
Benign prostatic hyperplasia
Prostatic abscess
Final cohort: 189 patients with 189 lesions

Figure 1. Flowchart of patient exclusion and inclusion. PI-RADS, Prostate Imaging Reporting and Data System; MRI, magnetic resonance imaging; PCa, prostate cancer.

Model diagnostic performance

Table 4 and Figure 4A demonstrate the diagnostic efficacy of different diagnostic models. Age and PSAD were integrated into the clinical model with diagnostic efficacy as follows: AUC of 84.6%, SEN of 97.4%, SPE of 16.0%, PPV of 95.7%, and NPV of 25.1%. The ADC texture model includes two parameters: the ADC median and signal COV. The diagnostic efficacy of this model is as follows: AUC of 86.5%, SEN of 88.1%, SPE of 44.4%, PPV of 96.8%, and NPV of 16.4%. The combined model shows a significant improvement in diagnostic performance compared with the two models mentioned above. The diagnostic efficacy is as follows: AUC of 93.1%, SEN of 98.7%, SPE of 60.0%, PPV of 97.9%, and NPV of 71.6%. There was no significant difference in diagnostic performance between the ADC texture model and clinical model (P = 0.7697). The decision curve showed that when the threshold was between 0.1 and 0.8 (Figure 4B), the combined model obtained significant clinical benefits when deciding whether to perform a biopsy. The bootstrap results showed that the performance of the combined model was robust [93.2% (95% CI: 85.6%–98.6%); Figure 5] in internal validation and superior to either the clinical model or the ADC texture model.

Discussion

The present study determined that texture analysis of ADC MRI combined with clinical parameters significantly improved the differentiation of prostatitis from PCa in PI-RADS 5 lesions. Improving the PPV of PI-RADS 5 lesions can potentially reduce unnecessary biopsies and the associated patient anxiety and morbidity. It might also lead to more appropriate treatment strategies based on a more accurate diagnosis.

Previous studies have demonstrated that PSAD identifies significant differences between prostatitis and PCa[20] and is significantly more effective than PSA in differentiating between benign and malignant histology. Integration of PSAD into decision-making for prostate biopsy may facilitate improved risk-adjusted care.[21-23] These findings are consistent with those of the present study. However, after using simple clinical parameters to build a prediction model, the AUC of the clinical model was significantly lower than that of the model constructed using ADC texture analysis, demonstrating the limitations of PSAD in identification. In addition, MRI not only measures prostate volume but also extracts texture features from ADC maps, making it a key supplementary source of data beyond clinical indicators. We found that in PI-RADS 5 lesions, the ADC parameters of the tumor were significantly lower than those of the inflammation group, which is consistent with previous studies.[24,25] However, our study only focused on PI-RADS 5 lesions, whereas other studies did not differentiate lesion PI-RADS scores; therefore, the numerical differences were greater than in this study. The narrowed data difference reflects the difficulty of the accurate diagnosis of PI-RADS 5 lesions and reflects the inadequacy of the PI-RADS scoring system, which is based on subjective qualitative evaluation of MRI. This reflects the importance of quantitative indicators for accurate diagnosis of prostate lesions, which may be of reference value for the future development of PI-RADS. In our study, there was no significant difference in histogram skewness, histogram kurtosis, or histogram entropy. This is inconsistent with previous findings and may be related to the small number of samples we included, especially the small number of prostate inflammation cases.

The study by Cheng et al.[26] found that combining clinical parameters and ADC values improved the PPV of PI-RADS 5 lesions, which is similar to our results, but they used the ADC mean and minimum without texture analysis to fully explore the possible morphological differences between PCa and inflammation. The study by Bonaffini et al.[16] used radiomics to distinguish PCa from non-cancer cases. However, with the deepening of radiomics research, it was found that the spatial resolution of MRI is still far behind that of computed tomography, and prostate lesions are usually small. Smaller ROIs may not fully utilize the advantages of radiomics, which, in turn, affects repeatability and limits its interpretability. Texture analysis goes beyond simple visual inspection and provides objective measures of tissue heterogeneity. Therefore, our expectation is that texture analysis can help distinguish between tumor and non-tumor PI-RADS 5 lesions. The results of our study validate our assumptions.

Based on previous research evidence, the findings of this study may be associated with the corresponding pathophysiological changes. Although PCa and prostate inflammation show a decrease in ADC signal, the mechanisms of the two are significantly different. The mechanism of ADC decrease in PCa is mainly caused by the increase in cell density per unit of volume caused by abnormal tumor proliferation,[27-29] which restricts the diffusion of water molecules between cells. Inflammation is related to cell infiltration in the acute phase and fibrous repair in the chronic phase.[28] Compared with the significant decrease in tumor ADC value, the decrease in inflammation ADC value is relatively mild. Histogram analysis of texture analysis can detect the above changes well on MRI.

The study's strengths include the use of appropriate, reasonable, and interpretable methods to perform texture analysis

Table 2. Patient and lesion characteristics

Lesions (n = 189)	Prostatitis (n = 25)	Prostate cancer (n = 164)	*P* value
Clinical features			
Age (years)	67.7 ± 7.2	72.0 ± 8.1	0.010
PSA (ng/mL)	8.6 (5.9–17.8)	15.1 (9.8–30.2)	0.002
PSAD (ng/mL^2)	0.18 (0.12–0.39)	0.48 (0.25–0.88)	< 0.001
Prostate volume (mL)	49.8 (35.0–60.5)	32.0 (26.5–42.9)	< 0.001
WBC ($\times 10^9$/L)	6.8 (5.5–8.6)	5.9 (5.2–7.3)	0.136
Location			
PZ (n, %)	12 (11.2)	96 (88.8)	
TZ (n, %)	10 (18.9)	44 (81.1)	
PZ + TZ (n, %)	3 (12.0)	24 (88.0)	
ISUP grade group (n/%)			
1 (GS: 3 + 3)		4 (2.4)	
2 (GS: 3 + 4)		34 (20.7)	
3 (GS: 4 + 3)		70 (42.6)	
4 (GS: 8)		3 (20.7)	
5 (GS ≥ 9)		22 (13.4)	
Specimen types (n/%)			
Biopsy	16 (64.0)	34 (20.7)	
TURP	1 (4.0)	7 (4.3)	
Radical surgery + biopsy	2 (8.0)	74 (45.1)	
TURP + biopsy	6 (24.0)	21 (12.8)	
Radical surgery		28 (17.0)	

PSA, prostate-specific antigen; PSAD, prostate-specific antigen density; WBC, white blood cell; GS, Gleason score; ISUP, International Society of Urological Pathology; TURP, transurethral resection of the prostate; PZ, peripheral zone; TZ, transition zone.

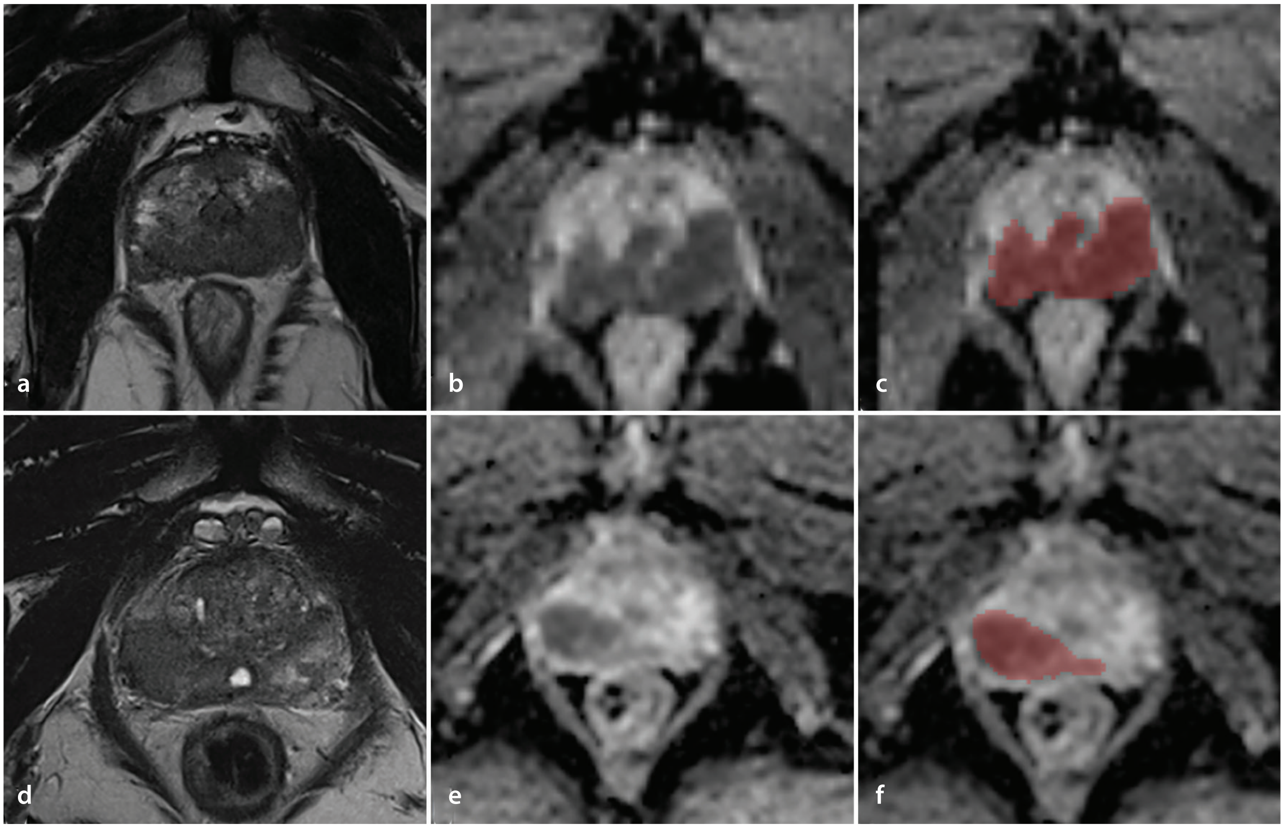


Figure 2. Magnetic resonance imaging (MRI) of prostate cancer (PCa) and prostatitis: **(a-c)** MRI images of patients with PCa; **(d-f)** MRI images of patients with prostatitis; **(a, d)** focal, low-signal lesions in the peripheral zone on T2-weighted imaging; **(b, e)** significant low signal on the apparent diffusion coefficient (ADC); **(c, f)** segmentation images on the ADC; both scored as Prostate Imaging Reporting and Data System 5.

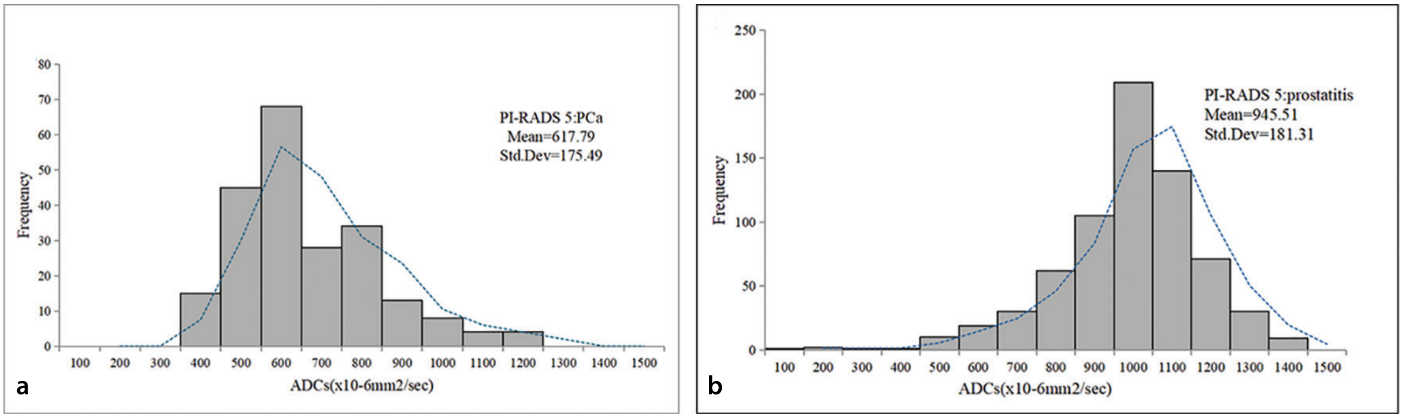


Figure 3. Comparison of whole-tumor histogram analysis of apparent diffusion coefficients (ADCs) between prostate cancer (PCa) and prostatitis: PCa foci **(a)** showing a higher relative frequency at low ADCs than **(b)** prostatitis foci, resulting in significant divergence between prostatitis and PCa at the high end of the cumulative ADC histogram. This suggests that PCa contained more pixels with low ADCs, indicating high cellularity. Std. Dev., Standard deviation.

Table 3. Clinical and apparent diffusion coefficient texture univariate and multivariate analyses

	Univariate analysis		Multivariate analysis	
Clinical	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Age (years)	1.069 (1.013–1.128)	0.015	1.081 (1.017–1.149)	0.012
PSA (ng/mL)	1.053 (1.009–1.098)	0.017		
Prostate volume (mL)	0.986 (0.961–0.991)	0.002	0.990 (0.971–1.009)	0.291
WBC ($\times 10^9$/L)	0.776 (0.619–0.972)	0.028		
PSAD (ng/mL^2)	68.469 (6.074–771.759)	0.001	35.540 (3.534–357.449)	0.002
ADC texture				
Histogram skewness	1.787 (0.617–5.177)	0.457		
Histogram kurtosis	1.050 (0.701–1.570)	0.971		
Histogram entropy	0.213 (0.012–3.743)	0.246		
ADC minimum	0.990 (0.986–0.995)	< 0.001		
ADC 5%	0.983 (0.977–0.989)	< 0.001		
ADC 25%	0.986 (0.981–0.991)	< 0.001		
ADC median	0.987 (0.982–0.992)	< 0.001	0.990 (0.985–0.995)	<0.001
ADC 75%	0.989 (0.984–0.993)	< 0.001		
ADC 95%	0.992 (0.988–0.996)	< 0.001		
ADC maximum	0.998 (0.995–1.000)	0.063		
ADC mean	0.986 (0.981–0.991)	< 0.001		
Signal COV	1.520 E+11 (1,674,931.931–1.379 E + 16)	< 0.001	1,587,241.411 (3.431–7.342 E + 11)	0.032
Signal SD	1.012 (0.997–1.026)	0.084		

PSA, prostate-specific antigen; PSAD, prostate-specific antigen density; WBC, white blood cell; ADC, apparent diffusion coefficient; COV, coefficient of variation; SD, standard deviation; CI, confidence interval; OR, odds ratio.

Table 4. Diagnostic performance of different diagnostic models

	AUC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (95%)	ACC
Clinical model[ab]	84.6 (78.5–89.4)	97.4 (93.6–99.3)	16.0 (4.5–36.1)	95.7 (94.8–96.3)	25.1 (8.2–55.6)	93.4 (88.8–96.5)
ADC texture model[ac]	86.5 (80.7–91.1)	88.1 (82.3–92.4)	44.4 (13.7–78.8)	96.8 (94.3–98.1)	16.4 (7.8–31.0)	85.9 (80.0–90.6)
Combined model[bc]	93.1 (88.4–96.3)	98.7 (95.5–99.8)	60.0 (38.6–78.8)	97.9 (96.6–98.7)	71.6 (38.0–91.2)	96.8 (93.1–98.8)

[a]: Clinical model vs. ADC texture model: *P* = 0.7697.

[b]: Clinical model vs. combined model: *P* = 0.0038.

[c]: ADC texture model vs. combined model: *P* = 0.0812.

AUC, area under the curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value; ADC, apparent diffusion coefficient; ACC, accuracy.

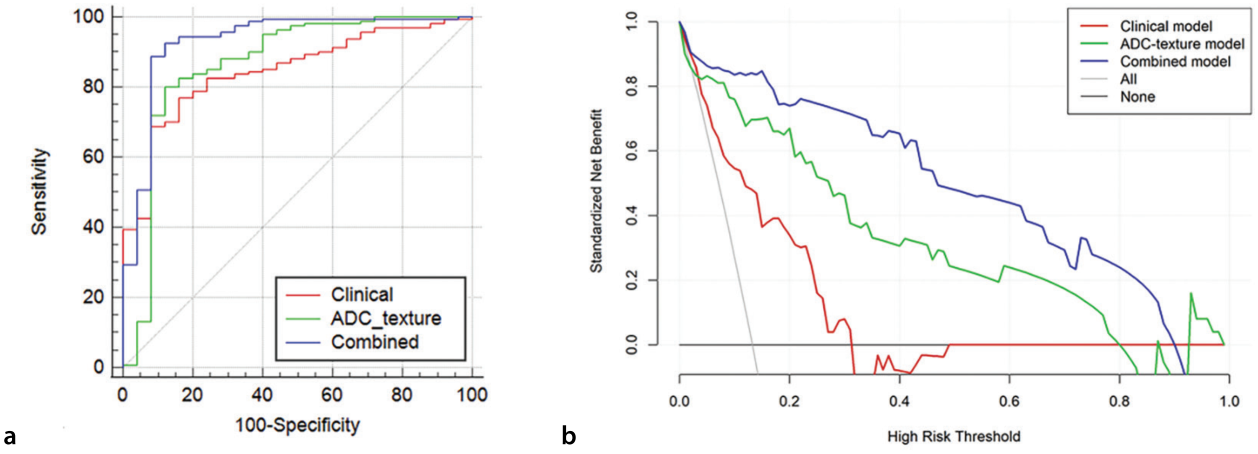


Figure 4. Receiver operating characteristic curve (ROC) and decision curve of different models: **(a)** ROC analysis demonstrates superior performance of the combined model (area under the curve: 0.931, sensitivity: 0.987, specificity: 0.600, positive predictive value: 0.979, negative predictive value: 0.716) to the apparent diffusion coefficient (ADC) texture and clinical models ($P < 0.05$). **(b)** Decision curve analysis reveals the combined model provides greater clinical net benefit across risk thresholds (0.1–0.8) than individual models.

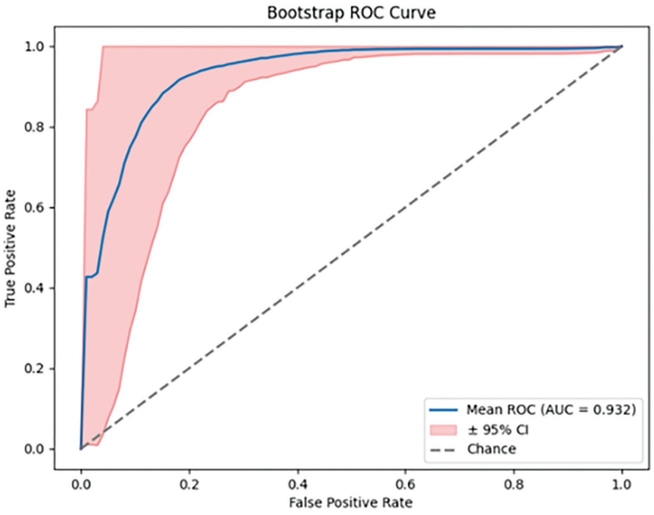


Figure 5. Internal validation. The bootstrap results demonstrated that the performance of the combined model was robust [93.2% (95% CI: 85.6%–98.6%)] in internal validation. ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval.

for accurate diagnosis of MRI-visible lesions assigned a score of PI-RADS 5. The volume of the prostate and the lesions within it are small, the ADC spatial resolution is not high, and the organs and lesions themselves and the scanning sequence parameters limit the full play of high-throughput, multi-dimensional image analysis methods such as radiomics. The combination of texture analysis parameters and clinical indicators for prediction improves the accuracy of the diagnosis of lesions classified as PI-RADS 5. This diagnostic model not only further distinguishes the nature of lesions based on the existing PI-RADS but also serves as a supplement and improvement to PI-RADS.

From a total of 829 consecutive cases screened as PI-RADS 5 in a single center, we selected 189 cases that had complete clinical, imaging, and pathological data. These cases were closer to real-world cases. The sample size of prostatitis cases in this study is limited, which may lead to potential overfitting in the constructed model. To address this issue, the bootstrap method was employed for internal validation, ultimately verifying the diagnostic accuracy of the newly developed model.

Our study is subject to certain limitations. It is a single-center investigation with a limited sample size of prostatitis cases, and external validation has not been conducted. We plan to conduct a multi-center, large-sample study in the future. Inter-reader variability in ROI drawing, although minimized by training, is still a potential factor influencing the model's accuracy. Although the machine scanning parameters are the same, we used three different machine models from two manufacturers. Existing studies have verified that different equipment models can lead to significant inconsistencies in radiomic feature extraction; for instance, Tocilă-Mătășel et al.[30] reported that texture features varied when acquired from different MRI scanners with different parameters. Hajianfar et al.[31] further verified that such equipment-induced variability could reduce the robustness of radiomics models in clinical applications. The impact of different machine models on texture analysis needs to be further studied, especially regarding how to mitigate its interference with the robustness of results.

In conclusion, ADC texture parameters (signal COV, ADC median), PSAD, and age are independent risk factors for distinguishing PCa and prostatitis in PI-RADS 5 lesions. The ADC texture analysis of lesions with a PI-RADS score of 5 combined with clinical parameters can effectively improve the accuracy of diagnosis to reduce unnecessary biopsies and improve patient management.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Funding

The authors declared that they have disclosed direct or indirect affiliation or financial interests in connection with the content of this paper: NSFC 62303235 (WCJ), NSFC 82302308 (YH).

References

1. Dickinson L, Ahmed HU, Allen C, et al. Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: recommendations from a European consensus meeting. Eur Urol. 2011;59(4):477-494. [Crossref]
2. Murphy G, Haider M, Ghai S, Sreeharsha B. The expanding role of MRI in prostate cancer. *AJR Am J Roentgenol.* 2013;201(6):1229-1238. [Crossref]
3. Johnson LM, Turkbey B, Figg WD, Choyke PL. Multiparametric MRI in prostate cancer management. *Nat Rev Clin Oncol.* 2014;11(6):346-353. [Crossref]
4. Sciarra A, Barentsz J, Bjartell A, et al. Advances in magnetic resonance imaging: how they are changing the management of prostate cancer. *Eur Urol.* 2011;59(6):962-977. [Crossref]
5. Barentsz JO, Richenberg J, Clements R, et al. ESUR prostate MR guidelines 2012. *Eur Radiol.* 2012;22(4):746-757. [Crossref]
6. Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic Performance of Prostate Imaging Reporting and Data System Version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *Eur Urol.* 2017;72(2):177-188. [Crossref]
7. Ahmed HU, El-Shater Bosaily A, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet.* 2017;389(10071):815-822. [Crossref]
8. Şeref C, Acar Ö, Kılıç M, et al. Histologically benign PI-RADS 4 and 5 lesions contain cancer-associated epigenetic alterations. Prostate. 2022;82(1):145-153. [Crossref]
9. Sheridan AD, Nath SK, Aneja S, et al. MRI-ultrasound fusion targeted biopsy of prostate imaging reporting and data system version 2 category 5 lesions found false-positive at multiparametric prostate MRI. *AJR Am J Roentgenol.* 2018; 210(5):W218-W225. [Crossref]
10. Apfelbeck M, Pfitzinger P, Bischoff R, et al. Predictive clinical features for negative histopathology of MRI/Ultrasound-fusion-guided prostate biopsy in patients with high likelihood of cancer at prostate MRI: analysis from a urologic outpatient clinic. *Clin Hemorheol Microcirc.* 2020;76(4):503-511. [Crossref]
11. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur Urol.* 2019;76(3):340-351. [Crossref]
12. Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS prostate imaging - reporting and data system: 2015, version 2. *Eur Urol.* 2016;69(1):16-40. [Crossref]
13. Wang X, Liu W, Lei Y, Wu G, Lin F. Assessment of prostate imaging reporting and data system version 2.1 false-positive category 4 and 5 lesions in clinically significant prostate cancer. *Abdom Radiol.* 2021;46(7):3410-3417. [Crossref]
14. Polanec SH, Helbich TH, Bickel H, et al. Quantitative apparent diffusion coefficient derived from diffusion-weighted imaging has the potential to avoid unnecessary MRI-guided biopsies of mpMRI-detected PI-RADS 4 and 5 lesions. *Invest Radiol.* 2018;53(12):736-741. [Crossref]
15. Wang J, Wu CJ, Bao ML, Zhang J, Wang XN, Zhang YD. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. *Eur Radiol.* 2017;27(10):4082-4090. [Crossref]
16. Bonaffini PA, De Bernardi E, Corsi A, et al. Towards the definition of radiomic features and clinical indices to enhance the diagnosis of clinically significant cancers in PI-RADS 4 and 5 lesions. *Cancers (Basel).* 2023;15(20):4963. [Crossref]
17. Ghalati MK, Nunes A, Ferreira H, Serranho P, Bernardes R. Texture analysis and its applications in biomedical imaging: a survey. *IEEE Rev Biomed Eng.* 2021;15:222-246. [Crossref]
18. Corsi A, De Bernardi E, Bonaffini PA, et al. Radiomics in PI-RADS 3 multiparametric MRI for prostate cancer identification: literature models re-implementation and proposal of a clinical-radiological model. *J Clin Med.* 2022;11(21):6304. [Crossref]
19. Doshi AM, Tong A, Davenport MS, et al. Assessment of renal cell carcinoma by texture analysis in clinical practice: a six-site, six-platform analysis of reliability. *AJR Am J Roentgenol.* 2021;217(5):1132-1140. [Crossref]
20. Iersel MPV, Witjes W, ROSETTE JD, Oosterhof G. Prostate-specific antigen density: correlation with histological diagnosis of prostate cancer, benign prostatic hyperplasia and prostatitis. *Br J Urol.*1995;76(1):47-53. [Crossref]
21. Wang S, Kozarek J, Russell R, et al. Diagnostic performance of prostate-specific antigen density for detecting clinically significant prostate cancer in the era of magnetic resonance imaging: a systematic review and meta-analysis. *Eur Urol Oncol.* 2024;7(2):189-203. [Crossref]
22. Wang C, Yuan L, Shen D, et al. Combination of PI-RADS score and PSAD can improve the diagnostic accuracy of prostate cancer and reduce unnecessary prostate biopsies. *Front Oncol.* 2022;12:1024204. [Crossref]
23. Stevens E, Truong M, Bullen JA, Ward RD, Purysko AS, Klein EA. Clinical utility of PSAD combined with PI-RADS category for the detection of clinically significant prostate cancer. *Urol Oncol.* 2020;38(11):846.e9-846.e16. [Crossref]
24. Nagel KN, Schouten MG, Hambrock T, et al. Differentiation of prostatitis and prostate cancer by using diffusion-weighted MR imaging and MR-guided biopsy at 3 T. *Radiology.* 2013;267(1):164-172. [Crossref]
25. Hoeks CM, Vos EK, Bomers JG, Barentsz JO, Hulsbergen-van de Kaa CA, Scheenen TW. Diffusion-weighted magnetic resonance imaging in the prostate transition zone: histopathological validation using magnetic resonance-guided biopsy specimens. *Invest Radiol.* 2013;48(10):693-701. [Crossref]
26. Cheng Y, Fan B, Fu Y, et al. Prediction of false-positive PI-RADS 5 lesions on prostate multiparametric MRI: development and internal validation of a clinical-radiological characteristics based nomogram. *BMC Urol.* 2024;24(1):76. [Crossref]
27. Koh DM, Collins DJ. Diffusion-weighted MRI in the body: applications and challenges in oncology. *AJR Am J Roentgenol.* 2007;188(6):1622-1635. [Crossref]
28. Gass A, Niendorf T, Hirsch JG. Acute and chronic changes of the apparent diffusion coefficient in neurological disorders--biophysical mechanisms and possible underlying histopathology. *J Neurol Sci.* 2001;186:S15-S23. [Crossref]
29. Herneth AM, Guccione S, Bednarski M. Apparent diffusion coefficient: a quantitative parameter for *in vivo* tumor characterization. *Eur J Radiol.* 2003;45(3):208-213. [Crossref]
30. Tocilă-Mătășel C, Dudea SM, Iana G. Addressing multi-center variability in radiomic analysis: a comparative study of image acquisition methods across two 3T MRI scanners. *Diagnostics.* 2025;15(4):485. [Crossref].
31. Hajianfar G, Hosseini SA, Bagherieh S, Oveisi M, Shiri I, Zaidi H. Impact of harmonization on the reproducibility of MRI radiomic features when using different scanners, acquisition parameters, and image pre-processing techniques: a phantom study. *Med Biol Eng Comput.* 2024;62(8):2319-2332. [Crossref]

DIR

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243152

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Systematic review of artificial intelligence competitions in radiology: a focus on design, evaluation, and trends

Muhammed Said Beşler[1]
Ural Koç[2]

[1]Necip Fazıl City Hospital, Clinic of Radiology, Kahramanmaraş, Türkiye

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

ABSTRACT

This article explores the characteristics and scope of artificial intelligence (AI) competitions in medical imaging. A retrospective evaluation of AI competitions related to medical imaging was conducted between 2017 and 2023. Relevant terms associated with AI and competitions were searched using the PubMed database and the grand-challenge website, and applicable studies were included in the review. The 26 AI competitions included in the review covered a wide range of topics, from brain imaging to extremities and from stroke detection to bone age estimation, with many organized through international collaborations between engineering and medical professionals. Various national screening and teleradiology databases, as well as university databases, were used. Teams from different regions worldwide participated in these competitions. These initiatives contribute to the global adoption of AI technologies in healthcare. Moreover, they help raise awareness among high school students, medical students, radiology trainees, and young radiologists of the intersection between AI and medical imaging. AI competitions play a crucial role in fostering collaboration between the medical field and AI, driving innovation, and increasing societal awareness of AI applications in healthcare.

KEYWORDS

Artificial intelligence, radiology, imaging, healthcare, competition

Artificial intelligence (AI) in healthcare is evolving through human–machine collaboration, with innovation driven by partnerships between academic healthcare institutions and industry. The proper validation of AI algorithms, effective data sharing, and training for radiologists is essential.[1] Fundamental requirements and quality standards applicable to all AI-related organizations have begun to be established.[2]

A study examining the impact of AI on radiology and medical imaging through web searches revealed a prevailing positive outlook, highlighting the leading role of radiologists in this discourse.[3] Radiology department chairs tend to be optimistic, believing that AI will be beneficial in areas such as quality, efficiency, healthcare costs, and interpretation workflow.[4] Although radiologists support the idea that AI will streamline workflow, medical students and surgeons approach it more cautiously.[5]

Despite potential biases and pitfalls in the use of AI technologies in medical imaging, their development and advancement are achievable through grand challenges. The expected benefits include creating code and trained datasets, openly sharing them, generating new work areas, and directly involving AI in patient care.[6]

With the widespread use of AI in the medical field, this systematic review aims to investigate the effectiveness of recently organized and popular radiological imaging competitions worldwide.

Methods

Ethical committee approval and patient consent are not required for this type of article. A search was conducted on the PubMed database using the terms “competition” or “contest”

Corresponding author: Ural Koç

E-mail: ukoc85@gmail.com

Received 03 December 2024; revision requested 24 January 2025; accepted 22 February 2025.

Epub: 07.04.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.243152

You may cite this article as: Beşler MS, Koç U. Systematic review of artificial intelligence competitions in radiology: a focus on design, evaluation, and trends. *Diagn Interv Radiol*. 2026;32(2):164-170.

added to the phrase "AI." The focus was on articles containing result reports of imaging-related competitions between 2019 and 2023. Completed competitions were identified using the "completed" filter on the grand-challenge website. Versions of identified competitions held in previous or subsequent years were also considered. A total of 26 competitions that provided sufficient information and had a substantial impact were included in the review (Figure 1).

Information recorded for each competition included the competition's name, year held, imaging modality, target region, search field, dataset source, dataset sample size, dataset accessibility, diversity of contributing institutions, derived academic publications (as of January 2024), citation count according to the Web of Science criteria (as of January 2024), competition location, evaluation criteria, and the number of participating individuals or teams.

Results

This review presents the characteristics of 26 AI and medical imaging-related competitions and datasets between 2017 and 2023 (Tables 1 and 2). These competitions were hosted by organizations such as the Annual Aviation, Space, and Technology Festival (TEKNOFEST), the Radiological Society of North America (RSNA), and the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), either individually or collectively. Final competitions or winner announcements were held either onsite or online.

Various imaging modalities, including magnetic resonance imaging, computed tomography, ultrasound, mammography, and digital breast tomosynthesis, were utilized. Competition themes covered different body regions, ranging from the head to the lower limb, with a focus on segmentation, cancer detection, and disease diagnosis. Most competitions used datasets from universities, but some also incorporated data from national teleradiology systems or screening programs. Although the majority of competition datasets were openly accessible, some required approval for access. One competition was conducted exclusively online, whereas others took place both online and onsite.

Main points

- In recent years, artificial intelligence (AI) competitions have become widespread in the field of medical imaging.
- Datasets are commonly shared openly, and competition results are published in prestigious journals, receiving substantial citations.
- AI competitions help shape perspectives on AI in radiology education and among aspiring radiologists.

In the TEKNOFEST competitions, high school students competed in a separate category, distinguishing them from other competitions. Studies derived from these competition datasets were predominantly published in high-impact journals.

Discussion

The current review aims to evaluate AI applications in medical imaging competitions, which are rapidly increasing in today's medical imaging landscape. High-participation competitions are organized online or onsite in different parts of the world. Collaboration in dataset preparation involves radiologists, clinicians, engineers, and data scientists from different countries and institutions. Studies produced after competitions are published in high-quality journals, and their citation potential is relatively high. Competitions play an effective role in increasing the positive impact and benefits of AI in medical imaging and in generating greater interest in this field.

Organizations such as RSNA, MICCAI, and TEKNOFEST, or online platforms such as the grand-challenge website, host these competitions.[7-35] Dataset organization teams have sometimes come together as multinational teams and are generally multi-institutional. AI competitions in medical imaging lead to the establishment of collaborations not only between interdisciplinary teams but also between institutions and countries, both for competition teams and data preparation teams. The robust infrastructure of national teleradiology systems and the strict preservation of imaging data enable the preparation of competition datasets and the generation of results closest to real-world data.

A study examining 2,517 clinical trials related to AI-associated medical devices revealed that research is generally conducted in specific countries at the national level, with studied populations limited to certain regions. In the last few decades, the development of AI technologies in the medical field has turned into a global competition led by China and the United States.[36] Allowing free participation from around the world in AI competitions in the health sector is increasing the momentum of innovation. The expansion of competitions to low-income countries will diversify the data population and facilitate the availability of developed software for the benefit of these countries.

In 2023, a competition format involving young radiologists and radiology trainees was first organized at the European Society of Medical Imaging Informatics Annual Meeting in Pisa, Italy; this marked a milestone in radiologists' orientation toward AI.[37] Participating in such competitions during the radiology training period can contribute to radiology education in the current era of strong momentum in AI and radiology collaboration.

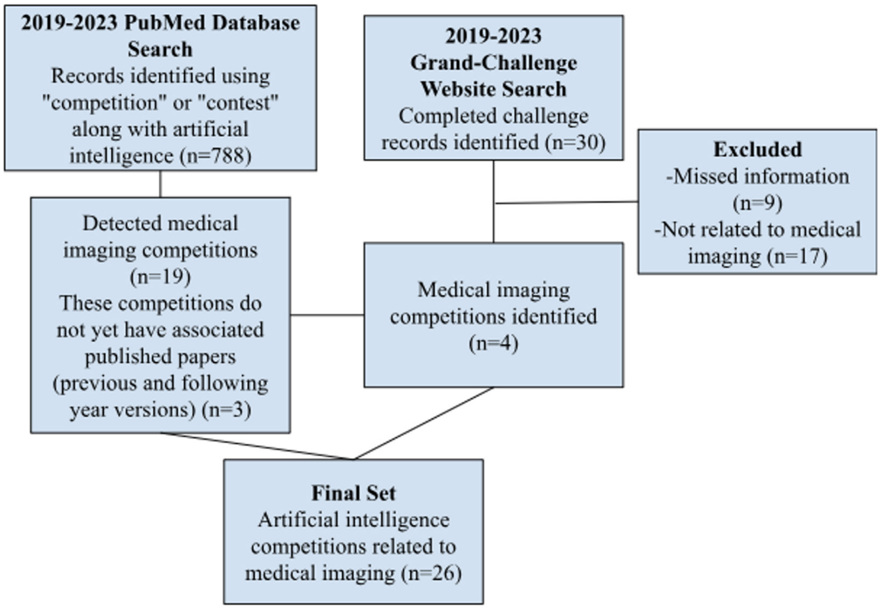

2019-2023 PubMed Database Search
Records identified using "competition" or "contest" along with artificial intelligence (n=788)
2019-2023 Grand-Challenge Website Search
Completed challenge records identified (n=30)
Excluded
-Missed information (n=9)
-Not related to medical imaging (n=17)
Detected medical imaging competitions (n=19)
These competitions do not yet have associated published papers (previous and following year versions) (n=3)
Medical imaging competitions identified (n=4)
Final Set
Artificial intelligence competitions related to medical imaging (n=26)

Figure 1. Flowchart of the selection process for AI competitions in medical imaging. AI, artificial intelligence.

Table 1. Features of the competitions and datasets

Competition	Date	Modality	Target structure	Search field	Dataset source	Sample size	Dataset access
TEKNOFEST 2021 artificial intelligence in health competition (stroke dataset)[7]	2021	CT	Brain	Stroke	National Teleradiology System, Türkiye	877 CT	Open
TEKNOFEST 2022 artificial intelligence in health competition[8]	2022	CT	Abdomen	Abdominal emergencies	National Teleradiology System, Türkiye	1,517 CT	Open
TEKNOFEST 2023 artificial intelligence in health competition[9]	2023	MG	Breast	Breast cancer	National Teleradiology System, Türkiye	N/A	Restricted
RSNA pediatric bone age challenge[10-12]	2017	X-ray	Hand	Bone age	Stanford University and University of Colorado	14,236 hand radiographs	Open
RSNA pneumonia detection challenge[13,14]	2018	X-ray	Lung	Pneumonia	Public NIH	26,684 radiographs	Open
RSNA intracranial hemorrhage detection challenge[15,16]	2019	CT	Head	Intracranial hemorrhage	Stanford University, Thomas Jefferson University, Unity Health Toronto, Universidade Federal de São Paulo, The American Society of Neuroradiology	27,861 unique CT	Open
RSNA pulmonary embolism challenge[17,18]	2020	CT	Lung	Pulmonary embolism	Multi-institutional	12,195 CT	Open
RSNA brain tumor AI challenge[19]	2021	MRI	Brain	Brain tumor segmentation/ radiogenomic classification	Multinational	8,000 MRI	Restricted
RSNA COVID-19 AI detection challenge (SIIM conference on machine intelligence in medical imaging)[20]	2021	X-ray	Lung	COVID-19 pneumonia	Multi-database	10,178 chest radiographs	Open
RSNA cervical spine fracture AI challenge[21]	2022	CT	Neck	Cervical spine fracture	Multinational	3,112 CT	Open
RSNA screening mammography breast cancer detection AI challenge[22]	2023	MG	Breast	Breast cancer	Mammography screening programs in Australia and the U.S.	8,000 MG	Open
RSNA abdominal trauma detection AI challenge[22]	2023	CT	Abdomen	Abdominal traumas	Multinational	>4,000 CT	Open
CHAOS - Combined (CT-MR) healthy abdominal organ segmentation[23]	2019	CT/MRI	Abdomen	Abdominal organ segmentation	Dokuz Eylül University	40 MRI and 40 CT	Open
Tumor detection, segmentation, and classification challenge on automated 3D breast ultrasound[24]	2023	Ultrasound	Breast	Breast cancer	Harbin Medical University Cancer Hospital	200 ultrasound	Upon request
KNee OsteoArthritis Prediction Challenge [25]	2020	X-ray/MRI	Knee	Knee osteoarthritis	Previous study data	423 X-ray/MRI	Open
Surface learning for clinical neuroimaging (MLCN workshop challenge, MICCAI)[24]	2022	MRI	Brain	Cortical development	Previous study data	514 MRI	Upon request
K2S: from undersampled k-space to automatic segmentation (MICCAI)[26]	2022	MRI	Knee	Knee joint degeneration	University of California	816 MRI	Upon request
1st Boston neonatal brain injury dataset for hypoxic ischemic encephalopathy lesion segmentation challenge (MICCAI)[27]	2023	MRI	Brain	Hypoxic ischemic encephalopathy	Massachusetts General Hospital	133 MRI	Open
DBTex Challenge[28]	2021	Digital breast tomosynthesis	Breast	Breast cancer	Duke University	22,032 digital breast tomosynthesis	Open

Table 1. Continued

Competition	Date	Modality	Target structure	Search field	Dataset source	Sample size	Dataset access
COVID-19 lung CT lesion segmentation challenge[29]	2020	CT	Lung	COVID-19 pneumonia	Previous study data from the Cancer Imaging Archive	295 CT	Partial
Kidney tumor segmentation challenge (MICCAI)[30]	2019	CT	Kidney	Kidney tumor	University of Minnesota Medical Center	300 CT	Partial
Kidney tumor segmentation challenge (MICCAI)[24]	2021	CT	Kidney	Kidney tumor/ cyst	M Health Fairview or Cleveland Clinic Medical Center	300 CT	Open
Kidney tumor segmentation challenge (MICCAI)[24]	2023	CT	Kidney	Kidney tumor/ cyst	M Health Fairview Medical Center	599 CT	Open
French Society of Radiology data challenge[31-33]	2018	MRI/CT/ Ultrasound	Knee/ Kidney/ Liver/ Breast	Meniscal tear, renal cortex segmentation, lesions of the liver, breast, and thyroid cartilage	Multi-institutional	5,170 images	N/A
French Society of Radiology data challenge[34]	2019	MRI/CT	Lung/ Brain/ Muscles	Pulmonary nodule, multiple sclerosis, sarcopenia	Multi-institutional	4,347 examinations	N/A
French Society of Radiology data challenge[35]	2020	CT/Ultrasound	Breast/ Neck/ Heart	Breast nodule, neck lymph node, coronary calcium score	Multi-institutional	2,076 examinations	N/A

TEKNOFEST, Annual Aviation, Space and Technology Festival; CT, computed tomography; MG, mammography; RSNA, Radiological Society of North America; NIH, National Institutes of Health; AI, artificial intelligence; MRI, magnetic resonance imaging; COVID-19, coronavirus disease 2019; MLCN, Machine Learning in Clinical Neuroimaging; MICCAI, International Conference on Medical Image Computing and Computer-Assisted Intervention.

In a survey conducted among medical students in Canada, it was observed that although radiology specialization was among the top choices, there were widespread concerns about the negative effects of AI on radiologists. Information meetings are suggested to address these concerns.[38] The negative effects of AI on radiology career development have also been noted among US medical students.[39] Public competitions involving medical students will contribute to a more realistic understanding of the relationship between AI and radiology expertise. Encouraging high school students to participate in some competitions strategically promotes AI development and raises social awareness among young individuals who have not yet made career choices. Technology teachers at the high school and even middle school levels can take the lead in encouraging participation in such activities during their training.

Additionally, such competitions can lead to the generation of new study topics on emerging issues and the establishment of new networks, facilitating the creation of start-ups. AI summer schools in medicine for high school students have begun to be established at universities.[40] Ethical dilemmas such as bias risk and data security, along with AI's potential to assist medical professionals, cannot be overlooked in the realm of AI in healthcare.[41] AI training programs should comprehensively address all these aspects.

The impact of AI-related medical imaging competitions on scientific publication conversion, citation potential, and integration into the literature was investigated. However, another crucial aspect—their clinical application and commercial utilization—lacks sufficient and effective information based on available datasets and publications. To bridge the gap between scientific innovation and clinical practice, it may be beneficial to increase awareness of these competitions among healthcare institutions, AI-related public organizations, and commercial entities while also expanding networking opportunities for competition participants.

Efforts have been made to standardize and enhance transparency in the evaluation of medical imaging competitions, from defining the competition's mission to dataset preparation methodologies and participant ranking metrics and criteria. However, substantial variations have been observed across these stages.[42] Proper competition design and interpretation can facilitate the validation of AI algorithms and promote their translation into clinical applications.[43] Several factors influence the outcomes of AI-related medical imaging competitions, including the dataset used, the reference annotations determined by annotators, and the scoring system applied for ranking.[44] Quality control at all stages of a competition enhances the validity and reliability of its results. In our review, a comprehensive framework has been established, detailing the design, execution, and outcomes of current AI-related medical imaging competitions.

This review has some limitations. Not all the databases where competitions could be included were examined for all dates. However, by focusing on recent competitions in the most well-known databases and platforms, an attempt was made to minimize selection bias. There are only a few studies in the literature examining competitions related to AI and radiology.[6,45] However, our review is the first to address the dataset, organization teams, and competition features.

In conclusion, as AI continues to play an increasing role in radiology, competitions related to AI and medical imaging contribute to quality dataset sharing, collaboration among experts, and increased awareness in this field.

Table 2. Characteristics of the competitions and publications

Competition	Studies derived from dataset	Citations	Dataset experts	Country	Number of individual participants or teams in the first application	Evaluation criteria
TEKNOFEST 2021 artificial intelligence in health competition	1	1	Multi-institutional radiologists and engineers	Türkiye	570 participants	F1 score, IoU
TEKNOFEST 2022 artificial intelligence in health competition	1	None	Multi-institutional radiologists and engineers	Türkiye	213 teams	Mean F1 scores computed across distinct threshold values for IoU
TEKNOFEST 2023 artificial intelligence in health competition	None	None	Multi-institutional radiologists and engineers	Türkiye	409 teams	F1 score
RSNA pediatric bone age challenge	3	271	Multi-institutional radiologists	U.S.	260 participants	Mean absolute distance in months
RSNA pneumonia detection challenge	2	39	Society for Thoracic radiology members and software	U.S.	1,400 teams	Mean average precision at different IoU thresholds
RSNA intracranial hemorrhage detection challenge	2	110	Multinational via commercial software	U.S.	1,345 teams	Weighted multi-label logarithmic loss
RSNA pulmonary embolism detection challenge	2	32	Society of Thoracic Radiology members	U.S.	784 teams	Weighted log loss
RSNA brain tumor AI challenge	1	1	Multinational	U.S.	1,555 teams	Dice similarity coefficient, Hausdorff distance (95%), AUC, accuracy, FScore (Beta), and Matthew's correlation coefficient
RSNA COVID-19 AI detection challenge	1	6	Multinational	U.S.	1,305 teams	Standard PASCAL VOC 2010 mean average precision at IoU >0.5
RSNA cervical spine fracture AI challenge	1	None	Spine radiology specialists from the American Society of Neuroradiology and the American Society of Spine Radiology	U.S.	883 teams	Weighted multi-label logarithmic loss
RSNA screening mammography breast cancer detection AI challenge	None	None	Via commercial software tools	U.S.	1,687 teams	Probabilistic F1 score
RSNA abdominal trauma detection AI challenge	None	None	Society of Abdominal Radiology and the American Society of Emergency Radiology members	U.S.	1,123 teams	Average of the sample weighted log losses from each injury type and an any-injury prediction generated by the metric
CHAOS - Combined (CT-MR) healthy abdominal organ segmentation	1	195	Engineers, radiologists, and PhD/MSc/BSc students from Türkiye	Italy	1,500 participants	Intra- and inter-annotator scores
Tumor detection, segmentation, and classification challenge on automated 3D breast ultrasound	None	None	Engineer/radiologist from China	Canada (MICCAI 2023)	503 participants	Dice similarity coefficient, Hausdorff distance, accuracy, AUC, free-response ROC
KNee OsteoArthritis Prediction Challenge	1	6	N/A	Netherlands/ Online	20 participants	ROC AUC and balanced accuracy
Surface learning for clinical neuroimaging (MLCN workshop challenge, MICCAI)	None	None	Engineers and radiologists from the U.K.	Singapore	91 participants	Mean absolute error

Table 2. Continued

Competition	Studies derived from dataset	Citations	Dataset experts	Country	Number of individual participants or teams in the first application	Evaluation criteria
K2S: from undersampled k-space to automatic segmentation (MICCAI)	1	2	Multinational engineers and radiologists	Singapore	87 teams	Dice similarity coefficient
1st Boston neonatal brain injury dataset for hypoxic ischemic encephalopathy lesion segmentation challenge (MICCAI 2023)	1	None	Single-center PhD and MD	Canada	131 participants	Dice, mean average surface distance, normalized surface distance
DBTex challenge	1	1	Multinational engineers and radiologists	U.S.	8 teams	Free-response ROC
COVID-19 lung CT lesion segmentation challenge	1	6	Automated segmentation and confirmation by single-center radiologists	Online	1,096 teams	Dice coefficient, normalized surface Dice, normalized absolute volume error
Kidney tumor segmentation challenge (MICCAI 2019)	3	173	Single radiologist and supervised students	China	106 teams	Sørensen–Dice coefficient
Kidney tumor segmentation challenge (MICCAI 2021)	None	None	Multi-institutional radiologists, urologists, and supervised students	France	N/A	Sørensen–Dice, surface Dice
Kidney tumor segmentation challenge (MICCAI 2023)	None	None	Multi-institutional radiologists, urologists, urologic oncologists, and supervised students	Canada	N/A	Sørensen–Dice, surface Dice
French Society of Radiology Data Challenge 2018	1	31	Multi-institutional radiologists and data scientists	France	323 participants	Dice score, binary AUC
French Society of Radiology Data Challenge 2019	1	18	Multi-institutional radiologists and data scientists	France	143 participants	Dice coefficient, AUC, mean square error
French Society of Radiology Data Challenge 2020	1	10	Multi-institutional radiologists and data scientists	France	39 participants	Concordance index, Dice score, AUROC

IoU, Intersection over Union; TEKNOFEST, Annual Aviation, Space and Technology Festival; RSNA, Radiological Society of North America; AI, artificial intelligence; AUC, area under the curve; COVID-19, coronavirus disease 2019; VOC, visual object classes; CT, computed tomography; MRI, magnetic resonance imaging; PhD, Doctor of Philosophy; MSc, master of science; BSc, bachelor of science; MICCAI, International Conference on Medical Image Computing and Computer-Assisted Intervention; ROC, receiver operating characteristic; MLCN, Machine Learning in Clinical Neuroimaging; MD, doctor of medicine; AUROC, area under the receiver operating characteristics.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol.* 2020;30(6):3576-3584. [Crossref]
2. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en Accessed: 05 January 2024. [Crossref]
3. Mulryan P, Ni Chleirigh N, O'Mahony AT, et al. An evaluation of information online on artificial intelligence in medical imaging. *Insights Imaging*. 2022;13(1):79. [Crossref]
4. Burnside ES, Grist TM, Lasarev MR, Garrett JW, Morris EA. Artificial intelligence in radiology: a leadership survey. *J Am Coll Radiol.* 2025:S1546-1440(25)00041-00049. [Crossref]
5. van Hoek J, Huber A, Leichtle A, et al. A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over. *Eur J Radiol.* 2019;121:108742. [Crossref]
6. Armato SG 3rd, Drukker K, Hadjiiski L. AI in medical imaging grand challenges: translation from competition to research benefit and patient care. *Br J Radiol.* 2023;96(1150):20221152. [Crossref]
7. Koç U, Akçapınar Sezer E, Özkaya YA, et al. Artificial intelligence in healthcare competition (TEKNOFEST-2021): stroke data set. *Eurasian J Med.* 2022;54(3):248-258. [Crossref]
8. Koç U, Sezer EA, Özkaya YA, et al. Elevating healthcare through artificial intelligence: analyzing the abdominal emergencies data set (TR_ABDOMEN_RAD_EMERGENCY) at TEKNOFEST-2022. *Eur Radiol*. 2024;34(6):3588-3597. [Crossref]

9. Artificial intelligence in health competition. Last Accessed: 10.03.2025. [Crossref]

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

11. Siegel EL. What can we learn from the RSNA pediatric bone age machine learning challenge? *Radiology*. 2019;290(2):504-505. [Crossref]

12. 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]

13. Pan I, Cadrin-Chênevert A, Cheng PM. Tackling the radiological Society of North America pneumonia detection challenge. *AJR Am J Roentgenol*. 2019;213(3):568-574. [Crossref]

14. Chang IY, Huang TY. Deep learning-based classification for lung opacities in chest X-ray radiographs through batch control and sensitivity regulation. *Sci Rep*. 2022;12(1):17597. [Crossref]

15. Flanders AE, Prevedello LM, Shih G, et al. Erratum: construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiol Artif Intell*. 2020;2(4):e209002. Erratum for: *Radiol Artif Intell*. 2020;2(3):e190211. [Crossref]

16. Danilov G, Kotik K, Negreeva A, et al. Classification of intracranial hemorrhage subtypes using deep learning on CT scans. *Stud Health Technol Inform*. 2020;272:370-373. [Crossref]

17. Colak E, Kitamura FC, Hobbs SB, et al. The RSNA pulmonary embolism CT dataset. *Radiol Artif Intell*. 2021;3(2):e200254. [Crossref]

18. Callejas MF, Lin HM, Howard T, et al. Augmentation of the RSNA pulmonary embolism CT dataset with bounding box annotations and anatomic localization of pulmonary emboli. *Radiol Artif Intell*. 2023;5(3):e230001. [Crossref]

19. Kim BH, Lee H, Choi KS, et al. Validation of MRI-based models to predict MGMT promoter methylation in gliomas: BraTS 2021 radiogenomics challenge. *Cancers (Basel)*. 2022;14(19):4827. [Crossref]

20. Lakhani P, Mongan J, Singhal C, et al. The 2021 SIIM-FISABIO-RSNA machine learning COVID-19 challenge: annotation and standard exam classification of COVID-19 chest radiographs. *J Digit Imaging*. 2023;36(1):365-372. [Crossref]

21. Lin HM, Colak E, Richards T, et al. The RSNA cervical spine fracture CT dataset. *Radiol Artif Intell*. 2023;5(5):e230034. [Crossref]

22. https://www.kaggle.com/search?q=rsna Accessed: 08 January 2024. [Crossref]

23. Kavur AE, Gezer NS, Barış M, et al. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Med Image Anal*. 2021;69:101950. [Crossref]

24. Grand challenge. Accessed: 08 Jan 2024. [Crossref]

25. Hirvasniemi J, Runhaar J, van der Heijden RA, et al. The KNee OsteoArthritis Prediction (KNOAP2020) challenge: an image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images. *Osteoarthritis Cartilage*. 2023;31(1):115-125. [Crossref]

26. Tolpadi AA, Bharadwaj U, Gao KT, et al. K2S challenge: from undersampled k-space to automatic segmentation. *Bioengineering (Basel)*. 2023;10(2):267. [Crossref]

27. Bao R, Song Y, Bates SV, et al. Boston neonatal brain injury dataset for hypoxic ischemic encephalopathy (BONBID-HIE): part I. MRI and manual lesion annotation. bioRxiv [Preprint]. 2023;2023.06.30.546841. Update in: *Sci Data*. 2025;12(1):53. [Crossref]

28. Konz N, Buda M, Gu H, et al. A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. *JAMA Netw Open*. 2023;6(2):e230524. [Crossref]

29. Roth HR, Xu Z, Tor-Díez C, et al. Rapid artificial intelligence solutions in a pandemic-The COVID-19-20 lung CT lesion segmentation challenge. *Med Image Anal*. 2022;82:102605. [Crossref]

30. Heller N, Isensee F, Maier-Hein KH, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Med Image Anal*. 2021;67:101821. [Crossref]

31. Sathianathen NJ, Heller N, Tejpaul R, et al. Automatic segmentation of kidneys and kidney tumors: the KiTS19 international challenge. *Front Digit Health*. 2022;3:797607. [Crossref]

32. Causey J, Stubblefield J, Qualls J, et al. An ensemble of U-Net models for kidney tumor segmentation with CT images. *IEEE/ACM Trans Comput Biol Bioinform*. 2022;19(3):1387-1392. [Crossref]

33. Lassau N, Estienne T, de Vomecourt P, et al. Five simultaneous artificial intelligence data challenges on ultrasound, CT, and MRI. *Diagn Interv Imaging*. 2019;100(4):199-209. [Crossref]

34. Lassau N, Bousaid I, Chouzenoux E, et al. Three artificial intelligence data challenges based on CT and MRI. *Diagn Interv Imaging*. 2020;101(12):783-788. [Crossref]

35. Lassau N, Bousaid I, Chouzenoux E, et al. Three artificial intelligence data challenges based on CT and ultrasound. *Diagn Interv Imaging*. 2021;102(11):669-674. [Crossref]

36. Serra-Burriel M, Miquel, Locher L, Kerstin N. Vokinger KN. Development pipeline and geographic representation of trials for artificial intelligence/machine learning–enabled medical devices (2010 to 2023). *NEJM AI*. 2023;1(1). [Crossref]

37. Akinci D'Antonoli T, Huisman M. EuSoMII 2023 Highlights and the EU AI Act. Accessed: 10 Jan 2024. [Crossref]

38. Gong B, Nugent JP, Guest W, Parker W, Chang PJ, Khosa F, Nicolaou S. Influence of artificial intelligence on Canadian Medical Students' preference for radiology specialty: anational survey study. *Acad Radiol*. 2019;26(4):566-577. [Crossref]

39. Reeder K, Lee H. Impact of artificial intelligence on US medical students' choice of radiology. *Clin Imaging*. 2022;81:67-71. [Crossref]

40. Center for Artifical Intelligence in Medicine & Imaging. Summer Research Internship. Accessed: 10 Jan 2024. [Crossref]

41. Korkmaz S. Artificial intelligence in healthcare: a revolutionary ally or an ethical dilemma? *Balkan Med J*. 2024;41(2):87-88. [Crossref]

42. Maier-Hein L, Reinke A, Kozubek M, et al. BIAS: transparent reporting of biomedical image analysis challenges. *Med Image Anal*. 2020;66:101796. [Crossref]

43. Reinke A, Tizabi MD, Eisenmann M, Maier-Hein L. Common pitfalls and recommendations for grand challenges in medical artificial intelligence. *Eur Urol Focus*. 2021;7(4):710-712. [Crossref]

44. Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun*. 2018;9(1):5217. Erratum in: *Nat Commun*. 2019;10(1):588. [Crossref]

45. Wagner DT, Tilmans L, Peng K, et al. Artificial intelligence in neuroradiology: a review of current topics and competition challenges. *Diagnostics (Basel)*. 2023;13(16):2670. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253314

ARTIFICIAL INTELLIGENCE AND INFORMATICS

INVITED COMMENTARY

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Reporting checklists: from a tool after publication to a tool before submission

Jingyu Zhong[1,2]

[1]Department of Imaging, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China

Corresponding author: Jingyu Zhong

E-mail: wal_zjy@163.com

Received 14 March 2025; revision requested 23 March 2025; accepted 24 March 2025.

Epub: 05.05.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253314

Artificial intelligence (AI) methods have attracted widespread interest in the field of medical imaging, and the increasing number of AI publications in radiological journals reflects this growing attention from researchers and journals alike. As the old saying goes, "interest is the best teacher," yet interest in AI does not automatically translate into proper use and adequate reporting of AI methods. Promising results in published articles do not necessarily ensure high methodological quality. More often than not, incomplete reporting of methodology and the lack of data and code sharing hinder study evaluation and model replication. To address this issue, the Checklist for AI in Medical Imaging (CLAIM) was developed as a guide for the complete reporting of AI studies in medical imaging.[1] CLAIM has been widely adopted, with more than 800 articles citing the guideline as of March 14, 2025, and has also been used as a tool for quality assessment in systematic reviews of AI studies.[2] However, these systematic reviews, which typically focus on medical imaging studies using AI, highlight the limited quality of current studies. This raises the question of whether CLAIM has been used appropriately and how the reporting and methodological quality of AI studies in medical imaging can be improved.

In this issue of Diagnostic and Interventional Radiology, the study by Koçak et al.[3] not only reveals a substantial gap between the current state of reporting and the ideal reporting of AI studies in medical imaging but also identifies factors influencing adherence to CLAIM. The study finds that CLAIM adherence is associated with the journal impact factor quartile, publication year, and specific radiology subfields. Not surprisingly, CLAIM adherence improved after CLAIM's publication, likely because authors became more familiar with standard practices. Higher adherence to CLAIM was observed in cardiovascular studies, suggesting a more mature use of AI methods in this subfield, from automated reconstruction tools for coronary computed tomography angiography to analysis software for cardiac magnetic resonance. Despite this progress, improving CLAIM adherence remains more important than identifying the sources of high adherence. High-impact journals might promote more transparent reporting practices through more rigorous peer review processes and encourage authors to follow AI guidelines and include them in submission requirements.[4] As the mandatory use of reporting guidelines has been shown to improve study quality,[5] the current study provides a clear and actionable recommendation to enhance the quality of AI studies: increase journal support for CLAIM use.[6]

In this study, Koçak et al.[3] follow a two-level analysis to address concerns regarding common CLAIM critiques. The study summarizes comments from systematic reviews and identifies two main critiques: concerns about the inapplicability of certain items to all study types and the subjective nature of reporting decisions. The concern regarding inapplicability has been addressed with the update of CLAIM 2024, which includes a "not applicable" option for item evaluation,[7] but the issue of subjectivity remains. These factors may all contribute to the unreliability of CLAIM evaluation, such as unclear item descriptions, subjective comprehension, and the complexity of AI methods.[8] When researchers use CLAIM for future systematic reviews, a greater focus on reproducibility may be necessary. CLAIM still needs updates, including more explanations and elaborations with examples, so that users can apply the tool with a better understanding.[9] Additionally, developing user-friendly online tools would enhance convenience.[10] The introduction of automatic tools, such as large language models, may also aid in optimizing the reproducibility of CLAIM evaluation. Furthermore, translated versions of the tools endorsed by the original authors may increase visibility and adaptability to local cultures.

You may cite this article as: Zhong J. Reporting checklists: from a tool after publication to a tool before submission. *Diagn Interv Radiol*. 2026;32(2):171-172.

Beyond the CLAIM tool itself, the quality of individual studies remains crucial. A previous study by Kocak et al.[9] investigated the use of CLAIM in individual studies. The study found that only a small percentage of publications used CLAIM along with a supplementary filled-out checklist, and many of the completed checklists contained errors. CLAIM is a useful tool for post-publication evaluation,[2] but it is not currently required before submission. It remains unclear whether the endorsement of CLAIM and other AI-specific guidelines can improve reporting and methodological quality. Furthermore, it is uncertain whether and how these AI-specific guidelines are used during the editorial process, as only a limited number of journals practice open peer review or publish articles with filled-out checklists. Instead of solely critiquing the adherence of published studies to CLAIM, it would be more valuable to investigate the influence of CLAIM on scientific publication practices. The primary intention of developing a checklist is not to evaluate existing studies retrospectively with strict standards but to guide ongoing research. The checklist can also serve as guidance for peer review before publication and as a tool for study design prior to submission.

In conclusion, the work by Koçak et al.[3] draws the community's attention to the lack of quality in reporting and methodology in AI studies in medical imaging. Although checklists may not resolve this problem overnight, they pave the way for a future of transparent reporting and high-quality methodology. Therefore, the use of reporting checklists is recommended before submission, during evaluation, and after publication.

Footnotes

Conflict of interest disclosure

Dr. Jingyu Zhong acknowledges his position as a member of the Scientific Editorial Board of European Radiology, Insights into Imaging, American Journal of Roentgenology, and BMC Medical Imaging.

Funding

This study has received funding by National Natural Science Foundation of China (82302183), Research Found of Health Commission of Shanghai Municipality (20244Y0214), Yangfan Project of Science and Technology Commission of Shanghai Municipality (22YF1442400), Laboratory Open Fund of Key Technology and Materials in Minimally Invasive Spine Surgery (2024JZWC-YBA07), and Research Fund of Tongren Hospital, Shanghai Jiao Tong University School of Medicine (TRKYRC-XX202204, TR2024RC16). They played no role in the study design, data collection or analysis, decision to publish, or manuscript preparation.

References

1. 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]
2. 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]
3. 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*. Epub 2025 Feb 10. [Crossref]
4. 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]
5. Dewey M, Levine D, Bossuyt PM, Kressel HY. Impact and perceived value of journal reporting guidelines among radiology authors and reviewers. *Eur Radiol*. 2019;29(8):3986-3995. [Crossref]
6. 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]
7. 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]
8. 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]
9. Kocak B, Borgheresi A, Ponsiglione A, et al. Explanation and elaboration with examples for CLEAR (CLEAR-E3): an EuSoMII radiomics auditing group initiative. *Eur Radiol Exp*. 2024;8(1):72. [Crossref]
10. 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]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253101

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties

Yasin Celal Güneş[1]
Turay Cesur[2]
Eren Çamur[3]

[1]Kırıkkale Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

[2]Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

[3]Ankara 29 May State Hospital, Clinic of Radiology, Ankara, Türkiye

Corresponding author: Yasin Celal Güneş

E-mail: gunesyasincelal@gmail.com

Received 28 October 2024; revision requested 10 December 2024; last revision received 21 February 2025; accepted 05 April 2025.

Epub: 12.05.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253101

PURPOSE

This study aimed to compare six large language models (LLMs) [Chat Generative Pre-trained Transformer (ChatGPT)o1-preview, ChatGPT-4o, ChatGPT-4o with canvas, Google Gemini 1.5 Pro, Claude 3.5 Sonnet, and Claude 3 Opus] in generating radiology references, assessing accuracy, fabrication, and bibliographic completeness.

METHODS

In this cross-sectional observational study, 120 open-ended questions were administered across eight radiology subspecialties (neuroradiology, abdominal, musculoskeletal, thoracic, pediatric, cardiac, head and neck, and interventional radiology), with 15 questions per subspecialty. Each question prompted the LLMs to provide responses containing four references with in-text citations and complete bibliographic details (authors, title, journal, publication year/month, volume, issue, page numbers, and PubMed Identifier). References were verified using Medline, Google Scholar, the Directory of Open Access Journals, and web searches. Each bibliographic element was scored for correctness, and a composite final score [(FS): 0-36] was calculated by summing the correct elements and multiplying this by a 5-point verification score for content relevance. The FS values were then categorized into a 5-point Likert scale reference accuracy score (RAS: 0 = fabricated; 4 = fully accurate). Non-parametric tests (Kruskal–Wallis, Tamhane's T2, Wilcoxon signed-rank test with Bonferroni correction) were used for statistical comparisons.

RESULTS

Claude 3.5 Sonnet demonstrated the highest reference accuracy, with 80.8% fully accurate references (RAS 4) and a fabrication rate of 3.1%, significantly outperforming all other models ($P < 0.001$). Claude 3 Opus ranked second, achieving 59.6% fully accurate references and a fabrication rate of 18.3% ($P < 0.001$). ChatGPT-based models (ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview) exhibited moderate accuracy, with fabrication rates ranging from 27.7% to 52.9% and <8% fully accurate references. Google Gemini 1.5 Pro had the lowest performance, achieving only 2.7% fully accurate references and the highest fabrication rate of 60.6% ($P < 0.001$). Reference accuracy also varied by subspecialty, with neuroradiology and cardiac radiology outperforming pediatric and head and neck radiology.

CONCLUSION

Claude 3.5 Sonnet significantly outperformed all other models in generating verifiable radiology references, and Claude 3 Opus showed moderate performance. In contrast, ChatGPT models and Google Gemini 1.5 Pro delivered substantially lower accuracy with higher rates of fabricated references, highlighting current limitations in automated academic citation generation.

CLINICAL SIGNIFICANCE

The high accuracy of Claude 3.5 Sonnet can improve radiology literature reviews, research, and education with dependable references. The poor performance of other models, with high fabrication rates, risks misinformation in clinical and academic settings and highlights the need for refinement to ensure safe and effective use.

KEYWORDS

Reference, citation, ChatGPT o1-preview, Claude 3.5 Sonnet, large language models

You may cite this article as: Güneş YC, Cesur T, Çamur E. Evaluating the reference accuracy of large language models in radiology: a comparative study across subspecialties. *Diagn Interv Radiol*. 2026;32(2):173-181.

The rapid advancement of large language models (LLMs) represents a key milestone in artificial intelligence (AI), offering unprecedented capabilities in text generation and comprehension.[1] These models, trained on extensive datasets, have shown promise in medical applications such as literature summarization, manuscript editing, and reference generation.[2,3] However, their reliability in reference generation remains a critical concern, particularly in radiology, where evidence-based practice depends on accurate and verifiable sources.[4,5] A key challenge is their tendency to generate "hallucinations" (fabricated or inaccurate references), which undermine their utility in clinical and academic settings.[5]

The issue of hallucinated references in LLMs is well documented in the literature.[6-16] Chelli et al.[7] reported hallucination rates of 39.6% for Chat Generative Pre-trained Transformer (ChatGPT)-3.5, 28.6% for ChatGPT-4, and an alarming 91.4% for Bard when generating references for systematic reviews. Walters and Wilder[8] found that although ChatGPT-4 exhibited a lower hallucination rate (18%) than ChatGPT-3.5 (55%), both models produced considerable inaccuracies, even among seemingly valid references. In radiology, Wagner et al.[9] observed that 63.8% of references generated by ChatGPT-3 were fabricated, with only 37.9% offering adequate support. These findings are particularly concerning in radiology, where inaccurate references could contribute to misinformation, potentially affecting clinical research, educational materials, and evidence-based decision-making.[9]

Retrieval-augmented LLMs combine traditional language models with external data retrieval mechanisms, grounding responses in current, domain-specific information.[17] Emerging solutions, such as retrieval-augmented LLMs and platforms like OpenEvidence, aim to address these limitations by integrating real-time access to credible sources.[18] OpenEvidence, for instance, delivers up-to-date, evidence-based answers with clearly labeled references, reducing the risk of misinformation.[18] However, its accessibility remains restricted, requiring a National Provider Identifier number, which is issued to U.S. healthcare providers, for unlimited access and is available only in certain regions. In contrast, advanced LLMs such as ChatGPT-4o with canvas, ChatGPT o1-preview, and Claude 3.5 Sonnet offer worldwide accessibility, making them versatile and inclusive tools for users across diverse geographies.[19] These models have the potential to overcome prior limitations by leveraging enhanced natural language processing capabilities and expanded datasets, ensuring broader applicability and impact.[20]

Despite the rapid advancements in LLMs, no systematic evaluation has been conducted to assess the accuracy of references generated by state-of-the-art LLMs across radiology subspecialties. To address this gap, this study aims to provide the first systematic evaluation of the reference-generation accuracy of advanced LLMs, with a focus on identifying the most reliable model and characterizing variability across eight radiology subspecialties. By highlighting their strengths and limitations, this research seeks to clarify the potential roles of LLMs in radiology and provide actionable guidance for improving AI-driven reference generation.

Main points

- Claude 3.5 Sonnet demonstrated the highest reference accuracy, significantly outperforming other large language models (LLMs) across all radiology subspecialties, making it the most reliable tool for generating medical references.
- Chat Generative Pre-trained Transformer (ChatGPT)-4o, ChatGPT-4o with canvas, and Google Gemini 1.5 Pro exhibited lower reference accuracy, with considerable inconsistencies in generating accurate references, highlighting the need for further improvements in these models for use in clinical settings.
- Accurate reference generation by Claude 3.5 Sonnet supports its potential to enhance literature reviews, research preparation, and educational content creation in radiology, improving the efficiency and quality of work in both clinical and academic domains.
- The study emphasizes the necessity of validating LLM-generated references, as errors and inconsistencies in models such as ChatGPT and Google Gemini could lead to serious risks in clinical decision-making and academic integrity.

Methods

Study design

This cross-sectional observational study evaluated the performance of six LLMs—ChatGPT o1-preview, ChatGPT-4o, ChatGPT-4o with canvas, Google Gemini 1.5 Pro, Claude 3.5 Sonnet, and Claude 3 Opus—in generating medical references for radiology questions across eight subspecialties. The study exclusively used publicly available, internet-based data without any identifiable patient information, eliminating the need for ethics committee approval. It was conducted in accordance with the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of LLMs in Healthcare guidelines.[21] An overview of the workflow is presented in Figure 1.

Question preparation

Eight radiology subspecialties—neuroradiology, abdominal imaging, musculoskeletal radiology, thoracic imaging, pediatric radiology, cardiac imaging, head and neck radiology, and interventional radiology—were selected to represent a broad range of clinical domains. For each subspecialty, 15 questions were developed, yielding a total of 120 questions. This sample size not only balances comprehensive coverage with the feasibility of manual reference verification but also exceeds the minimum requirement of approximately 96 questions—calculated using a standard sample size formula for estimating a 50% proportion with a 10% margin of error at the 95% confidence level—thus ensuring robust statistical power and enhancing the precision of our findings.

All questions were independently created by Radiologist 1 (Y.C.G.) without the use of any LLMs, thereby preventing any influence from the models' internal training data and minimizing potential bias from "leaked" context. All questions are provided in Supplementary Material 1.

Design of input–output procedures and performance evaluation for large language models

The input prompt was initiated as follows: "I am solving a radiology quiz and will provide you with open-ended, text-based questions. Please act as a radiology professor with 30 years of experience. Provide clear, comprehensive, and detailed answers to each question. Each answer must include four references to papers indexed in Medline. The references should include in-text citations as well as complete details, including the authors' names, title, journal, publication year, month, volume, issue, page numbers, and PubMed identifier (PMID)" (Figure 2). This prompt was presented in December 2024 on six distinct platforms with default parameters: OpenAI's ChatGPT o1-preview, ChatGPT-4o, ChatGPT-4o with canvas (https://chat.openai.com), Google Gemini 1.5 Pro (https://gemini.google.com), Claude 3.5 Sonnet, and Claude 3 Opus (https://claude.ai).

The allocation of tasks among the radiologists was as follows:

- Radiologist 2 (T.C.) conducted the questioning of ChatGPT-4o with canvas, Google Gemini 1.5 Pro, and ChatGPT o1-preview and recorded the responses.
- Radiologist 3 (E.Ç.) conducted the questioning of ChatGPT-4o, Claude 3.5 Sonnet, and Claude 3 Opus.

Due to resource limitations, the experiments were conducted with a single response per model per question to establish a standardized baseline. All LLMs were operated using their default parameters; only the first complete response generated by each model for each question was recorded. Notably, the LLMs were not pre-trained on any specific prompts, data, or question set prior to this study.

Reference evaluation

Validation of reference authenticity

Although the query requested Medline-indexed references, multiple databases were used for verification to account for possible indexing inconsistencies and to ensure a comprehensive assessment of reference accuracy. Each reference was verified across

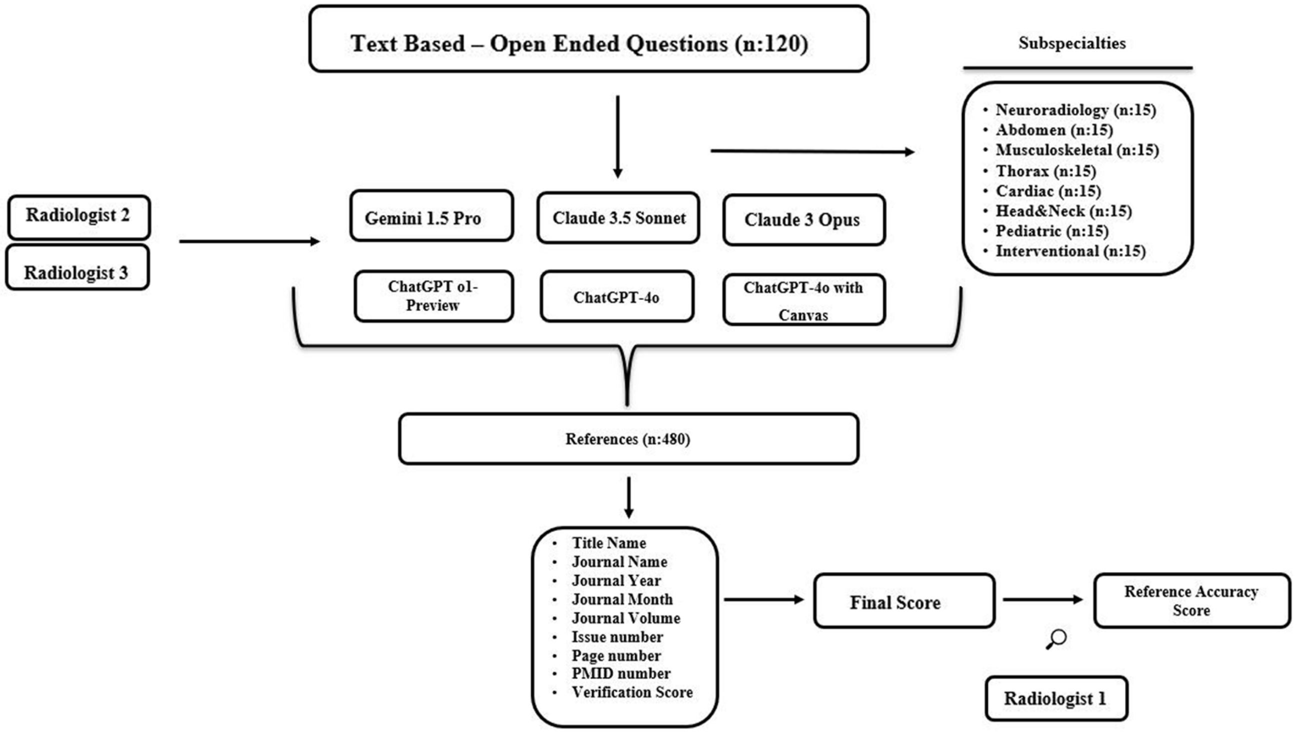
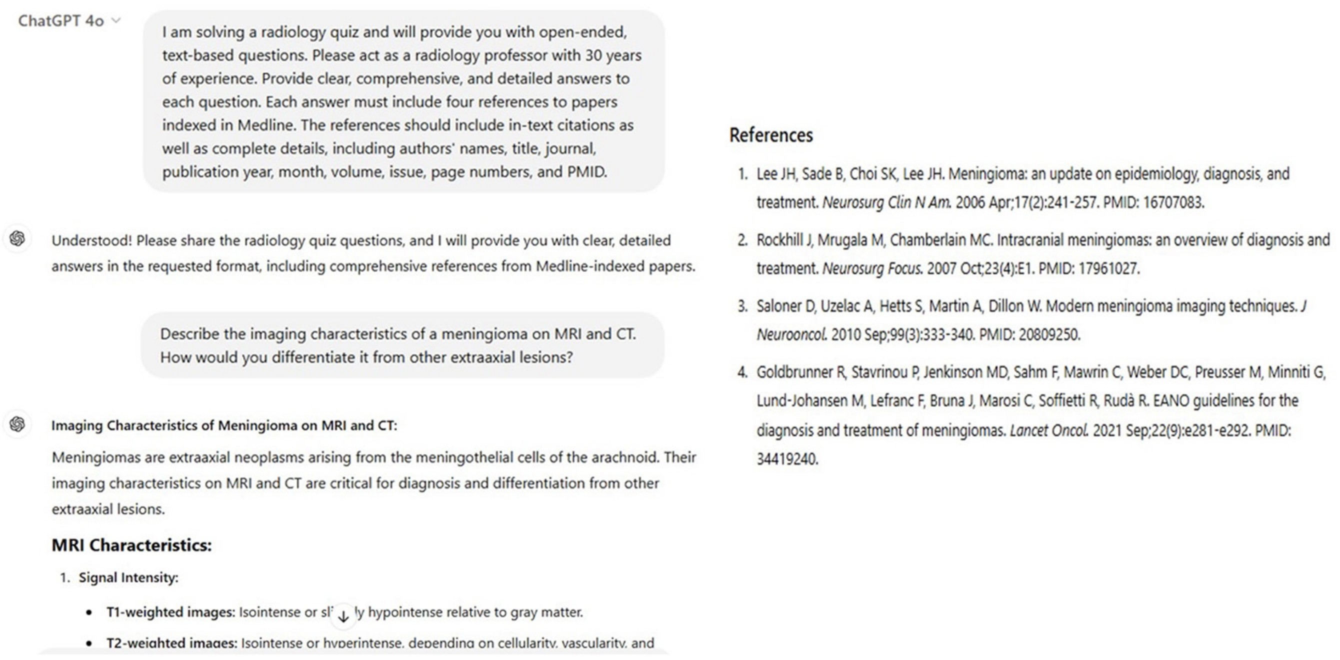


Figure 1. Overview of the study workflow.



I am solving a radiology quiz and will provide you with open-ended, text-based questions. Please act as a radiology professor with 30 years of experience. Provide clear, comprehensive, and detailed answers to each question. Each answer must include four references to papers indexed in Medline. The references should include in-text citations as well as complete details, including authors' names, title, journal, publication year, month, volume, issue, page numbers, and PMID.

Understood! Please share the radiology quiz questions, and I will provide you with clear, detailed answers in the requested format, including comprehensive references from Medline-indexed papers.

Describe the imaging characteristics of a meningioma on MRI and CT. How would you differentiate it from other extraaxial lesions?

Imaging Characteristics of Meningioma on MRI and CT:

Meningiomas are extraaxial neoplasms arising from the meningothelial cells of the arachnoid. Their imaging characteristics on MRI and CT are critical for diagnosis and differentiation from other extraaxial lesions.

MRI Characteristics:

1. **Signal Intensity:**
   - **T1-weighted images:** Isointense or sl↓ly hypointense relative to gray matter.
   - **T2-weighted images:** Isointense or hyperintense, depending on cellularity, vascularity, and

References

1. Lee JH, Sade B, Choi SK, Lee JH. Meningioma: an update on epidemiology, diagnosis, and treatment. *Neurosurg Clin N Am.* 2006 Apr;17(2):241-257. PMID: 16707083.
2. Rockhill J, Mrugala M, Chamberlain MC. Intracranial meningiomas: an overview of diagnosis and treatment. *Neurosurg Focus.* 2007 Oct;23(4):E1. PMID: 17961027.
3. Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imaging techniques. *J Neurooncol.* 2010 Sep;99(3):333-340. PMID: 20809250.
4. Goldbrunner R, Stavrinou P, Jenkinson MD, Sahm F, Mawrin C, Weber DC, Preusser M, Minniti G, Lund-Johansen M, Lefranc F, Bruna J, Marosi C, Soffietti R, Rudà R. EANO guidelines for the diagnosis and treatment of meningiomas. *Lancet Oncol.* 2021 Sep;22(9):e281-e292. PMID: 34419240.

Figure 2. Illustration of the prompts given to large language models and the corresponding responses they generated. MRI, magnetic resonance imaging; CT, computed tomography.

three databases—Medline, Google Scholar, and the Directory of Open Access Journals—and an internet search. If a reference could not be located in any of these databases, it was classified as fabricated.

Stylistic and bibliographic accuracy check

Although references were ultimately scored using a composite measure, each bibliographic element was explicitly examined:

- Authors' names (A), article title (T), journal name (J), publication year (Y), publication month (M), journal volume (V), issue number (I), page numbers (P), PMID number (PM).

Verification score

The verification score (VS) evaluates the accuracy and relevance of references generated by LLMs. Although LLMs may cite sources from the literature, it is crucial for authors to verify that the cited material precisely matches the phrase or statement being referenced. This ensures the accuracy and validity of the reference. To facilitate this evaluation, references are scored using a 5-point Likert scale:

- **0:** Reference is fabricated (not indexed).
- **1:** No pertinent information found in the source.
- **2:** Some pertinent information present.
- **3:** Largely pertinent information.
- **4:** Entirely pertinent information.

Reference accuracy score

The reference accuracy score (RAS) provides a unified metric for evaluating the bibliographic and verification accuracy of references. It is calculated using the following formula:

$$\text{RAS} = (A + T + J + Y + M + V + I + P + PM) \times VS$$

Each bibliographic element (A, T, etc.) is assigned 1 for a match or 0 for a mismatch. The VS, which reflects the alignment between the content and the cited source, is added to the total. This approach ensures a comprehensive evaluation, with scores ranging from 0 (fabricated) to 36 (fully accurate).

To facilitate interpretation, the RAS is categorized into a 5-point Likert scale:

- **RAS 0:** final score (FS) = 0 (fabricated)
- **RAS 1:** FS = 1–11 (weak accuracy)
- **RAS 2:** FS = 12–23 (moderate accuracy)
- **RAS 3:** FS = 24–35 (near accuracy)
- **RAS 4:** FS = 36 (fully accurate)

This categorization simplifies interpretation, offering a clear understanding of reference accuracy, from entirely fabricated to fully verified. Figure 3 provides a visual representation of the calculation and classification methods.

Radiologists' background

Three board-certified radiologists, each with 6 years of radiology experience, participated in this study. Radiologist 2 and radiologist 3 asked the questions to LLMs and recorded all answers. Radiologist 1 then evaluated all references and assessed the accuracy of the responses in a blinded manner, thereby minimizing the risk of bias.

Statistical analysis

Descriptive statistics, including medians, interquartile ranges (IQR), frequencies, and percentages, were calculated. The normality of variable distributions was assessed using the Kolmogorov–Smirnov test.

Due to the non-parametric distribution of the data, the Kruskal–Wallis test was employed to compare quantitative data across multiple groups (different LLMs). Following the Kruskal–Wallis test, Tamhane's T2 test was used for multiple post-hoc comparisons to identify specific group differences. Additionally, the Wilcoxon signed-rank test with a Bonferroni correction was applied to compare paired samples of RASs between LLMs. Statistical significance was set at $P < 0.003$ after applying the Bonferroni correction for 15 pairwise comparisons across six LLMs; otherwise, a P value < 0.05 was considered statistically significant. All statistical analyses were performed using SPSS version 28.0 (IBM Corp., Armonk, NY, USA).

Results

Reference accuracy by large language models

A total of 480 references were analyzed to compare the performance of the six LLMs. The evaluation focused on overall fabrication rates as well as stylistic and bibliographic accuracy across nine core components of each reference.

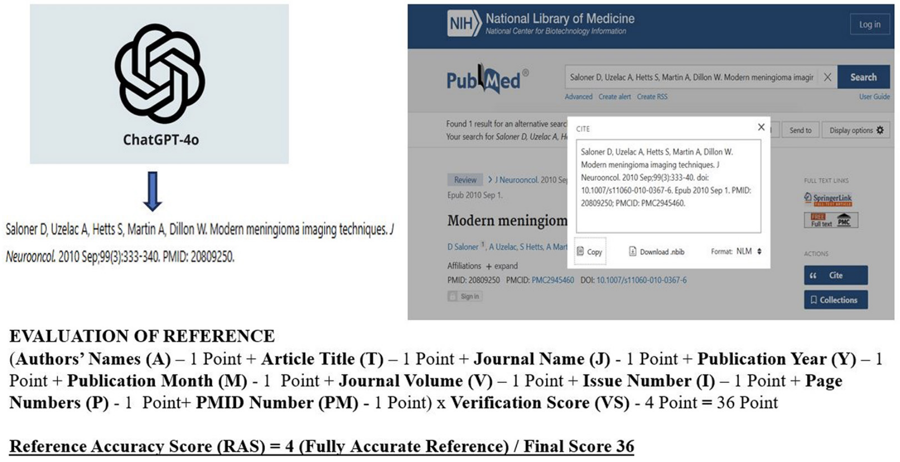
Stylistic and bibliographic accuracy

Authors' names and titles

Claude 3.5 Sonnet showed the highest accuracy for A (96.5%) and T (96.5%), followed by Claude 3 Opus at 81.7% for A and 81.3% for T. The ChatGPT-based models—ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview—generally fell in the mid-range, with accuracies between 44.8% and 58.5% for A and between 46.0% and 53.5% for T. Gemini 1.5 Pro performed the worst in both categories, reaching 38.5% for A and 40.2% for T.

Journal name, year and month

An analogous hierarchy appeared when evaluating J. Here, Claude 3.5 Sonnet again led at 95.6%, followed by Claude 3 Opus at 79.2%. The ChatGPT models ranged from 45.6% to 53.1%, and Gemini 1.5 Pro achieved 38.3%. For Y, Claude 3.5 Sonnet and Claude 3 Opus scored 95.6% and 77.7%, respec-



Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imaging techniques. *J Neurooncol.* 2010 Sep;99(3):333-340. PMID: 20809250.

Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imagir

Found 1 result for an alternative searc
Your search for *Saloner D, Uzelac A, H*

J Neurooncol. 2010 Se
Epub 2010 Sep 1.

Modern meningiom

D Saloner [1], A Uzelac, S Hetts, A Mart

Affiliations + expand

PMID: 20809250 PMCID: PMC2945460 DOI: 10.1007/s11060-010-0367-6

Saloner D, Uzelac A, Hetts S, Martin A, Dillon W. Modern meningioma imaging techniques. J Neurooncol. 2010 Sep;99(3):333-40. doi: 10.1007/s11060-010-0367-6. Epub 2010 Sep 1. PMID: 20809250; PMCID: PMC2945460.

EVALUATION OF REFERENCE
(**Authors' Names (A)** – 1 Point + **Article Title (T)** – 1 Point + **Journal Name (J)** - 1 Point + **Publication Year (Y)** – 1 Point + **Publication Month (M)** - 1 Point + **Journal Volume (V)** – 1 Point + **Issue Number (I)** – 1 Point + **Page Numbers (P)** - 1 Point+ **PMID Number (PM)** - 1 Point) x **Verification Score (VS)** - 4 Point **=** 36 Point

Reference Accuracy Score (RAS) = 4 (Fully Accurate Reference) / Final Score 36

Figure 3. The example showcases the formatting of a reference generated by ChatGPT-4o, followed by its verification on PubMed. Each reference component, including author names, article title, journal name, publication year, month, volume, issue number, page numbers, and PMID, contributes to the final reference accuracy score. ChatGPT, Chat Generative Pre-trained Transformer; PMID, PubMed identifier.

tively, whereas the ChatGPT group landed between 41.9% and 53.1%. Gemini 1.5 Pro showed a low 26.7%. In M, Claude 3.5 Sonnet recorded 95.6% versus Claude 3 Opus at 77.3%, with the ChatGPT models coming in between 13.8% and 23.1% and Gemini 1.5 Pro at 31.7%.

Journal volume, issue number, and page number

Performance remained consistent for V, where Claude 3.5 Sonnet reached 95.2% and Claude 3 Opus 78.1%. The ChatGPT series ranged from 40.0% to 44.4%, and Gemini 1.5 Pro again dipped to 8.8%. Assessing I revealed 94.6% accuracy for Claude 3.5 Sonnet and 77.7% for Claude 3 Opus, with ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview spanning 29.8% to 42.7% and Gemini 1.5 Pro at 18.5%. For P, Claude 3.5 Sonnet and Claude 3 Opus recorded 93.8% and 77.5%, respectively, whereas ChatGPT-based models came in between 26.3% and 44.0%. Gemini 1.5 Pro once more ranked lowest at 16.5%.

PubMed identifier number

A similar pattern was seen in the PM category. Claude 3.5 Sonnet scored 94.0%, followed by Claude 3 Opus at 77.5%. The ChatGPT-4o model reached 23.1%, ChatGPT-4o with canvas 9.8%, ChatGPT o1-preview 10.8%, and Gemini 1.5 Pro was placed last at 3.3%.

Verification scores

VS showed a clear ranking among the LLMs. Claude 3.5 Sonnet and Claude 3 Opus both achieved the highest median verification Likert score of 4, with an IQR of 4–4 for each. In contrast, ChatGPT-4o recorded a median score of 3 (IQR: 0–4). ChatGPT-4o with canvas, ChatGPT o1-preview, and Gemini 1.5 Pro all had lower VSs, each reporting a median of 0 (IQR: 0–4).

Final scores of large language models

Final scores, presented as median and IQR, confirmed the leading positions of Claude 3.5 Sonnet and Claude 3 Opus. Claude 3.5 Sonnet ranked first with a median score of 36 (IQR: 36–36), followed by Claude 3 Opus at 36 (IQR: 36–18). ChatGPT o1-preview and ChatGPT-4o recorded median scores of 16 (IQR: 28–0) and 8 (IQR: 28–0), respectively. The lowest-ranked models were ChatGPT-4o with canvas with 0 (IQR: 28–0) and Gemini 1.5 Pro with 0 (IQR: 16–0).

All scores and reference component accuracies are summarized in Table 1.

Comparison of reference accuracy score by large language models

Claude 3.5 Sonnet exhibited the smallest fabrication rate at 3.1% while also achieving the highest proportion of fully accurate references (80.8%). Although Claude 3 Opus showed a higher fabrication rate of 18.3%, it still produced 59.6% fully accurate references. In comparison, the ChatGPT-based models all generated significantly more fabricated references (27.7%–52.9%) and fewer fully accurate ones (5.6%–7.3%). Gemini 1.5 Pro stood out with the highest fabrication rate of 60.6% and the lowest rate of fully accurate references at 2.7% (Table 2) (Figure 4).

Claude 3.5 Sonnet emerged as the top-performing model, significantly outperforming all others, including Claude 3 Opus ($P < 0.001$). Claude 3 Opus demonstrated strong performance, ranking second, with significant differences observed against all other models ($P < 0.001$). No significant differences were observed among the ChatGPT models. Specifically, comparisons of ChatGPT o1-preview and ChatGPT-o4 against ChatGPT-4o with canvas yielded Bonferroni-corrected P values of 0.019 and 0.037, respectively—both above the significance threshold of 0.003. Additionally, the difference between ChatGPT-4o and ChatGPT o1-preview was not significant ($P = 0.456$). In contrast, Google Gemini 1.5 Pro recorded the lowest accuracy, significantly underperforming compared with the Claude and ChatGPT models ($P < 0.001$) (Table 3).

Performance analysis by subspecialty

In a performance analysis of reference accuracy across multiple radiology subspecialties, several LLMs demonstrated distinct patterns of variability. Claude 3.5 Sonnet, Claude 3 Opus, ChatGPT-4o, ChatGPT o1-preview, and ChatGPT-4o with canvas each showed notable fluctuations ($P < 0.05$), whereas Google Gemini 1.5 Pro exhibited uniformly lower performance across all subspecialties without any statistically significant differences ($P > 0.05$) (Table 4).

Table 1. Comparative performance of large language models in reference component accuracy and overall scores

	Reference (n = 480)					
	Claude 3.5 Sonnet	Claude 3 Opus	ChatGPT-4o	ChatGPT o1-preview	ChatGPT-4o with canvas	Gemini 1.5 Pro
Authors' names	463 (96.5%)	392 (81.7%)	281 (58.5%)	251 (52.3%)	215 (44.8%)	185 (38.5%)
Title name	463 (96.5%)	390 (81.3%)	257 (53.5%)	250 (52.1%)	221 (46.0%)	193 (40.2%)
Journal name	459 (95.6%)	380 (79.2%)	219 (45.6%)	255 (53.1%)	220 (45.8%)	184 (38.3%)
Journal year	459 (95.6%)	373 (77.7%)	201 (41.9%)	248 (51.7%)	212 (44.2%)	128 (26.7%)
Journal month	459 (95.6%)	371 (77.3%)	66 (13.8%)	111 (23.1%)	68 (14.2%)	152 (31.7%)
Journal volume	457 (95.2%)	375 (78.1%)	204 (42.5%)	213 (44.4%)	192 (40.0%)	42 (8.8%)
Issue number	454 (94.6%)	373 (77.7%)	183 (38.1%)	205 (42.7%)	143 (29.8%)	89 (18.5%)
Page number	450 (93.8%)	372 (77.5%)	126 (26.3%)	211 (44.0%)	172 (35.8%)	79 (16.5%)
PMID number	451 (94.0%)	372 (77.5%)	111 (23.1%)	52 (10.8%)	47 (9.8%)	16 (3.3%)
Verification Likert score* [median, IQR (Q3-Q1)]	4 (4–4)	4 (4–2)	3 (4–0)	3 (4–0)	0 (4–0)	0 (4–0)
Final score [median, IQR (Q3-Q1)]**	36 (36–36)	36 (36–18)	8 (28–0)	16 (28–0)	16 (0–0)	0 (32–0)

IQR: interquartile range, Q1: 25% quantile, Q3: 75% quantile.

*Verification Likert score: this is categorized into a 5-point Likert scale reference accuracy score (0 = fabricated; 4 = fully accurate).

**Final score: the final score provides an integrated metric that combines the bibliographic accuracy of references with their verification score (VS). For each bibliographic element—such as authors' names, article title, journal name, and others—a match was scored as 1, and a mismatch was scored as 0. The VS, which measures how well the content aligns with the cited source, was then multiplied by the sum of the matched elements. PMID, PubMed identifier; ChatGPT, Chat Generative Pre-trained Transformer.

The post-hoc Tamhane test revealed that the Claude 3.5 Sonnet model showed no significant differences in reference accuracy across subspecialties, indicating uniformly consistent performance without any specific category demonstrating clear outperformance or underperformance. Similarly, Google Gemini 1.5 Pro performed uniformly across all subspecialties but with overall lower accuracy than other models.

Within Claude 3 Opus, neuroradiology demonstrated consistent superiority over most categories ($P < 0.05$), except for abdominal, cardiac, and head and neck radiology, where no significant differences were observed. Additionally, cardiac radiology outperformed the pediatric radiology group ($P = 0.020$). No other significant differences were found among the remaining subgroups.

For ChatGPT-4o, cardiac radiology consistently emerged as the best-performing category ($P < 0.05$), except when compared with abdominal and interventional radiology, where performance was comparable. Conversely, pediatric radiology showed the weakest results, being significantly outperformed by other subspecialties, except for head and neck and musculoskeletal radiology ($P < 0.05$). No additional significant differences were detected.

In the case of ChatGPT-4o with canvas, thoracic radiology emerged as the highest-performing category, achieving significantly greater accuracy than most other subspecialties ($P < 0.05$), except for neuroradiology, cardiac, and musculoskeletal radiology. Conversely, head and neck radiology showed the weakest performance, being significantly outperformed by both thoracic radiology and cardiac radiology ($P < 0.05$). Additionally, cardiac radiology demonstrated superior performance to abdominal, pediatric, and interventional radiology ($P < 0.05$). No further significant differences were observed among the subgroups.

As for ChatGPT o1-preview, head and neck radiology exhibited the lowest performance, being significantly outperformed by all other categories ($P < 0.05$) except for interventional and pediatric radiology, where no significant differences were observed. No further significant differences were identified among the subgroups.

Discussion

The most striking finding of our study is the consistent superiority of the Claude 3.5 Sonnet model in generating accurate and reliable medical references across diverse radiology subspecialties. With a significantly higher RAS ($P < 0.001$), a notably low fabrication rate (3.1%), and 80.8% of its references being fully accurate, Claude 3.5 Sonnet demonstrates a remarkable ability to integrate comprehensive radiological literature into its outputs. Given the critical importance of accuracy in reference generation, where even minor errors can have serious implications, Claude 3.5 Sonnet's ability to produce such a high percentage of fully accurate references underscores its potential as a reliable reference generator compared with other advanced LLMs. This superior performance likely stems from several factors, including a broader and more specialized training dataset and algorithmic refinements aimed at reducing hallucination rates—a common limitation in other models.[20] The Claude models leverage constitutional AI, a framework that prioritizes accuracy, ethical reasoning, and factual integrity, which may contribute to its minimized hallucination rates and enhanced reliability.[22]

In contrast, the Claude 3 Opus model, although ranking second overall, displayed a higher fabrication rate (18.3%) and a reduced proportion of fully accurate references (59.6%). This difference suggests that the underlying architecture of the Claude models is promising, especially in subspecialties where the training data may be less robust, such as pediatric or interventional radiology.

The ChatGPT models (ChatGPT-4o, ChatGPT-4o with canvas, and ChatGPT o1-preview) exhibited only moderate performance. Their elevated rates of fabricated references—ranging from 27.7% to 52.9%—and recurrent inaccuracies in critical bibliographic components (such as PMID numbers and page details) indicate that these models have not yet achieved the precision required for reliable academic referencing. This result is consistent with prior studies on ChatGPT-generated medical content.[6-16] For instance, Bhattacharyya et al.[6] reported that nearly half the references produced by ChatGPT-3.5 were fabricated, with 47% being non-authentic and only 7% being both authentic and accurate. Similarly, Walters and Wilder[8] found that 55% of references from ChatGPT-3.5 were fabricated, and even in ChatGPT-4, the fabrication rate remained concerning at 18%, with 43% of authentic references from ChatGPT-3.5 and 24% from ChatGPT-4o containing substantive errors. Wagner et al.[9] evaluated ChatGPT-3's accuracy in answering 88 radiology questions and verifying references. Correct answers were provided for 67% of questions, and 33% contained errors. Of 343 references, 63.8% were fabricated, and only 37.9% of the verified references offered sufficient information.[9]

Gravel et al.[16] further observed that 69% of the 59 references generated by ChatGPT for medical questions were fabricated. In our study, ChatGPT-4o produced only 31 correct references out of 480, and ChatGPT o1-preview improved only modestly to 35 correct references, underscoring the persistent challenges in achieving accurate citation generation. These specific findings, along with the reported fabrication rates in our models, mirror the issues highlighted in the previous literature and indicate that even the upgraded versions of ChatGPT continue to fall short in reliably generating complete and verifiable academic references.

Table 2. Comparative evaluation of large language models based on reference accuracy score

	Reference (n = 480)					
RAS	Claude 3.5 Sonnet	Claude 3 Opus	ChatGPT-4o	ChatGPT-4o with canvas	ChatGPT o1-preview	Gemini 1.5 Pro
0 (fabrication)	15 (3.1%)	88 (18.3%)	133 (27.7%)	254 (52.9%)	226 (47.1%)	291 (60.6%)
1 (weak)	7 (1.5%)	18 (3.8%)	142 (29.6%)	22 (4.6%)	5 (1.0%)	21 (4.4%)
2 (moderate)	4 (0.8%)	21 (4.4%)	62 (12.9%)	32 (6.7%)	41 (8.5%)	74 (15.4%)
3 (near accurate)	66 (13.8%)	67 (14.0%)	112 (23.3%)	145 (30.2%)	173 (36.0%)	81 (16.9%)
4 (accurate)	388 (80.8%)	286 (59.6%)	31 (6.5%)	27 (5.6%)	35 (7.3%)	13 (2.7%)

Reference accuracy score: this evaluates the accuracy and relevance of references generated by large language models (LLMs). Although LLMs may cite sources from the literature, it is crucial for authors to verify that the cited material precisely matches the phrase or statement being referenced. This ensures the accuracy and validity of the reference. To facilitate this evaluation, references are scored using a 5-point Likert scale. ChatGPT, Chat Generative Pre-trained Transformer.

Google Gemini 1.5 Pro's performance was the poorest among the evaluated models, with a fabrication rate of 60.6% and only 2.7% of its references being fully accurate. The uniform underperformance of Google Gemini 1.5 Pro across all radiology subspecialties implies potential fundamental limitations—possibly stemming from a training dataset that underrepresents or insufficiently emphasizes medical literature or from an algorithmic framework that is less suited to the nuances of academic citation generation.

In our performance analysis by subspecialty, we highlighted that although Claude 3.5 Sonnet maintained uniformly high reference accuracy across all subspecialties, other models exhibited substantial variability. For example, Claude 3 Opus demonstrated superior performance in neuroradiology, whereas ChatGPT-4o achieved remarkable results in cardiac radiology and ChatGPT-4o with canvas showed exceptional performance in thoracic radiology. In contrast, Google Gemini 1.5 Pro consistently exhibited low accuracy across all subspecialties. These findings suggest that differences in data complexity and training representation may account for the inter-model and inter-subspecialty performance variations.

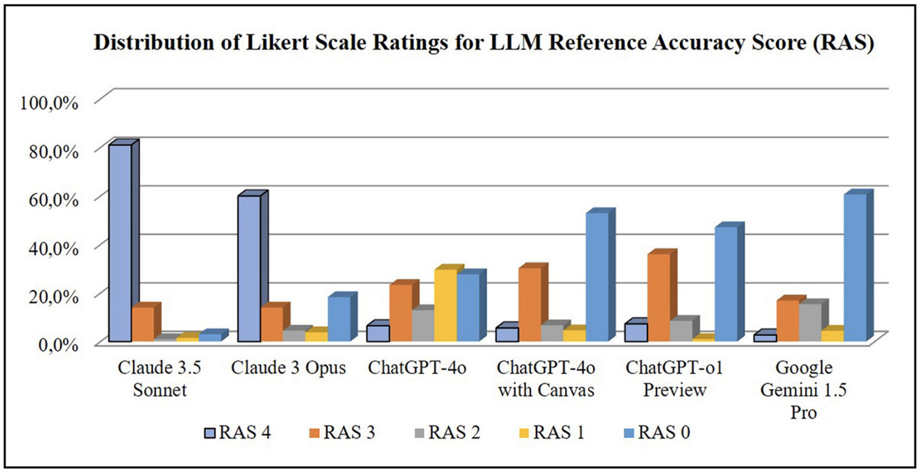


Figure 4. Distribution of Likert scale ratings for large language model reference accuracy scores. LLM, large language model.

Table 3. Comparison of accuracy of large language models with *P* values from the Wilcoxon test

	ChatGPT-4o	ChatGPT-4o with canvas	ChatGPT o1-preview	Google Gemini 1.5 Pro	Claude 3.5 Sonnet	Claude 3 Opus
ChatGPT-4o	-	0.037	0.456	<0.001	<0.001	<0.001
ChatGPT-4o with canvas	0.037	-	0.019	<0.001	<0.001	<0.001
ChatGPT o1-preview	0.456	0.019	-	<0.001	<0.001	<0.001
Google Gemini 1.5 Pro	<0.001	<0.001	<0.001	-	<0.001	<0.001
Claude 3.5 Sonnet	<0.001	<0.001	<0.001	<0.001	-	<0.001
Claude 3 Opus	<0.001	<0.001	<0.001	<0.001	<0.001	-

ChatGPT, Chat Generative Pre-trained Transformer.

Table 4. Reference accuracy score of large language models and classification by subspecialities

		Neuro	Abdomen	Musculoskeletal	Thorax	Cardiac	Head and Neck	Pediatric	Interventional	*P* value	
Claude 3 Opus	Median	4	4	4	3	4	4	3	4	<0.001	K*
	IQR (Q3–Q1)	(4–4)	(4–3.25)	(4–1)	(4–2)	(4–4)	(4–3)	(4–0)	(4–1)		
Claude 3.5 Sonnet	Median	4	4	4	4	4	4	4	4	0.008	K*
	IQR (Q3–Q1)	(4–4)	(4–4)	(4–4)	(4–3)	(4–4)	(4–3)	(4–4)	(4–3.25)		
ChatGPT-4o	Median	1	2	1	1	1	1	1	1	<0.001	K*
	IQR (Q3–Q1)	(2.75–0)	(3–0)	(2–0)	(3–1)	(3–1)	(3–0)	(1–0)	(3–1)		
ChatGPT-4o with canvas	Median	2	0	0	3	3	0	0	0	<0.001	K*
	IQR (Q3–Q1)	(3–0)	(3–0)	(3–0)	(3–0)	(3–0)	(2–0)	(2.75–0)	(2–0)		
ChatGPT o1-preview	Median	2	3	2.5	3	3	0	0.5	0	<0.001	K*
	IQR (Q3–Q1)	(3–0)	(3–0)	(3–0)	(3–0)	(3–0)	(1.5–0)	(3–0)	(3–0)		
Google Gemini 1.5 Pro	Median	0.5	0	0	0	0	0	0	0	0.244	K
	IQR (Q3–Q1)	(2.75–0)	(2–0)	(2–0)	(2–0)	(3–0)	(2–0)	(2–0)	(2.75–0)		

IQR, interquartile range; Q1, 25% quantile; Q3, 75% quantile; Neuro, neuroradiology; K, Kruskal–Wallis; *Tamhane's T2 test was used as a post-hoc comparison between each group, and the results of these comparisons are discussed in the results section. ChatGPT, Chat Generative Pre-trained Transformer.

Accurate reference generation is crucial in radiology, as evidence-based decision-making and scientific communication depend on verifiable and precise citations.[9] Inaccurate or fabricated references can lead to serious repercussions. For instance, misleading citations may result in clinicians basing diagnostic or treatment decisions on non-existent or irrelevant studies, ultimately affecting patient outcomes; in academic settings, reliance on erroneous citations can erode trust in literature reviews, undermine scholarly debates, and propagate errors in subsequent research.[23,24] Given these risks, the marked superiority of Claude 3.5 Sonnet has considerable practical implications, as this model could be integrated into workflows for manuscript preparation, automated literature retrieval, or even serve as an adjunct tool in clinical guideline development, provided that human experts continue to verify its outputs.

Additionally, our study observed that all the LLMs evaluated tend to favor references from the most well-known radiology papers. This tendency to prioritize widely cited papers can reinforce the "Matthew Effect," which refers to the phenomenon where frequently cited papers continue to gain references, overshadowing lesser-known but potentially important studies, in literature review processes.[25] This inclination of LLMs to rely on popular sources could narrow the scope of the literature being considered, limiting the diversity and range of research references. As a result, the use of these models may unintentionally contribute to reinforcing a limited set of references, reducing the overall richness of the academic discussion.

Although this study offers valuable insights into the capabilities of LLMs in generating medical references in radiology, several limitations must be noted. The dataset was relatively small, potentially limiting the generalizability of the findings across various radiological subspecialties and medical topics. Moreover, the use of a single standardized prompt may not capture the full variability of LLM responses arising from different prompting strategies or settings (e.g., temperature, top-K, top-P, and token limits). In addition, model performance was not assessed across multiple citation styles (e.g., AMA, Chicago), which restricts understanding of the broader applicability of these models in academic and clinical settings. The absence of repeated measurements for each LLM could introduce stochastic variability into the results, and the study evaluated only specific versions of LLMs available at the time, potentially misrepresenting the evolving capabilities of newer models. Future work may explore response consistency through multiple iterations per query.

In conclusion, Claude 3.5 Sonnet outperformed all other LLMs, demonstrating high accuracy and reliability in generating radiology references, making it well suited for tasks such as literature retrieval and manuscript preparation. This model holds great potential as a supportive tool for radiologic reference generation, offering a valuable resource to complement evidence-based practice. In contrast, other models exhibited higher fabrication rates and inconsistent accuracy, underscoring the need for substantial improvements. Future efforts should focus on enhancing performance in underperforming subspecialties and refining bibliographic accuracy to meet the rigorous demands of evidence-based radiology.

Acknowledgements

The authors thank Büşra Özal Güneş for his assistance with the statistical analysis of this manuscript. They utilized ChatGPT-4o, a large language model (January 2025 Version; OpenAI; https://chat.openai.com), to revise the grammar and English translation. The content of this publication is entirely the responsibility of the authors, who have reviewed and edited it as necessary.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary: https://d2v96fxpocvxx.cloudfront.net/c1dc3a38-51db-436b-af33-1bc7522029b3/content-images/a6628944-e199-444c-8b76-3c2a723824b2.pdf

References

1. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol.* 2024;34(5):2817-2825. [Crossref]
2. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]
3. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidiscip Healthc*. 2023;16:1513-1520. [Crossref]
4. Goktas P, Agildere AM. Transforming radiology with artificial intelligence visual chatbot: a balanced perspective. *J Am Coll Radiol*. 2024;21(2):224-225. [Crossref]
5. Bera K, O'Connor G, Jiang S, Tirumani SH, Ramaiya N. Analysis of ChatGPT publications in radiology: literature so far. *Curr Probl Diagn Radiol*. 2024;53(2):215-225. [Crossref]
6. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15(5):e39238. [Crossref]
7. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164. [Crossref]
8. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045. [Crossref]
9. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Can Assoc Radiol J*. 2024;75(1):69-73. [Crossref]
10. Athaluri SA, Manthena SV, Kesapragada VSRKM, et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*. 2023;15(4):e37432 [Crossref]
11. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology-head and neck surgery. *Eur Arch Otorhinolaryngol*. 2024;281(4):2159-2165. [Crossref]
12. Day T. A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. *Prof Geogr*. 2023;75(6):1024-1027. [Crossref]
13. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res*. 2023;326:115334. [Crossref]
14. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol*. 2023;280(11):5129-5133. [Crossref]
15. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *J Med Internet Res*. 2024;26:e52935. [Crossref]
16. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin Proc Digit Health*. 2023;1(3):226-234. [Crossref]
17. Steybe D, Poxleitner P, Aljohani S, et al. Evaluation of a context-aware chatbot

using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. *J Craniomaxillofac Surg*. 2025;53(4):355-360. **[Crossref]**

18. Patel N, Grewal H, Buddhavarapu V, Dhillon G. OpenEvidence: enhancing medical student clinical rotations with AI but with limitations. *Cureus*. 2025;17(1): e76867. **[Crossref]**
19. Temsah MH, Jamal A, Alhasan K, Temsah AA, Malki KH. OpenAI o1-preview vs. ChatGPT in healthcare: a new frontier in medical AI reasoning. *Cureus*. 2024;16(10):e70640. **[Crossref]**
20. Kurokawa R, Ohizumi Y, Kanzawa J, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in radiology's "Diagnosis Please" cases. *Jpn J Radiol*. 2024;42(12):1399-1402. **[Crossref]**
21. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865-868. **[Crossref]**
22. Aydın Ö, Karaarslan E. Is ChatGPT leading generative AI? What is beyond expectations? *APJESS*. 2023;11(3):118-134. **[Crossref]**
23. Dumas-Mallet E, Boraud T, Gonon F. Le mésusage des citations et ses conséquences en médecine [Citation misuse and its effects on public health]. *Med Sci (Paris*). 2021;37(11):1035-1041. **[Crossref]**
24. Peoples N, Østbye T, Yan LL. Burden of proof: combating inaccurate citation in biomedical literature. *BMJ*. 2023;383:e076441. **[Crossref]**
25. Larivière V, Gingras Y. 2010. The impact factor's Matthew Effect: a natural experiment in bibliometrics. J Assoc Information Sci Technol 2010;61(2):424-427. **[Crossref]**

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253352

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Findings of suspicious calcifications on contrast-enhanced mammography and their pathological correlation

Dilşah Oral[1]
İhsan Şebnem Örgüç[1]
Hanife Seda Mavili[2]
Teoman Coşkun[3]

[1]Manisa Celal Bayar University Hafsa Sultan Hospital, Department of Radiology, Manisa, Türkiye

[2]Manisa Celal Bayar University Hafsa Sultan Hospital, Department of Pathology, Manisa, Türkiye

[3]Manisa Celal Bayar University Hafsa Sultan Hospital, Department of General Surgery, Manisa, Türkiye

PURPOSE

This study aimed to determine the performance of contrast-enhanced mammography (CEM) in evaluating suspicious calcifications not associated with a mass.

METHODS

Patients with suspicious calcifications detected on CEM performed at our center between February 2021 and December 2023 were included in the study. Retrospectively, the morphology, distribution, and longest axis length of the calcifications were assessed on low-energy images, whereas contrast enhancement intensity, pattern, longest axis length, and enhancement curves were analyzed on recombined images. The pathological diagnosis, grade, Ki-67 index, and (if available) the longest lesion length in the surgical specimen were recorded. Using pathology as the gold standard, various CEM parameters were evaluated for their performance in assessing this group of calcifications. Primary and secondary analyses were performed based on combined low or no enhancement and no enhancement alone, respectively.

RESULTS

Our study includes 132 lesions in 114 patients,18 of whom had bilateral calcifications. Of the 132 lesions included in the study, 78 were benign, and 54 were malignant. Sensitivity, specificity, positive predictive value, and negative predictive value were determined as follows: 72.2%, 62.8%, 57.3%, and 76% in low-energy images; 79.6%, 80.8%, 74.1%, and 85.1% in the primary analysis of recombined images; and 98.2%, 47.4%, 56.4%, and 97.4% in the secondary analysis. Contrast enhancement intensity was identified as a significant parameter influencing malignancy risk. A strong statistical correlation was observed between lesion length measurements in both low-energy and recombined images compared with pathology ($r = 0.733$ and $r = 0.879$, $P < 0.001$ for both), with mean differences of -4.75 mm and +4.45 mm. No statistically significant relationship was found between contrast enhancement intensity and the distinction between invasive and *in situ* carcinoma ($P = 0.698$) or the differentiation of ductal carcinoma *in situ* grade ($P = 0.336$). A significant correlation was detected between pathology and dynamic contrast enhancement types adapted from magnetic resonance imaging (MRI) ($P = 0.019$). Although no statistically significant linear correlation was found between the Ki-67 index and contrast enhancement intensity, the P value was close to significance ($P = 0.057$).

CONCLUSION

CEM demonstrates strong performance in the assessment of suspicious calcifications by combining the morphological and distributional features of digital mammography with enhancement characteristics similar to MRI.

CLINICAL SIGNIFICANCE

The findings support that CEM exhibits effective performance in evaluating suspicious calcifications not associated with a mass and may have a potential role in routine clinical practice.

KEYWORDS

Contrast-enhanced mammography, suspicious calcifications, breast cancer

Corresponding author: Dilşah Oral

E-mail: dilsahoral@gmail.com

Received 17 March 2025; revision requested 13 April 2025; accepted 12 June 2025.

Epub: 21.07.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253352

You may cite this article as: Oral D, Örgüç İŞ, Mavili HS, Coşkun T. Findings of suspicious calcifications on contrast-enhanced mammography and their pathological correlation. *Diagn Interv Radiol*. 2026;32(2):182-190.

Breast cancer is the most common cancer among women, and mammography is the primary screening method.[1] Calcifications are frequently observed findings on mammograms.[2] According to the Breast Imaging Reporting and Data System (BI-RADS) atlas, based on their morphology and distribution, calcifications are typically classified as benign or suspicious.[3] Some suspicious calcifications may represent invasive cancer or ductal carcinoma *in situ* (DCIS), the earliest stage of breast cancer. Approximately one-third of breast cancers present solely as calcifications on mammography.[4] Histopathological sampling is recommended for BI-RADS 4 and BI-RADS 5 calcifications.[3]

In the diagnosis of breast cancer, ultrasonography (USG) and magnetic resonance imaging (MRI), which are used in addition to mammography, detect calcifications at a low rate or may not detect them at all. In recent years, contrast-enhanced mammography (CEM) has become increasingly utilized as an imaging technique that, comparable with MRI, demonstrates neovascularization in the breast through the use of iodinated contrast agents. In malignant lesions, vascular structures formed during angiogenesis exhibit increased permeability, allowing intravenously administered contrast agents to penetrate the tumor, resulting in enhancement.[5,6] In CEM, two image types are acquired: low-energy and high-energy. Low-energy images are considered equivalent to digital mammography and are suitable for the evaluation of calcifications.[7] Recombined images, generated using both low-and high-energy images, demonstrate enhancement. In CEM, although the morphology and distribution of calcifications are assessed in low-energy images, the functional characteristics of the same tissue can also be evaluated using recombined images.

Main points

- Contrast-enhanced mammography (CEM) is an imaging technique based on the dual-energy method and provides information on tissue function in addition to the structural evaluation of calcifications.
- According to our study, the presence of enhancement in the calcification region exhibited a sensitivity of 98.2% for malignancy.
- Recombined images demonstrating enhancement provided more accurate results in determining tumor length compared with low-energy mammographic equivalent images. This suggests that CEM may reduce the positive surgical margin rate in calcifications.

Our study aims to investigate the utility of CEM, a relatively new and increasingly adopted technique, in the characterization of suspicious calcifications not associated with a mass. Additionally, we aim to evaluate our hypothesis that CEM may help reduce unnecessary biopsies, contribute to determining the extent of disease in patients undergoing breast-conserving surgery, and decrease the rate of positive surgical margins.

Methods

Approval was obtained from Manisa Celal Bayar University Clinical Research Ethics Committee for this retrospective study (decision number: 20.478.486/2230, date: 31/01/2024). Between February 2021 and December 2023, CEM images of 910 patients with clinically suspected malignancy and/or suspicious findings detected by other imaging modalities were reviewed. Patients with suspicious calcifications not associated with a mass were selected. Patients without suspicious calcifications, those with suspicious calcifications associated with a mass, those with suspicious calcifications but accompanied by a mass in another focus, and those who had received neoadjuvant therapy before imaging were excluded from the study. With 96 patients having lesions (calcification focus) in a single breast and 18 patients having lesions in both breasts, in 114 patients a total of 132 lesions were included in this retrospective study (Figure 1).

Contrast-enhanced mammography examinations

All lesions included in this study were imaged using a digital mammography system with dual-energy capability (Pristina, GE, Rue De La Minière, Buc, France) available in the clinic. A low-osmolar, nonionic contrast agent (Opaxol 350 mg/mL, Kopaq 350 mg/mL) was administered at a dose of 1.5 mL/kg, with a maximum volume of 100 mL, using an automatic injector system (Medrad) at a rate of 3 mL/s. Imaging commenced 1.5–2 minutes after the start of contrast injection. Craniocaudal and mediolateral oblique (MLO) images were acquired in both breasts, starting with the target breast. Additionally, in cases suspicious for malignancy, around the 8th minute, a second MLO image (delayed phase) of the pathological breast was obtained after the MLO image of the non-target breast. The total procedure lasted approximately 8–9 minutes, depending on patient compliance. Low-energy images were acquired at 28–32 kVp, and high-energy images were obtained at 45–49 kVp. The filtration material and kVp values were automatically determined by the system. Recombined images were automatically generated from the low-and high-energy images.

Image analysis

The images were evaluated in consensus by two observers: one radiologist with 34 years of experience in breast imaging and a fourth-year radiology resident. During the assessment of the CEM images, the BI-RADS CEM lexicon published by the American College of Radiology in 2022 was used. Suspicious calcifications were classified based on their morphology and distribution using the low-energy images, which are equivalent to standard mammography. According to the BI-RADS atlas, calcifications with a morphol-

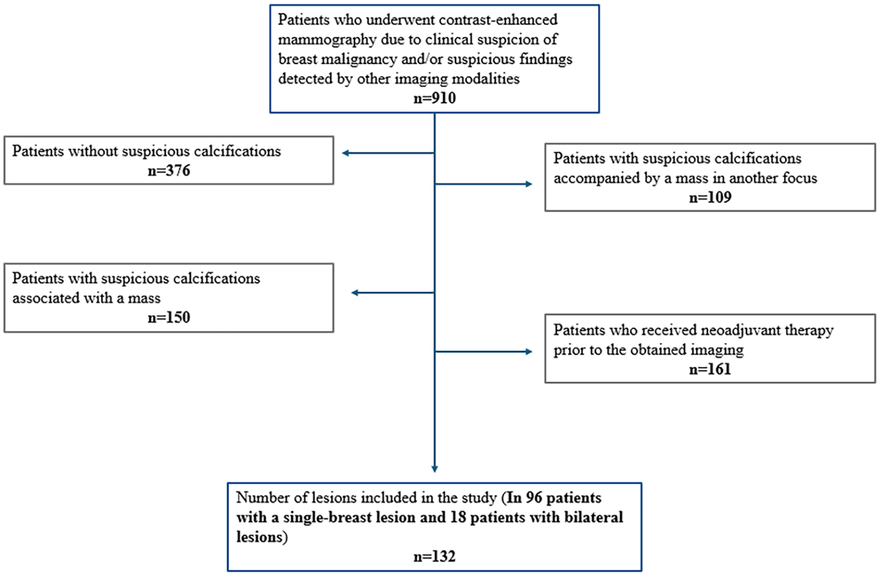

Patients who underwent contrast-enhanced mammography due to clinical suspicion of breast malignancy and/or suspicious findings detected by other imaging modalities
n=910
Patients without suspicious calcifications
n=376
Patients with suspicious calcifications accompanied by a mass in another focus
n=109
Patients with suspicious calcifications associated with a mass
n=150
Patients who received neoadjuvant therapy prior to the obtained imaging
n=161
Number of lesions included in the study (In 96 patients with a single-breast lesion and 18 patients with bilateral lesions)
n=132

Figure 1. Flowchart of the study.

ogy or distribution associated with <50% malignancy were categorized as low-suspicion (amorphous, coarse heterogeneous, fine pleomorphic and diffuse, regional, grouped), and those associated with ≥50% malignancy were classified as high-suspicion (fine linear/fine linear branching and linear, segmental). Cases were categorized into three risk groups: low-risk if both morphology and distribution were low-suspicion, intermediate-risk if one was high-suspicion, and high-risk if both were high-suspicion. In the analysis of low-energy images, the low-risk group was considered benign, whereas the intermediate-and high-risk groups were classified as malignant. On the recombined images, the relationship between the suspicious calcification area and the enhancing tissue was assessed in four categories: no enhancement observed, partial enhancement of the tissue, complete enhancement with overlapping characteristics, or enhancement extending beyond the calcification area. The enhancement pattern was classified as homogeneous, heterogeneous, or clumped. The intensity of enhancement was qualitatively graded as high, moderate, low, or absent. If delayed-phase images were available, the enhancement intensity was compared qualitatively between the early and delayed phases, and inspired by the dynamic contrast enhancement curves used in MRI, enhancement patterns were classified as type 1 (persistent) if there was an increase between the two phases, type 2 (plateau) if the signal remained stable, and type 3 (washout) if there was a decrease. The longest length of the calcifications and enhancing area was measured in a single plane using the projection where they appeared largest. For patients who did not undergo pathological sampling and/or were placed under follow-up, particularly those in whom the level of suspicion was lowered based on other imaging modalities (USG or MRI) and who declined histopathological sampling, prior and follow-up examinations were reviewed for dimensional changes. Cases classified as benign were required to demonstrate stability for ≥2 years.

Pathological analysis

The pathology reports of patients who underwent sampling were reviewed, and detailed pathological diagnosis was recorded. All sampled patients underwent tru-cut biopsy, and surgical specimen data were also available for a subset of these patients. The histopathological grades of *in situ* lesions were documented. In cases with surgical specimens, the longest tumor dimension and pathological diagnosis were included in the dataset. Additionally, the Ki-67 index was recorded for invasive carcinomas.

Statistical analysis

Statistical analyses were conducted using IBM Corp., Armonk, NY, USA SPSS Statistics version 27.0. Frequency tables and descriptive statistics were used for data interpretation. Parametric methods were applied for measurement values that followed a normal distribution. Specifically, analysis of variance (F-test) was used to compare measurement values among three or more independent groups, whereas the paired sample t-test (t-value) was used for comparisons between two dependent groups. For measurement values that did not follow a normal distribution, non-parametric methods were applied. In this context, the Wilcoxon test (*Z*-value) was used to compare two dependent groups. The Pearson chi-squared (χ^2) test was employed to assess relationships between two categorical variables. For correlations between two normally distributed quantitative variables, Pearson correlation and Bland–Altman plots were used, whereas Spearman's correlation coefficient was applied if at least one variable did not follow a normal distribution. To determine risk factors influencing malignancy risk, binary logistic regression analysis using the backward likelihood ratio model was performed. A *P* value of <0.05 was considered indicative of a statistically significant relationship.

Results

The study included a total of 132 lesions. All 114 patients were women (age range, 25–79 years, average 48.4 years). A total of 36 lesions were considered benign due to their stability for at least 2 years and the absence of additional malignant features. A total of 42 lesions were pathologically benign, and 54 were malignant. Among the malignant lesions, 25 were invasive carcinomas, and 29 were *in situ* (Table 1).

In low-energy images, where morphology and distribution were evaluated together, sensitivity was found to be 72.2%, specificity 62.8%, positive predictive value (PPV) 57.3%, negative predictive value (NPV) 76.5%, and accuracy 66.7%; 23.4% of the low-, 52.5% of the intermediate-, and 88.9% of the high-risk group were malignant.

In the evaluation based on contrast enhancement intensity in recombined images, of the lesions without contrast enhance-

Table 1. Pathological distributions

	n	Total
Pathologically malignant		
Invasive ductal carcinoma	22	**54**
Invasive lobular carcinoma	3	
Ductal carcinoma *in situ*	27	
Lobular carcinoma *in situ*	2	
Pathologically benign		**42**
Benign with atypia		
Atypical ductal hyperplasia	3	
Flat epithelial atypia	1	
Apocrine atypia	1	
Benign without atypia		
Papilloma	4	
Fibrocystic changes	4	
Fat necrosis	3	
Fibroadenomatous changes	2	
Ductal hyperplasia without atypia	9	
Apocrine metaplasia	4	
Adenosis	5	
Non-specific connective tissue	2	
Inflammatory process	3	
Ductal ectasia	1	
Classified as benign		**36**
		132

ment, 97.4% (37/38) were benign. In cases with low contrast enhancement, 72.2% (26/36) were benign. As an example, Figure 2 illustrates this finding. In contrast, moderate and high enhancement intensities were more frequently associated with malignancy, with 61.9% (13/21) and 81.1% (30/37) of the lesions being malignant, respectively. Among the non-enhancing lesions, only one was malignant. In the primary analysis, where non-enhancing and low-intensity enhancing lesions were grouped as benign and moderate and high-intensity enhancing lesions were grouped as malignant, a statistically significant relationship was found between pathology and contrast enhancement intensity categories ($P < 0.001$). In the secondary analysis, where non-enhancing lesions were classified as benign and any degree of contrast enhancement was classified as malignant, a statistically significant relationship was found between the categories ($P < 0.001$) (Table 2).

In the primary analysis, the sensitivity was 79.6%, specificity was 80.8%, PPV was 74.1%, NPV was 85.1%, and accuracy was 80.3%. In the secondary analysis, the sensitivity increased to 98.2%, but the specificity decreased to 47.4%. The PPV was 56.4%, NPV was 97.4%, and accuracy was 68.2%. In the receiver operating characteristics analysis, the area under the curve (AUC) for low-energy images was 0.67, whereas the AUC for recombined images was 0.80 in the primary analysis and 0.73 in the secondary analysis (Figure 3). According to regression analysis, malignancy risk is 14 times higher for low- 60 times for moderate-, and 159 times for high-intensity contrast than for non-enhancing lesions.

In the analysis evaluating the relationship between contrast enhancement intensity and invasiveness, 25 lesions (46.3%) were invasive, and 29 (53.7%) were *in situ* carcinoma. No statistically significant association was found between these two parameters ($P = 0.698$). Of the 29 *in situ* carcinomas, 2 were lobular carcinoma *in situ* and 27 were DCIS. All but one DCIS showed enhancement. Among the 26 DCIS lesions, 5 were low-grade (2 low, 2 moderate, 1 high enhancement intensity), 9 intermediate-grade (1 low, 4 moderate, 4 high enhancement intensity), and 12 high-grade (2 low, 2 moderate, 8 high enhancement intensity). No statistically significant association was found between the histopathological grade of DCIS and contrast enhancement intensity ($P = 0.336$).

Out of 132 lesions, 17 (12.9%) exhibited a homogeneous contrast enhancement pattern (10 benign, 7 malignant), 70 (53%) showed a heterogeneous pattern (25 benign, 45 malignant), and 7 (5.3%) demonstrated a clumped pattern (6 benign, 1 malignant). In 38 lesions (28.8%), no contrast enhancement was observed. Among the 54 malig-

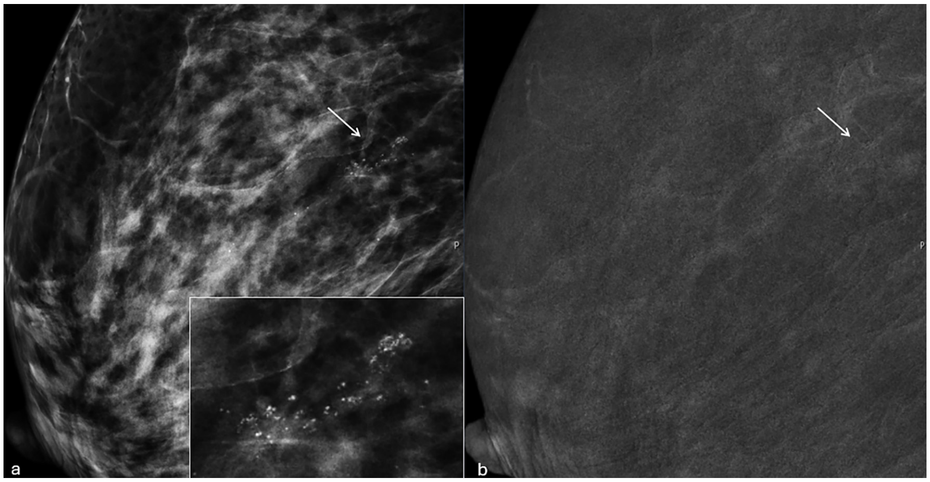


Figure 2. In the low-energy mediolateral oblique image **(a)** of a 53-year-old female patient, suspicious calcifications with a fine pleomorphic morphology and linear distribution are observed. In the recombined image **(b)**, low-intensity enhancement extending beyond the calcification site is noted. The pathological diagnosis was benign.

Table 2. Primary and secondary analyses of contrast enhancement intensities and pathological outcomes

Contrast enhancement intensity	Benign (n = 78)		Malignant (n = 54)		Pearson's chi-squared (χ^2) test *P* value
	n	%	n	%	
Primary analysis					
Favoring benign (no + low enhancement)	63	80.8	11	20.4	χ^2 = 47,256 **$P < 0.001$**
Favoring malignancy (moderate + high enhancement)	15	19.2	43	79.6	
Secondary analysis					
Favoring benign (no enhancement)	37	47.4	1	1.9	χ^2 = 32,343 **$P < 0.001$**
Favoring malignancy (all types of enhancement)	41	52.6	53	98.1	

Pearson's chi-squared (χ^2) test with cross-tabulations was used to evaluate the association between two categorical variables.

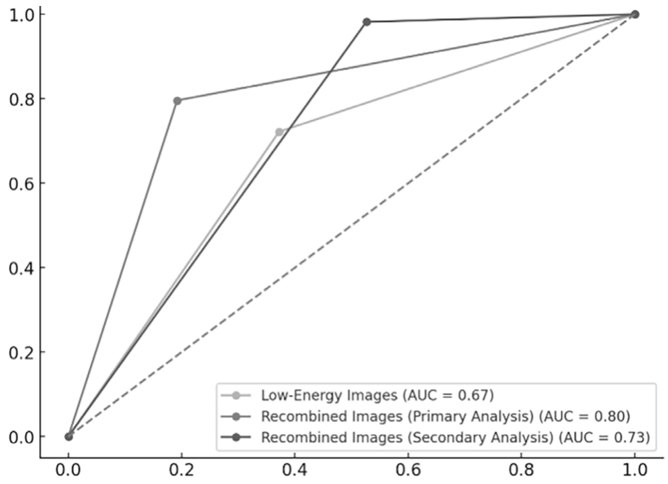


Figure 3. Receiver operating characteristic curves for both images and analyses. AUC, area under the curve.

nant lesions, 83% exhibited heterogeneous enhancement, whereas only 13% showed homogeneous enhancement. The statistical analysis yielded $P = 0.018$, indicating that heterogeneous contrast enhancement is significantly associated with malignancy.

In 22 cases (16.6%), partial contrast enhancement was observed in the calcification region, and in another 22 cases (16.6%), enhancement completely overlapped with this tissue. In 50 cases (37.9%), contrast enhancement extended beyond the calcified tissue. Among the 21 lesions that underwent excision in the center and did not receive neoadjuvant therapy before surgery, with surgical specimen data, 13 showed enhancement extending beyond the calcification, 4 had partial enhancement, and 3 exhibited complete overlap. One non-enhancing case was excluded. Figure 4 presents a malignant case with contrast enhancement extending beyond the calcifications; in this case, the lesion length measured on the recombined image closely matched the pathological excision size. Correlation analysis revealed a strong and statistically significant relationship between lesion length in both low-energy ($r = 0.733$, $P < 0.001$) and recombined images ($r = 0.879$, $P < 0.001$) and pathological length. The Wilcoxon signed-rank test revealed no statistically significant difference between the methods (Table 3). Bland–Altman analysis showed that recombined images tend to overestimate lesion length by 4.45 mm, whereas low-energy images tend to underestimate it by 4.75 mm (Figure 5).

In the 13 patients who exhibited contrast enhancement extending beyond the calcifications and constituted the majority of the pathologically sampled group, correlation analysis showed a strong positive and statistically significant relationship between pathological length and both calcification ($r = 0.822$, $P < 0.001$) and contrast-enhanced area length ($r = 0.825$, $P < 0.001$), similar to the overall group. This group was important for addressing the question of whether the pathological extent exceeded what was observed on low-energy images. The paired sample test revealed a significant difference between pathological and calcification length ($P = 0.012$), indicating that calcification length alone may underestimate the true extent of the lesion.

In the analysis of dynamic contrast-enhancement types, 8 lesions were excluded due to the absence of late-phase images, and 28 lesions were excluded due to a lack of contrast enhancement. Among the 46 benign lesions, 38 (82.6%) exhibited a type 1 enhancement. Among the 50 malignant lesions, 28 (56%) showed a type 1, 16 (32%) exhibited a type 2, and 6 (12%) demonstrated a type 3. Figure 6 presents a case demonstrating high-intensity contrast enhancement with a type 1 dynamic enhancement pattern, which was ultimately diagnosed as benign. The majority of lesions with type 2 (72.7%) and type 3 (75%) enhancement were malignant. Figure 7 illustrates an example of type 3 contrast enhancement with a washout pattern, corresponding to a malignant case. A statistically significant association was found between type and malignancy/benignity ($P = 0.019$). However, no statistically significant association was found between enhancement types and the distinction between invasive and *in situ* carcinoma ($P = 0.331$) (Table 4).

In the 25 invasive carcinoma cases with available Ki-67 index data, no statistically significant linear relationship was found between Ki-67 (%) values and contrast enhancement intensity ($P = 0.057$). The median (interquartile range) Ki-67 values were 20.0 (12.3), 8.5 (19.5), and 45.0 (60.0) for lesions with low-, moderate-, and high-intensity enhancement, respectively.

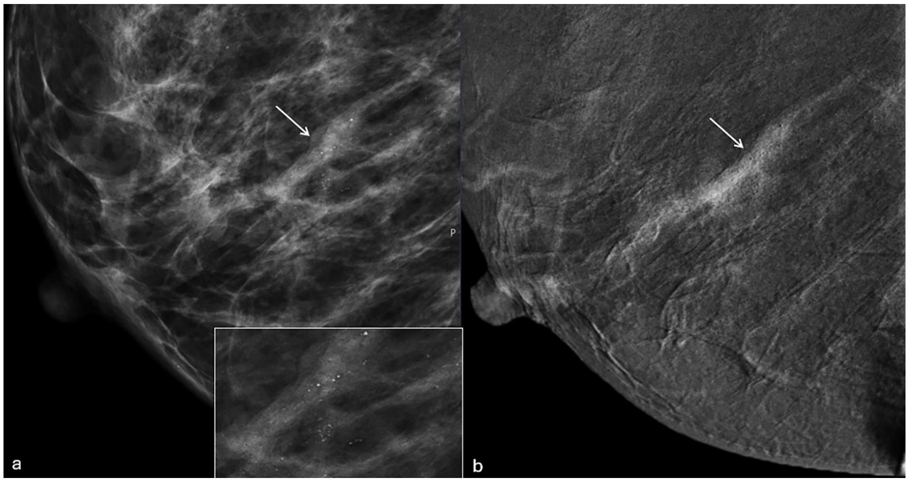


Figure 4. In the low-energy mediolateral oblique image **(a)** of a 42-year-old female patient, suspicious calcifications with a fine pleomorphic morphology and segmental distribution are observed. In the recombined image **(b)**, high-intensity heterogeneous enhancement extends beyond the calcification site, reaching the nipple. The pathological diagnosis was ductal carcinoma *in situ*, and the measured pathological length closely matched the lesion length observed in the recombined images.

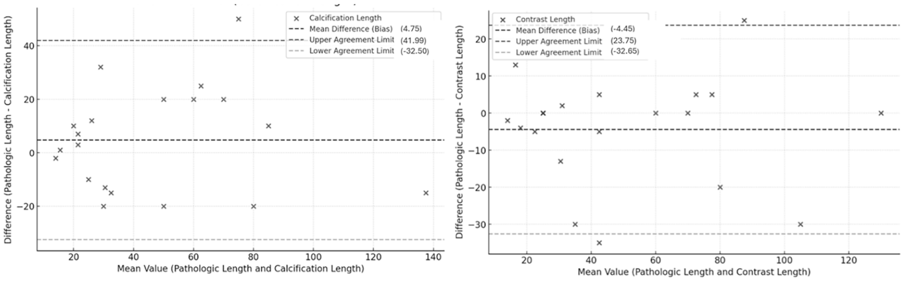


Figure 5. Bland–Altman plots visualizing the discrepancies between the pathological and imaging length. The left one shows results for low-energy images, and the right one is for recombined images.

Table 3. Comparison of pathological excision size and imaging-based lesion size (Wilcoxon and Spearman Analyses)

	Length (mm)		Wilcoxon signed-rank test *P* value	Spearman's ρ (rho) *P* value
		Median (IQR)		
Pathological length	49.15 ± 33.22	36.0 (50.5)		
Microcalcification length	44.40 ± 32.31	40.0 (39.0)	Z = −0.880 *P* = 0.379	r = 0.733 ***P* < 0.001**
Enhancement length	53.60 ± 33.39	47.5 (48.8)	Z = 1.084 *P* = 0.278	r = 0.879 ***P* < 0.001**

The Wilcoxon signed-rank test was used to assess paired differences between measurements. Spearman's correlation coefficient (ρ) was calculated to evaluate the relationship between imaging and pathological sizes. IQR, interquartile range; SD, standard deviation.

Discussion

In the analysis of groups categorized based on morphology and distribution, the findings indicate that although low-energy images (i.e., mammography) demonstrate good performance according to our classification, they may still lead to missed malignancy in 23.4% of patients with low-suspicion features. Therefore, according to the BI-RADS atlas, pathological sampling is recommended, even for patients we classified as having low suspicion. However, this approach resulted in unnecessary biopsy recommendations in 47.5% of the intermediate suspicion group and 11.1% of the high suspicion group. These findings on standard mammography support the need for functional assessment of the tissue using CEM, an alternative imaging modality.

Our study identified only one malignant case that did not exhibit contrast enhancement that was diagnosed as intermediate-grade DCIS. We attributed the absence of enhancement to the *in situ* nature and a low Ki-67 of only 1%. As the contrast enhancement intensity increased, the malignancy rate was observed to rise across groups. In the secondary analysis, a much higher sensitivity was achieved than with the primary analysis and low-energy images. If contrast enhancement presence alone is considered a malignancy indicator, CEM achieves 98.2% sensitivity, almost eliminating the risk of missing malignant cases. However, in the secondary analysis, specificity decreased, leading to a higher false-positive rate. On the other hand, the increase in NPV indicates that in the absence of enhancement, CEM can rule out malignancy with 97.4% accuracy. The primary analysis provided a more balanced and consistent performance, albeit with slightly lower sensitivity. From a breast cancer diagnostic perspective, missing a diagnosis can have fatal consequences, so avoiding false negatives should be prioritized. Although false positives may lead to unnecessary biopsies, the high NPV suggests that some unnecessary procedures can be avoided. Low-intensity contrast enhancement may have been confused with background enhancement in some cases, potentially affecting our findings. Additionally, the qualitative nature of our study presents another limitation: the lack of sharply defined classification boundaries.

Among various studies, where the presence of enhancement is considered a malignant feature, our secondary analysis demonstrated the highest sensitivity and NPV.[8-11]

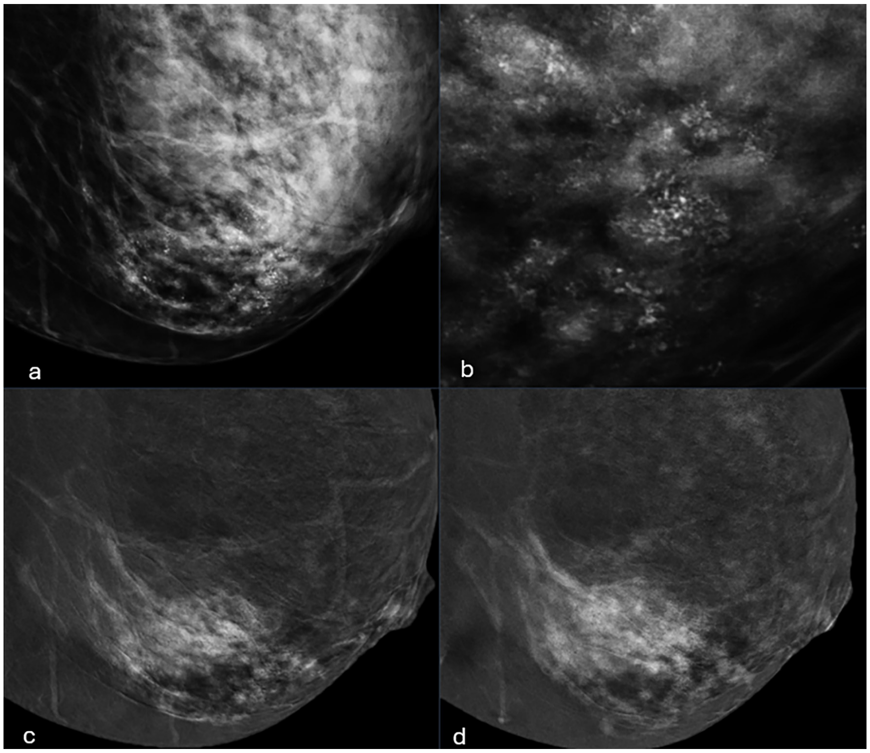


Figure 6. In the low-energy mediolateral oblique (MLO) image of a 44-year-old female patient **(a)** and the corresponding magnified view **(b)**, suspicious calcifications with coarse heterogeneous morphology and segmental distribution are observed. In the recombined early-phase **(c)** and recombined delayed-phase **(d)** MLO images, heterogeneous, high-intensity contrast enhancement overlapping the calcification area is noted, with an increase in intensity from the early to the delayed phase, consistent with a type 1 (persistent) enhancement pattern. The pathological diagnosis was papillomatous changes with benign characteristics.

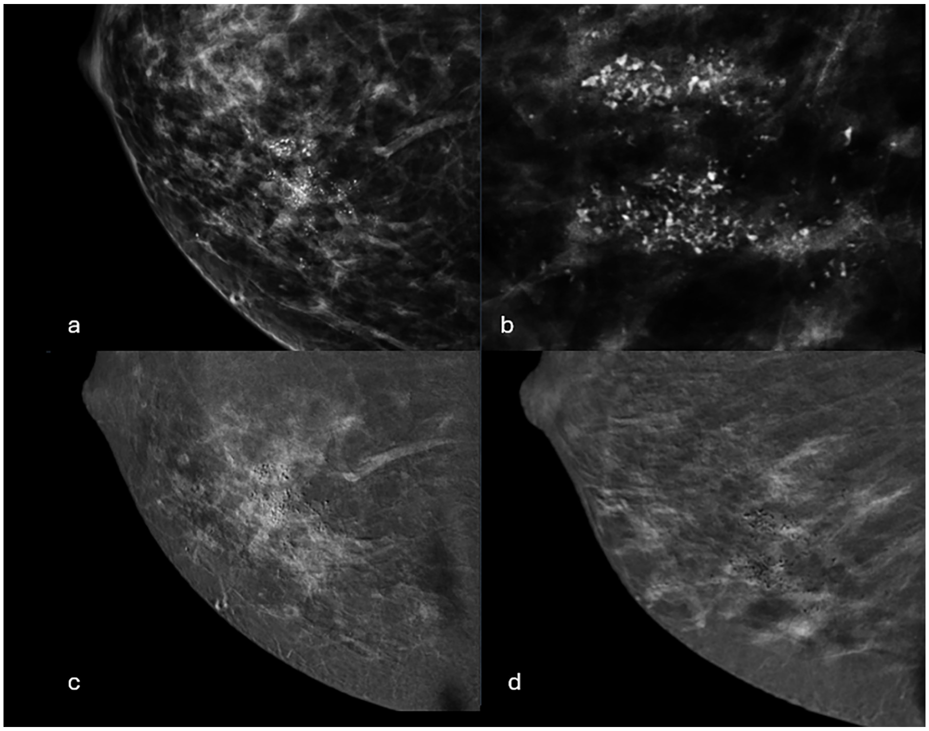


Figure 7. In the low-energy mediolateral oblique (MLO) image **(a)** and its magnified view **(b)** of a 48-year-old female patient, suspicious calcifications with a coarse heterogeneous morphology and grouped distribution are observed. In the recombined early-phase MLO image **(c)** and late-phase MLO image **(d)**, high-intensity heterogeneous enhancement extending beyond the calcification site is noted, with more intense enhancement in the early phase, followed by a decrease in the late phase, which was a type 3 (washout) pattern. The pathological diagnosis was intermediate-grade ductal carcinoma *in situ*.

Table 4. Distribution of dynamic enhancement types in benign and malignant and invasive and *in situ* Lesions

Dynamic enhancement type	Benign (n = 46)		Malignant (n = 50)		Pearson's chi-squared (χ^2) test *P* value
	n	%	n	%	
1	38	82.6	28	56.0	χ^2 = 7,907 *P* = **0.019**
2	6	13	16	32.0	
3	2	4.4	6	12.0	
	Invasive malignant (n = 24)		***In situ* malignant (n = 26)**		
	n	**%**	**n**	**%**	
1	11	45.8	17	65.4	χ^2 = 2,210 *P* = 0.331
2	10	41.7	6	23.1	
3	3	12.5	3	11.5	

Pearson's chi-squared (χ^2) test with cross-tabulations was used to evaluate the association between two categorical variables.

In the majority of studies, NPV was consistently found to be high. This supports the role of CEM as a valuable imaging tool in reducing unnecessary biopsies for suspicious calcifications, which are relatively more challenging to sample than mass lesions.

In the study by Nicosia et al.[11], a primary analysis similar to ours was also conducted. Their findings reported a sensitivity of 53.3%, specificity of 95.8%, PPV of 84.2%, NPV of 82.9%, and an accuracy of 83.2%. Similar to our study, in the secondary analysis, sensitivity increased but specificity decreased. In another study among 24 cases, all malignant lesions exhibited contrast enhancement at varying intensities. All *in situ* carcinomas and one benign lesion showed low-intensity enhancement, whereas the remaining benign lesions showed no enhancement.[12] Quantitative studies that include various breast lesions, rather than being specific to suspicious calcifications, have also identified a trend of lower contrast enhancement intensity in benign lesions and higher in malignant lesions.[13-17]

In studies evaluating suspicious calcifications not associated with a mass with MRI, a 2016 meta-analysis reported that studies assessing the presence of enhancement found MRI to have an average sensitivity of 92%, specificity of 75%, PPV of 78%, and NPV of 93%.[18] In a study by Taskin et al.[19], which included 444 cases of suspicious calcifications detected on mammography, MRI demonstrated a sensitivity of 95.2%, specificity of 40.2%, PPV of 49.2%, and NPV of 93.3%. These findings are highly comparable with our secondary analysis. The effectiveness of MRI in assessing tissue function is undeniable, but the primary advantage of CEM over MRI is its ability to overlay functional information directly onto the low-energy image, which already provides a detailed assessment of calcification morphology and distribution. MRI does not visualize calcifications and only allows for the interpretation of tissue function. As a result, evaluation requires additional digital mammography, necessitating the integration of two separate imaging modalities. Moreover, differences in patient positioning–prone for MRI and upright for mammography–can make three-dimensional localization and anatomical correlation between calcifications and enhancement more challenging. Additionally, MRI's longer acquisition time, limited accessibility, lower patient tolerance, and high cost further highlight the need for an alternative imaging method in this patient group.[20]

In the comparison of both images, the sensitivity, NPV, accuracy, and AUC of recombined images alone were found to be higher than those of low-energy images in both analyses. Although these values were compared separately in our study, CEM inherently combines both imaging types. Therefore, it can be anticipated that assessing findings together would significantly enhance the overall diagnostic performance.

Studies in the literature evaluating specifically suspicious calcifications have typically assessed the distinction between invasive and *in situ* carcinoma or the grading of DCIS based solely on the presence or absence of contrast enhancement.[9,10,21,22] In our study, however, all but 1 malignant lesion exhibited contrast enhancement. In one study including 15 lesions, all DCIS cases exhibited low contrast enhancement; however, the variation in enhancement intensity among invasive ductal carcinoma cases introduces inconsistency, limiting its reliability in clearly differentiating these two entities.[12] Our results suggest that there is a predominance of high-intensity contrast enhancement in invasive carcinomas; however, since half of the *in situ* cases also exhibited high-intensity enhancement, this prevented us from obtaining a statistically significant distinction. Similarly, although no statistically significant relationship was found between contrast enhancement intensity and the histopathological grade of DCIS, 66.6% of high-grade cases demonstrated high-intensity enhancement. Although the relationship was not statistically significant, a trend toward higher enhancement in high-grade lesions was observed. Larger-scale studies are needed to obtain more definitive conclusions.

When evaluating contrast enhancement patterns, it was observed that heterogeneous enhancement was more prominent in malignant cases, whereas more than half of the homogeneously enhancing lesions were benign. These results are consistent with the existing literature.[16,17] Clumped enhancement did not provide meaningful data due to the limited number of cases in this group.

Recombined images outperformed low-energy images, exhibiting a smaller mean difference and a higher correlation coefficient. The group with contrast enhancement extending beyond the calcification area was particularly important in evaluating our hypothesis: "Could the pathological tissue be larger than what is observed in low-energy images?" In this subgroup, the correlation coefficient for recombined images was higher than for low-energy images, indicating a stronger association. Additionally, a statistically significant difference was observed between low-energy images and pathological length. In the study by Cheung et al.[8], low-energy images overestimated pathological length by an average of 4.2 mm, whereas recombined images showed a smaller mean difference of 0.5 mm, indicating a more accurate length estimation. However, Houben et al.[9] reported a mean difference of 0.3 mm for low-energy and 4.5 mm for recombined images, although the correlation coefficients in their study were comparable with those in our study. To our knowledge, there is no other study specifically comparing the lesion length between CEM and pathology in this group. Most of the studies mentioned above, as well as other studies including various breast lesions in the literature, share a common finding: contrast-enhanced images tend to overestimate lesion length, which is consistent with our study's results.[12,23-25] In some studies, the mean difference between pathology and low-energy images was smaller, but in all cases, recombined images exhibited a higher correlation coefficient, indicating a stronger association between recombined images and pathological length. The overestimation in CEM can be attributed to contrast agent extravasation into the surrounding tissue due to increased vascular permeability and compression applied during imaging, which may further disperse tissues and enhance lesion dimensions. Some studies have identified that low-energy images may also overestimate lesion length, which could be attributed to breast compression.[8,9,23] However, contrary to this, in our study, low-energy images tended to underestimate lesion length. This finding supports our hypothesis that relying solely on mammographic images in breast-conserving surgery may increase the risk of positive surgical margins.

To our knowledge, no prior study has specifically analyzed dynamic enhancement patterns in suspicious calcifications without a mass. A significant association was found between enhancement types and malignancy. Compared with a previous study that evaluated various breast lesions, the main difference in our findings is that approximately half of our malignant lesions also exhibited a type 1 enhancement.[15] We attribute this difference to the higher metabolic activity of mass-forming lesions in previous studies compared with the suspicious calcifications not associated with a mass in our study, which may result in faster contrast uptake and washout kinetics.

The other parameter we evaluated was the Ki-67 index, a marker of tumor cell proliferation, which has been associated with higher relapse rates and lower survival.[26] In recent years, the potential impact of pre-treatment prognosis prediction on therapy selection has led to increasing interest in assessing whether Ki-67 levels–and consequently prognosis–can be inferred from imaging findings. To our knowledge, the only two studies in the literature investigating CEM and Ki-67 were conducted by Depretto et al.[27] Their findings showed that most non-enhancing lesions had a Ki-67 index <20%, whereas more than half of the malignant, contrast-enhancing calcifications had a Ki-67 index >20%.[10] In our study, only invasive carcinomas were analyzed. Although the median Ki-67 index in moderately enhancing lesions was unexpectedly lower than in low-enhancing lesions, disrupting a clear linear relationship, the *P* value was very close to statistical significance (0.05). The high-intensity enhancement group had the highest median Ki-67 values. Although answering the question of whether higher contrast enhancement at diagnosis could indirectly indicate a worse prognostic factor requires larger studies with broader patient populations, our findings suggest promising potential for further investigation in this area.

Our study was conducted retrospectively, which presents certain limitations. Some benign lesions were considered benign without histopathological confirmation. Even if stability is observed over a 2-year period, considering such lesions as benign–particularly in cases of *in situ* carcinomas, which may progress slowly–may not be a controversial approach in clinical practice. Retrospective evaluation reveals that in the majority of these patients, the suspicion of malignancy was significantly reduced through additional imaging methods and clinical experience, and the decision for follow-up without biopsy was made in accordance with patient preference. Additionally, surgical tumor length data of some patients were not available. For comparisons with pathology, only the longest dimension in a single imaging plane was used. Other measurements in our study were qualitative. Furthermore, this was a single-center study, which may limit the generalizability of the findings.

In conclusion, CEM offers a significant advantage over MRI in the evaluation of suspicious calcifications, as it allows for the simultaneous assessment of both calcification morphology and distribution as well as tissue functionality. This dual capability, which combines the mammographic equivalent low-energy images with MRI-like recombined images, enhances diagnostic accuracy beyond mammography alone and provides valuable insights into tumor extent, potentially aiding surgical decision-making. Through various analyses, we found that the presence of enhancement demonstrated high sensitivity for malignancy, and increasing enhancement intensity correlated with a higher malignancy risk. Additionally, CEM outperformed mammography alone in detecting malignancy and could assist in surgical planning by better reflecting disease extent. However, despite its potential, contrast enhancement intensity did not achieve statistical significance for distinguishing invasive from *in situ* carcinoma or for grading DCIS. Dynamic time-intensity analysis, similar to MRI kinetics, may aid lesion characterization, and the potential role of enhancement intensity as a prognostic factor warrants further investigation. Given the limited number of studies focusing on this subgroup, larger-scale, multicenter research is needed for more robust conclusions. CEM remains a promising imaging modality for suspicious calcifications, a category where radiologists have yet to reach a consensus on standard practice, routine approaches vary, and pathological sampling remains relatively challenging. It has the potential to provide more reliable and clinically useful results, making it a valuable candidate for wider implementation.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209-249. [Crossref]
2. Lorena Arancibia Hernández P, Taub Estrada T, López Pizarro A, Lorena Díaz Cisternas Carla Sáez Tapia M. Breast calcifications: description and classification according to BI-RADS 5th edition. Published online 2016. [Crossref]
3. Sickles E, D'Orsi CJ, Bassett LW, et al. ACR BI-RADS® Mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA: *American College of Radiology*; 2013. [Crossref]
4. Horvat JV, Keating DM, Rodrigues-Duarte H, Morris EA, Mango VL. Calcifications at digital breast tomosynthesis: imaging features and biopsy techniques. *RadioGraphics.* 2019;39(2):307-318. [Crossref]
5. Patel BK, Lobbes MBI, Lewin J. Contrast enhanced spectral mammography: a review. *Semin Ultrasound CT MR.* 2018;39(1):70-79. [Crossref]
6. Kuhl C. The current status of breast MR imaging Part I. Choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice. *Radiology.* 2007;244(2):356-378. [Crossref]
7. Lalji UC, Jeukens CRLPN, Houben I, et al. Evaluation of low-energy contrast-enhanced spectral mammography images by comparing them to full-field digital mammography using EUREF image quality criteria. *Eur Radiol.* 2015;25(10):2813-2820. [Crossref]
8. Cheung YC, Tsai HP, Lo YF, Ueng SH, Huang PC, Chen SC. Clinical utility of dual-energy contrast-enhanced spectral mammography

for breast microcalcifications without associated mass: a preliminary analysis. *Eur Radiol*. 2016;26(4):1082-1089. [Crossref]

9. Houben IP, Vanwetswinkel S, Kalia V, et al. Contrast-enhanced spectral mammography in the evaluation of breast suspicious calcifications: diagnostic accuracy and impact on surgical management. *Acta Radiol*. 2019;60(9):1110-1117. [Crossref]
10. Depretto C, D'Ascoli E, Della Pepa G, et al. Assessing the malignancy of suspicious breast microcalcifications: the role of contrast enhanced mammography. *Radiol Med*. 2024;129(6):855-863. [Crossref]
11. Nicosia L, Bozzini AC, Signorelli G, et al. Contrast-enhanced spectral mammography in the evaluation of breast microcalcifications: controversies and diagnostic management. *Healthcare*. 2023;11(4):511. [Crossref]
12. Łuczyńska E, Niemiec J, Hendrick E, et al. Degree of enhancement on contrast enhanced spectral mammography (CESM) and Lesion type on mammography (MG): comparison based on histological results. *Med Sci Monit*. 2016;22:3886-3893. [Crossref]
13. Rudnicki W, Heinze S, Piegza T, Pawlak M, Kojs Z, Łuczyńska E. Correlation between enhancement intensity in contrast enhancement spectral mammography and types of kinetic curves in magnetic resonance imaging. *Med Sci Monit*. 2020;26:920742. [Crossref]
14. Rudnicki W, Heinze S, Niemiec J, et al. Correlation between quantitative assessment of contrast enhancement in contrast-enhanced spectral mammography (CESM) and histopathology-preliminary results. *Eur Radiol*. 2019;29(11):6220-6226. [Crossref]
15. Deng CY, Juan YH, Cheung YC, et al. Quantitative analysis of enhanced malignant and benign lesions on contrast-enhanced spectral mammography. *Br J Radiol*. 2018;91(1086). [Crossref]
16. Mohamed Kamal R, Hussien Helal M, Wessam R, Mahmoud Mansour S, Godda I, Alieldin N. Contrast-enhanced spectral mammography: Impact of the qualitative morphology descriptors on the diagnosis of breast lesions. *Eur J Radiol*. 2015;84(6):1049-1055. [Crossref]
17. Chi X, Zhang L, Xing D, Gong P, Chen Q, Lv Y. Diagnostic value of the enhancement intensity and enhancement pattern of CESM to benign and malignant breast lesions. *Medicine*. 2020;99(37):e22097. [Crossref]
18. Bennani-Baiti B, Baltzer PA. MR Imaging for diagnosis of malignancy in mammographic microcalcifications: a systematic review and meta-analysis. *Radiology*. 2017;283(3):692-701. [Crossref]
19. 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]
20. Hobbs MM, Taylor DB, Buzynski S, Peake RE. Contrast-enhanced spectral mammography (CESM) and contrast enhanced MRI (CEMRI): patient preferences and tolerance. *J Med Imaging Radiat Oncol*. 2015;59(3):300-305. [Crossref]
21. Cheung YC, Juan YH, Lin YC, et al. Dual-energy contrast-enhanced spectral mammography: enhancement analysis on BI-RADS 4 non-mass microcalcifications in screened women. *PLoS One*. 2016;11(9):0162740. [Crossref]
22. Shetat OMM, Moustafa AFI, Zaitoon S, Fahim MII, Mohamed G, Gomaa MM. Added value of contrast-enhanced spectral mammogram in assessment of suspicious microcalcification and grading of DCIS. *Egypt J Radiol Nucl Med*. 2021;52(1):186. [Crossref]
23. Luczyńska E, Heinze-Paluchowska S, Dyczek S, Blecharz P, Rys J, Reinfuss M. Contrast-enhanced spectral mammography: comparison with conventional mammography and histopathology in 152 women. *Korean J Radiol*. 2014;15(6):689-696. [Crossref]
24. Patel BK, Garza SA, Eversman S, Lopez-Alvarez Y, Kosiorek H, Pockaj BA. Assessing tumor extent on contrast-enhanced spectral mammography versus full-field digital mammography and ultrasound. *Clin Imaging*. 2017;46:78-84. [Crossref]
25. Fallenberg EM, Dromain C, Diekmann F, et al. Contrast-enhanced spectral mammography: does mammography provide additional clinical benefits or can some radiation exposure be avoided? *Breast Cancer Res Treat*. 2014;146(2):371-381. [Crossref]
26. Petrelli F, Viale G, Cabiddu M, Barni S. Prognostic value of different cut-off levels of Ki-67 in breast cancer: a systematic review and meta-analysis of 64,196 patients. *Breast Cancer Res Treat*. 2015;153(3):477-491. [Crossref]
27. Depretto C, Borelli A, Liguori A, et al. Contrast-enhanced mammography in the evaluation of breast calcifications: preliminary experience. *Tumori*. 2020;106(6):491-496. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253376

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Magnetic resonance imaging-based artificial intelligence model predicts neoadjuvant therapy response in triple-negative breast cancer

Raşit Eren Büyüktoka[1]
Zehra Hilal Adıbelli[2,9]
Murat Sürücü[3]
Özlem Özdemir[4]
Yalçın İşler[5]
Ali Murat Koç[6]
Aslı Dilara Büyüktoka[2]
Demet Kocatepe Çavdar[7]
Özge Aslan[8]
Serhat Değer[2]
Ayşenur Oktay[8]

[1]İzmir Foça State Hospital, Clinic of Radiology, İzmir, Türkiye

[2]University of Health Sciences Türkiye, İzmir City Hospital, Department of Radiology, İzmir, Türkiye

[3]Burdur Mehmet Akif Ersoy University Faculty of Bucak Computer and Informatics, Department of Software Engineering, Burdur, Türkiye

[4]University of Health Sciences Türkiye, İzmir City Hospital, Department of Medical Oncology, İzmir, Türkiye

[5]Alanya Alaaddin Keykubat University Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Antalya, Türkiye

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

[7]University of Health Sciences Türkiye, İzmir City Hospital, Department of Pathology, İzmir, Türkiye

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

[9]University of Health Sciences Türkiye, İzmir Faculty of Medicine, İzmir, Türkiye

Corresponding author: Raşit Eren Büyüktoka

E-mail: rasiterenbuyuktoka@hotmail.com

Received 04 June 2025; revision requested 10 August 2025; last revision received 01 September 2025; accepted 05 October 2025.

Epub: 27.10.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253376

PURPOSE

Triple-negative breast cancer (TNBC) is an aggressive subtype of breast cancer with limited treatment options and poorer overall survival than other subtypes. Neoadjuvant chemotherapy (NACT) is often used to reduce tumor size and improve surgical outcomes. However, predicting patients' response to NACT remains challenging, and non-responding patients risk unnecessary chemotoxicity. This study aimed to develop a deep learning-based artificial intelligence (AI) model using pre-treatment magnetic resonance imaging (MRI) to predict pathological complete response (pCR) in patients with TNBC undergoing NACT.

METHODS

This retrospective, double-centered study included 49 lesions from 43 patients with TNBC. Data from MRI, including T2-weighted, T1-weighted, and diffusion-weighted imaging, were segmented and processed to train a residual convolutional neural network model.

RESULTS

The AI model achieved an accuracy of 0.82 and an area under the receiver operating characteristic curve of 0.75 in differentiating pCR from non-pCR cases. The model's performance was validated through intra- and inter-reader agreement metrics, with Dice similarity coefficients ranging from 0.821 to 0.915.

CONCLUSION

Our results demonstrate that AI models can effectively predict NACT responses in patients with TNBC using only pre-treatment MRI data.

CLINICAL SIGNIFICANCE

This proof-of-concept study supports the potential for AI-based tools to aid clinical decision-making and reduce the risks associated with ineffective therapies. Future research with larger datasets and additional imaging modalities is needed to improve model generalizability and clinical applicability.

KEYWORDS

Breast cancer, artificial intelligence, neoadjuvant chemotherapy, magnetic resonance imaging, residual convolutional neural network

Breast cancer (BC) is a common health problem worldwide and remains the most common cancer type among women. Despite its high incidence, mortality rates have consistently decreased over the last decades due to technological advancements in imaging and novel therapeutic options.[1] BC has different subtypes, and each subtype has a different prognosis. It is crucial to evaluate the tumor molecularly to assess the patient's treatment options and clinical outcomes.[2]

Triple-negative BC (TNBC) is characterized by the lack of estrogen receptors, progesterone receptors, and expression of human epidermal growth factor receptor 2. It is the most aggressive subtype and has the least favorable overall survival (OS); it is diagnosed in almost

You may cite this article as: Büyüktoka RE, Adıbelli ZH, Sürücü M, et al. Magnetic resonance imaging-based artificial intelligence model predicts neoadjuvant therapy response in triple-negative breast cancer. *Diagn Interv Radiol*. 2026;32(2):191-199.

15%–20% of all patients with BC. In contrast to other subtypes, TNBC has limited hormonal and target-specific treatment options.[2-4]

Neoadjuvant chemotherapy (NACT) for BC is increasingly used to decrease the tumor volume and to downstage the disease to create a bridge to surgery.[5] Early TNBC is commonly treated with surgery and adjuvant chemotherapy.[6] Furthermore, unresectable and locally advanced TNBC treatment is mainly based on NACT.[6,7] Compared with adjuvant chemotherapy, preoperative systemic therapy for BC has no advantages in disease-free survival or OS.[8,9] However, there is a survival advantage in patients who achieve pathological complete response (pCR) after NACT compared with those with residual disease.[10,11] With NACT becoming the standard treatment, clinicians have focused on patients who do not achieve pCR. This is because patients without pCR show poorer survival outcomes than those with pCR, and post-NACT has been applied to patients without pCR to achieve long-term survival outcomes.[11] Imaging studies and physical exams have provided early response assessments, helping distinguish non-responders. This allows for alternative treatments to overcome resistance, aiming to improve pCR rates and forming the basis for post-neoadjuvant treatment strategies.[12-14]

Assessment of disease stage is mainly based on radiological examinations. Imaging modalities include mammography, ultrasound (US), magnetic resonance imaging (MRI) of the breast, and positron emission tomography/computed tomography.[15] Evaluation of the response after completion of NACT is based on radiological examinations. Mammography and US are routinely used to assess the response to NACT.[16] However, after the initiation of NACT, it is impossible to predict the patient's response status with conventional radiological methods.[17,18]

Main points

- Triple-negative breast cancer (TNBC) is the most aggressive and least common breast cancer subtype.
- Pre-treatment magnetic resonance imaging (MRI) may contain helpful information for artificial intelligence (AI) models to predict neoadjuvant chemotherapy response in advance to individualize treatment.
- Our AI model predicts therapy response in TNBC using pre-treatment MRI data and achieved accuracy of 0.82 and area under the curve of 0.75 in predicting pathological complete response (pCR) compared with non-pCR.

In this proof-of-concept study, we introduce a deep learning-based artificial intelligence (AI) model using pre-treatment MRIs to predict the NACT response status before the initialization of NACT. Convolutional neural networks (CNNs) are artificial neural networks composed of multiple layers, specifically designed to evaluate datasets that contain grid-like (coordinate-based) information such as radiological images.[19-21] We hypothesized that tumor appearances in different MRI sequences, as reflected by different gray-level pixel presentations and tumor features, can be deciphered by a residual CNN-based AI model using pre-treatment MRIs.

Methods

Study design and patient population

Our study was a retrospective double-center study conducted in accordance with the Declaration of Helsinki, and this retrospective study was approved by the University of Health Sciences Türkiye, İzmir Bozyaka Training and Research Hospital Clinical Research Ethics Committee (decision number: 2023/19, date: 08.02.2023). Due to the retrospective design of the study, informed consent was waived by the local ethics committee.

Patients with biopsy-proven TNBC underwent and completed NACT between 2018 and 2023. These patients had pathology data at the time of initial diagnosis and underwent breast MRI before NACT. A flowchart of the patient selection, inclusion, and exclusion criteria is presented in Figure 1.

Magnetic resonance imaging acquisition

MRIs of the patients were acquired at two different centers using 1.5 Tesla MRI units (Magnetom Amira and Symphony, Siemens Healthineers, Erlangen, Germany / Philips Achieva, Philips Medical Systems, Drachten, Netherlands) and a 3-Tesla MRI unit (Magnetom Verio, Siemens Healthineers, Erlangen, Germany). All patients were imaged in the prone position using a breast coil. The MRI sequences included fast spin echo (FSE) T2-weighted images (T2WIs), b800 diffusion-weighted images (DWIs), and fat-suppressed pre- and post-contrast images at 180 seconds, which were used for segmentation. For contrast-enhanced images, 0.1 mmoL/kg of gadobutrol (Gadovist®, Bayer, Germany) or gadoteric acid (Clariscan®, GE Healthcare, Norway) was injected as a rapid bolus, followed by a 10-mL saline flush at 2-mL/s. The 180-second post-contrast images were used to feed the AI algorithm.

Definition of pathological complete response

After completion of NACT, pathological response data from surgical specimens were classified as pCR and non-pCR. Pathological classifications were made according to the Miller-Payne grading system, with Grade 5 classified as a complete response and Grade 4 or below classified as no pCR.[22]

Lesion segmentation

During data collection, the leading researcher (R.E.B.) included 49 lesions from 43 patients based on the inclusion and exclusion criteria. The images were anonymized using local software, all image labels were removed, and new patient numbers were assigned post-anonymization. After anonymization, the FSE-T2WIs, DWIs, and pre- and post-contrast fat-suppressed T1-weighted images (T1WIs) were selected for annotation. The researcher evaluated the images along with pathological data. The lesions were segmented volumetrically in three-dimensional (3D) polygon mode using ITK-SNAP 4.x open-source software in FSE-T2WIs, DWIs, and post-contrast images.

After the initial segmentation, following an interval of at least 1 month, the researcher randomly selected 20% of the lesions from each sequence for re-segmentation to calculate intra-observer agreement using different metrics. Moreover, 20% of the lesions in each sequence were re-segmented by a second radiologist (A.D.B.) with similar experience, and inter-observer agreement was calculated.

Artificial intelligence model

Data preprocessing

Before entering the data into the deep learning network, several preprocessing steps were applied.

• Segmentation: Lesions identified by radiologists were annotated on the imaging sequences.

• Image cropping: Images were cropped to include only the annotated lesions.

• Image scaling: Lesions were resized to a fixed 50 × 50 × 50 scale, with zero padding used for any gaps.

• Normalization: The pixel values of the 3D tumor slices were normalized between 0 and 1.

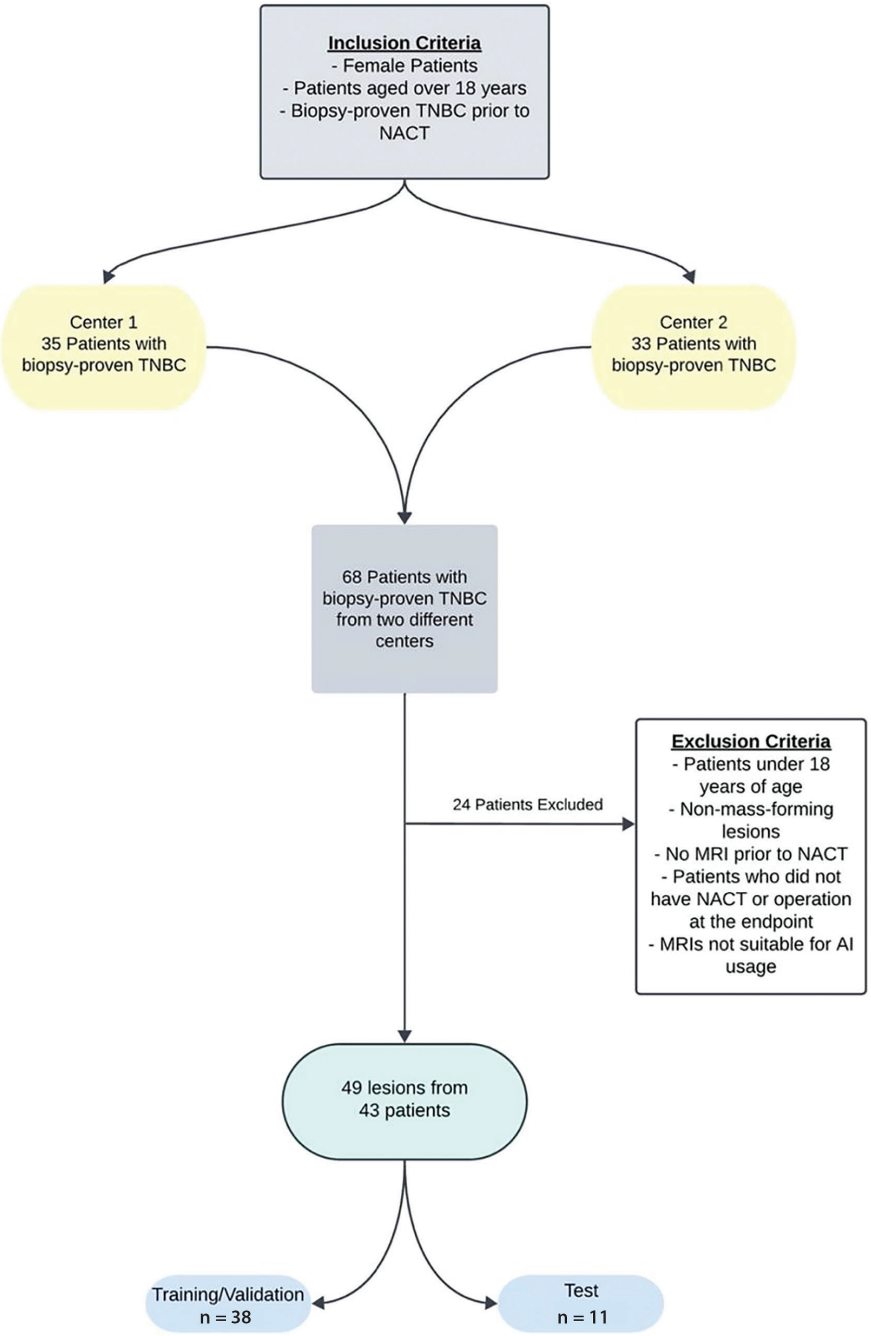

Inclusion Criteria
- Female Patients
- Patients aged over 18 years
- Biopsy-proven TNBC prior to NACT
Center 1
35 Patients with biopsy-proven TNBC
Center 2
33 Patients with biopsy-proven TNBC
68 Patients with biopsy-proven TNBC from two different centers
24 Patients Excluded
Exclusion Criteria
- Patients under 18 years of age
- Non-mass-forming lesions
- No MRI prior to NACT
- Patients who did not have NACT or operation at the endpoint
- MRIs not suitable for AI usage
49 lesions from 43 patients
Training/Validation
n = 38
Test
n = 11

Figure 1. Flowchart of patient selection. TNBC, triple-negative breast cancer; NACT, neoadjuvant chemotherapy; MRI, magnetic resonance imaging; AI, artifical intelligence.

- Data splitting: The normalized tumor slices were randomly split into "Training," "Validation," and "Test" sets (random state: 42). After splitting, 30 tumor slices were selected for training, 8 for validation, and 11 for testing.

Data augmentation

Due to the relatively small dataset and imbalanced data distribution, data augmentation was applied to the training set. There were 13 lesions in the "pCR" class and 17 in the "non-pCR" class. To address this imbalance, data augmentation was first applied to underrepresented classes. Each class was then further augmented by randomly rotating the 3D MRI slices on the two-dimensional (2D; x, y) axis.

Deep learning model

Residual CNNs were used for their advantages in processing limited data and achieving better generalization. A residual CNN layer was designed in accordance with the ResNet architecture (Figure 2). The network input consisted of a 50 × 50 × 50 lesion image. A 2D CNN layer with 64 channels, followed by batch normalization and MaxPooling (MP) layers, reduced the data to 25 × 25 × 64. After two residual blocks and a 128-channel 2D CNN layer, the data were further reduced to 12 × 12 × 128 through another MP layer. Finally, a flattening layer produced a feature pool of 18,432 attributes. Similar processes were applied to other imaging sequences, and the features were combined after they had passed through the residual CNN layers. Although lesions were segmented volumetrically, the implemented architecture functions as a 2D CNN, operating on individual axial slices.

Due to their low count, T2WI sequences were excluded from the study. The features extracted from the pre-contrast T1WI, post-contrast T1WI, and DWI sequences were combined for each lesion, and classification was performed through a fully connected network with 1,792,896, and 256 neurons, respectively, in three dense layers (Figure 3). However, we evaluated multiple input configurations: (i) post-contrast T1WIs alone and (ii) multi-sequence inputs (pre-contrast T1WIs, post-contrast T1WIs, DWIs) using the same backbone. Due to sequence availability and performance on the test set, the final model reported in the Results section utilizes post-contrast T1WIs only. Despite the limited training data, accuracy values comparable to those in the literature were achieved. Residual CNNs offer key advantages, such as easier learning, robustness to model complexity, and training efficiency.

Statistical analysis

All statistical analyses were performed using R statistical software (version 3.6.0, Posit Software, PBC). Descriptive statistics were calculated to summarize patient and lesion characteristics. Continuous variables were expressed as mean ± standard deviation (SD) for normally distributed data and median with quartile values (Q1, Q3) for non-normally distributed data. Categorical variables were presented as absolute frequencies and percentages. Between-group comparisons for continuous variables were conducted using the Student's t-test or the Mann–Whitney U test according to distributional assumptions,

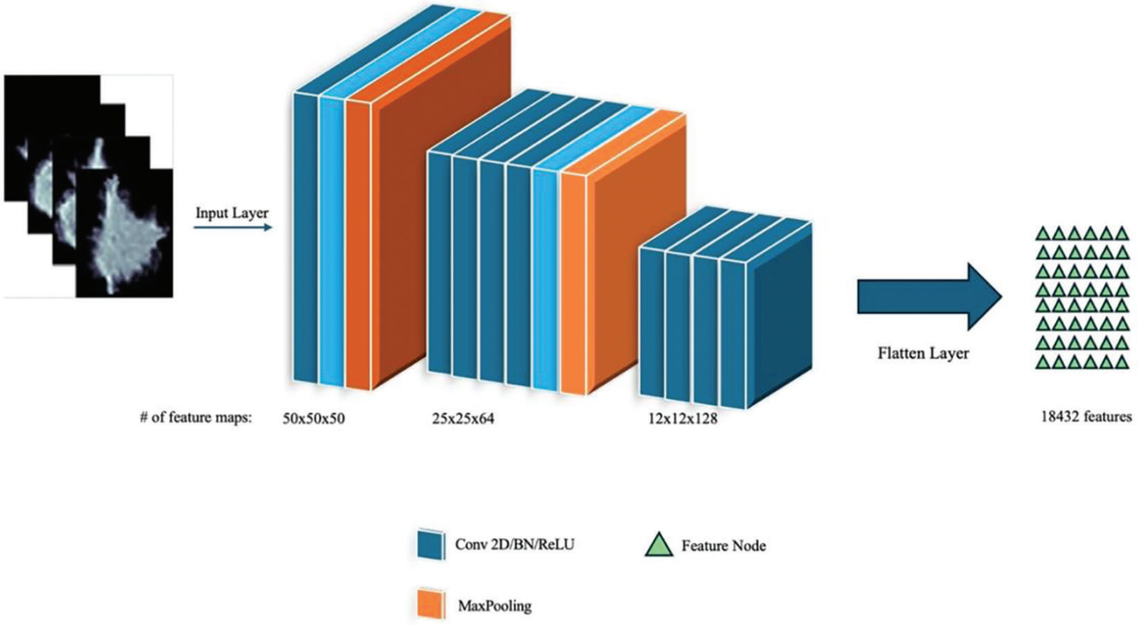


Figure 2. Residual convolutional neural network (CNN) layers. The network input consisted of a 50 × 50 × 50 lesion image. A two-dimensional (2D) CNN layer with 64 channels, followed by batch normalization and MaxPooling (MP) layers, reduced the data to 25 × 25 × 64. After two residual blocks and a 128-channel 2D CNN layer, the data were further reduced to 12 × 12 × 128 through further MP layer. Finally, a flattening layer produced a feature pool of 18,432 attributes.

whereas categorical variables were compared using the Fisher–Freeman–Halton test, as appropriate. It was considered statistically significant when $P < 0.05$. Variables found to be statistically significant in univariable analyses were included in a multivariable logistic regression model to identify independent predictors of pCR following NACT. The results were expressed as odds ratios (ORs) with 95% confidence intervals (CIs). The overall model fit was assessed using the likelihood ratio test, and predictive performance was evaluated with Nagelkerke's pseudo R^2. Model calibration was tested using the Hosmer–Lemeshow goodness-of-fit test. Model performance on the test set was evaluated by calculating accuracy from the confusion matrix (true-positive, true-negative, false-positive, and false-negative counts). Receiver operating characteristic (ROC) curves were plotted, and the area under the curve (AUC) was computed directly from the classification results. Intra-reader agreement was assessed by comparing repeated segmentations from the same reader on the same dataset using the Dice similarity coefficient formula. Pairwise Dice values were computed between all readers, and the mean (± SD) Dice value was reported to summarize inter- and intra-reader agreement.

All analyses were performed using functions from the readxl, dplyr, compareGroups, broom, and ResourceSelection packages in R.

Results

Descriptive results

The patient and lesion characteristics of the 43 patients included in the study are summarized in Table 1.

The study includes a total of 49 lesions, with 20 (40.82%) achieving pCR and 29 (59.18%) not achieving pCR. The mean age of patients in the pCR group was 50.1 ± 10.9 years, slightly older than the non-pCR group, which had a mean age of 48.9 ± 13.3 years. Patients who achieved pCR had significantly smaller median tumor sizes at baseline ($P = 0.034$), with a median of 28.5 mm (Q1–Q3: 22.5–32.0), than those who did not achieve pCR, who had a median tumor size of 35.5 mm (Q1–Q3: 24.25–56.5). The median tumor volume on post-contrast T1WIs was significantly less ($P = 0.045$) in the pCR group [median 9,243 mm^3 (Q1–Q3 3,714–13,665 mm^3)] than in the non-pCR group [median 19,453 mm^3 (Q1–Q3: 5,029–58,595 mm^3)].

In the multivariable logistic regression analysis, tumor volume measured on post-contrast T1WIs and the Ki-67 proliferation index were found to be independent predictors of achieving pCR after NACT. Tumor volume was associated with pCR (adjusted OR: 1.00; 95% CI: 1.00–1.00; $P = 0.040$), and higher Ki-67 levels were significantly associated with increased odds of pCR (adjusted OR: 1.04; 95% CI: 1.01–1.07; $P = 0.018$). Tumor size did not reach statistical significance (adjusted OR: 1.05; 95% CI: 0.96–1.17; $P = 0.299$). The overall model demonstrated a good fit (Nagelkerke's pseudo $R^2 = 0.422$, Hosmer–Lemeshow test $P = 0.702$) and was statistically significant according to the likelihood ratio test ($P < 0.001$). These results are summarized in Table 2.

Intra-reader and inter-reader agreement results

The AI algorithm was fed with 3D volumetric segmentations, and its reliability was evaluated by assessing intra-reader agreement using different scores for each sequence. Accordingly, the average Dice coefficient for segmentations performed on DWIs was 0.841 ± 0.075, and for segmentations performed on post-contrast T1WI sequences, the average Dice coefficient was 0.915 ± 0.046. Considering the inter-reader agreement between the radiologists based on different segmentations, the average Dice coefficient was 0.821 ± 0.050 for segmentations performed on DWIs and 0.890 ± 0.059 for segmentations performed on post-contrast T1WI sequences. These data demonstrate that the segmentations performed by the primary researcher at different times and those performed by the second researcher were highly consistent.

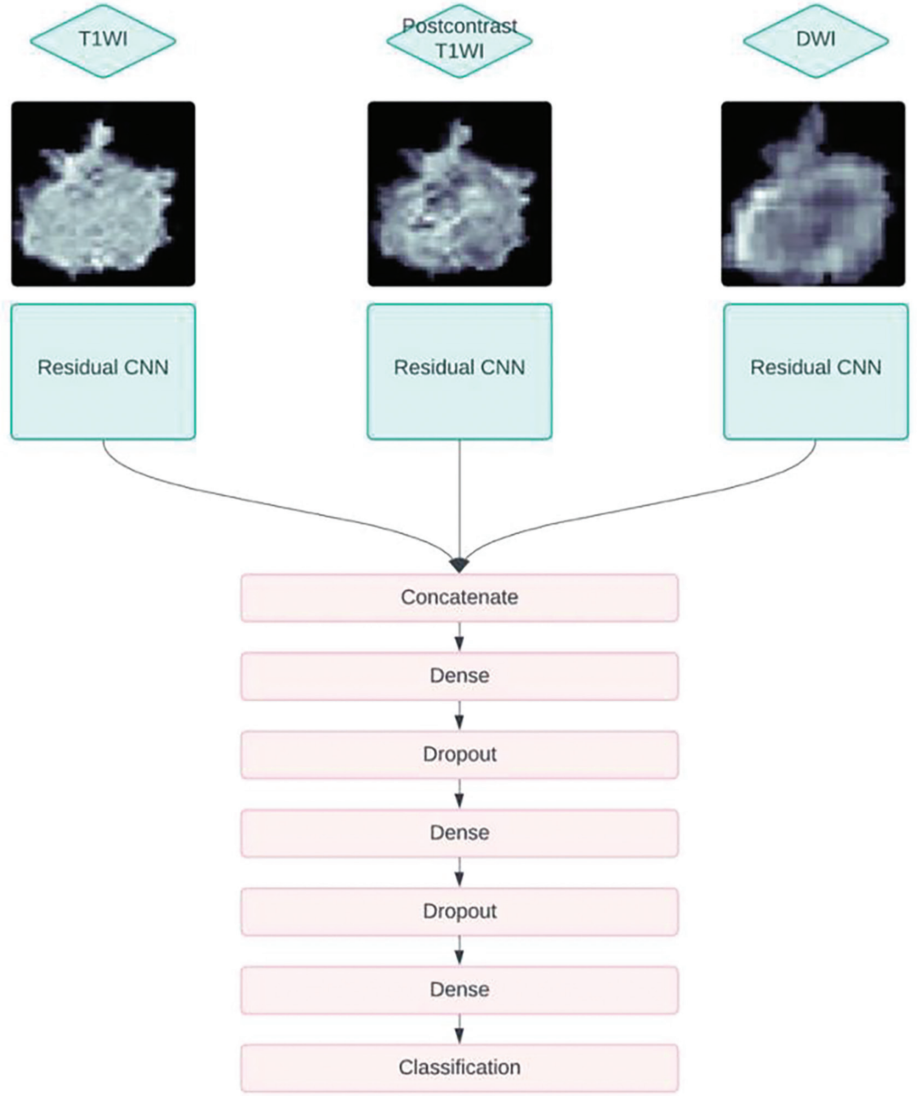


Figure 3. Deep learning model. The features extracted from the pre-contrast T1-weighted image (T1WI), post-contrast T1WI, and diffusion-weighted image (DWI) sequences were combined for each lesion, and classification was performed through a fully connected network with 1,792,896, and 256 neurons, respectively, in three dense layers.

*Note: We tested alternative model configurations using multiple sequences (pre-T1WI, post-T1WI, DWI); however, the final reported model uses post-contrast T1WI only. CNN, convolutional neural network.

Artificial intelligence model results

The best AI model for differentiating pCRs from non-pCRs on the test set revealed an accuracy of 0.82 (95% CI: 0.545–1.000) and AUC ROC of 0.75 (Figure 4). The best-performing model used only post-contrast T1WI data. These results demonstrate that the CNN-based AI model can predict response status with high performance. True-positive and true-negative examples predicted by the model are presented in Figure 5.

Discussion

TNBC is a rare but aggressive subtype of BC with a higher risk of metastasis than other subtypes.[2-4] Neoadjuvant therapy improves surgical outcomes, but its success is still unpredictable. If patients do not respond to this therapy, they face unnecessary toxicities. Therefore, predicting NACT response would help optimize treatment, reduce chemotoxicity risks, and improve clinical decision-making.

Despite advances in radiology, there is still a lack of data for accurately predicting NACT outcomes. Although AI is increasingly applied in radiology, few studies have focused on predicting NACT response in BC, especially for the TNBC subtype. This study aims to fill this gap by using only pre-treatment MRIs to predict responses in patients with TNBC. Our AI model, based on a CNN, achieved an accuracy of 0.82 in distinguishing patients who achieved pCR. Our model has several advantages over the previous studies that tried to predict or detect the response status of NACT in patients with BC. First, our model used only pre-treatment MRIs to classify patients as pCR or non-pCR. This extends the period for clinicians to modify the treatment plan and enhances their decision-making when concluding NACT early. Second, our model tried to predict responses using more sequences than previously used, which may add additional value to the AI model by using the different features of the various sequences. However, the best-performing model was identified as that using only post-contrast T1WI data to predict NACT response status. This might be because tumor heterogeneity is best determined in this sequence, and other sequences, such as DWIs and pre-contrast T1WIs, might lack sufficient data for the AI model to extract. Therefore, we also segmented the tumors in 3D volumetrically, enhancing the information that is acquired from the tumors. Moreover, selecting only a few slices of the tumor might create selection bias. Finally, unlike in earlier studies,[23-26] our model is based on the biopsy-proven TNBC subtype. This is because different types of BC behave differently to NACT, and studies including various types of BC might have heterogeneity that influences the results of the AI model in the future.

In terms of conventional analysis, after logistic regression analysis, we found that the parameters that might help identify NACT predictors in TNBC were the tumor proliferation index (Ki-67) and volume. Previous studies have shown that tumor Ki-67 values can predict response to NACT.[27] In a study by Penault-Llorca et al.,[28] which examined the predictive performance of various pathological markers for NACT in different types of BC with 710 patients, high Ki-67 values were found to be significant in predicting complete response, consistent with our findings. Similarly, MacGrogan et al.[29] identified high Ki-67 as an independent predictor of NACT response in patients with BC (n = 128). By contrast, Petit et al.[30] observed higher Ki-67 values in the complete response group but reported that the difference was not statistically significant. Additionally, studies by Bottini et al.[31] and Estévez et al.[32] found that Ki-67 was not a key predictor of NACT response. These different results are thought to arise from variations in patient groups and treatment protocols.

Table 1. Descriptive results of the dataset indicate that tumor size (mm), tumor volume (mm^3), and proliferation index (Ki67-) are significantly different between pCR and non-pCR groups

		pCR	Non-pCR	
Total		20	29	
Age (years) mean (± SD)		50.1 (10.9)	48.9 (13.3)	*P* = 0.732
Mammographic density BIRADS	A (%)	2 (50%)	2 (50%)	*P* = 0.556
	B (%)	10 (50%)	10 (50%)	
	C (%)	5 (27.7%)	13 (72.3%)	
	D (%)	3 (42.8%)	4 (57.2%)	
Tumor size (mm) median (Q1; Q3)		28.5 (22.5; 32.0)	35.5 (24.25; 56.5)	***P* = 0.041**
Tumor volume post-contrast T1WI (mm^3) median (Q1; Q3)		9,243 (3,714; 13,665)	19,453 (5,029; 58,595)	***P* = 0.030**
Background parenchymal enhancement BIRADS	Minimal (%)	13 (44.8%)	16 (55.2%)	*P* = 0.782
	Mild (%)	5 (38.5%)	8 (61.5%)	
	Moderate (%)	2 (40%)	3 (60%)	
	Marked (%)	0 (0%)	2 (100%)	
Proliferation (Ki-67) median (Q1; Q3)		80 (50; 80)	50 (30; 77.5)	***P* = 0.027**

SD, standard deviation; pCR, pathological complete response; T1WI, T1-weighted image.

Table 2. Multivariable logistic regression analysis of predictors for pathological complete response

	Adjusted OR (95% CI)	*P*
Tumor size (mm)	1.05 (0.96–1.17)	0.299
Tumor volume, post-contrast T1WI (mm^3)	1.00 (1.00–1.00)	0.040
Proliferation (Ki-67)	1.04 (1.01–1.07)	0.018

Pseudo R^2 (Nagelkerke): 0.422, Hosmer–Lemeshow *P* = 0.702, likelihood ratio test *P* < 0.001; OR, odds ratio; CI, confidence interval; T1WI, T1-weighted image.

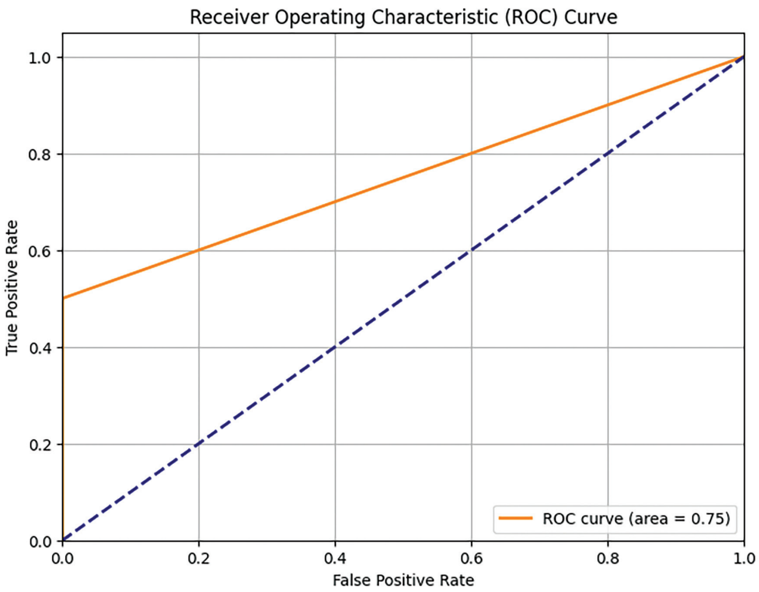


Figure 4. Receiver operating characteristic curve for the best-performing convolutional neural network model using post-contrast T1-weighted imaging to differentiate pathological complete response (pCR) from non-pCR in the test set (n = 11).

Besides conventional analysis, several studies support the potential of AI in predicting NACT outcomes. However, most of these studies used either one imaging method or one sequence, both pre-treatment and post-treatment images in combination, or all subtypes of the BC for the dataset. For instance, Herrero Vicent et al.[24] combined multiparametric MRIs and clinical data to create a machine learning model. This study, conducted on a small group of 58 patients, achieved an accuracy of 0.87 using only radiological imaging features.[24] Similarly, our study achieved high accuracy despite using a small patient group, demonstrating that AI models can perform well even with limited data.

Skarping et al.[23] developed an AI model using pre-treatment digital mammograms to predict pCR in all BC subtypes. This model was applied to 453 lesions, and an AUC score of 0.71[23] was achieved. Although their model used pre-treatment mammographic data, ours focused on MRI, demonstrating the versatility of imaging modalities in AI applications. In addition, we used different sequences to better understand the information in each sequence. In another study, Qu et al.[26] tested deep learning models on different imaging sets, including pre- and post-neoadjuvant T1WIs. Their model using only pre-treatment images had a lower AUC score of 0.55, but the combined model achieved a high AUC of 0.97.[26] This suggests that combining imaging datasets could substantially improve prediction accuracy; however, our study achieved stronger performance by only using pre-treatment images. Ha et al.[25] developed a CNN using pre-neoadjuvant MRI data from 141 patients, achieving an impressive AUC score of 0.98. This high accuracy highlights the potential of deep learning methods in predicting therapy responses. These findings suggest that with more data and re-

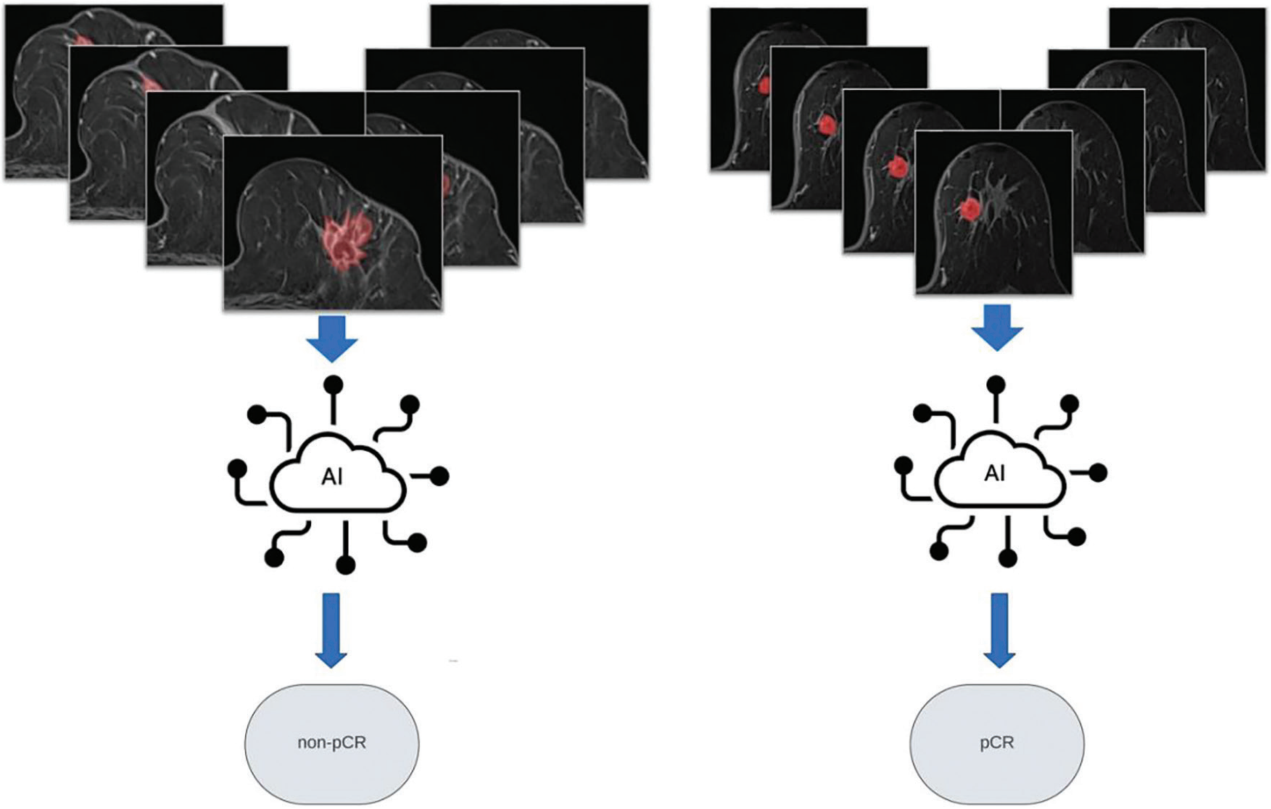


Figure 5. Examples of the artificial intelligence (AI) model's prediction. The upper-left tumor has a median size of 70 mm, a volume of 68,870 mm^3, and a Ki-67 value of 35, with no response. By contrast, the upper-right tumor has a median size of 25 mm, a volume of 3,714 mm^3, and a Ki-67 value of 50, with a complete response. pCR, pathological complete response.

fined methodologies, the performance of AI models such as that presented in this study could be further enhanced. Zhou et al.[33] developed an AI model focusing solely on TNBC, using MRI datasets collected before and after four cycles of NACT. This study achieved an accuracy of 0.77, and using open-source data allowed them to expand their patient group to 162. Furthermore, this study used both pre-treatment and post-treatment images to increase the performance, but they failed to note whether they tried only pre-treatment images for any model.[33] Previous studies are summarized in Table 3.

Overall, these studies highlight the promise of AI-based models in predicting NACT responses. As seen in the literature and our study, AI can provide high accuracy in predicting therapy outcomes, although larger patient groups and refined methodologies are necessary to enhance performance. Integrating clinical and radiological data and AI can substantially aid clinical decision-making processes.

This study has several limitations. First, the cohort size and small test set constrain statistical power and widen uncertainty around performance estimates. Second, clinical staging at diagnosis was not consistently available across centers, which precluded stage-stratified analyses and may introduce clinical heterogeneity. Third, although we initially evaluated multiple MRI sequences, T2WIs were excluded because complete, high-quality series were insufficiently available across patient cohorts and centers; moreover, in other models within our sample, adding DWIs and/or pre-contrast T1WIs did not improve discrimination over post-contrast T1WIs alone. These factors may limit generalizability and should be addressed in larger, prospectively curated, multi-institutional cohorts. Moreover, patients with non-mass enhancement were excluded due to difficulties in tumor segmentation. Although rare, this exclusion limits the model's applicability to specific patient subgroups. Finally, this study evaluates a single residual CNN backbone without head-to-head comparisons against alternative deep learning architectures or classic machine learning approaches using hand-crafted radiomics, which restricts the scope for architectural comparison.

Future research focusing on external validation across multiple institutions and scanners, prospective enrollment to ensure complete clinical staging and acquisition protocols, and development of multimodal models that fuse imaging-derived representations with clinical biomarkers (e.g., Ki-67) might improve discrimination, calibration, and decision utility. With larger datasets and more complete sequence availability, we will revisit multi-sequence inputs and explore end-to-end 3D architectures and other architectural designs to test whether additional sequences and different architectures (e.g., T2WIs, DWIs) provide incremental value beyond post-contrast T1WIs.

In conclusion, AI-based models hold considerable potential in predicting NACT responses, particularly for aggressive subtypes such as TNBC. These models can improve clinical outcomes by optimizing treatment plans and personalizing care. However, expanding research with larger, multicenter datasets is necessary to enhance the models' generalizability and ensure broader clinical application. With continued advancements, AI can play a crucial role in the future of personalized BC treatment.

Table 3. Comparison of the results of different studies, indicating that artificial intelligence-assisted models can predict the neoadjuvant therapy response status in different categories for patients with breast cancer

Study	Data used	Method	Total number	Test number	Subtypes of breast cancer	Performance result
Herrero Vicent et al.[24]	Multiparametric MRI and clinical data	Machine learning	58	24	All types of breast cancer	0.87 of accuracy (only with radiological images)
Skarping et al.[23]	Digital mammograms	Deep learning	453	53	All types of breast cancer	0.71 AUC
Qu et al.[26]	Pre-treatment and post-treatment post-contrast T1WI	Deep learning	302	58	All types of breast cancer	0.55 AUC (only with pre-treatment images) 0.97 AUC (combined)
Ha et al.[25]	Pre-treatment post-contrast T1WI	Deep learning	141	28	All types of breast cancer	0.98 AUC
Zhou et al.[33]	Pre-treatment and post-treatment post-contrast T1WI and DWI	Deep learning	162	32	Triple-negative breast cancer	0.77 of accuracy
This study	**Pre-treatment post-contrast T1WI**	**Deep learning**	**49**	**11**	**Triple-negative breast cancer**	**0.82 of accuracy**

MRI, magnetic resonance imaging; T1WI, T1-weighted image; AUC, area under the curve; DWI, diffusion-weighted image.

Footnotes

Conflict of Interest

The authors declared no conflicts of interest.

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12-49. [Crossref]
2. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. *Lancet*. 2021;397(10286):1750-1769. [Crossref]
3. Cho N. Imaging features of breast cancer molecular subtypes: state of the art. *J Pathol Transl Med*. 2021;55(1):16-25. [Crossref]
4. Yin L, Duan JJ, Bian XW, Yu SC. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res*. 2020;22(1):61. [Crossref]
5. Curigliano G, Burstein HJ, Gnant M, et al. Understanding breast cancer complexity to improve patient outcomes: The St Gallen International Consensus Conference for the Primary Therapy of Individuals with Early Breast Cancer 2023. *Ann Oncol*. 2023;34(11):970-986. [Crossref]
6. Cardoso F, Kyriakides S, Ohno S, et al. Electronic address: clinicalguidelines@esmo.org. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. *Ann Oncol*. 2019;30(8):1194-1220. [Crossref]
7. Golshan M, Loibl S, Wong SM, et al. Breast conservation after neoadjuvant chemotherapy for triple-negative breast cancer: surgical results from the brightness randomized clinical trial. *JAMA Surg*. 2020;155(3):e195410. [Crossref]
8. Bear HD, Anderson S, Brown A, et al. The effect on tumor response of adding sequential preoperative docetaxel to preoperative doxorubicin and cyclophosphamide: preliminary results from National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *J Clin Oncol*. 2003;21(22):4165-4174. [Crossref]
9. Fisher B, Bryant J, Wolmark N, et al. Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *J Clin Oncol*. 1998;16(8):2672-2685. [Crossref]
10. Rastogi P, Anderson SJ, Bear HD, et al. Preoperative chemotherapy: updates of National Surgical Adjuvant Breast and Bowel Project Protocols B-18 and B-27. *J Clin Oncol*. 2008;26(5):778-785. [Crossref]
11. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384(9938):164-172. [Crossref]
12. Caparica R, Lambertini M, Pondé N, Fumagalli D, de Azambuja E, Piccart M. Post-neoadjuvant treatment and the management of residual disease in breast cancer: state of the art and perspectives. *Ther Adv Med Oncol*. 2019;11:1758835919827714. [Crossref]
13. Matuschek C, Jazmati D, Bölke E, et al. Post-neoadjuvant treatment strategies in breast cancer. *Cancers (Basel)*. 2022;14(5):1246. [Crossref]
14. Bozer A, Yilmaz C, Çetin Tunçez H, Kocatepe Çavdar D, Adıbelli ZH. Correlation of histopathological and radiological response patterns and their prognostic implications in breast cancer after neoadjuvant chemotherapy. *Breast Cancer (Dove Med Press)*. 2024;16:1005-1017. [Crossref]
15. Expert Panel on Breast Imaging:; Slanetz PJ, Moy L, et al. ACR appropriateness criteria® monitoring response to neoadjuvant systemic therapy for breast cancer. *J Am Coll Radiol*. 2017;14(11S):S462-S475. [Crossref]
16. Ollivier L, Balu-Maestro C, Leclère J. Imaging in evaluation of response to neoadjuvant breast cancer treatment. *Cancer Imaging*. 2005;5(1):27-31. [Crossref]
17. Kong X, Moran MS, Zhang N, Haffty B, Yang Q. Meta-analysis confirms achieving pathological complete response after neoadjuvant chemotherapy predicts favourable prognosis for breast cancer patients. *Eur J Cancer*. 2011;47(14):2084-2090. [Crossref]
18. Houssami N, Macaskill P, von Minckwitz G, Marinovich ML, Mamounas E. Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. *Eur J Cancer*. 2012;48(18):3342-3354. [Crossref]
19. Li J, Jiang P, An Q, Wang GG, Kong HF. Medical image identification methods: a review. *Comput Biol Med*. 2024;169:107777. [Crossref]
20. Torres AD, Yan H, Aboutalebi AH, Das A, Duan L, Rad P. Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration. *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications. Elsevier*. 2018:61-89. [Crossref]
21. Surucu M, Isler Y, Kara R. Diagnosis of paroxysmal atrial fibrillation from thirty-minute heart rate variability data using convolutional neural networks. *Turkish Journal of Electrical Engineering and Computer Sciences*. 2021;29:2886-2900. [Crossref]
22. Wang W, Liu Y, Zhang H, et al. Prognostic value of residual cancer burden and Miller-Payne system after neoadjuvant chemotherapy for breast cancer. *Gland Surg*. 2021;10(12):3211-3221. [Crossref]
23. Skarping I, Larsson M, Förnvik D. Analysis of mammograms using artificial intelligence

to predict response to neoadjuvant chemotherapy in breast cancer patients: proof of concept. *Eur Radiol.* 2022;32(5):3131-3141. [Crossref]

24. Herrero Vicent C, Tudela X, Moreno Ruiz P, et al. Machine learning models and multiparametric magnetic resonance imaging for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Cancers (Basel)*. 2022;14(14):3508. [Crossref]
25. Ha R, Chin C, Karcich J, et al. Prior to initiation of chemotherapy, can we predict breast tumor response? Deep learning convolutional neural networks approach using a breast MRI tumor dataset. *J Digit Imaging*. 2019;32(5):693-701. [Crossref]
26. Qu YH, Zhu HT, Cao K, Li XT, Ye M, Sun YS. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (DL) method. *Thorac Cancer*. 2020;11(3):651-658. [Crossref]
27. Yerushalmi R, Woods R, Ravdin PM, Hayes MM, Gelmon KA. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol*. 2010;11(2):174-183. [Crossref]
28. Penault-Llorca F, Abrial C, Raoelfils I, et al. Changes and predictive and prognostic value of the mitotic index, Ki-67, cyclin D1, and cyclo-oxygenase-2 in 710 operable breast cancer patients treated with neoadjuvant chemotherapy. *Oncologist*. 2008;13(12):1235-1245. [Crossref]
29. MacGrogan G, Mauriac L, Durand M, et al. Primary chemotherapy in breast invasive carcinoma: predictive value of the immunohistochemical detection of hormonal receptors, p53, c-erbB-2, MiB1, pS2 and GST pi. *Br J Cancer.* 1996;74(9):1458-1465. [Crossref]
30. Petit T, Wilt M, Velten M, et al. Comparative value of tumour grade, hormonal receptors, Ki-67, HER-2 and topoisomerase II alpha status as predictive markers in breast cancer patients treated with neoadjuvant anthracycline-based chemotherapy. *Eur J Cancer.* 2004;40(2):205-211. [Crossref]
31. Bottini A, Berruti A, Bersiga A, et al. Relationship between tumour shrinkage and reduction in Ki67 expression after primary chemotherapy in human breast cancer. *Br J Cancer*. 2001;85(8):1106-1112. [Crossref]
32. Estévez LG, Cuevas JM, Antón A, et al. Weekly docetaxel as neoadjuvant chemotherapy for stage II and III breast cancer: efficacy and correlation with biological markers in a phase II, multicenter study. *Clin Cancer Res*. 2003;9(2):686-692. [Crossref]
33. Zhou J, Lu J, Gao C, et al. Predicting the response to neoadjuvant chemotherapy for breast cancer: Wavelet transforming radiomics in MRI. *BMC Cancer.* 2020;20:100. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253371

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ORIGINAL ARTICLE

Magnetic resonance imaging findings in Ménière's disease: the impact of radiologist experience on hydrops imaging

Çağatay Cihan[1]
Uğur Toprak[1]
Emre Emekli[1,2]
Armağan İncesulu[3]
Hamit İpek[3]

[1]Eskişehir Osmangazi University Faculty of Medicine, Department of Radiology, Eskişehir, Türkiye

[2]Eskişehir Osmangazi University, Translational Medicine Application and Research Center, Eskişehir, Türkiye

[3]Eskişehir Osmangazi University Faculty of Medicine, Department of Ear Nose Throat, Eskişehir, Türkiye

Corresponding author: Emre Emekli

E-mail: emreemekli90@gmail.com

Received 08 April 2025; revision requested 08 May 2025 ; last revision received 07 July 2025; accepted 27 July 2025.

Epub: 18.08.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253371

PURPOSE

This study investigates the competence of a newly certified radiologist in reporting hydrops imaging and examines the role of magnetic resonance imaging (MRI) findings in diagnosing definite and probable Ménière's disease (MD).

METHODS

Sixty-four cases were retrospectively evaluated–blinded to clinical data–by a senior radiologist (O-1) and a newly certified radiologist (O-2) using 3D heavily T2-weighted and delayed contrast-enhanced three-dimensional fluid-attenuated inversion recovery sequences. The posterior fossa–posterior semicircular canal (P–P) distance, endolymphatic hydrops (EH), perilymphatic enhancement (PE), and the round window sign (RWS) were assessed.

RESULTS

Interobserver agreement was moderate for cochlear ($\kappa = 0.591$) and vestibular hydrops ($\kappa = 0.566$), good for PE ($\kappa = 0.663$), and excellent for the RWS ($\kappa = 0.817$). O-1 demonstrated good intraobserver agreement for the RWS ($\kappa = 0.787$) and excellent agreement for the other parameters. O-2 showed lower intraobserver agreement for cochlear hydrops, vestibular hydrops, and the RWS ($\kappa = 0.366$, $\kappa = 0.332$, and $\kappa = 0.398$, respectively). The P–P distance showed excellent interobserver [intraclass correlation coefficient (ICC) = 0.932] and intraobserver agreement (ICC = 0.978 for O-1; ICC = 0.886 for O-2). The P–P distance was significantly shorter in definite MD (dMD) than in probable MD (pMD) (1.23 ± 1.07 mm vs. 2.17 ± 1.79 mm, $P = 0.021$). The rate and grade of hydrops were higher in dMD ($P < 0.050$), whereas the RWS was more frequent in pMD. Hydrops and PE were more often observed on the symptomatic side ($P < 0.001$). Cochlear hydrops was identified in 14.3% and vestibular hydrops in 31.2% of asymptomatic sides.

CONCLUSION

The newly certified radiologist's intraobserver agreement for hydrops imaging was insufficient. In dMD, the retrolabyrinthine bone is thinner, hydrops is more frequent and advanced, and the RWS is less common. Approximately one in five patients with MD may have a perilymphatic fistula. Close monitoring of asymptomatic contralateral ears is essential.

CLINICAL SIGNIFICANCE

Accurate MRI evaluation of EH in MD strongly depends on the radiologist's expertise. This study highlights that newly certified radiologists may show lower reliability in assessing hydrops imaging, underscoring the need for targeted training programs.

KEYWORDS

Endolymphatic hydrops, magnetic resonance imaging , Meniere's disease, perilymphatic enhancement, perilymphatic fistula, round window sign

You may cite this article as: Cihan Ç, Toprak U, Emekli E, İncesulu A, İpek H. Magnetic resonance imaging findings in Ménière's disease: the impact of radiologist experience on hydrops imaging. *Diagn Interv Radiol*. 2026;32(2):200-207.

Ménière's disease (MD) is a clinical syndrome characterized by spontaneous vertigo, fluctuating low-frequency sensorineural hearing loss, tinnitus, and aural fullness.[1] The 2015 diagnostic criteria classify MD into two categories: definite MD (dMD) and probable MD (pMD), based on the duration of vertigo attacks and the presence of low- to mid-frequency hearing loss.[1] However, in the early stages of the disease, key symptoms–such as vertigo, tinnitus, and hearing loss–may not occur simultaneously, complicating diagnosis.[2] The etiopathogenesis of MD is not fully understood, but it is associated with an excessive accumulation of endolymph, resulting in endolymphatic hydrops (EH). Due to overlapping clinical features, MD is often misdiagnosed as other conditions, such as vestibular migraine or vestibular schwannoma.[3-6]

Advancements in endolymphatic imaging have been made possible by 3 Tesla (3T) magnetic resonance imaging (MRI) systems. Nakashima et al.[7] conducted the first notable study in 2007, using intratympanic administration of a contrast agent. In 2010, the same group introduced intravenous contrast administration for endolymphatic imaging.[8] Imaging is typically performed approximately 4 hours after intravenous contrast injection, when the agent reaches peak concentration in the perilymph but does not enter the endolymph, allowing the endolymph to appear as negative contrast.[8]

Since Nakashima's pioneering studies, numerous investigations have focused on EH imaging. However, most rely on subjective visual assessments, raising concerns about reliability and reproducibility. A review of the literature shows few studies examining interobserver agreement and even fewer evaluating intraobserver agreement. These studies typically involve neuroradiologists or head and neck radiologists, often senior-level experts.[9-12] At our institution, hydrops imaging has been integrated into routine MRI scans for head and neck radiology since 2019. However, it is not yet standard practice in many countries, including ours, where it remains primarily a research topic. This study aims to evaluate the competence of a newly certified radiologist in hydrops imaging and to assess the diagnostic role of MRI findings in dMD and pMD.

Main points

- Magnetic resonance imaging (MRI) evaluation of endolymphatic hydrops (EH) in Ménière's disease (MD) is highly dependent on the radiologist's experience; newly certified radiologists may require additional training to achieve adequate diagnostic consistency.
- EH and perilymphatic enhancement on MRI are considerably more common and advanced on the symptomatic side in patients with definite MD (dMD) than in those with probable MD (pMD).
- Reduced retrolabyrinthine bone thickness (i.e., shorter posterior fossa–semicircular canal distance) measured by MRI may serve as a supportive imaging marker for dMD.
- The round window sign, suggestive of perilymphatic fistula, can mimic symptoms of MD–particularly in pMD cases–and was observed in approximately one-fifth of symptomatic patients.
- MRI may detect EH even on clinically asymptomatic sides, emphasizing the importance of bilateral evaluation and long-term follow-up.

Methods

Approval was obtained from the Eskişehir Osmangazi University Non-Interventional Clinical Research Ethics Committee (16.05.2023/58). As this was a retrospective study, the ethics committee waived the requirement for informed consent from the patients.

Magnetic resonance imaging protocol

MRI scans were performed using a 3T scanner (GE Discovery 750W, General Electric, Milwaukee, WI, USA) equipped with a 32-channel head coil. Fifty-four cases were scanned using the older version of the scanner, and 10 cases were scanned using the upgraded version. For the older version, 0.2 mmol/kg of a gadolinium-based contrast agent was administered intravenously, whereas 0.1 mmol/kg was used for the upgraded version.

The protocol for the delayed contrast-enhanced CUBE three-dimensional fluid-attenuated inversion recovery (3D-FLAIR) sequence on the older version was as follows: field of view (FOV) 260 mm, slice thickness 0.8 mm, repetition time (TR) 6,800 ms, echo time (TE) 115 ms, number of excitations (NEX) 1, inversion time (TI) 1,769 ms, matrix 320 × 288, bandwidth 42 Hz/pixel, echo train length (ETL) 200, voxel size 0.8 × 0.8 × 0.8 mm, and scan time 8 minutes. The protocol for the upgraded version was as follows: FOV 160 mm, slice thickness 0.4 mm, TR 8,500 ms, TE 168 ms, NEX 2, TI 2,185 ms, flip angle 142°, matrix 224 × 224, bandwidth 50 Hz/pixel, ETL 220, voxel size 0.7 × 0.7 × 0.8 mm, and scan time 6 minutes. The protocol for the heavily T2-weighted fast imaging employing steady-state acquisition (FIESTA) sequence was as follows: FOV 220 mm, slice thickness 0.8 mm, TR 5.4 ms, TE 2.1 ms, NEX 2, flip angle 55°, matrix 320 × 320, and scan time 5 minutes. This 3D-FLAIR sequence is T1-weighted and optimized for delayed post-contrast imaging to assess inner ear fluid compartments.

Patient selection

Between September 2019 and April 2023, MRI scans were performed. Patients were excluded from the study due to images of insufficient quality (motion artifacts, low contrast-to-noise ratio) or clinical diagnoses of vertigo or hearing loss (sensorineural and sudden sensorineural), leaving 86 patients with MD. Among these, 22 cases with a clinical diagnosis of bilateral MD were included in the radiological evaluation but excluded from statistical analysis to avoid bias related to disease laterality. Finally, 64 patients were included in the study (Figure 1).

The following patient characteristics were recorded: age, sex, symptomatic side (right or left), diagnostic classification (pMD or dMD), and history of intratympanic therapy (e.g., gentamicin). MRI evaluations included cochlear and vestibular hydrops grading, presence of perilymphatic enhancement (PE), the round window sign (RWS), and measurement of the posterior fossa–posterior semicircular canal (P–P) distance.

Interobserver and intraobserver agreement

The 64 patients included in the study were evaluated independently, twice each, by a radiologist with 20 years of experience in head and neck radiology and 5 years of experience in EH imaging (XX, O-1), and a newly certified radiologist with 5 years of general radiology experience (XX, O-2). During residency, O-2 completed three separate 3-month head and neck radiology rotations, during each of which they reviewed approximately 30 cases involving EH imaging. The observers were blinded to clinical findings, disease laterality, and whether symptoms were unilateral or bilateral. Each patient was evaluated twice by each observer, with a 2–4-week interval between the two evaluations. The assessments were performed using the GE Advantage Workstation VolumeShare 5 (General Electric, Milwaukee, WI, USA).

Image evaluation

Vestibular hydrops was evaluated on delayed contrast-enhanced 3D-FLAIR series using the Bernaerts classification (grade 0: normal-sized saccule and utricle; grade 1: the saccule is equal in size to or larger than the utricle; grade 2: confluence of saccule and utricle encompassing >50% of the vestibule; grade 3: total effacement of the perilymphatic space) (Figure 2). Cochlear hydrops (Figure 3 and 4) was evaluated using the Baráth classification (grade 1: mild dilatation of the non-enhancing cochlear duct; grade 2: uniform obstruction of the scala vestibuli by the severely distended cochlear duct).[13,14] Asymmetric PE (Figure 5)[15] and the presence of the RWS[16] were also investigated (Figure 6). In heavily T2-weighted images, the P–P distance (Figure 7) was measured as the distance from the posterior border of the vertical part of the posterior semicircular canal to the posterior cortex of the petrous bone, used as the reference.[10,17] In cases of discrepancy between the two evaluations, the radiologists jointly re-evaluated the imaging findings in a consensus session, during which both examiners reviewed the images together on the same workstation and reached an agreement through discussion. If disagreement persisted, the finding was recorded as "non-consensus" and excluded from the final agreement analysis.

Statistical analysis

Data were analyzed using (IBM, Armonk, NY, USA). A value of $P < 0.05$ was considered statistically significant. The Shapiro–Wilk test was used to assess the normality of continuous variables. McNemar's test was applied to compare findings between symptomatic and asymptomatic sides. The Wilcoxon test was used to compare P–P distances, whereas the Mann–Whitney U test compared the dMD and pMD groups. Chi-square and Fisher's exact tests were used for categorical comparisons. Interobserver and intraobserver agreement for P–P measurements was assessed using the intraclass correlation coefficient (ICC), and Cohen's kappa was used to analyze agreement for categorical variables.

Results

Interobserver and intraobserver agreement

The mean age of the 64 cases was 45.49 ± 10.81 years, with 35 women (54.7%) and 29 men (45.3%). A total of 43 cases (67.2%) were classified as pMD and 21 cases as dMD.

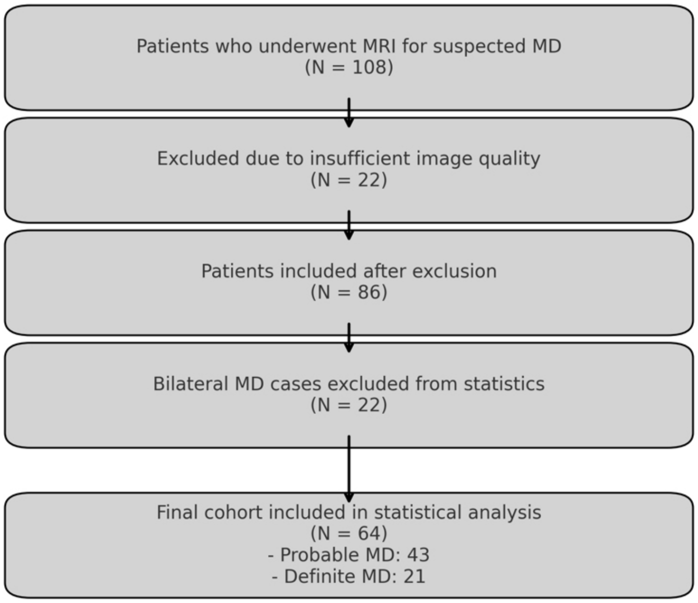

Patients who underwent MRI for suspected MD
(N = 108)
Excluded due to insufficient image quality
(N = 22)
Patients included after exclusion
(N = 86)
Bilateral MD cases excluded from statistics
(N = 22)
Final cohort included in statistical analysis
(N = 64)
- Probable MD: 43
- Definite MD: 21

Figure 1. Patient selection flowchart. MRI, magnetic resonance imaging; MD, Ménière's disease.

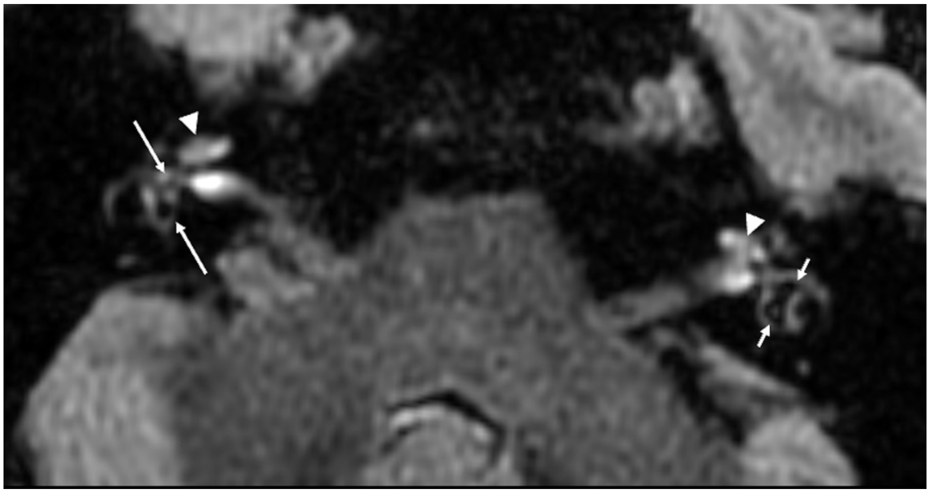


Figure 2. Axial delayed contrast-enhanced 3D-FLAIR images showing vestibular hydrops evaluation. On the right side, the saccule and utricle (long arrows) are normal in size and well separated (grade 0, Bernaerts classification).[19] On the left side, the saccule and utricle are confluent (short arrows), but the perilymphatic space remains partially visible, consistent with grade 2. Bilateral cochleae are marked with arrowheads. 3D-FLAIR, three-dimensional fluid-attenuated inversion recovery.

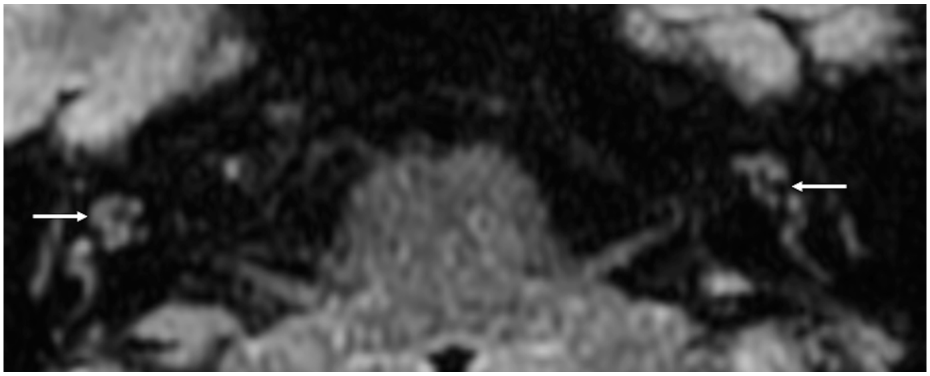


Figure 3. Axial delayed contrast-enhanced 3D-FLAIR image showing cochlear hydrops evaluation. On the left side, punctate non-enhancing areas within the cochlear duct (arrows) indicate grade 1 hydrops, according to the Baráth classification.[13] The right cochlea appears normal. 3D-FLAIR, three-dimensional fluid-attenuated inversion recovery.

In 36 patients (56.3%), the right side was symptomatic, whereas in 28 (43.8%), the left side was symptomatic.

Interobserver agreement was moderate for cochlear and vestibular hydrops, good for PE, and very good for the RWS (Table 1). O-1's intraobserver agreement was good for the RWS and very good for the other criteria. O-2's intraobserver agreement was low (Table 2). For the P–P distance (n = 128), O-1 measured a mean of 1.87 ± 1.57 mm, whereas O-2 measured 1.67 ± 1.55 mm, with very good agreement [interobserver agreement: ICC = 0.932 (95% confidence interval; CI: 0.905–0.952), $P < 0.001$; intraobserver agreement: ICC = 0.978 for O-1 (95% CI: 0.960–0.989), ICC = 0.886 for O-2 (95% CI: 0.796–0.938), $P < 0.001$]. The P–P distance was shorter in cases of dMD than in pMD (1.23 ± 1.07 mm vs. 2.17 ± 1.79 mm, $P = 0.021$), but the difference between the normal sides was not significant (1.36 ± 1.14 mm vs. 2.13 ± 1.61 mm, $P = 0.056$).

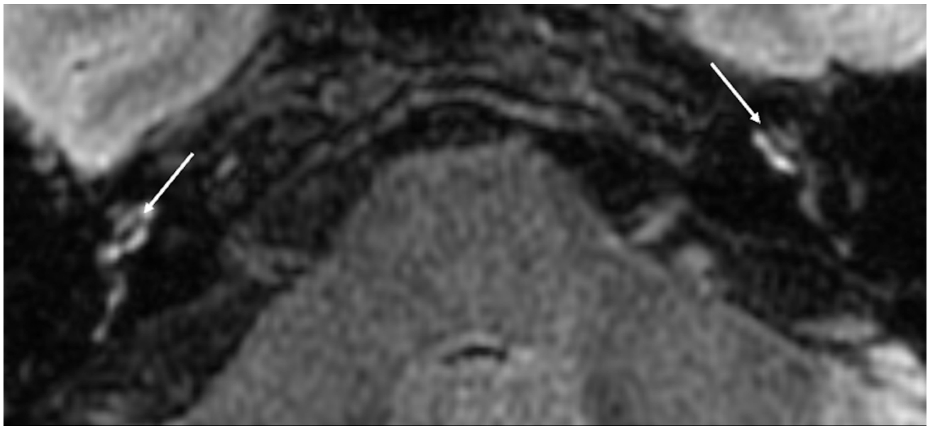


Figure 4. Axial delayed contrast-enhanced 3D-FLAIR image demonstrating bilateral grade 2 cochlear hydrops. Uniform dilation of the cochlear ducts (arrows) causes linear filling defects within the scala vestibuli, as defined by the Baráth classification.[13] 3D-FLAIR, three-dimensional fluid-attenuated inversion recovery.

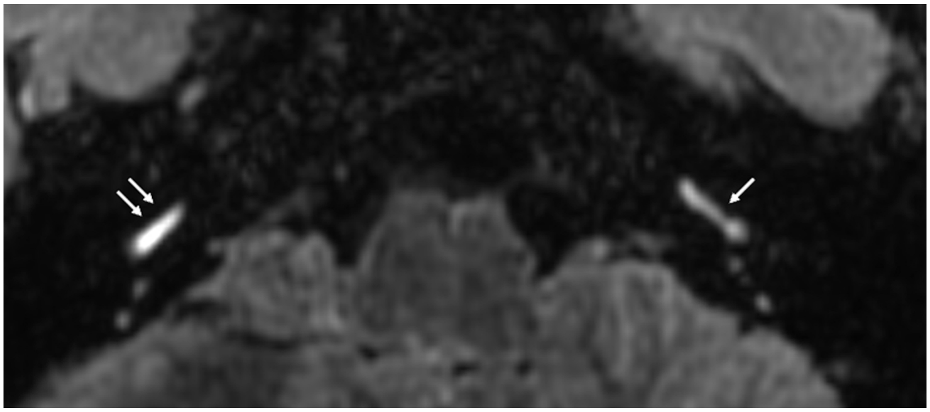


Figure 5. Axial delayed contrast-enhanced 3D-FLAIR image showing asymmetric perilymphatic enhancement. Increased contrast uptake is noted in the cochlear basal turn on the right side (arrows), compared with the left, consistent with asymmetric PE as described by Bernaerts et al.[19] 3D-FLAIR, three-dimensional fluid-attenuated inversion recovery; PE, perilymphatic enhancement.

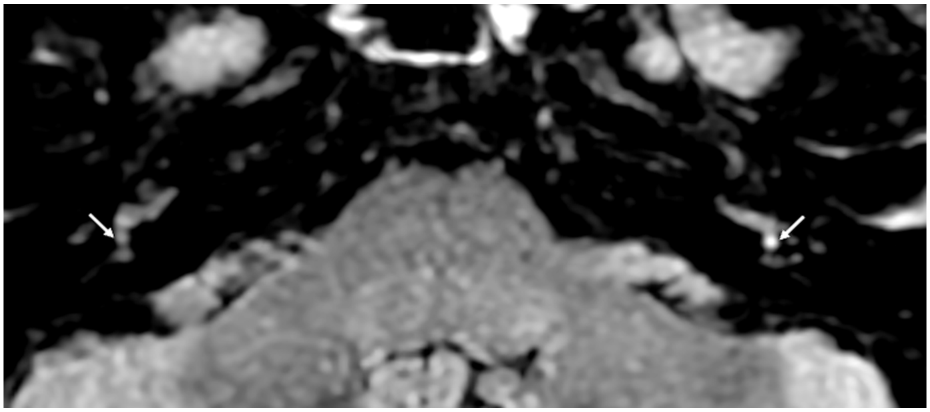


Figure 6. Axial delayed contrast-enhanced 3D-FLAIR image demonstrating high signal intensity in the left round window niche (arrow) in a patient with probable left-sided Ménière's disease, consistent with the round window sign suggestive of perilymphatic fistula, as described by Dubrulle et al.[16] The right round window niche appears normal (arrow). 3D-FLAIR, three-dimensional fluid-attenuated inversion recovery.

Radiologic evaluation

On the asymptomatic side, grade 1 cochlear hydrops was observed in 5/64 cases (7.8%) and grade 2 in 4/64 cases (6.3%). Vestibular hydrops was grade 1 in 7/64 cases (10.9%) and grade 2 in 13/64 cases (20.3%). No grade 3 vestibular hydrops was observed on the asymptomatic side. PE was observed in 2/64 cases (3.1%), and the RWS was seen in 8/64 cases (12.5%), equal to the symptomatic side.

The comparison of radiological findings between the symptomatic and asymptomatic sides is presented in Table 3. On the symptomatic side, cochlear hydrops, vestibular hydrops, and PE were significantly more frequent than on the asymptomatic side ($P < 0.001$). The presence of the RWS was similar on both sides ($P > 0.05$).

On the symptomatic side, in cases of dMD, the rate of grade 2 cochlear hydrops (7/21 vs. 3/43; $P = 0.032$) and grade 3 vestibular hydrops (10/21 vs. 3/43; $P < 0.001$) was higher than in pMD. The rates of the RWS (3/21 vs. 5/43; $P = 1$) and PE (9/21 vs. 12/43; $P = 0.232$) were similar between the groups.

In 15 symptomatic ears (13 pMD, 2 dMD), neither EH nor PE was observed. In 3 of these 15 cases, the RWS was present (2 pMD, 1 dMD). In total, the RWS was detected in 8 symptomatic ears (5 pMD, 3 dMD) and on the asymptomatic side, including 6 cases of bilateral MD. On the symptomatic side, PE accompanied the RWS in 3 out of 8 ears. Among the 8 symptomatic ears with the RWS, EH (3 cochlear, 5 vestibular) was observed in 3 cases, along with PE. On the asymptomatic side, EH was observed in 4 cases (1 cochlear, 3 vestibular), but no PE was detected. No soft tissue or effusion indicating hyperintensity in the area corresponding to the RWS was observed on the heavily T2-weighted sequences.

Bilateral RWS was observed in one patient with dMD who had undergone intratympanic gentamicin therapy, whereas bilateral grade 1 vestibular hydrops was noted in another patient. None of the other patients included in the study had a history of intratympanic therapy.

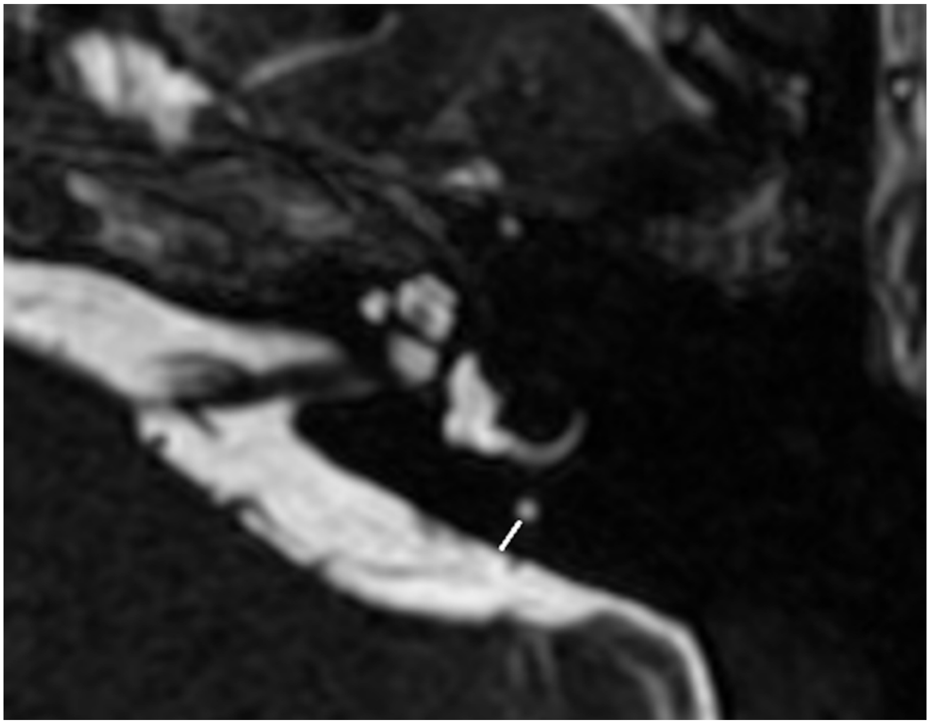


Figure 7. Coronal heavily T2-weighted FIESTA image showing measurement of the posterior fossa–posterior semicircular canal distance. The line indicates the shortest distance between the posterior border of the vertical limb of the posterior semicircular canal and the posterior cortical surface of the petrous temporal bone, used as an indicator of retrolabyrinthine bone thickness, as described by Lei et al.[10] FIESTA, fast imaging employing steady-state acquisition.

Discussion

In our study, interobserver agreement was moderate for cochlear and vestibular hydrops and good for visual PE evaluation. The 2022 study by Bernaerts et al.[18] reported similar agreement rates among senior neuroradiologists for cochlear and vestibular hydrops, as well as for visual PE evaluation using turbo spin-echo (TSE) FLAIR sequences. However, agreement was considerably higher in their study when using SPACE FLAIR sequences for hydrops and PE assessment. Their 2019 study also reported higher interobserver agreement, reaching a good level.[19] In the 2022 study by Deng et al.[20], interobserver agreement for vestibular hydrops using TSE FLAIR was good, and for cochlear hydrops, it was very good. That study also demonstrated even higher agreement when using real inversion recovery (IR) sequences. As previously noted, IR sequences developed after TSE FLAIR improved geometrical resolution, resulting in higher agreement rates. Studies using these sequences have reported near-perfect agreement among senior researchers.[9-12]

Table 1. Interobserver agreement for categorical data

Observer 1\observer 2		None	Present/grade 1	Grade 2	Grade 3	Kappa	*P*
Vestibular hydrops	**None**	53	5	6	0	0.566	<0.001
	Grade 1	8	7	4	0		
	Grade 2	4	4	23	1		
	Grade 3	0	0	4	9		
Cochlear hydrops	**None**	83	7	5		0.591	<0.001
	Grade 1	7	11	1			
	Grade 2	0	3	11			
Perilymphatic enhancement	**None**	95	10			0.663	<0.001
	Present/grade 1	4	19				
Round window sign	**None**	110	2			0.817	<0.001
	Present/grade 1	3	13				

Table 2. Intraobserver agreement for categorical data

	Observer-1		Observer-2	
	Kappa	*P*	Kappa	*P*
Cochlear hydrops	851	<0.001	366	0.001
Vestibular hydrops	964	<0.001	332	<0.001
Perilymphatic enhancement	908	<0.001	590	<0.001
Round window sign	787	<0.001	398	0.002

Table 3. Comparison of magnetic resonance imaging findings between symptomatic and asymptomatic sides of the cases

Symptomatic side/asymptomatic side		None	Present	*P*
Cochlear hydrops	**None**	38 (59.4)	2 (3.0)	<0.001
	Present	17 (26.6)	7 (11.0)	
Vestibular hydrops	**None**	18 (28.1)	2 (3.0)	<0.001
	Present	26 (40.7)	18 (28.2)	
Perilymphatic enhancement	**None**	41 (64.2)	2 (3.0)	<0.001
	Present	21 (32.8)	0	
Round window sign	**None**	54 (84.6)	2 (3.0)	0.001
	Present	2 (3.0)	6 (9.4)	

Good agreement was achieved for the RWS, which had not been evaluated for interobserver agreement in previous studies. The acceptable level of agreement in identifying the RWS–a clinically important abnormality that can mimic MD and is treatable through surgery–is noteworthy. Since the measurement site is defined and quantitative, P–P measurements also demonstrated very good agreement.

The experienced radiologist's intraobserver agreement for EH and PE was higher in our study than in the study by Bernaerts et al.[18], which used the TSE FLAIR sequence, achieving very good agreement. Agreement for the RWS was good. In contrast, the general radiologist's intraobserver agreement was low for EH and the RWS and moderate for PE, indicating that MR evaluation of hydrops in general radiology practice–without more quantitative criteria or higher-resolution sequences–remains a challenge.

Generally, kappa values above 0.60 are considered sufficient, but ideally, a value above 0.70 is targeted, as this indicates "good agreement" in medical studies.[21-23] When good agreement is considered the acceptable threshold, the values obtained in this study are insufficient. The low intraobserver agreement for hydrops imaging by the newly certified radiologist (O-2) suggests that limited experience in diagnosing MD may lead to clinically important discrepancies. Since hydrops imaging is not yet routinely integrated into clinical practice, the need for specialized training to ensure accurate application of this technique becomes apparent. In this context, structured training or mentorship programs led by experienced radiologists in centers that perform hydrops assessments, along with technological solutions to enhance reliability–such as automatic classification algorithms or artificial intelligence (AI)-supported analysis systems–should be developed. These findings highlight the importance of targeted education and structured training for radiologists, particularly in interpreting EH imaging. The relatively low intraobserver agreement observed in the newly certified radiologist underscores a gap that could be addressed through formalized curricula. Recent national-level data also indicate that radiology residents, despite having high awareness of technological terms such as AI and advanced imaging methods, often lack formal training and hands-on experience.[24] International standards–such as the European Training Curriculum developed by the European Society of Radiology and the curriculum of the European Society of Head and Neck Radiology–advocate for subspecialty-level training in head and neck imaging. Incorporating EH imaging into such curricula, particularly with standardized assessment protocols and hands-on case review, could enhance diagnostic reproducibility and clinical confidence among radiologists at various stages of training. This approach would support the routine use of hydrops imaging by improving diagnostic accuracy and reliability.

In this study, cochlear and vestibular hydrops and PE were considerably more frequent on the symptomatic side than on the asymptomatic side. Bernaerts et al.[19] demonstrated that cochlear PE and vestibular EH are the two most distinguishing features for differentiating symptomatic ears from asymptomatic ones. This finding is confirmed by the study of Van Steekelenburg et al.[25] In cases of dMD, the rate of grade 2 cochlear hydrops was higher on the affected side than in cases of pMD. Similarly, the rate of grade 3 vestibular hydrops was higher in dMD than in pMD, consistent with previous studies.[9]

EH does not always cause MD symptoms, and not all patients diagnosed with MD have EH.[16] In this study, 15 of 64 (23.4%) symptomatic ears (13 pMD, 2 dMD) had neither EH nor PE. Previous studies also failed to detect EH by MRI in 10%–33% of patients with MD.[13,26,27] Of these cases, 3 of 15 (20%) had the RWS (2 pMD, 1 dMD). On the asymptomatic side, 14.3% had cochlear hydrops, and 31.2% had saccular hydrops. These rates are higher than those reported in previous studies, which showed a maximum of 6.7% and 8.3%, respectively.[9,28] Differences in grading systems, such as those used by Ito et al.[28], may explain this discrepancy. The relatively high rate of hydrops detected on asymptomatic sides in our study, compared with previous reports, may also be explained by differences in patient selection criteria, sample composition, or hydrops grading systems. Nevertheless, it is notable that no cases of grade 3 hydrops were identified on the asymptomatic side in our cohort.

Perilymphatic fistula (PLF) can mimic MD symptoms, and it is extremely difficult to differentiate it from MD clinically, especially in the case of pMD.[29] The RWS has been previously described as a localized high signal covering the round window on 3D-FLAIR, indicating PLF.[16,30] PLF's clinical symptoms are highly variable and non-specific. Motion intolerance, fluctuating hearing loss, dizziness with or without true vertigo, tinnitus, and aural fullness are the most common symptoms. Symptoms can worsen with changes in pressure (such as during air travel, mountain climbing, rapid elevator rides, bending, lifting heavy objects, coughing, or sneezing) due to increased cerebrospinal fluid pressure. Although initial symptoms may be either entirely auditory or vestibular, many patients develop both types of symptoms over time.[16] Unlike most other causes of sensorineural hearing loss and dizziness, PLF can be corrected surgically by repairing the fistula.[31,32] Based on the results of this study, the RWS, potentially representing PLF, was recorded in 8/64 (12.5%) cases, mostly in pMD cases (11.6%, 5/43). The majority of RWS findings in the study by Dubrulle et al.[16] were also seen in patients with pMD.

In this study, the RWS was also observed in 8 asymptomatic sides (6 bilateral MD cases). However, EH accompanied the RWS in 4 asymptomatic sides (3 vestibular), with no PE in any of them. In the study by Attyé et al.[30], 2/30 healthy volunteers also had the RWS.[29] Our study lacked a healthy volunteer group, so asymptomatic sides were used as controls, and EH was observed in 34% of asymptomatic sides, mostly low-grade. No soft tissue or hyperintensity suggesting inflammation in the round window niche was observed on the FIESTA sequence corresponding to the RWS in any case.

Unilateral defunction therapy is increasingly being used in patients requiring deep vestibular deafferentation to effectively control vertigo symptoms.[33] In our study, two patients with dMD had a history of intratympanic gentamicin application. One patient had bilateral RWS (on both the symptomatic side where the application was performed and the asymptomatic side), whereas the other had bilateral grade 1 vestibular hydrops. Attyé et al.[30] reported a correlation between the presence of PLF and a history of intratympanic gentamicin application. However, in our case, the RWS was observed on the asymptomatic side where no injection had been administered. Additionally, their study reported that three patients with vestibular hydrops also had PLF, and all had a history of intratympanic gentamicin application. In our second case, a similar vestibular hydrops was observed, but it was also present on the asymptomatic side. Therefore, contrary to the findings of their study, in both of our cases, the RWS and hydrops were detected on the asymptomatic side. Further studies with larger patient and control groups are needed on this topic.

In MD, aside from EH, retrolabyrinthine bone thickness has also been investigated. One study found that the distance between the vertical portion of the posterior semicircular canal and the posterior fossa was shorter in patients with unilateral MD than in patients with ipsilateral delayed EH and healthy controls.[10] In our study, in contrast to the study by Lei et al.[10], hydrops imaging was performed alongside heavily T2-weighted anatomical sequences, and dMD and pMD were compared rather than delayed EH cases and healthy controls. The P–P distance was found to be considerably shorter in dMD than in pMD. Recent literature has shown that this distance is associated with the hypoplastic MD endotype.[34,35] This finding supports the role of the hypoplastic endolymphatic sac in the pathogenesis of MD. Hypoplastic retrolabyrinthine bone thickness is proposed as a radiological marker with the potential to specifically identify the hypoplastic endotype of MD. Given the high interobserver agreement rates, incorporating retrolabyrinthine bone thickness measurement into the routine radiological evaluation for diagnosing the hypoplastic endotype of MD may enhance diagnostic accuracy.

There are several limitations to this study. First, the number of patients with dMD was relatively small. Second, the use of different imaging parameters may have complicated the evaluation. Third, the absence of sequences other than 3D-FLAIR may have limited interobserver agreement for hydrops evaluation. However, acquiring additional sequences involves higher costs, and such sequences are not yet widely available. Fourth, this study used clinically normal contralateral ears as the control group, and these control ears showed cochlear hydrops in 9 cases and vestibular hydrops in 20 cases (a total of 22 cases, 34.3%). Given that the 2015 criteria of the Bárány Society[1] still regard clinical, auditory, and vestibular function tests as the gold standard for diagnosing MD–and that MD is clinically limited to 1 ear in most cases–we used normal-appearing contralateral ears as controls. In addition, we employed two different classification systems: the Baráth classification for cochlear hydrops and the Bernaerts classification for vestibular hydrops. Although this approach is consistent with the existing literature and allows for structure-specific grading, it may reduce interpretive consistency and complicate reproducibility.

Lastly, the specificity of the RWS on delayed post-contrast 3D-FLAIR imaging remains a concern. Although the RWS has been proposed as a radiologic indicator of PLF, similar signal enhancement may occur in other inner ear conditions due to alterations in the blood–labyrinth barrier, particularly near the basal turn of the cochlea. As surgical confirmation was not available in our cases, definitive correlation with PLF could not be established. Therefore, the RWS findings should be interpreted with caution and always in the context of clinical data.

In conclusion, the agreement coefficients of the newly certified radiologist trained in hydrops imaging using current criteria were insufficient for evaluating hydrops MRI scans. When considered alongside existing literature, the findings suggest that higher-resolution sequences and more quantitative diagnostic criteria may improve evaluation accuracy. In dMD, the retrolabyrinthine bone is thinner, hydrops is more frequent and advanced, and the RWS is less common. One in five patients clinically diagnosed with MD may have PLF. Asymptomatic contralateral ears, which may also be hydropic, should be closely monitored.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Lopez-Escamez JA, Carey J, Chung WH et al. Diagnostic criteria for Menière's disease. *J Vestib Res.* 2015;25(1):1-7. [CrossRef]
2. Nakashima T, Pyykkö I, Arroll MA, et al. Meniere's disease. *Nature Reviews Disease Primers*. 2016;2:1-18. [CrossRef]
3. Pyykkö I, Manchaiah V, Färkkilä M, Kentala E, Zou J. Association between Ménière's disease and vestibular migraine. *Auris Nasus Larynx*. 2019;46(5):724-733. [CrossRef]
4. Neff BA, Staab JP, Eggers SD, et al. Auditory and vestibular symptoms and chronic subjective dizziness in patients with Ménière's disease, vestibular migraine, and Ménière's disease with concomitant vestibular migraine. *Otol Neurotol.* 2012;33(7):1235-1244. [CrossRef]
5. Brantberg K, Baloh RW. Similarity of vertigo attacks due to Meniere's disease and benign recurrent vertigo, both with and without migraine. *Acta Otolaryngol.* 2011;131(7):722-727. [CrossRef]
6. Kentala E, Pyykkö I. Vestibular schwannoma mimicking Ménière's disease. *Acta Otolaryngol Suppl*. 2000;120:17-19. [CrossRef]
7. Nakashima T, Naganawa S, Sugiura M, et al. Visualization of endolymphatic hydrops in patients with Meniere's disease. *Laryngoscope.* 2007;117:415-420. [CrossRef]
8. Nakashima T, Naganawa S, Teranishi M, et al. Endolymphatic hydrops revealed by intravenous gadolinium injection in patients with Ménière's disease. *Acta Otolaryngol.* 2010; 130(3):338-343. [CrossRef]
9. Li J, Wang L, Hu N, et al. Improving diagnostic accuracy for probable and definite Ménière's disease using magnetic resonance imaging. *Neuroradiology.* 2023;65:1371-1379. [CrossRef]
10. Lei P, Leng Y, Li J, Zhou R, Liu B. Anatomical variation of inner ear may be a predisposing factor for unilateral Ménière's disease rather than for ipsilateral delayed endolymphatic hydrops. *Eur Radiol.* 2022;32:3553-3564. [CrossRef]
11. Sousa R, Lobo M, Cadilha H, Eça T, Campos J, Luis L. Is there progression of endolymphatic hydrops in Ménière's disease? Longitudinal magnetic resonance study. *Eur Arch Otorhinolaryngol.* 2023;280(5):2225-2235. [CrossRef]
12. Li J, Jin X, Kong X, et al. Correlation of endolymphatic hydrops and perilymphatic enhancement with the clinical features of Ménière's disease. *Eur Radiol.* 2024;34;6036-6046. [CrossRef]
13. Baráth K, Schuknecht B, Monge Naldi A, Schrepfer T, Bockisch CJ, Hegemann SCA. Detection and grading of endolymphatic hydrops in Menière disease using MR imaging. *AJNR Am J Neuroradiol.* 2014;35(7):1387-1392. [CrossRef]

14. Nakashima T, Naganawa S, Pyykkö I, et al. Grading of endolymphatic hydrops using magnetic resonance imaging. *Acta Otolaryngol Suppl*. 2009;560:5-8. [CrossRef]

15. Bernaerts A, De Foer B. Imaging of Ménière disease. *Neuroimaging Clin N Am*. 2019;29(1):19-28. [CrossRef]

16. Dubrulle F, Chaton V, Risoud M, Farah H, Charley Q, Vincent C. The round window sign: a sensitive sign to detect perilymphatic fistulae on delayed postcontrast 3D-FLAIR sequence. *Eur Radiol*. 2020;30(11):6303-6310. [CrossRef]

17. Osman S, Hautefort C, Attyé A, Vaussy A, Houdart E, Eliezer M. Increased signal intensity with delayed post contrast 3D-FLAIR MRI sequence using constant flip angle and long repetition time for inner ear evaluation. *Diagn Interv Imaging*. 2022;103(4):225-229. [CrossRef]

18. Bernaerts A, Janssen N, Wuyts FL, et al. Comparison between 3D SPACE FLAIR and 3D TSE FLAIR in Menière's disease. *Neuroradiology*. 2022;64:1011-1020. [CrossRef]

19. Bernaerts A, Vanspauwen R, Blaivie C, et al. The value of four stage vestibular hydrops grading and asymmetric perilymphatic enhancement in the diagnosis of Menière's disease on MRI. *Neuroradiology*. 2019;61:421-429. [CrossRef]

20. Deng W, Lin X, Su Y, Cai Y, Zhong J, Ou Y. Comparison between 3D-FLAIR and 3D-real IR MRI sequences with visual classification method in the imaging of endolymphatic hydrops in Meniere's disease. *Am J Otolaryngol*. 2022;43(6):103557. [CrossRef]

21. Sun S. Meta-analysis of Cohen's kappa. *Health Serv Outcomes Res Methodol*. 2011;11:145-163. [CrossRef]

22. Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC Cancer*. 2023;23(1):799. [CrossRef]

23. Sim J, Wright CC. The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257-268. [CrossRef]

24. Emekli E, Coşkun O, Budakoğlu Iİ. The role of the role of artificial intelligence in radiology residency training: a national survey study. *Eur J Ther*. 2024;30(6):844-849 [CrossRef]

25. Van Steekelenburg JM, Van Weijnen A, De Pont LMH, et al. Value of endolymphatic hydrops and perilymph signal intensity in suspected Ménière disease. *AJNR Am J Neuroradiol*. 2020;41(3):529-534. [CrossRef]

26. Kim WB, Pope AR, Sepahdari MN, et al. Blood-labyrinth barrier permeability in Menière disease and idiopathic sudden sensorineural hearing loss: findings on delayed postcontrast 3D-FLAIR MRI. *AJNR Am J Neuroradiol*. 2016;37(10):1903-1908. [CrossRef]

27. Pyykkö I, Nakashima T, Yoshida T, Zou J, Naganawa S. Ménière's disease: a reappraisal supported by a variable latency of symptoms and the MRI visualisation of endolymphatic hydrops. *BMJ Open*. 2013;3(2):001555. [CrossRef]

28. Ito T, Kitahara T, Inui H, et al. Endolymphatic space size in patients with Meniere's disease and healthy controls. *Acta Otolaryngol*. 2016;136:879-882. [CrossRef]

29. Lopez-Escamez JA, Carey J, Chung WH, et al. Diagnostic criteria for Menière's disease. *J Vestib Res*. 2015;25(1):1-7. [CrossRef]

30. Attyé A, Eliezer M, Galloux A, et al. Endolymphatic hydrops imaging: differential diagnosis in patients with Meniere disease symptoms. *Diagn Interv Imaging*. 2017;98(10):699-706. [CrossRef]

31. Foster PK. Autologous intratympanic blood patch for presumed perilymphatic fistulas. *J Laryngol Otol*. 2016;130(12):1158-1161. [CrossRef]

32. Deveze A, Matsuda H, Elziere M, Ikezono T. Diagnosis and treatment of perilymphatic fistula. *Adv Otorhinolaryngol*. 2018;81:133-145. [CrossRef]

33. Junet P, Karkas A, Dumas G, Quesada JL, Schmerber S. Vestibular results after intratympanic gentamicin therapy in disabling Menière's disease. *Eur Arch Otorhinolaryngol*. 2016;273(10):3011-3018. [CrossRef]

34. Bächinger D, Filidoro N, Naville M, et al. Radiological feature heterogeneity supports etiological diversity among patient groups in Meniere's disease. *Sci Rep*. 2023;13(1):10303. [CrossRef]

35. Juliano AF, Lin KY, Shekhrajka N, Shin D, Rauch SD, Eckhard AH. Retrolabyrinthine bone thickness as a radiologic marker for the hypoplastic endotype in Menière disease. *AJNR Am J Neuroradiol*. 2024;45(9):1363-1369. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253406

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Physiological indices for evaluating balloon angioplasty outcomes in below-the-knee artery lesions of patients with chronic limb-threatening ischemia

Murat Canyiğit[1]
Muhammed Said Beşler[2]
Turan Kaya[3]

[1]Ankara Yıldırım Beyazıt University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

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

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

PURPOSE

To assess the potential use of resting distal pressure/aortic pressure (Pd/Pa) and constant resistance ratio (cRR) physiological indices in the treatment of tibial artery lesions with balloon angioplasty in patients with chronic limb-threatening ischemia (CLTI).

METHODS

In this single-center retrospective study, resting Pd/Pa and cRR measurements were performed using a pressure microcatheter after balloon angioplasty. Procedures were conducted using balloons with diameters of 3 and/or 3.5 mm. The optimal group was defined as patients with either resting Pd/Pa or cRR ≥0.9, whereas the acceptable group included those with both values between 0.8 and 0.9. Clinical improvement in patients with rest pain (Rutherford 4) was defined as at least a 1-point category improvement, indicating a reduction or resolution of rest pain.

RESULTS

The study population consisted of 40 patients (75% men; mean age 64 ± 11.2 years), with a follow-up duration of 92 ± 40.5 days. Foot ulcers were present in 90% of the patients. During follow-up, wound healing was observed in 69.7% of patients. The optimal group exhibited higher rates of wound healing and clinical improvement than the acceptable group, although the difference was not statistically significant (80% vs. 50%, $P = 0.151$). No patient required target vessel revascularization. The overall limb salvage rate during follow-up was 94.6%.

CONCLUSION

Short-term follow-up demonstrated favorable rates of wound healing, patency, and limb salvage. The optimal group showed a trend toward improved wound healing and clinical improvement.

CLINICAL SIGNIFICANCE

This study highlights the utility of resting Pd/Pa and cRR as reproducible physiological indices for objectively evaluating the success of balloon angioplasty in below-the-knee arteries in patients with CLTI. Physiological assessment can guide procedural decisions, contributing to improved limb salvage and high patency rates.

KEYWORDS

Below-the-knee, artery, balloon angioplasty, physiological index, constant resistance ratio

Corresponding author: Muhammed Said Beşler

E-mail: msbesler@gmail.com

Received 14 April 2025; revision requested 10 May 2025; last revision received 28 June 2025; accepted 02 July 2025.

Epub: 08.07.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253406

Peripheral artery disease (PAD), observed in nearly 30% of the elderly population, includes a severe form characterized by ischemic rest pain and tissue loss, known as chronic limb-threatening ischemia (CLTI), which carries high risks of mortality and major amputation. In a considerable proportion of these cases, below-the-knee (BTK) arteries are the primary culprit.[1] The predominant treatment approach for BTK artery disease is percutaneous transluminal angioplasty alone.[2] Although technical success in BTK interventions has tradi-

You may cite this article as: Canyiğit M, Beşler MS, Kaya T. Physiological indices for evaluating balloon angioplasty outcomes in below-the-knee artery lesions of patients with chronic limb-threatening ischemia. *Diagn Interv Radiol*. 2026;32(2):208-213.

tionally been assessed using conventional angiography, recent studies have demonstrated the potential utility of evaluating lumen patency with intravascular ultrasound (IVUS).[3,4] The search for new methods to improve the efficacy of endovascular treatment and to enhance patency and limb salvage rates in BTK artery disease remains ongoing.

In coronary interventions, anatomical optimization methods such as IVUS and functional optimization methods such as fractional flow reserve (FFR) are successfully used for post-procedural assessment. Patients with optimal angiographic results and high FFR values tend to experience better clinical outcomes.[5] In the physiological evaluation of stenoses, the gold-standard FFR–obtained under hyperemia induced by vasodilator agents such as adenosine–has shown good correlation with non-hyperemic parameters such as resting distal pressure/aortic pressure (Pd/Pa), which can be measured without pharmacological agents.[6] Additionally, a novel resting physiological index, the constant resistance ratio (cRR), has shown high diagnostic consistency. It is calculated as the average Pd/Pa value measured during constant-resistance periods across five consecutive resting cardiac cycles.[7]

There are only a few studies related to PAD. Physiological indices obtained after endovascular treatment of iliac and superficial femoral artery stenoses have been found to correlate with clinical improvement.[8] Furthermore, in BTK arteries, physiological parameters have been reported to align with standard morphological parameters.[9]

The present study investigates the potential of resting Pd/Pa and cRR measurements in BTK arteries as effective criteria for assessing the success of balloon angioplasty procedures. We hypothesized that the physiological indices resting Pd/Pa and cRR can objectively reflect the functional assessment of post-procedural vessel patency in the BTK arteries of patients with CLTI, thereby indicating procedural success.

Main points

- Resting distal pressure/aortic pressure and constant resistance ratio measurements appear to be feasible and potentially effective physiological indices for assessing procedural success following below-the-knee (BTK) artery balloon angioplasty in patients with chronic limb-threatening ischemia.
- Optimal post-procedural values (≥0.9) of either physiological index may be associated with improved wound healing and clinical outcomes.
- Performing effective balloon angioplasty with the guidance and support of functional assessment may have contributed to improved wound healing in the present study population, which predominantly consisted of severely calcified and complex BTK artery lesions.

Methods

Study design

This single-center, retrospective study was approved by the Ankara Bilkent City Hospital Medical Research Scientific Ethics Review Board (decision number: TABED 2-25-841, date: 22/01/2025) and conducted in accordance with the principles of the Declaration of Helsinki. Due to its retrospective nature, the requirement for written informed consent was waived by the Institutional Review Board. Patients with CLTI (Rutherford category 4–6) who underwent balloon angioplasty for ≥70% stenosis in BTK tibial arteries, with resting Pd/Pa and cRR measurements between June 2024 and November 2024, were included in the study. Two patients were excluded because they were lost to follow-up. During the same period, participants who underwent digital subtraction angiography (DSA) with resting Pd/Pa and cRR measurements for claudication or foot ulcers but had no major BTK artery stenosis were included in the healthy tibial arteries group (Figure 1).

Procedural steps

A 5F sheath was inserted into the ipsilateral common femoral artery using an antegrade approach, followed by the administration of 70 IU/kg of heparin through the sheath. In patients with an estimated glomerular filtration rate of <60 mL/min/1.73 m^2, imaging from the groin to the ankle was performed using carbon dioxide (CO_2). Stenoses or occlusions in the BTK arteries were crossed using a 0.018-inch guidewire (Gladius, Asahi Intecc, Aichi, Japan). Balloon angioplasty was performed using semi-compliant balloons with diameters of 3 and/or 3.5 mm and lengths of 150 or 200 mm, inflated at pressures ranging from 6 to 12 mmHg for 30 seconds (Minerva, Guangdong, China). In patients with lower body habitus, 3 mm balloons were initially preferred.

In all patients, following pressure microcatheter equalization, the microcatheter was advanced to the ankle over the same guidewire and then slowly pulled back at a constant speed to the tibial artery origin. Physiological indices were measured during this process (TruePhysio pressure microcatheter, Insight Lifetech, Shenzhen, China). No vasodilator agents were administered, and all data were obtained under resting conditions.

Study parameters and follow-up examination

In BTK arteries, the resting Pd/Pa was defined as the ratio of the lowest pressure value

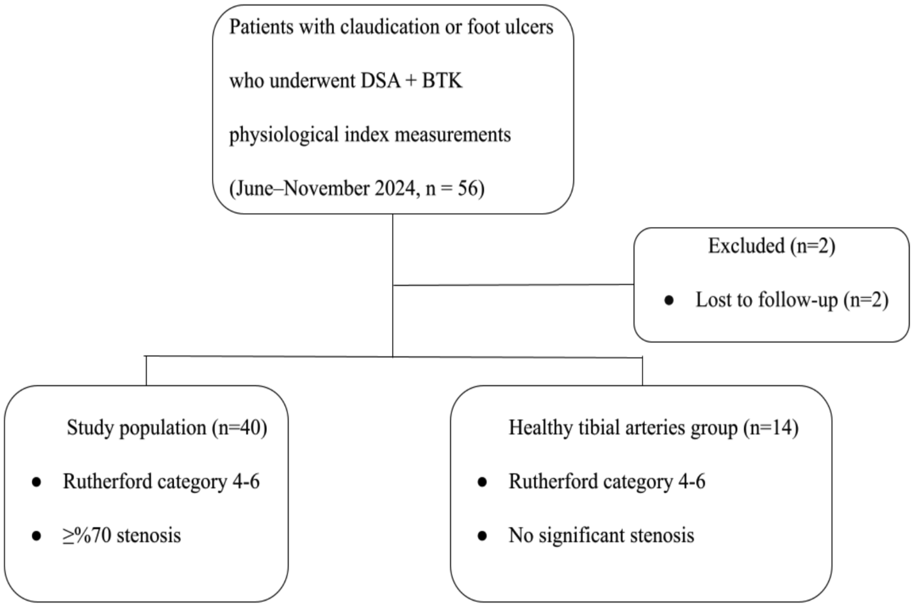


Figure 1. Patient selection flowchart for the study group and the healthy tibial arteries group. DSA, digital subtraction angiography; BTK, below-the-knee.

obtained while the pressure microcatheter was pulled back from the distal tibial artery to its origin, relative to the popliteal artery pressure. A post-procedural resting Pd/Pa and cRR value of >0.8 was targeted. The optimal group was defined as patients with at least one index (resting Pd/Pa or cRR) with a value ≥0.9. A failed procedure was defined as the presence of at least one value ≤0.8. The acceptable group included patients with both resting Pd/Pa and cRR values between 0.8 and 0.9.

The primary endpoint was a composite of complete wound healing in patients with foot ulcers (Rutherford category 5–6) and clinical improvement in patients with rest pain (Rutherford category 4), defined as at least a 1-point improvement in Rutherford category. The secondary endpoint was vessel patency, assessed by Doppler ultrasound during follow-up. At 1st-and 3rd-month follow-up examinations, a peak systolic velocity of <200 cm/s on Doppler ultrasound was considered indicative of no substantial restenosis.

Statistical analysis

Continuous variables were expressed as mean ± standard deviation or median (range). Normality was assessed using the Shapiro–Wilk test. Comparisons of physiological index values in tibial arteries between the study group and the healthy group were performed using the Student's t-test. Wound healing and clinical improvement between patients with optimal and acceptable outcomes were compared using the chi-square test or Fisher's exact test, as appropriate. Student's t-test, Mann–Whitney U test, the chi-square test, or Fisher's exact test was applied, depending on the type and distribution of each variable, to compare baseline characteristics between the optimal and acceptable groups. Three patients who underwent planned amputation and one patient who died before the 1-month follow-up examination were excluded from follow-up analyses. Statistical analyses were performed using SPSS version 26.0 (BM, Armonk, NY, USA), and a *P* value of <0.05 was considered statistically significant.

Results

The study population consisted of 40 patients (75% men; mean age 64 ± 11.2 years), with a mean follow-up duration of 92 ± 40.5 days (Table 1). The healthy tibial arteries group included 14 participants, comprising 18 arteries (85.7% men; mean age 55.5 ± 9.9 years), with 66.7% involving the anterior tibial artery (ATA) and 33.3% the posterior tibial artery (PTA). Foot ulcers were present in 90% of the study patients.

Table 1. Baseline characteristics of the study population (n = 40)

Characteristic	Value
Age, years	64 ± 11.2
Gender, male	30 (75%)
Hypertension	27 (67.5%)
Diabetes mellitus	35 (87.5%)
Chronic kidney disease	18 (45%)
Dialysis	4 (10%)
Coronary artery disease	19 (47.5%)
Cerebrovascular ischemic event	2 (5%)
Dyslipidemia	19 (47.5%)
Smoking, pack × years	20 (range: 0–50)
Current smoker	6 (15%)
Chronic total occlusion	27 (62.8%)
Lesion length, mm	270 (51–335)
Lesion location (PTA/ATA)	17/26
TASC C–D lesions	39 (90.7%)
PACSS 3–4 lesions	30 (69.8%)
Procedure time, min	57.4 ± 15.6
Adjunctive femoropopliteal artery balloon angioplasty	6 (15%)
Iliac artery balloon angioplasty	0 (0%)

Data are presented as mean ± standard deviation, median (range), or number (percentage). PTA, posterior tibial artery; ATA, anterior tibial artery; TASC, Trans-Atlantic Inter-Society Consensus; PACSS, peripheral artery calcification scoring system.

Baseline characteristics were similar between the optimal and acceptable groups (Table 2). At least acceptable outcome values were achieved in all treated cases. Optimal results could not be obtained in 16.3% of the treated arteries due to patients experiencing intolerable pain during balloon angioplasty, which limited the use of larger balloon diameters or higher inflation pressures. A 3.5 mm balloon was directly used in 53.5% of the target vessels, whereas only a 3 mm balloon was used in 14%. In the remaining cases, balloon angioplasty was performed sequentially with 3 mm and then 3.5 mm balloons.

Following balloon angioplasty, resting Pd/Pa and cRR values reached levels comparable to those observed in the healthy tibial arteries group (Table 3). No procedure-related access complications, vascular rupture, flow-limiting dissection, or distal embolization were observed. Final DSA confirmed full patency of the pedal arch in all cases. In 47.5% of the study population, CO_2 angiography was utilized, and only 2 mL of iodinated contrast material–diluted at a 1:4 ratio with saline–was administered for final pedal arch imaging.

Among patients with foot ulcers, complete wound healing was observed in 69.7% during short-term follow-up. The median time to complete wound healing was 45 days (range: 20–150 days). In patients classified as Rutherford category 4, clinical improvement was observed. Although higher rates of wound healing and clinical improvement were noted in the optimal group than in the acceptable group, the difference was not statistically significant (80% vs. 50%, *P* = 0.151).

In three patients, both the ATA and PTA were treated with balloon angioplasty. In one of these patients, the physiological outcome was classified as acceptable, whereas optimal outcomes were achieved in the remaining two. In all cases, the strategy of maintaining at least one straight flow was applied, toward the wound area in patients with ulcers, and toward the foot in those with rest pain.

No substantial restenosis was detected on follow-up Doppler ultrasound, and no patient required target vessel revascularization. The overall limb salvage rate during follow-up was 94.6%. One patient died of a myocardial infarction unrelated to the procedure 3 weeks after discharge.

Table 2. Comparison of baseline characteristics between the optimal and acceptable groups

Characteristic	Optimal group (n = 34)	Acceptable group (n = 6)	*P* value
Age, years	63.9 ± 11.3	64 ± 11.5	0.991
Gender, male	25 (73.5%)	5 (83.3%)	1.000
Hypertension	23 (67.6%)	4 (66.7%)	1.000
Diabetes mellitus	29 (85.3%)	6 (100%)	1.000
Chronic kidney disease	15 (44.1%)	3 (50%)	1.000
Dialysis	2 (5.9%)	2 (33.3%)	0.100
Coronary artery disease	15 (44.1%)	4 (66.7%)	0.398
Cerebrovascular ischemic event	1 (2.9%)	1 (16.7%)	0.281
Dyslipidemia	15 (44.1%)	4 (66.7%)	0.398
Smoking, pack × years	20 (0–50)	10 (0–36)	0.532
Current smoker	4 (11.8%)	2 (33.3%)	0.215
Chronic total occlusion	23 (63.9%)	4 (57.1%)	1.000
Lesion length, mm	270 (51–335)	290 (85–322)	0.964
TASC C–D lesions	33 (91.7%)	6 (85.7%)	0.523
PACSS 3–4 lesions	23 (63.9%)	6 (100%)	0.082

Data are presented as mean ± standard deviation, median (range), or number (percentage). TASC, Trans-Atlantic Inter-Society Consensus; PACSS, peripheral artery calcification scoring system.

Table 3. Comparison of resting Pd/Pa and cRR measurements after balloon angioplasty between the study group and the healthy tibial arteries group

	Study group (n = 40)	Healthy tibial arteries group (n = 14)	*P* value
ATA			
Resting Pd/Pa	0.92 ± 0.04 (n = 26)	0.93 ± 0.04 (n = 12)	0.429
cRR	0.90 ± 0.06 (n = 26)	0.93 ± 0.04 (n = 12)	0.130
PTA			
Resting Pd/Pa	0.94 ± 0.04 (n = 17)	0.95 ± 0.06 (n = 6)	0.828
cRR	0.92 ± 0.05 (n = 17)	0.92 ± 0.03 (n = 6)	0.834
Total			
Resting Pd/Pa	0.93 ± 0.04 (n = 43)	0.94 ± 0.05 (n = 18)	0.524
cRR	0.90 ± 0.06 (n = 43)	0.93 ± 0.03 (n = 18)	0.165

Data are presented as mean ± standard deviation. ATA, anterior tibial artery; Pd/Pa, distal pressure/aortic pressure; PTA, posterior tibial artery; cRR, constant-resistance ratio.

Discussion

The assessment of procedural success following balloon angioplasty for BTK artery lesions is critical for evaluating clinical outcomes. The present study population primarily consisted of patients with CLTI, more than 60% of whom had chronic total occlusions. Over 90% were classified as Trans-Atlantic Inter-Society Consensus C–D, and approximately 70% had peripheral artery calcification scoring system 3–4 complex BTK artery lesions. The majority presented with foot ulcers. Baseline characteristics were comparable between the optimal and acceptable groups. During short-term follow-up, high rates of wound healing, patency, and limb salvage were observed. Efforts were made to achieve optimal resting Pd/Pa and cRR values through balloon angioplasty using high inflation pressures and larger balloon diameters when necessary. The findings suggest that optimizing physiological indices–which enable hemodynamic functional assessment–may contribute to improved wound healing and clinical outcomes.

Physiological index measurements serve as quantitative parameters that eliminate inconsistencies often associated with the visual assessment of angiography in evaluating vessel patency and stenosis.[10] In the TARGET-FFR study, a final FFR value of ≤0.8 was considered indicative of an inadequate procedure, whereas a value ≥0.9 signified an optimal outcome in coronary artery disease.[11] Similarly, in a study on coronary artery lesions, Li et al.[7] reported a cRR threshold of 0.89 as indicative of physiologically relevant stenosis. Lei et al.[12] demonstrated a considerable correlation between resting Pd/Pa and gold-standard FFR measurements in coronary artery evaluations. Another coronary artery study found that resting Pd/Pa independently predicted long-term clinical outcomes.[13] Although standard cut-off values have not yet been established for PAD, the present study adopted cut-off values derived from coronary literature to define optimal and acceptable outcomes.

Adenosine, which induces maximal hyperemia and allows for optimal pressure measurements, is considered the gold standard for FFR assessment. However, it may cause side effects such as bradycardia and hypotension, and it can prolong procedure time.[14] In this study, resting pressure measurements were preferred as they provided effective and appropriate assessments while avoiding potential side effects and procedural delays.

Recent studies have begun to explore FFR measurements in PAD. However, due to the limited number of studies, no standardization or consensus has yet been established regarding hemodynamic physiological assessment in PAD.[15] Kobayashi et al.[16] used FFR to grade residual dissections following balloon angioplasty in the superficial femoral artery, identifying patients who did not require bailout stenting. In a study on the iliac and superficial femoral arteries, FFR values were shown to strongly correlate with calf oxygenation, with high FFR values associated with successful revascularization.[8] Ruzsa et al.[9] reported that high final FFR values in BTK arteries were associated with freedom from re-intervention, major amputation, and mortality. In the present study, the principle of achieving better outcomes with high Pd/Pa and cRR values was adopted, and procedural success was determined based on physiological indices.

In BTK artery angioplasties, approaches involving oversized balloon use–up to 4 mm based on angiographic assessment–and gradual balloon diameter escalation up to 3.5 mm under IVUS guidance have yielded effective outcomes.[3,17] In a vascular reference diameter study conducted in older men, similar to the present study population, tibial artery diameters were reported as 3.8–4.2 mm.[18] In this study, optimal vessel patency in BTK arteries was evaluated using final physiological index measurements. Balloon inflation pressure and diameter were increased until the optimal threshold was achieved. In all patients without small body habitus, direct angioplasty with a 3.5 mm balloon was performed, achieving physiological values comparable to those in healthy tibial arteries for both the anterior and PTAs, without any vascular complications.

Following BTK angioplasty, complete wound healing within 3–4 months has been reported in approximately 60% of cases.[4,19] In the present study, nearly 70% complete wound healing was observed by the mean 3-month follow-up, suggesting that high wound healing rates may reflect the effectiveness of physiological vessel patency assessment. A meta-analysis of BTK lesions reported a 1-year major amputation rate of 5.5% and a 1-year primary patency rate of 50.6% in the balloon angioplasty group.[20] In the present study, despite the majority of lesions being anatomically complex and severely calcified, no target vessel revascularization was required during the short-term (mean 3-month) follow-up after achieving optimal resting Pd/Pa and cRR values.

Despite successful revascularization of large vessels, the presence of residual microvascular disease is known to increase the risk of amputation.[21] In two patients, major amputation was required due to worsening wound progression before the first follow-up examination, which may indicate that revascularization alone was insufficient for wound healing. During follow-up, mortality occurred in only one patient, who suffered a myocardial infarction after discharge, unrelated to the procedure.

In coronary artery disease, microcatheter-derived FFR measurements have been associated with better post-procedural outcomes than conventional angiography. Given the need for repeated pullbacks, microcatheter-based pressure measurements are considered safer than wire-based systems.[22] Unlike other PAD studies, in this study, measurements were performed using a pressure microcatheter advanced over the existing guidewire rather than a pressure wire.[8,9] In nearly half of the study population, the initial assessment was conducted using CO_2 angiography, followed by pressure measurements for procedural success determination and pedal arch imaging with iodinated contrast material. In BTK artery endovascular interventions, iodinated contrast volumes exceeding 100 mL may be required.[4] IVUS-guided anatomical imaging has enabled a near-zero contrast approach in BTK artery balloon angioplasties.[3] The present study demonstrated that pressure measurements contributed to a reduction in iodinated contrast material usage as a secondary outcome.

This study had several limitations. As a single-center, retrospective study, the follow-up period was relatively short. Wound, ischemia, foot infection classification, and body mass index were not included due to incomplete documentation in the retrospective records, which may limit the assessment of wound severity and nutritional status. The identification of constant-resistance periods used in the calculation of cRR is based on hemodynamic assumptions that have not yet been specifically validated in tibial arteries, potentially limiting its physiological applicability in this vascular territory. Because only patients with Rutherford category 4–6 were included, and lower categories were excluded, selection bias was unavoidable. Vasodilator agents were not used to avoid potential side effects; however, future studies incorporating FFR measurements with vasodilator agents may yield more definitive results.

Among the strengths of this study are the inclusion of healthy tibial arteries as a reference group, the determination of normal physiological values, and the objective evaluation of procedural success through classification based on physiological indices. Future multicenter, prospective, randomized controlled trials with long-term follow-up will be essential to further validate these findings. To the best of our knowledge, this is the first study to physiologically evaluate the efficacy and follow-up clinical outcomes of endovascular treatment in BTK arteries using resting Pd/Pa and cRR measurements.

In conclusion, high rates of wound healing and clinical improvement were observed in the optimal group. These findings suggest that high physiological index values may be associated with favorable clinical outcomes, although additional studies are necessary to confirm this association. Resting Pd/Pa and cRR may serve as objective parameters for evaluating the outcomes of endovascular treatment in BTK arteries.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Zilinyi RS, Alsaloum M, Snyder DJ, et al. Surgical and endovascular therapies for below-the-knee peripheral arterial disease: a contemporary review. *J Soc Cardiovasc Angiogr Interv*. 2024;3(3Part A):101268. [Crossref]
2. Singh N, Ding L, Magee GA, Shavelle DM. Contemporary treatment of below-the-knee peripheral arterial disease in patients with chronic limb threatening ischemia: observations from the vascular quality initiative. *Catheter Cardiovasc Interv*. 2022;99(4):1289-1299. [Crossref]
3. Beşler MS, Sözeri A, Canyiğit M. Effectiveness of balloon angioplasty under intravascular ultrasound guidance in calcified below-the-knee tibial arteries. *Diagn Interv Radiol*. [Crossref]
4. Beşler MS, Sözeri A, Canyiğit M. Effectiveness of balloon angioplasty under intravascular ultrasound guidance in calcified below-the-knee tibial arteries. *J Endovasc Ther*. 2020;27(4):565-574. [Crossref]

5. Hakeem A, Uretsky BF. Role of postintervention fractional flow reserve to improve procedural and clinical outcomes. *Circulation*. 2019;139(5):694-706. [Crossref]

6. Leone AM, Campo G, Gallo F, et al. Adenosine-free indexes vs. fractional flow reserve for functional assessment of coronary stenoses: systematic review and meta-analysis. *Int J Cardiol*. 2020;299:93-99. [Crossref]

7. Li C, Wu J, Lin J, et al. Validation of a new non-hyperemic physiological index: the constant-resistance ratio (cRR). *J Invasive Cardiol*. 2024;36(8). [Crossref]

8. Albayati MA, Patel A, Modi B, et al. Intra-arterial fractional flow reserve measurements provide an objective assessment of the functional significance of peripheral arterial stenoses. *Eur J Vasc Endovasc Surg*. 2024;67(2):332-340. [Crossref]

9. Ruzsa Z, Róna S, Tóth GG, et al. Fractional flow reserve in below the knee arteries with critical limb ischemia and validation against gold-standard morphologic, functional measures and long term clinical outcomes. *Cardiovasc Revasc Med*. 2018;19(2):175-181. [Crossref]

10. Malmberg S, Lauermann J, Karlström P, Gulin D, Barmano N. Resting full-cycle ratio versus fractional flow reserve: A SWEDEHEART-registry-based comparison of two physiological indexes for assessing coronary stenosis severity. *J Interv Cardiol*. 2023;2023:6461691. [Crossref]

11. Collison D, Didagelos M, Aetesam-Ur-Rahman M, et al. Post-stenting fractional flow reserve vs coronary angiography for optimization of percutaneous coronary intervention (TARGET-FFR). *Eur Heart J*. 2021;42(45):4656-4668. [Crossref]

12. Lei Y, Liu X, Jiang M, et al. Correlation and consistency between resting full-cycle ratio and fractional flow reserve in assessing coronary artery function in a Chinese real-world cohort with non-ST-segment elevation acute coronary syndrome: a retrospective observational study. *BMJ Open*. 2024;14(8):e082913. [Crossref]

13. Boerhout CKM, de Waard GA, Lee JM, et al. Combined use of hyperemic and non-hyperemic pressure ratios for revascularization decision-making: From the ILIAS registry. *Int J Cardiol*. 2023;370:105-111. [Crossref]

14. Gill GS, Gadre A, Kanmanthareddy A. Comparative efficacy and safety of adenosine and regadenoson for assessment of fractional flow reserve: a systematic review and meta-analysis. *World J Cardiol*. 2022;14(5):319-328. [Crossref]

15. Mangi MA, Kahloon R, Elzanaty A, Zafrullah F, Eltahawy E. The use of fractional flow reserve for physiological assessment of indeterminate lesions in peripheral artery disease. *Cureus*. 2019;11(4):e4445. [Crossref]

16. Kobayashi T, Fujiwara T, Hamamoto M, et al. Mean pressure gradient and fractional flow reserve at a superficial femoral artery dissection after drug-coated balloon angioplasty. *Vasc Endovascular Surg*. 2024;58(8):818-824. [Crossref]

17. Teymen B, Emin Öner M, Erdağ Y. Compensating for angiographic underestimation with oversized balloon angioplasty in patients with chronic limb-threatening ischemia and occluded below-the-knee vessels. *J Endovasc Ther*. 2024:15266028241283534. [Crossref]

18. Lorbeer R, Grotz A, Dörr M, et al. Reference values of vessel diameters, stenosis prevalence, and arterial variations of the lower limb arteries in a male population sample using contrast-enhanced MR angiography. *PLoS One*. 2018;13(6):e0197559. [Crossref]

19. Soga Y, Takahara M, Ito N, et al. Clinical impact of intravascular ultrasound-guided balloon angioplasty in patients with chronic limb threatening ischemia for isolated infrapopliteal lesion. *Catheter Cardiovasc Interv*. 2021;97(3):376-384. [Crossref]

20. Matsuoka EK, Hasebe T, Ishii R, Miyazaki N, Soejima K, Iwasaki K. Comparative performance analysis of interventional devices for the treatment of ischemic disease in below-the-knee lesions: a systematic review and meta-analysis. *Cardiovasc Interv Ther*. 2022;37(1):145-157. [Crossref]

21. Li J, Varcoe R, Manzi M, Kum S, et al. Below-the-knee endovascular revascularization: a position statement. *JACC Cardiovasc Interv*. 2024;17(5):589-607. [Crossref]

22. Biscaglia S, Verardi FM, Erriquez A, et al. Coronary physiology guidance vs conventional angiography for optimization of percutaneous coronary intervention: the AQVA-II Trial. *JACC Cardiovasc Interv*. 2024;17(2):277-287. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253309

INTERVENTIONAL RADIOLOGY

REVIEW

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Short to mid-term outcomes of flow re-direction endoluminal device X (FRED™ X) in the management of intracranial aneurysms: a meta-analysis

Alperen Elek[1]
Gülfem Nur Dindar[2]
Sidar Karagöz[3]
Semra Yücel[4]
Eda Teomete[5]
Celal Çınar[1]
Mahmut Küsbeci[1]
Egemen Öztürk[1]
İsmail Oran[1]

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

[2]Gazi University Faculty of Medicine, Ankara, Türkiye

[3]Dicle University Faculty of Medicine, Diyarbakır, Türkiye

[4]Küre Public Hospital, Kastamonu, Türkiye

[5]University of Michigan, Pre-Medicine, Ann Arbor, USA

ABSTRACT

This meta-analysis evaluates the clinical and angiographic outcomes of the flow re-direction endoluminal device X (FRED™ X) in treating intracranial aneurysms. A systematic review was performed across Medline, Scopus, and Web of Science databases from inception to March 2025. Eligible studies included those reporting clinical and angiographic results of FRED X treatment. Favorable outcomes were defined as those stated explicitly in the studies or a modified Rankin scale score of 0–2. Pooled estimates were calculated using a random-effects model in R. A total of nine studies encompassing 780 patients with 869 aneurysms were included. The weighted mean age was 56.28 years, with 19.1% of patients being men. Most aneurysms were saccular (85.7%), unruptured (92.52%), and located in the anterior circulation (73.6%), primarily in the internal carotid artery. The average aneurysm size was 13.12 mm. All studies employed dual antiplatelet therapy, with antiplatelet response testing performed in eight studies. The mean clinical follow-up period was 9.27 months. The meta-analysis demonstrated favorable neurological outcomes in 97.71% of cases and complete or near-complete occlusion in 86.9%. Procedure-related complications were reported in 9.28% of cases, while in-stent thrombosis or intimal hyperplasia occurred in 4.29%. Overall mortality was low at 0.60%. Subgroup analysis revealed that unruptured aneurysms had a 100% rate of favorable neurological outcomes and an 84.76% rate of complete or near-complete occlusion. Complication and mortality rates were 7.76% and 0.25%, respectively. In addition, favorable outcomes were seen in 100% of ruptured aneurysm cases; however, complete occlusion was achieved in only 59.65%, and the mortality rate was higher at 9.19%. Therefore, FRED X demonstrated high efficacy and procedural safety in the treatment of intracranial aneurysms, offering improved outcomes compared with earlier-generation flow diverters.

KEYWORDS

Intracranial aneurysms, flow re-direction endoluminal device X (FRED X), flow diverter, X technology

In 2011, the approval of a pipeline embolization device by the United States (US) Food and Drug Administration (FDA) marked a significant milestone in the treatment of intracranial aneurysms. Following the success of the pipeline embolization device, flow diverter (FD) stents have become an increasingly integral part of clinical practice. The flow re-direction endoluminal device (FRED™) and its smaller counterpart, FRED junior (FRED Jr), manufactured by MicroVention, now Terumo Neuro (Aliso Viejo, CA, USA), are notable examples of these devices. Designed with a dual-layer construction, they employ a self-expanding braided nitinol mesh to ensure superior wall apposition and facilitate safe delivery to distal aneurysms.[1,2]

An important advancement in flow-diversion technology has been the development of antithrombotic coatings, which aim to reduce the risk of thrombus-related complications. Building on this trend and the original FRED system, FRED X represents the latest generation of FDs. While retaining the structural design of its uncoated predecessor, the FRED X incorporates innovative X technology surface coating. This coating consists of an amphiphilic polymer, poly

Corresponding author: İsmail Oran

E-mail: ismailoran@gmail.com

Received 25 February 2025; revision requested 20 April 2025; accepted 08 June 2025.

Epub: 21.07.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253309

You may cite this article as: Elek A, Dindar GN, Karagöz S, et al. Short to mid-term outcomes of flow re-direction endoluminal device X (FRED™ X) in the management of intracranial aneurysms: a meta-analysis. *Diagn Interv Radiol*. 2026;32(2):214-220.

(2-methoxyethyl acrylate), which features a hydrophobic segment facing the device's surface and a hydrophilic segment oriented toward the vascular lumen.[3,4] This unique design creates a boundary layer along the stent struts, minimizing protein denaturation and subsequent platelet adhesion. These attributes suggest a potentially improved safety profile for the device. The FRED X received FDA premarket approval in September 2021, and the first clinical use was reported in February 2022. However, while preliminary studies suggest high procedural safety with the FRED X, the clinical benefits of these surface coatings remain uncertain and warrant further investigation.[5]

To address this gap, this study provides a comprehensive systematic review and meta-analysis of the existing literature on the clinical outcomes of using FRED X in the treatment of intracranial aneurysms. This meta-analysis aims to identify knowledge gaps and offer valuable insights to inform clinical decision-making regarding this novel therapeutic approach by synthesizing the current evidence and critically analyzing the findings.

Methods

This systematic review and meta-analysis followed the recommendations of the Cochrane Collaboration Handbook for Systematic Review of Interventions[6] and the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) 2020 statement guidelines.[7] The NeuroComp Meta-Analysis Checklist was adopted to assess the outcomes related to complications.[8]

Eligibility criteria

This meta-analysis included studies that met the following criteria: (1) case series, prospective, or retrospective cohort studies, or randomized controlled trials; (2) studies providing follow-up data on clinical and angiographic outcomes; and (3) studies specifically involving the FRED X. Studies using other FRED stents (e.g., FRED, FRED Jr) were excluded from this analysis. There were no restrictions regarding aneurysm type, location, rupture status, or other related factors.

Main points

- The flow re-direction endoluminal device X (FRED™ X) device demonstrated favorable neurological outcomes in 97.71% of cases, with a low overall mortality and complication rate of 0.6% and 9.28%, respectively, highlighting its safety and efficacy in treating intracranial aneurysms.
- Complete or near-complete aneurysm occlusion was achieved in 86.9% of cases, indicating strong angiographic efficacy, even with relatively short follow-up periods.
- The study highlights the potential benefits of FRED X's surface-modifying X technology coating in reducing platelet adhesion and thrombogenicity, although its direct clinical impact remains under investigation.

Search strategy, data extraction, and quality assessment

A comprehensive literature search was conducted in the Medline, Scopus, and Web of Science databases from inception to March 8, 2025. The search strategy employed keywords such as "FRED X," "FREDX," and "FRED™" in various combinations using "AND" and "OR" to ensure a broad capture of relevant studies.

Two independent researchers performed the data extraction and quality assessment; any discrepancies were resolved through consensus with two of the study's researchers. The risk of bias was evaluated using a modified Newcastle–Ottawa Scale.[9-11] Studies were classified as having a "low risk of bias" if they provided satisfactory angiographic and clinical follow-up data together with clear outcome reporting. Studies with unsatisfactory follow-up were categorized as "medium risk of bias," while those lacking satisfactory follow-up and clear outcome reporting were deemed to have a "high risk of bias" (Supplementary Table 1).

Endpoints, subanalysis, and definitions

Clinical and angiographic outcomes assessed included the following: (1) favorable neurological outcomes; (2) minor and major complications; (3) occlusion rate; (4) adjunctive coiling; (5) aneurysm presentation (ruptured or unruptured); (6) technical success; and (7) mortality. Additional subanalyses examined outcomes specifically for ruptured and unruptured aneurysms. Favorable neurological outcomes were defined as either those reported directly or as a modified Rankin scale score of 0–2 for aneurysms. Complete occlusion was directly reported or, if not, classified as Raymond–Roy class 1 (indicating complete obliteration) or O'Kelly–Morata grade D. Near-complete occlusions were defined as Raymond–Roy class 2 or O'Kelly–Morata grade C, with other outcomes classified as incomplete occlusions.

Statistical Analysis

Analyses were conducted using pooled estimates with a 95% confidence interval (CI) under a random-effects model. Heterogeneity was assessed using the I^2 statistic, with values above 50% indicating substantial heterogeneity. To explore the robustness of the pooled estimates and identify potential sources of heterogeneity, leave-one-out sensitivity analyses were performed. Assessment of publication bias was not performed using funnel plots or Egger's regression tests because the meta-analysis comprised fewer than 10 studies, as these methods are considered unreliable with a small sample size. Pearson's χ^2 test examined the relationship between aneurysm localizations. A *P* value <0.05 was considered statistically significant. All statistical analyses were performed using R (version 4.2.3, R Foundation for Statistical Computing, Vienna, Austria), applying inverse variance and restricted maximum likelihood methods.

Results

Literature review

Nine studies were included in the final analysis.[1-5,12-15] The search process is illustrated in Figure 1. Two studies employed a prospective design,[2,12] whereas the remaining seven were conducted retrospectively. Additionally, two studies were conducted at a single center,[12,13] while the other seven were multicenter studies.

Patient population and study characteristics

The nine studies included in the final analysis represented 780 patients with a weighted mean age of 56.28 years and encompassed 869 aneurysms. Among the 780 patients with reported sex information, 149 (19.1%) were men.

Morphologically, saccular aneurysms were the most common type, comprising 680 of 793 aneurysms (85.7%). The mean aneurysm size across the studies was 13.12 mm. Analysis of the rupture status of the 869 aneurysms evaluated determined that 65 (7.4%) were ruptured and 804 (92.5%) were unruptured. The mean clinical follow-up duration of the nine studies included was 9.27 months (Table 1). Detailed information for each study is provided in Table 2.

Aneurysms were predominantly located in the anterior circulation (n = 640, $P < 0.001$), with the internal carotid artery (n = 542) being the most common site. Among posterior circulation aneurysms (n = 103), the vertebral (n = 42) and basilar (n = 26) arteries were the primary locations. Additionally, 126 aneurysms lacked specific location information (Table 3).

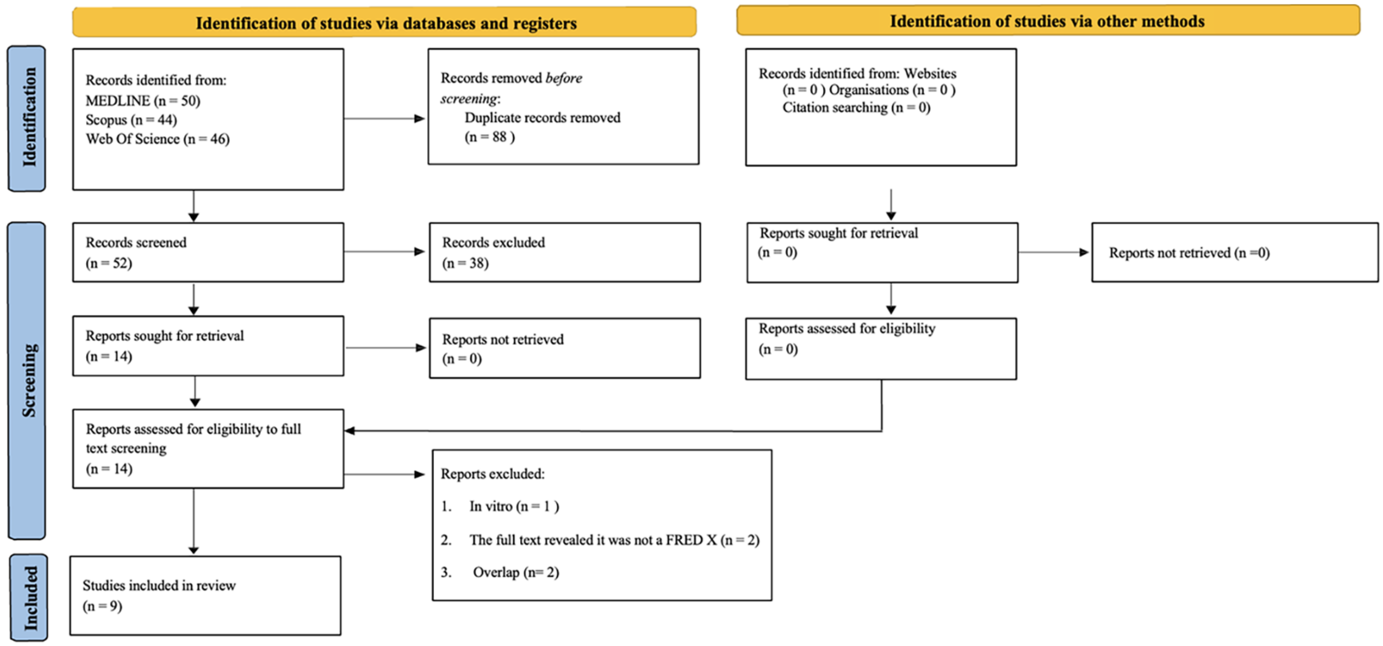

Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from:
MEDLINE (n = 50)
Scopus (n = 44)
Web Of Science (n = 46)
Records removed before screening:
Duplicate records removed (n = 88)
Records identified from: Websites (n = 0) Organisations (n = 0) Citation searching (n = 0)
Records screened (n = 52)
Records excluded (n = 38)
Reports sought for retrieval (n = 0)
Reports not retrieved (n =0)
Reports sought for retrieval (n = 14)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 0)
Reports assessed for eligibility to full text screening (n = 14)
Reports excluded:
1. In vitro (n = 1)
2. The full text revealed it was not a FRED X (n = 2)
3. Overlap (n= 2)
Studies included in review (n = 9)

Figure 1. Prisma flow diagram.

Table 1. Baseline clinical and aneurysm characteristics*

Variable	Raw number (%)*	Number of studies
Number of patients	780	9
Number of aneurysms	869	9
Mean age[†]	56.28	9
Proportion of men	149/780 (19.1%)	9
Previous treatment history	125/772 (16.1%)	7
Mortality	11/780 (1.4%)	9
Clinical status at presentation		
mRS ≤2/good neurologic	594/780 (76.1%)	6
mRS >2/bad neurologic	14/780 (1.79%)	6
NA	87	3
Morphology		
Saccular	680/793 (85.7%)	9
Fusiform	40/793 (5%)	8
Other	73/793 (9.2%)	8
Mean aneurysm size (mm)	13.12	9
Ruptured	65/869 (7.4%)	6
Unruptured	804/869 (92.5%)	9
Mean follow-up (in months)[††]	9.27	9

*Reported as pooled aggregate data and not based on random-effects inverse-variance meta-analysis. [†]Represents weighted average. [††]Based on the last longitudinal clinical follow-up scan. mRS, modified Rankin scale.

Antiplatelet regiment

Dual antiplatelet therapy was utilized in all nine studies. The antiplatelet response was assessed in eight studies, with five using the VerifyNow™ system (Werfen, Bedford, MA, USA), while the method of assessment was not specified in three studies (Supplementary Table 2).

Angiographic, clinical outcomes, and complications

The meta-analysis of the nine studies demonstrated favorable neurological outcomes in 97.7% of cases (95% CI 95.42–100). Complete or near-complete occlusion was achieved in 86.9% of aneurysms (95% CI 79.84–93.60), based on the data reported in eight studies. Complications occurred in 9.2% of cases (95% CI 4.94–13.61), based on the data derived from nine studies. In-stent thrombosis or intimal hyperplasia was observed in 4.2% of cases (95% CI 1.16–7.42) in eight of the studies. Mortality was low, at 0.60% (95% CI 0.00–1.31), based on nine studies (Figure 2).

Subanalysis of ruptured and unruptured aneurysms

For unruptured aneurysms, five studies representing 379 patients reported favorable neurological outcomes in 100% of cases (95% CI 99.31–100) (Supplementary Figure 1). Complete or near-complete occlusion was achieved in 84.7% of cases (95% CI 78.81–90.71), based on five studies representing 331 patients. Complications occurred in 7.7% of cases (95% CI 4.09–11.43), derived from seven studies involving 581 patients. In-stent thrombosis or intimal hyperplasia occurred in 2.1% of cases (95% CI 0–4.21), based on five studies. Mortality was 0.25% (95% CI 0–0.85), based on data from eight studies (Table 4).

For ruptured aneurysms, two studies with 13 patients reported favorable neurological outcomes in 100% of cases (95% CI 87.61–100) (Supplementary Figure 2). Complete or near-complete occlusion was observed in 59.6% of cases (95% CI 25.17–94.13), based on two studies with 23 patients. Procedure-related complications occurred in 0.52% of cases (95% CI 0–6.18), based on four studies involving 41 patients. Mortality was 9.19% (95% CI 0–25.95), based on five studies (Table 4). Details of complications and their

Table 2. Summary of included studies

Study and patient characteristics				Aneurysm characteristics						Follow-up	
Study	Design	Patients/ aneurysms (n)	Sex (M/W)	Aneurysm location (n)	Covered branches (n)	Type (n)	Mean diameter (range), mm	Mean neck width (range), mm	Parent vessel diameter (range), mm	Clinical mean (range), months	Radiological mean (range), months
Abbas et al.[3]	RMC	44/44	7/37	ICA (34), MCA (2), ACA (6), PCOM (2), AICA (1)	15, arising from SAC/ neck of aneurysm	SAC (41) Blister (2) DIS (2)	5.6 ± 4.6	5.6 ± 4.6	NA	6 (1–12)	6 (1–12)
Goertz et al.[1]	RMC	156/156	31/125	ICA (121), ACOM (3), ACA (2), MCA (9), VA (10), BA (7), PICA (3), PCA (1)	4	SAC (138) FU (10) DIS (3) Blister (5)	8.1*5.0	5.0 ± 3.8	Proximal 3.6 ± 0.7 Distal 3.2 ± 0.7	6 (1–12)	6 (1–12)
Hendrix et al.[4]	RMC	101/117	21/90	ICA (102), MCA (3), ACA (8), VA (4)	NA	SAC (112) FU (3) DIS (2)	4.6*3.5	3.5	3.3	6 (1–12)	6 (1–12)
Wen et al.[15]	RMC	22/24	2/20	ICA (10), PCOM (5), ACOM (3), MCA (2), others (4)	NA	SAC (24)	5.7	3.6	NA	21.5 (18–30)	21.5 (18–30)
Clausen et al.[13]	RSC	77/85	21/56	VA (19), ICA (24), BA (3), SHA (5), PCOM (18), Opht (7), ACOM (4), ACA (1) MCA (1), PICA (2), PCA (1)	NA	Saccular (42) Side wall (16) FU (8) DIS (6) Pseudoaneurysm (4) Blister (2) Multiple (2)	0–5 (28) 5–1 (26) >10 (19)	NA	NA	6 (1–6)	6 (1–6)
Roy et al.[14]	RMC	154/162	28/126	ICA (122), MCA (5), ACA (17) V4 (5), basilar (3), P1 (4), AICA (1), SCA (1)	38, arising from SAC/ neck of aneurysm	SAC (140) FU (6) Blister (8) DIS (8)	5.9	3.8	Proximal 3.5 ± 0.9 Distal 3.1 ± 0.7	6 (1–12)	6 (1–12)
Vollherbst et al.[5]	RMC	161/184	36/125	ICA (117), BA (13), MCA (10), VA (9), other (12)	NA	SAC (131) Blister (14) FU (11) DIS (5)	7.8 × 4.7 × 1.7	4.7	Proximal 3.5 ± 0.8 (1.5–7.0) Distal 3.1 ± 0.7 (1.2–5.2)	7 (1–2)	7 (1–12)
Cherednychenko et al.[12]	PSC	7/12	3/4	ICA (12)	NA	SAC (2) Blister (2) FU (1) NS (7)	6 × 8 × 5	NA	NA	7 (1–12)	7 (1–12)
Guimaraens et al.[2]	PMC	58/85	10/48	Carotid artery (44), ACA (5), MCA (13), PCA (3), choroidal artery (6), posterior (14)	NA	Blister (6) DIS (4) FU (2) SAC (60) Pre-treated (13)	7 × 13 × 52	NA	NA	6 (1–12)	6 (1–12)

ACA, anterior cerebral artery; ACOM, anterior communicating artery; BA, basilar artery; CMA, callosal marginal artery; DACA, distal anterior cerebral artery; DIS, dissecting; FU, fusiform; ICA, internal carotid artery; Inf, inferior; IQR, interquartile range; MCA, middle cerebral artery; PCA, posterior cerebral artery; PCOM, posterior communicating artery; PICA, posterior inferior cerebellar artery; M, men; NA, not available; NS, not specified; Opht, ophthalmic artery; RMC, retrospective multicenter; RSC, retrospective single-center; PSC, primary sclerosing cholangitis; SAC, saccular; SHA, superior hypophyseal artery; Sup, superior; VA, vertebral artery; W, women.

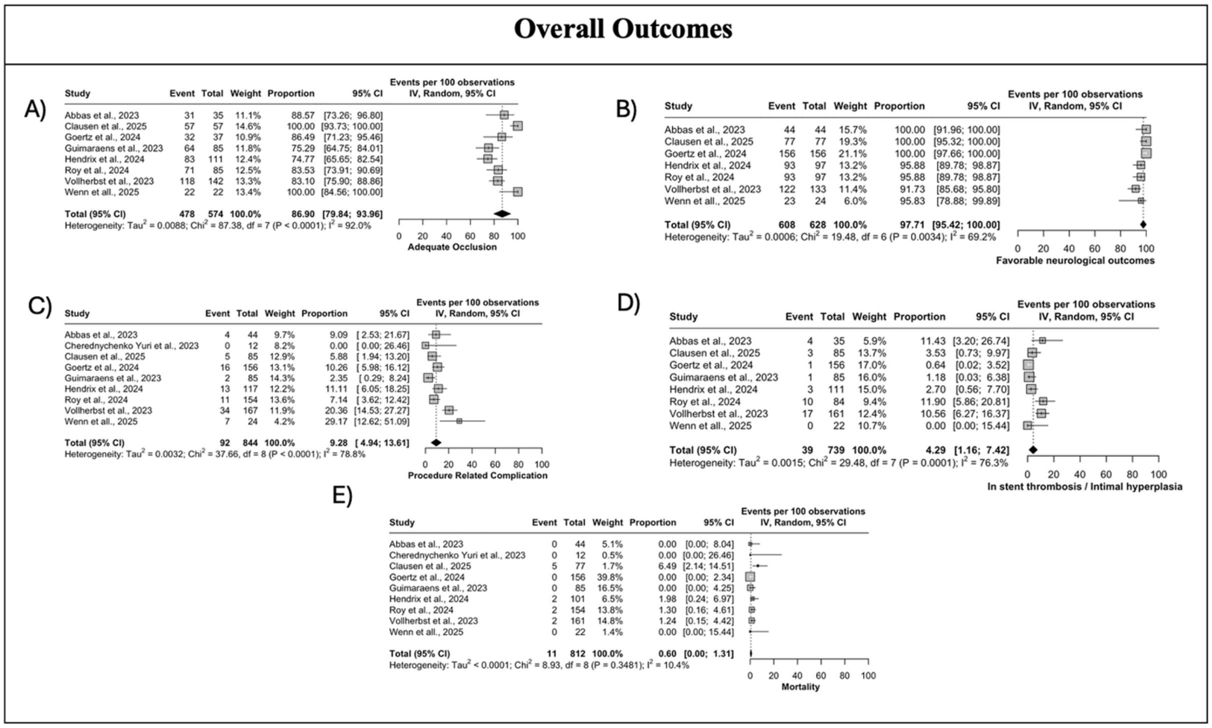


Figure 2. Overall outcomes of the meta-analysis of single-arm studies. Forest plots present pooled event proportions per 100 observations with 95% confidence intervals (CI) using a random-effects model (inverse variance). **(a)** Adequate occlusion. **(b)** Favorable neurological outcomes. **(c)** Procedure-related complication. **(d)** In-stent thrombosis and/or intimal hyperplasia. (e) Mortality. Heterogeneity was assessed using Tau^2, Chi^2, and I^2 statistics. IV, inverse variance.

management are presented in Supplementary Tables 3, 4. Leave-one-out results are presented in Supplementary Figures 3-5.

Table 3. Distribution of aneurysms by location

	Location	*P* value
Anterior circulation (n = 640)	Middle cerebral artery (MCA) (n = 44)	**< 0.001***
	Internal carotid artery (ICA) (n = 542)	
	Choroidal artery (n = 6)	
	Anterior cerebral artery (ACA) (n = 38)	
	Anterior communicating artery (AcomA) (n = 10)	
	Posterior cerebral artery (PCA) (n = 5)	
Posterior circulation (n = 103)	Basilar artery (n = 26)	**< 0.001****
	Vertebral artery (n = 42)	
	Posterior inferior cerebellar artery (PICA) (n = 5)	
	Posterior communicating artery (PcomA) (n = 25)	

*Within the anterior circulation, the number of internal carotid artery aneurysms is significantly higher. **Within the posterior circulation, the number of vertebral artery aneurysms is significantly higher.

Discussion

This meta-analysis, encompassing data for 780 patients with 869 aneurysms reported in nine studies, underscores several key findings. Favorable neurological outcomes were achieved in 97.71% of cases, with a low overall mortality rate of 0.60%. Most aneurysms were unruptured (92.52%) and predominantly located in the anterior circulation, especially in the internal carotid artery. Complete or near-complete occlusion was achieved in 86.9% of aneurysms, while complications occurred in only 9.28% of cases. Although outcomes for ruptured and unruptured aneurysms were generally comparable, unruptured aneurysms exhibited significantly lower complication rates. Despite the relatively short follow-up periods reported in the nine studies, their findings highlight the efficacy and safety of the FRED X device, particularly in achieving high rates of complete occlusion.

While FD technology has revolutionized the treatment of intracranial aneurysms, its use is not without limitations and potential complications. Since FDA approval, multiple studies have assessed the safety and efficacy of FDs, frequently reporting high occlusion rates. However, the overall complication rate, including major and minor events, has been reported to reach up to 17%, with ischemic complications being the most prevalent.[16] Notably, compared with the pipeline embolization device, the FRED device has been associated with higher rates of in-stent stenosis, potentially elevating the risk of ischemic events.[17]

To address these concerns, MicroVention, now Terumo Neuro, developed the next-generation FRED X, incorporating advanced X technology. This technology introduces a protective hydration layer across the stent's surface, aiming to reduce platelet adhesion and enhance endothelialization. By minimizing thrombogenicity and promoting natural vascular healing, the FRED X endeavors to improve safety and clinical outcomes. Furthermore, the novel surface coating facilitates improved device delivery without altering the core stent design. Importantly, the FRED X retains the same dual-layer design as the FRED, which features 16 + 48 wires, achieving 35%–40% metal coverage. In contrast, the FRED Jr, another variant, uses 16 + 36 wires with approximately 30% metal

Table 4. Meta-analysis outcome findings

Overall					
Outcomes	Proportion (95% CI)	I^2 (%)	Q statistic, *P* value	Number of studies	Events/total
Adequate occlusion	86.90 (79.84–93.60)	92	97.38, *P* < 0.001	8	478/574
Favorable neurological outcomes	97.71 (95.42–100)	69.2	19.48, *P* = 0.003	7	608/628
Procedure-related complications	9.28 (4.94–13.61)	78.8	37.66, *P* < 0.001	9	92/844
In-stent thrombosis/intimal hyperplasia	4.29 (1.16–7.42)	76.3	29.48, *P* < 0.001	8	39/739
Mortality	0.60 (0.00–1.31)	10.4	8.93, *P* = 0.348	9	11/812
Adequate occlusion	84.76 (78.81–90.71)	51.6	8.27, *P* = 0.082	5	273/331
Favorable neurological outcomes	100 (99.31–100)	0	0, *P* = 1	5	379/379
Procedure-related complications	7.76 (4.09–11.43)	67.9	18.70, *P* = 0.005	7	49/581
In-stent thrombosis/intimal hyperplasia	2.10 (0–4.21)	44	7.14, *P* = 0.129	5	11/387
Mortality	0.25 (0–0.85)	0	2.36, *P* = 0.937	8	3/633
Subanalysis of ruptured aneurysms					
Adequate occlusion	59.65 (25.17–94.13)	63.8	2.17, *P* = 0.096	2	12/23
Favorable neurologic outcomes	100 (87.61–100)	0	0, *P* = 1	2	13/13
Procedure related complications	0.52 (0–6.18)	0	1.22, *P* = 0.748	4	1/41
Mortality	9.19 (0–25.95)	69.4	13.07, *P* = 0.012	5	6/56

CI, confidence interval.

coverage. The specific impact of the surface coating, however, remains a subject of ongoing investigation.

Retrospective analyses, such as that by Cortez et al.,[18] comparing the Pipeline™ Flex with and without Shield Technology™, found no significant differences in diffusion-weighted imaging lesions. However, the study's small sample size and retrospective design limit the generalizability of these findings, emphasizing the need for further research. An *in vitro* blood loop model study by Yoshizawa et al.[19] demonstrated reduced platelet adhesion on the FRED X surface compared with the uncoated FRED. Additionally, multicenter trials suggest that the FRED X may have a lower complication rate than the uncoated FRED. For example, the SAFE study[20] reported thromboembolic events in 4.9% of cases and a morbidity rate of 3.0%. Moreover, Guimaraens et al.[2] highlighted that the FRED X achieved higher medium-term occlusion rates while maintaining a favorable safety profile.

Despite the widely accepted necessity of dual antiplatelet therapy (typically aspirin and clopidogrel/prasugrel/ticagrelor) following FD implantation, the optimal dosage and duration of post-treatment antiplatelet therapy remain unstandardized when using surface-modified FDs. The reduced thrombogenicity of surface-modified FDs compared with uncoated FDs raises the question of whether adjustments to antiplatelet regimens are warranted. The study by Goertz et al.[21] investigated single therapy following surface-modified FD implantation; aspirin, ticagrelor, and prasugrel were utilized as monotherapies, with prasugrel being the most frequently chosen agent. Evidence suggests that prasugrel and ticagrelor have a superior safety profile for ischemic events compared with aspirin in single antiplatelet therapy (SAPT) regimens. According to a meta-analysis by Ma et al.,[22] the rate of thromboembolic complications following surface-modified FD implantation was reported as 20% for aspirin, compared with 2% and 4% for prasugrel and ticagrelor monotherapies, respectively. These findings suggest that SAPT with prasugrel or ticagrelor is favored over aspirin in specific clinical scenarios when using surface-modified FDs.

Current evidence supports the reasonable safety and efficacy of SAPT in treating aneurysms with modified FDs; however, the lack of studies directly comparing modified FDs with dual antiplatelet therapy and uncoated FDs with SAPT, coupled with the scarcity of large, prospective trials, precludes definitive conclusions. The studies included in this meta-analysis reveal a persistent tendency to favor dual therapy. Results from registries and trials investigating new-generation FDs, such as the coating study,[23] have aimed to optimize aneurysm treatment and contribute valuable insights to refine treatment strategies. Notably, most studies included in this meta-analysis assessed platelet function before the procedure. However, no consensus exists on whether such evaluations are necessary for modified FDs. Further research is required to address this uncertainty.

Despite the short follow-up periods, the meta-analysis established that the FRED X had a satisfactory occlusion rate, with a combined major and minor complication rate of 9.2%. The average aneurysm size was 13.12 mm, a noteworthy finding given that the FRED X was used in small parent vessels and more proximal aneurysms, achieving relatively low complication rates. Larger aneurysms are generally associated with higher complication risks, as demonstrated by the study from Sweid et al.,[24] which identified aneurysm size >10–15 mm as a statistically significantly independent predictor of major ischemic stroke. Additionally, Sweid et al.[24] reported a statistically significant association between the time from treatment and the development of in-stent stenosis.

This meta-analysis did not include sufficient data to directly compare outcomes between FRED X with and without adjunctive coiling. However, Goertz et al.[25] investigated the impact of coiling in their study and found no significant differences in clinical or angiographic outcomes between coiled and non-coiled cases. In the studies included in this meta-analysis, only one patient experienced technical failure.

Although the preliminary safety and efficacy results for FRED X are promising, long-term follow-up studies, particularly those comparing coated and uncoated FDs, are needed to further evaluate its clinical performance.

This meta-analysis has several limitations, despite efforts to address heterogeneity and publication bias. The retrospective design of the included studies introduces inherent constraints, such as selection bias, missing data, and variability in reporting practices, which may affect the reliability and generalizability of the findings. In comparing ruptured and unruptured aneurysms, differences in aneurysm characteristics, such as location and morphology, and variations in treatment approaches, including the use of adjunctive coiling, may further complicate interpretations and reduce comparability. Additionally, the short- to mid-term follow-up periods in most of the included studies may not sufficiently capture long-term treatment outcomes. These factors should be carefully considered when interpreting the conclusions of this analysis.

In conclusion, this study demonstrates that the FRED X offers high feasibility and procedural safety, surpassing the performance of first-generation devices. While the short-term occlusion rates appear satisfactory, long-term and comparative studies are needed to fully evaluate the potential of the FRED X and other coated FDs.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary: https://d2v96fxpocvxx.cloudfront.net/c1dc3a38-51db-436b-af33-1bc7522029b3/content-images/f45ada21-f05d-4d4f-9a55-ece6d42af588.pdf

References

1. Goertz L, Hohenstatt S, Vollherbst DF, et al. Lower-ischemic-risk profile of coated flow redirection endoluminal device x compared with uncoated flow redirection endoluminal device flow diverter in the treatment of unruptured intracranial aneurysms. *Neurosurgery*. 2024. [Crossref]
2. Guimaraens L, Saldaña J, Vivas E, et al. Flow diverter stents for endovascular treatment of aneurysms: a comparative study of efficacy and safety between FREDX and FRED. *J Neurointerv Surg*. 2024;17(e1):e159-e165. [Crossref]
3. Abbas R, Lan M, Naamani KE, et al. First United States multicenter experience with the new-generation FRED X surface-modified flow diversion stent: feasibility, safety, and short-term efficacy. *J Neurosurg*. 2023;140(4):1054-1063. [Crossref]
4. Hendrix P, Hemmer S, Sioutas GS, et al. FRED X flow diversion stenting for unruptured intracranial aneurysms: US multicenter post-market study. *J Neurointerv Surg*. 2025:jnis-2024-022523. [Crossref]
5. Vollherbst DF, Lücking H, DuPlessis J, et al. The FRESH study: treatment of intracranial aneurysms with the new FRED X flow diverter with antithrombotic surface treatment technology-first multicenter experience in 161 patients. *AJNR Am J Neuroradiol*. 2023;44(4):474-480. [Crossref]
6. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane, 2024. Available from: www.training.cochrane.org/handbook. [Crossref]
7. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. [Crossref]
8. Ferreira MY, Cardoso LJC, Günkan A, et al. Challenges and limitations in meta-analyses of complications in neurosurgery: systematic review with proposed approach and checklist to mitigate errors and improve the assessment of the real-world experience. *Neurosurg Rev*. 2024;47(1):722. [Crossref]
9. Cinar C, Elek A, Allahverdiyev I, et al. Comprehensive evaluation of serpentine aneurysms: a systematic review and meta-analysis with a subanalysis for treatment approaches. *Clin Neuroradiol*. 2024;34(4):749-760. English. [Crossref]
10. Elek A, Allahverdiyev I, Ozcinar KK, et al. Comprehensive evaluation of management strategies and rupture status in partially thrombosed aneurysms: a systematic review and meta-analysis. *J Neurointerv Surg*. 2024:jnis-2024-022571. [Crossref]
11. Elek A, Karagoz S, Dindar GN, et al. Safety and efficacy of low profile flow diverter stents for intracranial aneurysms in small parent vessels: systematic review and meta-analysis. *J Neurointerv Surg*. 2025:jnis-2024-022834. [Crossref]
12. Cherednychenko YV, Armonda RA, Sirko AH, Zorin MO, Miroshnychenko AY, Perepelytsia VA. Intracranial aneurysms treatment using new generation FRED X flow diverters with antithrombotic coating and preoperative PreSize Neurovascular software simulation: literature review and own clinical observations analysis. *Ukr Neurosurg J*. 2023;29(3):43-57. [Crossref]
13. Clausen TM, Nakamura R, Conching A, et al. Flow diversion in the treatment of intracranial aneurysms using the novel FRED X device: An early experience from a single high-volume center. *Interv Neuroradiol*. 2025:15910199251319059. [Crossref]
14. Roy JM, El Naamani K, Amaravadi C, et al. Long-term safety and efficacy of the FRED X flow diverter for intracranial aneurysms: a multicenter study of 154 patients. *J Neurosurg*. 2025;1-11. [Crossref]
15. Wen DW, Ayre J, Puthuran M, et al. Treatment of intracranial aneurysms with the FRED X flow diverter stent: mid-term angiographic and safety results. *Clin Neuroradiol*. 2025. [Crossref]
16. Cagnazzo F, Lefevre PH, Derraz I, et al. Flow-diversion treatment for unruptured nonsaccular intracranial aneurysms of the posterior and distal anterior circulation: a meta-analysis. *AJNR Am J Neuroradiol*. 2020;41(1):134-139. [Crossref]
17. El Naamani K, Saad H, Chen CJ, et al. Comparison of flow-redirection endoluminal device and pipeline embolization device in the treatment of intracerebral aneurysms. *Neurosurgery*. 2023;92(1):118-124. [Crossref]
18. Cortez GM, Benalia VHC, Sauvageau E, Aghaebrahim A, Pederson JM, Hanel RA. Diffusion-weighted imaging lesions after intracranial aneurysm treatment with pipeline flex and pipeline flex with shield technology: a retrospective cohort analysis. *J Neurointerv Surg*. 2024;16(4):385-391. [Crossref]
19. Yoshizawa K, Kobayashi H, Kaneki A, et al. Poly(2-methoxyethyl acrylate) (PMEA) improves the thromboresistance of FRED flow diverters: a thrombogenic evaluation of flow diverters with human blood under flow conditions. *J Neurointerv Surg*. 2023;15(10):1001-1006. [Crossref]
20. Pierot L, Spelle L, Berge J, et al. Feasibility, complications, morbidity, and mortality results at 6 months for aneurysm treatment with the flow re-direction endoluminal device: report of SAFE study. *J Neurointerv Surg*. 2018;10(8):765-770. [Crossref]
21. Goertz L, Hohenstatt S, Vollherbst DF, et al. Safety and efficacy of coated flow diverters in the treatment of cerebral aneurysms during single antiplatelet therapy: a multicenter study. *Interv Neuroradiol*. 2024;30(6):819-826. [Crossref]
22. Ma Y, Madjidyar J, Schubert T, Thurner P, Barnaure I, Kulcsar Z. Single antiplatelet regimen in flow diverter treatment of cerebral aneurysms: the drug matters. a systematic review and meta-analysis. *Interv Neuroradiol*. 2023:15910199231177745. [Crossref]
23. Pierot L, Lamin S, Barreau X, et al. Coating (coating to optimize aneurysm treatment in the new flow diverter generation) study. The first randomized controlled trial evaluating a coated flow diverter (p64 MW HPC): study design. *J Neurointerv Surg*. 2023;15(7):684-688. [Crossref]
24. Sweid A, Starke RM, Herial N, et al. Predictors of complications, functional outcome, and morbidity in a large cohort treated with flow diversion. *Neurosurgery*. 2020;87(4):730-743. [Crossref]
25. Goertz L, Styczen H, Siebert E, et al. FRED X flow diverter for the treatment of intracranial aneurysms: two-center experience and mini-review of the literature. *Interv Neuroradiol*. 2024:15910199241246018. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253478

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Efficacy and safety of ultrasound-guided bedside percutaneous cholecystostomy using the transhepatic approach and trocar technique in patients with acute cholecystitis

Ahmet Cem Demirşah[1]
Berat Acu[2]
Çiğdem Öztunalı[2]

[1]Dr. Halil İbrahim Özsoy Bolvadin State Hospital, Clinic of Radiology, Afyonkarahisar, Türkiye

[2]Osmangazi University Faculty of Medicine, Department of Radiology, Eskişehir, Türkiye

PURPOSE

Despite the large number of patients requiring percutaneous cholecystostomy (PC) for acute cholecystitis (AC), no definitive results exist on the optimal imaging guidance modality, technique (Seldinger vs. trocar), or approach [transhepatic (TH) vs. transperitoneal]. This study evaluates the outcomes of ultrasound (US)-guided bedside PC using solely the TH approach and trocar technique in patients with AC.

METHODS

A single-center retrospective study was conducted at a tertiary university hospital between 2018 and September 2023. The study included 81 patients with AC treated with US-guided bedside PC using the TH approach and trocar technique alone. Patients were diagnosed through clinical, laboratory, and radiological examinations, and an experienced interventional radiologist performed the procedures. Outcomes and complication rates were then evaluated.

RESULTS

Technical and clinical success rates were 100% and 93%, respectively. No procedure-related complications occurred. Catheter dislodgement occurred in 4.9% (4/81). The catheter sizes used were 6 F (12.3%), 7 F (40.7%), 8 F (37%), and 10 F (9.9%). The median catheter dwell time was 42 days. Catheters were successfully removed in the majority of surviving patients following resolution of cholecystitis. At the end of the follow-up, 10 patients (12.3%) underwent elective cholecystectomy, and 12 patients (14.8%) died due to comorbidities with the catheter in place.

CONCLUSION

US-guided bedside PC using the TH approach and trocar technique is safe and effective for managing AC in high-risk patients. The study found no significant complications, highlighting the importance of thorough preprocedural evaluation and technique optimization. Further studies with larger, homogeneous patient groups are needed to compare outcomes across different PC techniques and approaches.

CLINICAL SIGNIFICANCE

Despite the growing adoption of PC in the management of AC, the definitive optimal access route and procedural technique remain unresolved. The current body of literature is limited by considerable heterogeneity across studies, including variability in technical approach, operator experience, patient coagulation profiles, and outcome definitions. This study exclusively employed bedside US-guided PC using the TH approach and trocar technique, and observed no procedure-related complications, including hemorrhage, bile leakage, infection, or abscess formation.

KEYWORDS

Acute cholecystitis, percutaneous cholecystostomy, transhepatic, trocar technique, ultrasonography

Corresponding author: Berat Acu

E-mail: beratacu@gmail.com

Received 22 June 2025; revision requested 07 July 2025; accepted 06 August 2025.

Epub: 29.09.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253478

You may cite this article as: Demirşah AC, Acu B, Öztunalı Ç. Efficacy and safety of ultrasound-guided bedside percutaneous cholecystostomy using the transhepatic approach and trocar technique in patients with acute cholecystitis. *Diagn Interv Radiol*. 2026;32(2):221-227.

Acute cholecystitis (AC) is one of the most common causes of emergency department admissions and carries high morbidity and mortality.[1,2] Although the standard treatment for AC is laparoscopic cholecystectomy, surgery carries a high-risk in patients with advanced age or with existing comorbidities. In the advanced age patient group, the surgical complication and mortality rates increase to 14%–30%.[3] Percutaneous cholecystostomy (PC) provides catheter-assisted gallbladder decompression under imaging guidance in high-risk patients. This approach can be used as a temporary or definitive treatment alternative to surgery in a bedside setting, is safe and rapid, and does not require general anesthesia.[3,4]

Ultrasonography, computed tomography (CT), fluoroscopy, or a combination of these modalities can provide the imaging guidance needed in PC procedures. Catheter insertion into the gallbladder lumen can be achieved using the trocar or Seldinger technique. The trocar technique involves directly placing a drainage catheter into the gallbladder cavity under imaging guidance. The Seldinger technique consists of initially accessing the gallbladder lumen with a thin needle and advancing a guidewire through the needle. The access route is dilated using consecutive dilators, and a larger PC drainage catheter is finally placed into the gallbladder lumen. The Seldinger technique is considered more reliable for initial access to the gallbladder because a fine needle is used. However, since this technique requires multiple dilations and over-the-wire exchanges, it is more time consuming than the one-step trocar technique.[5] The technique has also been deemed to carry a higher risk of bile leak and peritonitis.[6-9] The trocar technique, mainly when guided by ultrasound (US), is more operator dependent. It is a single-step, simpler, and quicker technique; however, it uses a larger diameter PC drainage catheter for the initial puncture of the gallbladder, which means it is believed to carry a greater risk.[5,8]

Main points

- The optimal access route for percutaneous cholecystostomy (PC) remains a subject of ongoing debate. This study contributes to the literature by providing outcome data from the exclusive use of the transhepatic (TH) approach and trocar technique in bedside ultrasound (US)-guided PC procedures.
- Bedside US-guided PC using the TH approach and the trocar technique alone is a safe procedure; the study found no procedure-related complications, including bleeding, bile leakage, infection, or abscess formation.
- The technical and clinical success rates of bedside US-guided PC using the TH approach and the trocar technique were 100% and 93%, respectively.
- A thorough preprocedural evaluation of the liver parenchyma and the hemostasis status of the patient, choosing the optimal TH route and the gallbladder puncture site, avoiding re-entries, and using small-caliber catheters may decrease the complication rates when performing US-guided PC using the TH approach and the trocar technique.
- Studies with large sample sizes involving homogeneous study groups regarding the operator experience level, PC technique and approach, patient coagulation status, and catheter sizes are needed to compare well-defined outcomes between different PC procedures.

The gallbladder lumen can be accessed using a transhepatic (TH) or transperitoneal approach (TP).[10] Of the two methods, the TH approach is considered to have a lower likelihood of bile leakage, a lower risk of catheter dislodgement, and a quicker maturation of the drainage route. Traversing the liver parenchyma in the TH approach has been reported to be associated with a higher risk of bleeding, especially when an underlying hepatic pathology is present.[6-9] However, no consensus exists on the optimal PC route. A recent meta-analysis of retrospective studies comparing TH and TP routes in PC in terms of complications concluded that there were confounding factors between these studies, such as the use of both Seldinger and trocar techniques, the variations in the catheter sizes, and the variations in definitions of outcomes and of complications.[6]

Despite the increasing use of PC as a temporary or definitive treatment method for AC, there is a paucity of literature on the optimal approach and technique.[5,6] The choice of PC access route has traditionally depended on operator preference and anatomical considerations; two recent Delphi consensus studies have addressed this issue. The 2024 international Delphi study led by Ramia et al.[11] recommended the TH route as the preferred approach. However, the 2025 Delphi consensus by Pesce et al.[12] accepted both TH and TP routes as viable, emphasizing the role of center-specific expertise and patient anatomy in decision-making. Despite these efforts, no definitive agreement has been reached, and access route selection remains controversial, as highlighted in a recent commentary calling for stronger leadership from interventional radiologists in resolving this debate.[13] Considering the ongoing debate surrounding the optimal access route for PC, this study aimed to contribute to the literature by evaluating the outcomes of US-guided bedside PC performed exclusively using the TH approach and the trocar technique in patients with AC, focusing on technical success, clinical efficacy, complication rates, and clinical and laboratory findings.

Methods

This single-center retrospective study was conducted at Eskişehir Osmangazi University Hospital, a tertiary care university hospital, between 2018 and September 2023. The Eskişehir Osmangazi University Ethics Committee approved the study on June 20, 2023, with decision no.: 24. Written informed consent was obtained from all patients included in the study.

The study included patients with AC treated with US-guided bedside PC using the trocar technique and the TH approach. All patients included were consecutive cases meeting the inclusion criteria during the study period. All patients presented to the emergency department with signs and symptoms of AC. The diagnosis was made through clinical, laboratory, and radiological examinations. Each patient had American Society of Anesthesiology (ASA) scores of 3 or 4, and the consultant general surgeon decided the indication for PC. Patients undergoing PC procedures performed for reasons other than AC (performed during transarterial chemoembolization or ablation procedures) were excluded from the study, as were individuals aged under 18 and pregnant women.

Preprocedural diagnostic abdominal US and CT were performed on all patients included in the study. Volumetric measurements of the gallbladder were obtained through CT and US. For the definition of gallbladder hydrops, a criterion of a transverse diameter greater than 4 cm and a longitudinal diameter greater than 9 cm was used.[14] The presence or absence of gallbladder stones was noted. Procalcitonin, leukocyte, alanine transaminase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and gamma-glutamyl transferase (GGT) values were recorded. Anticoagulant and antiplatelet use was determined in all patients before the procedure. A hemostasis panel was obtained, which included activated partial thromboplastin time (aPTT), international normalized ratio (INR), and platelet count. Care was taken to ensure the platelet count was over 50,000, the INR value was below 1.5, and the aPTT value was within normal limits. Anticoagulant and antiplatelet drugs were discontinued at appropriate intervals according to the Society of Interventional Radiology (SIR) guidelines if suitable.[15,16] In patients with unsuitable hemostasis panels, abnormal coagulation parameters were corrected with fresh frozen plasma and thrombocyte suspension infusion.

Procedures

A single interventional radiologist with 25 years of experience in interventional radiology performed all the procedures, which were conducted at the bedside with US guidance. All PC procedures at our institution are routinely performed using the trocar technique via the TH route, regardless of anatomical variation or complexity. The TH route and trocar technique were deliberately selected based on the interventional radiologist's experience and preference at our institution. To date, the Seldinger technique and the TP approach have not been utilized as part of institutional practice in conjunction with US-guided bedside interventions. Accordingly, all cases included in this study represent the total population of PC procedures performed at our center during the study period. No cases were excluded based on access route or technique.

The patient was placed in a supine or semi-decubitus position. After applying 10 mL of prilocaine to the skin and the liver capsule, access to the gallbladder was achieved transhepatically using an intercostal approach. This approach was specifically chosen to access the gallbladder because all patients in the study were monitored in the intensive care unit (ICU) and used abdominal muscle support during breathing, which may increase the risk of catheter dislodgement. Additionally, this approach is considered by the operator to provide a more appropriate intraparenchymal course, thereby improving catheter stability and decreasing the risk of dislodgement.

A convex US probe (Samsung HS 50™, Samsung Medison Co., ,Seoul, South Korea) was used for imaging guidance, and an out-of-plane freehand technique was used for the access. A 6–10-F trocar-locked pigtail catheter (SKATER™, Argon Medical Devices™, Frisco, TX, USA) was placed into the gallbladder lumen. Catheter caliber was selected at the discretion of the interventional radiologist, considering gallbladder distension, the composition of biliary contents, and catheter availability at the time of the procedure. A 10-F catheter was specifically reserved for cases with an anticipated risk of clogging due to the presence of thick, viscous, or tumefactive content-sludge.

Following confirmation of the pigtail shape of the catheter within the lumen with US guidance, the catheter was locked (Figure 1). A 5-mL lumen sample was obtained for bacterial culture, and the catheter was secured to the skin. Spontaneous catheter drainage was allowed following active aspiration of the lumen contents.

No sedation other than local anesthesia (prilocaine, CİTANEST®, AstraZeneca PLC, İstanbul, Türkiye) was administered during the procedures. Intravenous analgesics were administered for post-procedural pain management in all patients, who were monitored in the ICU throughout the post-intervention period.

Follow-up

All patients were followed up for bleeding by monitoring hemoglobin levels. Procalcitonin, leukocyte, ALT, AST, ALP, and GGT values were obtained at least weekly in all patients following the procedure. The catheter was removed in all patients after 4–6 weeks to allow the inflammation to subside and for the catheter tract to mature. This waiting period was necessary for the safe withdrawal of the catheter.[6] The exceptions were patients who underwent cholecystectomy, in whom the catheter was removed intraoperatively, and those who died before catheter removal. Abdominal US was performed in all patients before the removal of the catheter. The catheter removal was only performed if the imaging findings of AC were no longer present, the catheter flow was less than 10 mL per 24 hours, and the patient's clinical and laboratory inflammation findings had subsided. Decisions to remove the catheter were based solely on clinical and imaging criteria, and a clamping test or fistulography was not performed before removing the catheter.

Definitions of outcomes

Technical success was defined as ultrasonography verification of correct PC catheter placement within the gallbladder lumen with subsequent bile aspiration.[17] Clinical success was defined as the gradual subsidence of signs, symptoms, and inflammatory markers during the first 72-hour post-procedural follow-up.[17] Based on the SIR classification, the complications were categorized as minor or major.[18] Tube dislodgement was defined as the dislodgement of the pigtail catheter from the gallbladder lumen, whether or not remaining in the patient. Catheter removal caused by the patient pulling it out was not included in the definition of catheter dislodgement.[6] Bleeding was defined as fluid or hematoma in the extracapsular, subcapsular, or subcutaneous area at the level of the insertion or in the gallbladder bed in immediate post-procedural ultrasonography. Bile leakage was defined as fluid around the catheter, gallbladder, or liver on immediate or any follow-up post-procedural ultrasonography. A wound infection was defined as a skin infection of the PC insertion site, and an abscess was defined as a localized skin infection requiring incision and drainage. Mild skin erythema at the PC insertion site was not defined as a skin infection.[6]

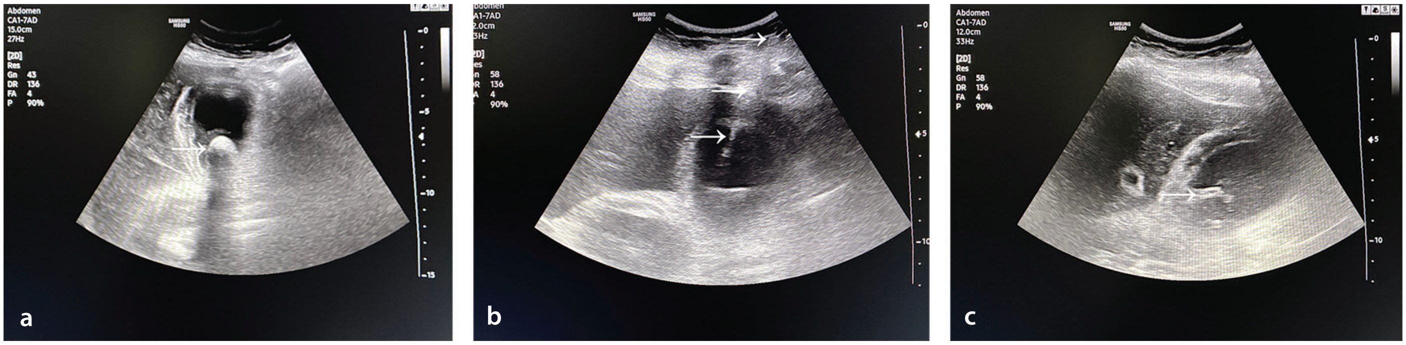


Figure 1. Representative grayscale ultrasonographic images showing the transhepatic route of the catheter during trocar technique-based percutaneous cholecystostomy. **(a)** Hyperechoic focus with posterior acoustic shadowing is visible within the gallbladder lumen, consistent with a gallstone (white arrow). **(b)** The echogenic catheter (white arrows) is seen traversing the liver parenchyma toward the gallbladder, confirming the transhepatic access. **(c)** The pig-tail catheter tip (arrow) is visualized within the gallbladder lumen, indicating successful placement.

Statistical analysis

The data was analyzed using IBM SPSS for Windows 11 (IBM, Armonk, NY, USA). The Shapiro–Wilk test was used to determine the variables' suitability for normal distribution. In summarizing the data, number and percentage statistics were used for qualitative data and the median for quantitative data.

Results

The study included 81 patients (40 women and 41 men) treated with US-guided PC using the TH approach and trocar technique alone (mean age: 75.3). Table 1 summarizes the demographic characteristics, preprocedural laboratory findings, catheter specifications, microbiological culture results, clinical outcomes, and complication rates of the study population.

The technical and clinical success rates for the PC procedures in the study were 100% and 93%, respectively. No procedure-related complications—including bleeding, bile leak, skin infection, or abscess formation—occurred during the immediate post-procedural period or the patients' follow-up. In 4 patients (4.9%), dislodgement of the cholecystostomy catheter was observed. The 30-day and 90-day mortality rates in the study population were 9.8% (8/81) and 14.8% (12/81), respectively.

Fifty-five patients had calculous AC, whereas 26 had acalculous AC. Notably, all of our patients were hospitalized in the ICU at the time of PC, a condition known to increase the risk of developing acalculous AC.

Seventy-four patients had a hydropic gallbladder at the time of admission. The median gallbladder volume was 165 mL. The microbiological culture results for the gallbladder aspiration material were available for 60 of 81 patients. The results showed *Escherichia coli* in 20 of 60 patients and bacteria other than *E. coli* in 26. In 16 of 60 patients, cultures of the aspiration material did not show any microbiological agent.

A 6-F catheter was used in 10 patients, a 7-F catheter in 33 patients, an 8-F catheter in 30 patients, and a 10-F catheter in 8 patients. In patients discharged and followed up in outpatient clinics, the median duration of the catheter stay was 42 days.

The median time between the emergency department admission and the PC procedure was 2 days (min: 1 day; max: 34 days). In 10 patients, the PC procedure was used as a bridge treatment before cholecystectomy. The average time between PC and cholecystectomy was 30.5 days (min: 1 day; max: 52 days).

Twelve patients (14.8%) died during the ward follow-up period. The median post-procedural survival of the patient group who died during this period was 13.5 days. Sixty-nine patients (79.2%) were discharged after a ward stay period and were followed up in outpatient clinics.

Table 1. Patient demographics and laboratory findings

Age	77 (19–94)
Sex	
Male (%)	41 (50.6%)
Female (%)	40 (49.6%)
Median catheter dwell time	42 (1–86)
Mean gallbladder volume (mL)	130 (37–625)
Preprocedural laboratory (mean)	
CRP value (mg/L)	150.5 (0.5–512)
Procalcitonin value (ng/mL)	1.885 (0.04–60.53)
Leucocyte count (10^3/uL)	12.900 (4.200–32.600)
ALT value (U/L)	20.5 (3–3.250)
AST value (U/L)	34 (7–9.560)
ALP value (U/L)	111.5 (51–1.379)
GGT value (U/L)	62 (8–1.799)
Median days of treatment before PC	1 (1–34)
American Society of Anesthesiology grade (median)	3
Catheter diameter frequencies	
6 F	10 (12.3%)
7 F	33 (40.7%)
8 F	30 (37%)
10 F	8 (9.9%)
Microbiological culture results	
Microbiological culture data not available	21 (26%)
Sterile microbiological culture	17 (21%)
Escherichia coli	15 (19%)
Bacterial growth other than *E. coli*	28 (35%)
Cholecystitis	
Calculus	55 (67%)
Acalculous	26 (32%)
Perforation	4 (5%)
Ascites	0 (0%)
Clinical success	
Yes	76 (94%)
No	5 (6%)
Complications	
Total	4 (4.9%)
Catheter dislodgement	4 (4.9%)

ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALP, alkaline phosphatase; CRP, C-reactive protein; GGT, gamma-glutamyl transferase; PC, percutaneous cholecystostomy.

Discussion

PC for calculous or acalculous AC has proven effective and safe in patients with old age or multiple and significant comorbidities.[5,6,8] The procedure's technical and clinical success rates range between 98.9% and 100%, and 85.6% and 97.5%, respectively,[17,19-22] and its complications are minor, with low occurrence rates.[5,6] Despite the large number of patients requiring PC for AC and the increasing use of PC in these patients, no definitive results exist regarding the optimal imaging guidance modality (US, CT, or fluoroscopy), technique (Seldinger vs. trocar), or approach (TH vs. TP).

In the PC procedures in the present study, ultrasonography was preferred as the sole imaging guidance modality because it allows for the procedure to be performed entirely at the bedside and provides urgent and quicker treatment without patient transportation and mobilization issues.

Although both the trocar and the Seldinger techniques are widely used for PC, recent prospective randomized trials have demonstrated that the former is at least non-inferior, and in some outcomes possibly superior, to the latter technique in terms of complication rates, procedure time, and bile leakage risk.[5,6,23] In the present study, the trocar technique was deliberately selected due to its compatibility with US-only guidance, which allows for bedside application without the need for patient transport or fluoroscopy. Additionally, in the setting of AC, where inflammation and wall necrosis increase vulnerability to injury, avoiding multiple tract dilations—as required in the Seldinger method—may offer a technical advantage. Therefore, this study aimed to further evaluate the safety and efficacy of the US-guided bedside trocar technique combined with the TH approach, without comparison to the Seldinger method.

Despite the larger diameter of the initial puncture in the trocar technique, we did not observe any minor or major bleeding complications in any PC procedures. This was in accordance with one prospective study comparing the complication rates between US-guided trocar and US-guided Seldinger PC techniques.[5] The study found the minor bleeding (bile mixed with blood) rate to be as low as 2% (1 out of 50 patients) in each technique, and there was no statistically significant difference in the occurrence of minor bleeding events between the two techniques.[5] The size of the drainage catheter used may be a factor affecting the bleeding complication rates in PC.[24] Using a small-caliber 6-F or 7-F catheter in more than half of the patients in the present study (6 F in 10 patients and 7 F in 33 patients) may have contributed to the lack of bleeding complications. However, the approach (TH vs. TP) used in PC procedures and the operator experience may also affect the bleeding complication rates.[5,6] In PC procedures performed via the TH route, choosing the optimal tract—such that the tract is short enough to avoid bleeding and long enough to allow for tract maturation—may depend on the operator's experience.

Several retrospective observational studies have reported the effect of the approach (TH vs. TP) on the complication rates in PC procedures.[7,25,26] A recent meta-analysis found that although the overall rate of bleeding complications was higher in the TH approach, the studies included in the analysis had significant differences in the technique used (trocar vs. Seldinger), the catheter size used, the number of patients, the number and the experience of the operators, and the definition of the bleeding (e.g., bleeding as visible hemorrhage at the tube site occurring following discharge, bleeding requiring immediate intervention, bleeding as gallbladder hemorrhage occurring in the immediate periprocedural period).[6] A recent multicenter retrospective study, the MACAFI study, comparing the results of the TH and TP approaches in PC in a total of 913 patients, found a significantly greater rate of intraprocedural bleeding in the TH approach than in the TP approach (2.6% vs. 0.3%).[8] However, the MACAFI study also had a heterogeneous study population regarding the technique used; most cases were performed using the Seldinger technique due to safety preferences. The study did not find a significant association between tube size and intraprocedural bleeding rates; however, most cases in both TH and TP groups were performed using an 8 F or larger catheter, with the catheter size ranging between 6 and 12 F. Moreover, no records were available on the risk factors for bleeding (e.g., underlying liver disease, abnormal hemostasis panel, anticoagulant use, decision or not to correct periprocedural coagulopathy). The periprocedural coagulation status of the patients, the presence of any underlying liver disease, and the number of re-entries may significantly affect the bleeding outcome when performing PC via the TH approach. Therefore, the present study's lack of bleeding complications may be related to the fact that there were no patients with underlying liver disease, abnormal coagulation parameters were corrected pre-procedurally in all patients, and no re-entries were performed during the procedures.

Regarding the risk of bile leakage, the TH approach has been associated with less risk than the TP approach, mainly due to the tampon effect exerted by the liver parenchyma.[25] However, the retrospective studies comparing the bile leakage rates between the TH and TP approaches were not homogeneous in the technique used. In Beland et al.'s[7] study, the trocar technique was used in 69.5% and 34.9% of the cases performed via the TP and TH approaches, respectively. In Bennett et al.'s[9] study, 79 of the 165 cases were performed using the trocar technique; however, no information was given on how many of the cases were managed with the trocar technique in conjunction with the TH approach. The MACAFI study reported using the Seldinger technique in "most" cases.[8] Although confounded by using two different techniques and different operators with varying levels of experience, previous retrospective studies found no statistically significant difference in bile leakage rates between the TH and TP approaches.[6,8]

Two prospective studies compared the bile leakage rates between the trocar and the Seldinger techniques in PC. Reppas et al.[23] found a higher rate of bile leakage in US- and fluoroscopy-guided PC procedures performed with the use of the Seldinger technique than in US-guided PC procedures performed with the use of the trocar technique (bile leakage occurred in 4 of 52 cases performed using the Seldinger technique vs. 0 of 53 cases performed using the trocar technique). Arkoudis et al.[5] reported one biloma in 50 patients who underwent US-guided PC using the Seldinger technique. In contrast, no cases of bile leakage were observed in the 50 patients who underwent US-guided PC using the trocar technique.[5] The authors of the two studies concluded that the US-guided trocar technique in PC is as safe as the Seldinger technique, if not safer.[24] It is worth noting that in PC procedures performed via the TH approach using the trocar technique, the gallbladder puncture site can also affect the risk of bile leakage. To use the tamponage effect of the liver parenchyma, puncturing the gallbladder wall at its corpus rather than at its fundus or infundibulum may reduce the risk of bile leakage.

The TH approach in PC has been considered less prone to catheter dislodgement than the TP approach due to the liver's support and traction effect.[8] Few studies evaluating the outcome of catheter dislodgement varied in their definitions of "dislodgement," and some included pulled-out catheters in the category of dislodgement. Excluding the pulled-out catheters, a meta-analysis of four studies on the incidence of catheter dislodgement in PC procedures found no statistically significant difference in catheter dislodgement rates between the TH and the TP approaches. Dislodgement was reported in a total of 15 of 361 cases performed using the TH approach (4.1%) compared with a total of 17 of 311 cases performed via the TP approach (5.4%).[6] The catheter dislodgement rate in the present study was 4.9%.

The present study found the 30-day and 90-day mortality rates following PC to be 9.8% and 14.8%, respectively. The MACAFI study also reported similar outcomes, with a 30-day mortality rate of 8.7% and a 90-day mortality rate of 13.8% for the TH group.[8] It is important to note that the mortality rate observed should not be directly attributed to PC but rather to the patient's pre-existing health conditions, morbid conditions, advanced age, and the presence of associated sepsis. Additionally, the presence of a PC catheter during the patients' ward follow-ups or at the time of death should not be considered a complication of PC or indicative of treatment failure. These patients, classified as ASA 3 and 4, are not typically planned for surgery, and the PC catheter is present as a definitive treatment at the time of death.[5]

Gandhi et al.[27] conducted a retrospective study involving ICU patients who underwent bedside PC under US guidance. In their cohort, all procedures were performed via the TH route using the Seldinger technique, in contrast to our study, where the trocar technique was employed. Smaller-caliber pigtail catheters (7–8 F) were utilized, and tract dilatation due to the catheter size was omitted. The authors reported a technical success rate of 100% and a clinical success rate of 92.1%, closely aligning with our outcomes. However, 1 patient (1.9%) developed a bile leak, likely due to multiple puncture attempts, and required surgical intervention. Importantly, no major complications were observed. The mean catheter dwell time in Gandhi et al.'s[27] study was 13 days (range: 3–45), which was shorter than in our series. A clamping trial was performed in 3 patients before elective tube removal. These findings underscore notable procedural differences—particularly the choice of access method and catheter management strategies—which may influence complication profiles and clinical outcomes. Comparative studies are warranted to evaluate further the impact of trocar versus Seldinger techniques in critically ill patients requiring bedside PC.[27]

Based on previous experience, the operator in the present study did not perform a clamping test or fistulography before catheter removal, and no recurrent AC cases were detected in any patients. However, the authors of this study consider the clamping test and fistulography to be safer and more objective procedures compared with clinical and imaging findings, and they suggest using them to confirm bile flow before catheter removal.

The main limitations of the present study are its retrospective nature and single-center design with a limited number of patients. All PC procedures at our center are performed using the same approach and technique; thus, a comparative outcome analysis of different approaches or techniques could not be presented. Most of the patients involved in the study had impaired consciousness due to their systemic severe illnesses. Consequently, obtaining valid and objective visual analog scores for pain assessment was impossible, meaning no pain-related data were collected.

In this study, the technical and clinical success rates of US-guided PC were 100% and 93%, respectively, which are at the higher end of the ranges reported in the literature. Differences in clinical success rates across studies may result from how clinical success is defined (e.g., subsidence of imaging and/or laboratory parameters), the timing of the assessments, the interval from patient admission to PC procedure, and the antibiotic regimens used. Additionally, comorbidities and the overall condition of the patients included in the study can influence clinical success rates. However, this study was not designed to assess the factors that could impact clinical success.

In conclusion, despite the increasing use of PC for the treatment of AC, current literature data on the optimal PC technique and the approach are indefinite; most retrospective studies on the subject are heterogeneous in terms of technique and approach, the number and experience level of the operator(s), and the coagulation status of the patients, and have variations in their definitions of outcomes. Therefore, the choice of the PC technique and the approach remains at the operator's discretion on a case-by-case basis.[5,6] The present study on 81 consecutive patients with AC treated by a single operator with bedside US-guided PC using the TH approach and the trocar technique alone found no procedure-related complications, including bleeding, bile leakage, infection, or abscess formation. A thorough preprocedural evaluation of the liver parenchyma and the hemostasis status of the patient, choosing the optimal TH route and the gallbladder puncture site, avoiding re-entries, and using small-caliber catheters may decrease the complication rates when performing US-guided PC using the TH approach and the trocar technique. Further studies with large sample sizes involving homogeneous study groups regarding operator experience level, PC technique and approach, patient coagulation status, and catheter sizes are needed to compare well-defined outcomes between different PC procedures.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Adachi T, Eguchi S, Muto Y. Pathophysiology and pathology of acute cholecystitis: a secondary publication of the Japanese version from 1992. *J Hepatobiliary Pancreat Sci*. 2021;29(2):212-216. **[CrossRef]**
2. Elwood DR. Cholecystitis. *Surgical Clinics of North Am*. 2008;88(6):1241-1252. **[CrossRef]**
3. Little MW, Briggs JH, Tapping CR, et al. Percutaneous cholecystostomy: the radiologist's role in treating acute cholecystitis. *Clin Radiol*. 2013;68(7):654-660. **[CrossRef]**
4. Spira RM, Nissan A, Zamir O, Cohen T, Fields SI, Freund HR. Percutaneous transhepatic cholecystostomy and delayed laparoscopic cholecystectomy in critically ill patients with acute calculus cholecystitis. *Am J Surg*. 2002;183(1):62-66. **[CrossRef]**
5. Arkoudis NA, Moschovaki-Zeiger O, Grigoriadis S, et al. US-guided trocar versus Seldinger technique for percutaneous cholecystostomy (TROSELC II trial). *Abdom Radiol*. 2023;48(7):2425-2433. **[CrossRef]**
6. Abdelhalim G, MacCormick A, Jenkins P, Ghauri S, Gafoor N, Chan D. Transhepatic versus transperitoneal approach in percutaneous cholecystostomy: a meta-analysis. *Clin Radiol*. 2023;78(6):459-465. **[CrossRef]**
7. Beland MD, Patel L, Ahn SH, Grand DJ. Image-guided cholecystostomy tube placement: short- and long-term outcomes of transhepatic versus transperitoneal placement. *AJR Am J Roentgenol*. 2019;212(1):201-204. **[CrossRef]**

8. Jenkins PE, MacCormick A, Zhong J, Makris GC, Gafoor N, Chan D. Transhepatic or transperitoneal technique for cholecystostomy: results of the multicentre retrospective audit of cholecystostomy and further interventions (MACAFI). *Br J Radiol*. 2023;96(1143):20220279. [CrossRef]
9. Bennett S, Shaida N, Godfrey E, Safranek P, O'Neill JR. A comparison of transhepatic versus transperitoneal cholecystostomy for acute calculous cholecystitis: a 5-year experience. *J Surg Case Rep*. 2021;2021(9):410. [CrossRef]
10. Venara A, Carretier V, Lebigot J, Lermite E. Technique and indications of percutaneous cholecystostomy in the management of cholecystitis in 2014. *J Visc Surg*. 2014;151(6):435-439. [CrossRef]
11. Ramia JM, Serradilla-Martín M, Villodre C, et al. International Delphi consensus on the management of percutaneous choleystostomy in acute cholecystitis (E-AHPBA, ANS, WSES societies). *World J Emerg Surg*. 2024;19(1):32. [CrossRef]
12. Pesce A, Ramírez-Giraldo C, Arkoudis NA, et al. Management of high-surgical-risk patients with acute cholecystitis following percutaneous cholecystostomy: results of an international Delphi consensus study. *Int J Sur*. 2025;111(5):3185-3192. [CrossRef]
13. Arkoudis NA, Moschovaki-Zeiger O, Spiliopoulos S. Transhepatic versus transperitoneal access for percutaneous cholecystostomy remains controversial: interventional radiologists must lead the discussion. *Cardiovasc Intervent Radiol*. 2025;48(8):1223-1225. [CrossRef]
14. Sebastian S, Araujo C, Neitlich JD, Berland LL. Managing incidental findings on abdominal and pelvic CT and MRI, Part 4: white paper of the ACR Incidental Findings Committee II on Gallbladder and Biliary Findings. *J Am Coll of Radiol*. 2013;10(12):953-956. [CrossRef]
15. Davidson JC, Rahim S, Hanks SE, et al. Society of Interventional Radiology Consensus Guidelines for the periprocedural management of thrombotic and bleeding risk in patients undergoing percutaneous image-guided interventions—part i: review of anticoagulation agents and clinical considerations. *J Vasc Interv Radiol*. 2019;30(8):1155-1167. [CrossRef]
16. Patel IJ, Rahim S, Davidson JC, et al. Society of Interventional Radiology Consensus Guidelines for the periprocedural management of thrombotic and bleeding risk in patients undergoing percutaneous image-guided interventions—part ii: recommendations. *J Vasc Interv Radiol*. 2019;30(8):1168-1184. [CrossRef]
17. Devane AM, Annam A, Brody L, et al. Society of interventional radiology quality improvement standards for percutaneous cholecystostomy and percutaneous transhepatic biliary interventions. *J Vasc Interv Radiol*. 2020;31(11):1849-1856. [CrossRef]
18. Sacks D, McClenny TE, Cardella JF, Lewis CA. Society of interventional radiology clinical practice guidelines. *J Vasc Interv Radiol*. 2003;14(9):199-202. [CrossRef]
19. Ahmed O, Rogers AC, Bolger JC, et al. Meta-analysis of outcomes of endoscopic ultrasound-guided gallbladder drainage versus percutaneous cholecystostomy for the management of acute cholecystitis. *Surg Endos*c. 2018;32(4):1627-1635. [CrossRef]
20. Pang KW, Tan CHN, Loh S, et al. Outcomes of percutaneous cholecystostomy for acute cholecystitis. *World J Surg*. 2016;40(11):2735-2744. [CrossRef]
21. Atar E, Bachar GN, Berlin S, et al. Percutaneous cholecystostomy in critically ill patients with acute cholecystitis: complications and late outcome. *Clin Radiol*. 2014;69(6):247-252. [CrossRef]
22. Winbladh A, Gullstrand P, Svanvik J, Sandström P. Systematic review of cholecystostomy as a treatment option in acute cholecystitis. *HPB*. 2009;11(3):183-193. [CrossRef]
23. Reppas L, Arkoudis NA, Spiliopoulos S, et al. Two-center prospective comparison of the trocar and seldinger techniques for percutaneous cholecystostomy. *AJR Am J Roentgenol*. 2020;214(1):206-212. [CrossRef]
24. Arkoudis NA, Reppas L, Spiliopoulos S. Image-guided percutaneous cholecystostomy: challenging the norms. *Abdom Radiology*. 2024;49(3):939-941. [CrossRef]
25. Kallini JR, Patel DC, Linaval N, Phillips EH, Van Allan RJ. Comparing clinical outcomes of image-guided percutaneous transperitoneal and transhepatic cholecystostomy for acute cholecystitis. *Acta Radiol*. 2020;62(9):1142-1147. [CrossRef]
26. Loberant N, Notes Y, Eitan A, Yakir O, Bickel A. Comparison of early outcome from transperitoneal versus transhepatic percutaneous cholecystostomy. *Hepatogastroenterology*. 2010;57(97):12-17. [CrossRef]
27. Gandhi R, Gala K, Shariq M, Gandhi A, Gandhi M, Shah A. Bedside ultrasound-guided percutaneous cholecystostomy in critically ill patients-outcomes in 51 patients. *Indian J Radiol Imaging*. 2024;34(2):262. [CrossRef]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253503

INTERVENTIONAL RADIOLOGY

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Pleural tail sign in computed tomography-guided lung biopsy: an imaging predictor of severe pneumothorax requiring chest tube placement

Jacob Jalil Hassan[1]
Jakob Leonhardi[1]
Timm Denecke[1]
Anne Beeskow[1]
Manuel Florian Struck[2]
Anne-Kathrin Höhn[3]
Sebastian Krämer[4]
Armin Frille[5]
Hans-Jonas Meyer[1]

[1]University of Leipzig Faculty of Medicine, Department of Diagnostic and Interventional Radiology, Leipzig, Germany

[2]University Hospital Leipzig Faculty of Medicine, Department of Anesthesiology and Intensive Care Medicine, Leipzig, Germany

[3]University of Leipzig Faculty of Medicine, University Hospital Leipzig, Department of Pathology, Leipzig, Germany

[4]University of Leipzig Faculty of Medicine, University Hospital Leipzig, Department of Thoracic Surgery, Leipzig, Germany

[5]Leipzig University Medical Center, Department of Medicine II, Division of Respiratory Medicine, Leipzig, Germany

PURPOSE

Pneumothorax is the most common complication following computed tomography (CT)-guided percutaneous transthoracic needle biopsy. In severe cases, it may require chest tube placement, which is associated with increased morbidity. The aim of this study was to evaluate the prognostic value of the pleural tail sign (PTS) as an imaging marker for predicting pneumothorax incidence and severity after lung biopsy.

METHODS

A total of 477 patients (mean age 65 ± 11.7 years, 37.2% women) undergoing CT-guided lung biopsies between 2012 and 2021 were retrospectively analyzed in this study. The presence and morphological subtype of PTS-classified as thin PTS or triangular PTS-were assessed on pre-interventional CT imaging. Associations between PTS and pneumothorax outcomes were evaluated using univariate and multivariate binary logistic regression analyses.

RESULTS

No statistically significant association was found between the overall presence of PTS and the incidence of pneumothorax ($P = 0.052$). However, patients with a triangular PTS showed a significantly increased risk of severe pneumothorax requiring chest tube placement (odds ratio: 2.092, 95% confidence interval: 1.097–3.990, $P = 0.025$), whereas a thin PTS did not show a statistically significant effect ($P = 0.456$).

CONCLUSION

Although PTS does not reliably predict overall pneumothorax risk after CT-guided lung biopsy, its triangular subtype may serve as a prognostic imaging marker for identifying patients at increased risk of severe pneumothorax requiring chest tube placement.

CLINICAL SIGNIFICANCE

The identification of a triangular PTS on pre-interventional CT imaging may help to stratify patients at higher risk of severe pneumothorax following CT-guided lung biopsy. This could enable more informed procedural planning, potentially leading to improved patient outcomes.

KEYWORDS

CT, image-guided biopsy, lung biopsy, pleural tail sign, pneumothorax

Corresponding author: Hans-Jonas Meyer

E-mail: hans-jonas.meyer@medizin.uni-leipzig.de

Received 05 June 2025; revision requested 21 July 2025; accepted 10 August 2025.

Epub: 19.11.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.253503

Computed tomography (CT)-guided percutaneous transthoracic needle biopsy (PTNB) is a well-established and minimally invasive procedure for acquiring tissue samples for the histopathologic evaluation of pulmonary lesions.[1] This procedure is especially suitable for peripheral lung lesions that are challenging to access via bronchoscopy.[2] It offers high diagnostic reliability, with a sensitivity ranging from 85.7% to 97.4% and a specificity from 88.6% to 100%, for correct histology tissue sampling.[3]

You may cite this article as: Hassan JJ, Leonhardi J, Denecke T, et al. Pleural tail sign in computed tomography-guided lung biopsy: an imaging predictor of severe pneumothorax requiring chest tube placement. *Diagn Interv Radiol*. 2026;32(2):228-232.

Despite its clinical value, PTNB is associated with various procedure-related complications, of which pneumothorax is the most frequent.[4] A recent meta-analysis reported an average pneumothorax rate of 25.9% following CT-guided lung needle biopsy, with a range between 4.3% and 52.4%.[5] Although many cases of pneumothorax are self-limiting, a subset of patients requires chest tube placement, which can lead to increased morbidity, prolonged hospital stays, and additional healthcare costs. The incidence of post procedural chest tube placements ranges from 2% to 15%.[6]

Various risk factors have been associated with an increased incidence of pneumothorax following CT-guided biopsies, including lesion depth, lesion size, the presence and severity of emphysema, the number of pleural punctures, and patient positioning, as well as lesion heterogeneity described by CT texture analysis.[7-10] Recent investigations have proposed the pleural tail sign (PTS) as a potential predictor of pneumothorax following PTNB.[11,12] The PTS is defined as a linear extension of a pulmonary lesion toward the visceral pleura and is histopathologically associated with interlobular septal thickening caused by tumor infiltration, fibrosis, or desmoplastic reaction.[13,14] It has been hypothesized that lesions with a PTS are more likely to exert traction on the pleural surface or form tethered points between the tumor and pleura.[11,12] During biopsy, this may increase susceptibility to pleural tears or air leakage due to mechanical distortion or reduced pleural compliance.

The aim of this study is to evaluate whether the PTS is associated with the overall risk of pneumothorax and, specifically, with the risk of severe pneumothorax requiring chest tube placement after PTNB. We hypothesized that the PTS represents a high-risk imaging biomarker for clinically significant post-biopsy pneumothorax.

Main points

- The presence of a pleural tail sign (PTS) is not significantly associated with the overall risk of pneumothorax after computed tomography-guided lung biopsy.
- The triangular PTS subtype is significantly associated with an increased risk of severe pneumothorax requiring chest tube placement, whereas the thin subtype is not.
- One in three patients exhibited a PTS (33.3%). Postinterventional pneumothorax occurred in 47.2% of cases, and 11.5% required chest tube placement.

Methods

Patient selection

This retrospective analysis was approved by the Ethics Committee of the University of Leipzig, Germany (register number: 344, approval date: 01.11.2007). A total of 487 patients who underwent CT-guided lung biopsies at our institution between January 2012 and November 2021 were screened for inclusion in the study. Of these, 10 patients were excluded, 8 because the intervention was stopped prior to lung puncture and 2 because of insufficient image quality. A total of 477 patients (37.2% women) with a mean age of 65 ± 11.7 years were included in the final analysis. A flowchart illustrating the patient screening and exclusion for the CT-guided lung biopsies is presented in Figure 1. All CT scans of patients who underwent CT-guided PTNB for pulmonary lesions were systematically reviewed to assess the presence or absence of a PTS. A PTS was defined as an extension from the lung lesion toward the visceral pleura. For cases in which a PTS was identified, further classification was performed based on morphological characteristics. Although various PTS classification systems have been described in the literature,[15-17] we adopted a simplified, clinically applicable two-tier model: a thin PTS, characterized by a linear tail without associated pleural retraction, and a triangular PTS, with pleural retraction at the attachment site. Image evaluation was performed by experienced radiologists with 4 years of general experience, blinded to the clinical outcomes. In addition, lesion size, lesion depth, the number of biopsy samples, and whether the biopsy tract passed through the PTS line were assessed. In a random selection of 30 patients, a senior radiologist with 10 years' experience in radiology and a board-certification in interventional radiology performed a second reading for interreader variability assessment of the PTS classification. Figure 2 provides representative images from the patient cohort.

Computed tomography-guided biopsy procedure

Written informed consent was obtained from all patients at least 1 day prior to CT-guided biopsy. Biopsies were only performed when there was no elevated risk of hemorrhage, as indicated by a platelet count of at least 50,000/mm^3, a partial thromboplastin time of ≤1.5 times the normal value, and a prothrombin time >50%.

Interventions were all performed using the same CT scanner (16-slice CT scanner, Brilliance Big Bore, Philips, Hamburg, Germany). Typical CT parameters were set as follows: 100 kVp; 125 mAs; slice thickness, 1 mm; pitch, 0.9).

Biopsies and chest tube placements, where necessary, were performed by radiologists with at least 2 years' experience in interventional radiology. The procedures, namely positioning and needle pathway, were planned using the latest available CT images. For the biopsies, a needle was inserted in the upper part of the ribs to minimize the risk of hematoma at an angle vertical to the parietal pleura, avoiding lung fissures and large bronchovascular structures, with the needle pathway minimized.

Prior to the biopsies, the skin was disinfected and local anesthesia (10 mL of lidocaine 1%; Xylocitin, Jenapharm, Jena, Germany) was applied. In all cases included in this study, a coaxial 18-gauge biopsy system with a 2-cm-long needle was used (Bard Mission, Bard Medical, Covington, GA, USA,

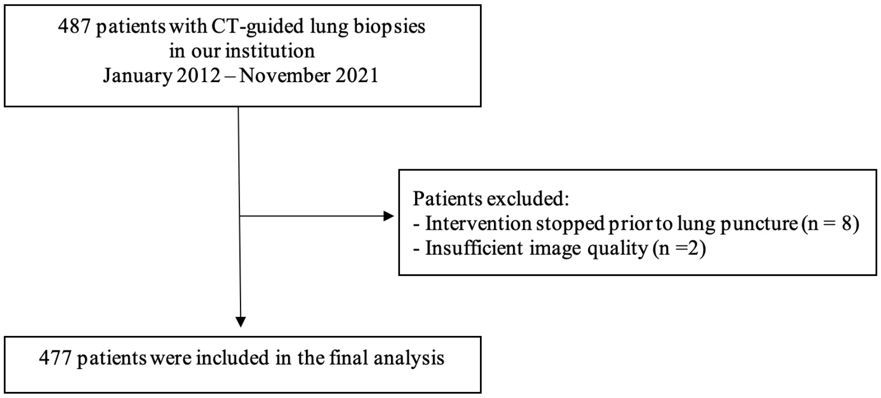

487 patients with CT-guided lung biopsies
in our institution
January 2012 – November 2021
Patients excluded:
- Intervention stopped prior to lung puncture (n = 8)
- Insufficient image quality (n =2)
477 patients were included in the final analysis

Figure 1. Flowchart illustrating the screening and exclusion criteria for patients undergoing computed tomography (CT)-guided lung biopsies between January 2012 and November 2021.

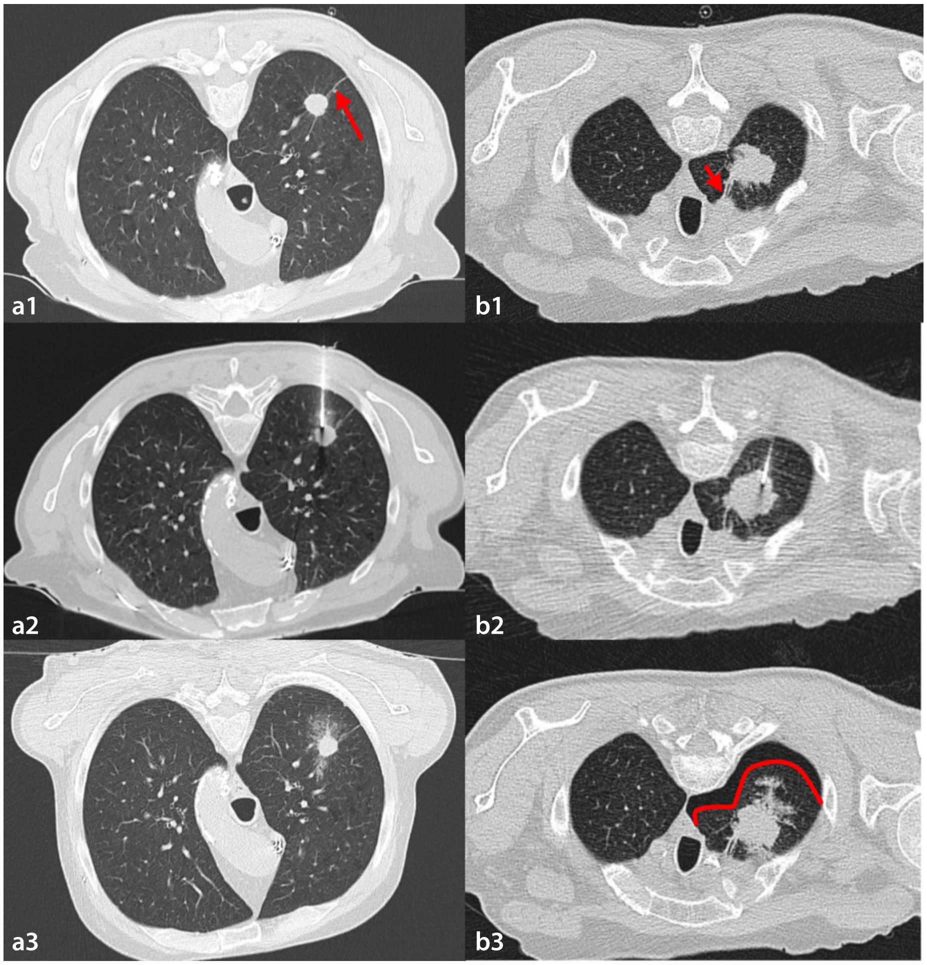


Figure 2. Representative computed tomography (CT) images of patients with pulmonary nodules and an associated thin or triangular pleural tail sign (PTS) undergoing CT-guided lung biopsy. On the left, a patient with primary lung cancer and a thin PTS without postinterventional pneumothorax **(a1-a3).** On the right, a patient with primary lung cancer and a triangular PTS with a severe postinterventional pneumothorax requiring chest tube placement **(b1-b3)**. The red arrows indicate the PTS. The red line demarcates the pneumothorax contour.

or Biopince, Argon Medical Devices, Athens, TX, USA). During the procedures, CT images were retrieved to validate accurate localization of the needle tip. After removal of the biopsy needle, CT images of the whole lung parenchyma were acquired to detect post-interventional complications. To minimize the occurrence of post-biopsy pneumothorax, patients were left to rest without eating or drinking for 2 hours after the procedure. In addition, a plain chest radiograph was obtained 2 hours after the biopsy to detect complications, particularly pneumothorax. Patients with pneumothorax on immediate post-biopsy CT images were labeled as "instant pneumothorax." Cases were classified as "delayed pneumothorax" when a new pneumothorax was identified on the following radiograph. Patients with a pneumothorax >10 mm in width and/or newly occurring symptoms, such as shortness of breath, an increased heart rate, and declining oxygen saturation, received chest tube placement.

Statistical analysis

Statistical analysis was performed using SPSS software (version 29.0; IBM, Armonk, NY, USA). Demographical statistics were provided by mean values with standard deviation. Group differences were analyzed using the Mann–Whitney U test and analysis of variance (ANOVA). Interreader variability was assessed using Cohen's kappa. Univariate and multivariate logistic regression analyses were performed to evaluate the association between the PTS and the occurrence of pneumothorax and pneumothorax requiring chest tube placement. The multivariate analysis included lesion size, lesion depth (distance to pleura), and the number of biopsies as covariates. In a separate univariate analysis limited to patients with a visible PTS, we also tested whether crossing through the PTS with the biopsy needle was associated with pneumothorax severity. The results are reported as odds ratios (ORs) with corresponding 95% confidence intervals (CIs). In all instances, *P* values <0.05 were interpreted as statistically significant.

Results

Prevalence and distribution of the pleural tail sign

A total of 159 out of all 477 cases (33.3%) were classified as PTS positive. Among these, 64 cases (13.4%) exhibited a thin PTS, and 95 cases (19.9%) had a triangular PTS.

Pleural tail sign and pneumothorax

Of all the cases (n = 477), 47.2% (n = 225) developed a pneumothorax following CT-guided PTNB. Of the 159 cases with a PTS, 53.5% (n = 85) developed a pneumothorax. In the 64 cases classified as having a thin PTS, 51.6% (n = 33) developed a pneumothorax. In the 95 cases with a triangular PTS, 54.7% (n = 52) developed a pneumothorax. In cases without a PTS (n = 318), 44.0% (n = 140) developed a pneumothorax. A Mann–Whitney U test was performed to compare the occurrence of pneumothorax between cases with a PTS and those without. The analysis revealed no statistically significant difference between the two groups (OR: 1.20, 95% CI: 0.99–1.46, *P* = 0.052). Moreover, ANOVA revealed no statistically significant difference in the PTS grading scale (0–2) between no pneumothorax and pneumothorax (*P* = 0.067).

Pleural tail sign and instant pneumothorax

For patients with an instant occurrence of pneumothorax (n = 154, of which n = 101 with no PTS, n = 23 with a thin PTS, and n = 30 with a triangular PTS), there was no statistically significant difference between "no PTS" and "any PTS" (OR: 0.93, 95% CI: 0.62–1.40, *P* = 0.73). Moreover, there was no significant difference in the distribution of the different PTS subcategories among cases with instant pneumothorax occurrence and those without instant pneumothorax (*P* = 0.82). Analysis of interreader reliability demonstrated good reliability, with Cohen's kappa of 0.617 (*P* < 0.001).

Pleural tail sign and pneumothorax with necessity of chest tube placement

Of all the cases, 11.5% (n = 55) developed a pneumothorax requiring chest tube placement following CT-guided PTNB. In the PTS group, 15.7% (n = 25) of patients developed a pneumothorax requiring chest tube placement, with 12.5% (n = 8) in the thin PTS subgroup and 17.9% (n = 17) in the triangular subgroup, compared with 9.4% (n = 30) in cases without a PTS. Univariate logistic regression revealed a statistically significant association between the presence of any PTS and the need for chest tube placement (OR: 1.791, 95% CI: 1.014–3.164, *P* = 0.043). Subgroup analysis showed no significant association between chest tube placement and a thin PTS (OR: 1.371, 95% CI: 0.598–3.148, *P* = 0.456), whereas a triangular PTS was significantly associated with an increased risk of requiring a chest tube (OR: 2.092, 95% CI: 1.097–3.990, *P* = 0.025).

In multivariate logistic regression, the presence of a triangular PTS remained an independent predictor of pneumothorax requiring chest tube placement after adjustment for lesion size, depth, and the number of biopsies (OR: 2.02, 95% CI: 1.08–3.79, *P* = 0.029).

In a subgroup analysis of patients with a visible PTS, no significant association was found between the biopsy needle passing through the PTS and the risk of either general pneumothorax (OR: 0.55, *P* = 0.159) or pneumothorax requiring chest tube placement (OR: 0.40, *P* = 0.230). Figure 3 provides an overview of the incidence of pneumothorax with the necessity of chest tube placement stratified by the PTS subgroups, along with the corresponding ORs.

Histopathological entities and pneumothorax and pleural tail sign

Among the histopathological categories, ANOVA revealed significant differences in the rate of PTS positive cases: primary lung cancer (n = 214), metastasis (n = 121), benign lesions (n = 76), and cases with no or an indeterminate histological outcome (n = 66) (*P* = 0.005).

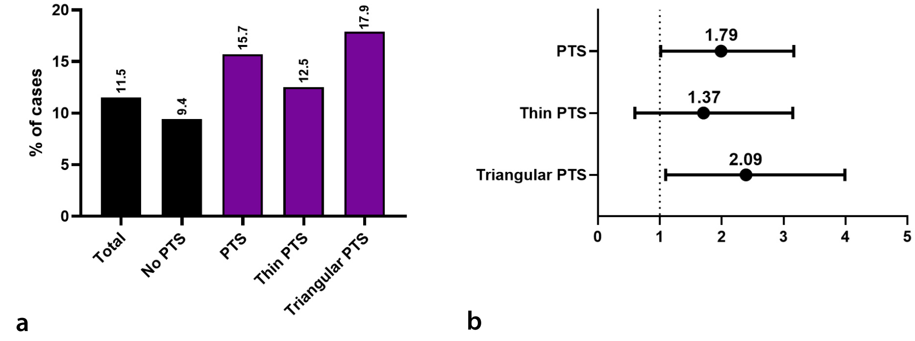


Figure 3. (a) Bar chart showing the percentage of patients who developed a severe pneumothorax requiring chest tube placement, presented for the total cohort as well as stratified by pleural tail sign (PTS) subgroups: no PTS, any PTS, thin PTS, and triangular PTS. **(b)** Forest plot illustrating the corresponding odds ratios for each group. Error bars represent 95% confidence intervals.

In univariate logistic regression analysis, primary lung cancer was significantly associated with higher odds of exhibiting a PTS than all other histopathological groups combined (metastases, benign lesions, and non-diagnostic cases), with an OR of 1.53 (95% CI: 1.04–2.25, *P* = 0.031).

To assess whether the underlying pathology was associated with the occurrence of pneumothorax, patients were categorized into the same four groups. The ANOVA results revealed no statistically significant differences in the overall pneumothorax rate (*P* = 0.646) or in the rate of severe pneumothorax requiring chest tube placement (*P* = 0.192).

Discussion

Multiple studies have been conducted to identify key risk factors for postinterventional pneumothorax after CT-guided lung biopsy.[18] These studies have consistently identified smaller lesion size, greater distance from the lesion to the pleura, needle paths passing through pulmonary fissures, and emphysema adjacent to the target lesion as factors associated with an increased pneumothorax risk.[10,19] Moreover, our previous investigation identified CT radiomics features of the target lesion and the lung-lobe CT-emphysema score as predictive biomarkers for the occurrence of pneumothorax and the need for chest tube placement after CT-guided PTNB.[10] Technical factors such as patient positioning, entry point relative to gravity, and needle angulation also play a key role and have been shown to significantly influence the risk of pneumothorax.[20-22] Two recent studies identified the PTS as a novel imaging marker associated with an increased risk of pneumothorax following CT-guided lung biopsy.[11,12] One study, based on 311 procedures, reported that the presence of a PTS was an independent risk factor for immediate pneumothorax,[11] whereas another, analyzing 775 cases, found that needle paths passing through the pleural tail significantly increased the pneumothorax rate.[12] However, neither study distinguished between morphological subtypes of the PTS or evaluated complication severity. By contrast, our study applied a two-tier PTS classification (thin vs. triangular) and specifically assessed the severity of pneumothorax, defined by the need for chest tube placement.

The PTS itself is referred to as a strip connected to a lung lesion, propagating to the pleura.[18] Histological analysis revealed that pleural tails are caused by interlobular thickening caused by lymphatic obstruction, inflammation, desmoplastic reaction, or tumor infiltration.[18,19]

In the current study, no statistically significant difference was found in the distribution of lesions with a PTS and those without a PTS regarding the occurrence of pneumothorax in CT-guided biopsies, although the significance level was close to the threshold of 0.05. In addition, differences in the prominence of PTS (classified as either thin or triangular) and in the timing of pneumothorax occurrence (immediate vs. delayed) did not reach statistical significance. These results differ from those of earlier studies.[11,12]

Nonetheless, we observed statistically significant differences in the occurrence of severe pneumothorax, as defined by the necessity of chest tube placement, both for the mere presence of a PTS and for the triangular subtype. Notably, in our multivariate logistic regression analysis, the presence of a triangular PTS remained an independent predictor for pneumothorax requiring chest tube placement after adjustment for lesion size, depth, and the number of biopsies. To our knowledge, this is the first study demonstrating the clinical significance of the PTS in the context of interventional risk stratification.

In contrast to previous reports, our subgroup analysis among patients with a visible PTS did not demonstrate a significant association between the biopsy needle passing through the PTS and the risk of pneumothorax or the need for chest tube placement.[11] Our study suggests that the pleural vulnerability associated with the PTS may primarily result from intrinsic fibrotic or infiltrative changes rather than from mechanical disruption caused by traversing the PTS.

Notably, the PTS was significantly more common in primary lung cancer than in metastatic or benign nodules. However, neither the overall pneumothorax rate nor the rate of severe, chest tube-requiring pneumothorax differed significantly across histopathological categories, indicating that the histopathological subtype is unlikely to be a major confounder.

The present study has several limitations. First, it is a retrospective study, which poses the risk of a possible known inherent bias. However, the imaging analysis was performed in a blinded fashion to the clinical outcome to reduce possible bias. Second, the identification of a pleural tail and the grading of its features may impose some interreader variability. Third, clinical indications for chest tube placement can vary, as decisions often hinge on the patient's symptoms, the pneumothorax size, and the operator's personal preference.

In conclusion, in this study, we found that the triangular PTS is an independent prognostic imaging marker for identifying patients at higher risk of clinically significant pneumothorax, defined by the need for chest tube placement. Additionally, in our cohort, the PTS was significantly more likely to occur in lesions reflecting primary lung cancer than in other histopathological entities.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Magnini A, Fissi A, Cinci L, et al. Diagnostic accuracy of imaging-guided biopsy of peripheral pulmonary lesions: a systematic review. *Acta Radiol*. 2024;65(10):1222-1237. [Crossref]
2. Bhatt KM, Tandon YK, Graham R, et al. Electromagnetic navigational bronchoscopy versus CT-guided percutaneous sampling of peripheral indeterminate pulmonary nodules: a cohort study. *Radiology*. 2018;286(3):1052-1061. [Crossref]
3. Yao X, Gomes MM, Tsao MS, et al. Fine-needle aspiration biopsy versus core-needle biopsy in diagnosing lung cancer: a systematic review. *Curr Oncol*. 2012;19(1):e16-27. [Crossref]
4. Heerink WJ, de Bock GH, de Jonge GJ, et al. Complication rates of CT-guided transthoracic lung biopsy: meta-analysis. *Eur Radiol*. 2017;27(1):138-148. [Crossref]
5. Huo YR, Chan MV, Habib AR, Lui I, Ridley L. Pneumothorax rates in CT-Guided lung biopsies: a comprehensive systematic review and meta-analysis of risk factors. *Br J Radiol*. 2020;93(1108):20190866. [Crossref]
6. Boskovic T, Stanic J, Pena-Karan S, et al. Pneumothorax after transthoracic needle biopsy of lung lesions under CT guidance. *J Thorac Dis*. 2014;6 Suppl 1(Suppl 1):S99-S107. [Crossref]
7. Topal U, Ediz B. Transthoracic needle biopsy: factors effecting risk of pneumothorax. *Eur J Radiol*. 2003;48(3):263-267. [Crossref]
8. Ozturk K, Soylu E, Gokalp G, Topal U. Risk factors of pneumothorax and chest tube placement after computed tomography-guided core needle biopsy of lung lesions: a single-centre experience with 822 biopsies. *Pol J Radiol*. [Crossref]
9. Theilig D, Petschelt D, Mayerhofer A, et al. Impact of quantitative pulmonary emphysema score on the rate of pneumothorax and chest tube insertion in CT-guided lung biopsies. *Sci Rep*. 2020;10(1):10978. [Crossref]
10. Leonhardi J, Dahms U, Schnarkowski B, et al. Impact of radiomics features, pulmonary emphysema score and muscle mass on the rate of pneumothorax and chest tube insertion in CT-guided lung biopsies. *Respir Res*. 2024;25(1):320. [Crossref]
11. Peng B, Deng Z, Wang Y, et al. The risk of immediate pneumothorax after CT-guided lung needle biopsy: pleural tail sign as a novel factor. *Quant Imaging Med Surg*. 2023;13(2):707-719. [Crossref]
12. Deng XB, Xie L, Zhu HB, et al. The nodule-pleura relationship affects pneumothorax in CT-guided percutaneous transthoracic needle biopsy: avoiding to cross pleural tail sign may reduce the incidence of pneumothorax. *BMC Pulm Med*. 2024;24(1):490. [Crossref]
13. Han J, Xiang H, Ridley WE, Ridley LJ. Pleural tail sign: pleural tags. *J Med Imaging Radiat Oncol*. 2018;62 (Suppl 1):37. [Crossref]
14. Gruden JF. What is the significance of pleural tags? *AJR Am J Roentgenol*. 1995;164(2):503-504. [Crossref]
15. Hsu JS, Han IT, Tsai TH, et al. Pleural Tags on CT scans to predict visceral pleural invasion of non-small cell lung cancer that does not abut the pleura. *Radiology*. 2016;279(2):590-596. [Crossref]
16. Onoda H, Higashi M, Murakami T, et al. Correlation between pleural tags on CT and visceral pleural invasion of peripheral lung cancer that does not appear touching the pleural surface. *Eur Radiol*. 2021;31(12):9022-9029. [Crossref]
17. Meng Y, Gao J, Wu C, et al. The prognosis of different types of pleural tags based on radiologic-pathologic comparison. *BMC Cancer*. 2022;22(1):919. [Crossref]
18. Sargent T, Kolderman N, Nair GB, Jankowski M, Al-Katib S. Risk factors for pneumothorax development following CT-guided core lung nodule biopsy. *J Bronchology Interv Pulmonol*. 2022;29(3):198-205. [Crossref]
19. Zhao Y, Bao D, Wu W, et al. Development and validation of a prediction model of pneumothorax after CT-guided coaxial core needle lung biopsy. *Quant Imaging Med Surg*. 2022;12(12):5404-5419. [Crossref]
20. Brönnimann MP, Barroso MC, Manser L, et al. The role of gravitational effects and pre-puncture techniques in reducing pneumothorax during CT-guided lung biopsies. *Radiol Med*. 2025;130(7):1024-1038. [Crossref]
21. Maalouf N, Abou Mrad M, Lavric D, et al. Safe zone to avoid pneumothorax in a CT-guided lung biopsy. *J Clin Med*. 2023;12(3):749. [Crossref]
22. Drumm O, Joyce EA, de Blacam C, et al. CT-guided lung biopsy: effect of biopsy-side down position on pneumothorax and chest tube placement. *Radiology*. 2019;292(1):190-196. [Crossref]

Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.242950

PEDIATRIC RADIOLOGY

ORIGINAL ARTICLE

Copyright @ 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License.

Diagnostic accuracy of convolutional neural network algorithms to distinguish gastrointestinal obstruction on conventional radiographs in a pediatric population

Ercan Ayaz[1]
Hasan Güçlü[2]
Ayşe Betül Oktay[3]

[1]Diyarbakır Children's Hospital, Radiology Clinic, Diyarbakır; Current: University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

[2]İstanbul Medeniyet University Faculty of Engineering and Natural Sciences, Department of Biostatistics and Medical Informatics, İstanbul; Current: TOBB University of Economics and Technology, Department of Artificial Intelligence Engineering, Ankara, Türkiye

[3]Yıldız Technical University Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

Corresponding author: Ercan Ayaz

E-mail: ercan.ayaz1@gmail.com

Received 31 October 2024; revision requested 12 December 2024; accepted 13 January 2025.

Epub: 28.02.2025

Publication date: 02.03.2026

DOI: 10.4274/dir.2025.242950

PURPOSE
Gastrointestinal (GI) dilatations are frequently observed in radiographs of pediatric patients who visit emergency departments with acute symptoms such as vomiting, pain, constipation, or diarrhea. Timely and accurate differentiation of whether there is an obstruction requiring surgery in these patients is crucial to prevent complications such as necrosis and perforation, which can lead to death. In this study, we aimed to use convolutional neural network (CNN) models to differentiate healthy children with normal intestinal gas distribution in abdominal radiographs from those with GI dilatation or obstruction. We also aimed to distinguish patients with obstruction requiring surgery and those with other GI dilatation or ileus.

METHODS
Abdominal radiographs of patients with a surgical, clinical, and/or laboratory diagnosis of GI diseases with GI dilatation were retrieved from our institution's Picture Archiving and Communication System archive. Additionally, abdominal radiographs performed to detect abnormalities other than GI disorders were collected to form a control group. The images were labeled with three tags according to their groups: surgically-corrected dilatation (SD), inflammatory/infectious dilatation (ID), and normal. To determine the impact of standardizing the imaging area on the model's performance, an additional dataset was created by applying an automated cropping process. Five CNN models with proven success in image analysis (ResNet50, InceptionResNetV2, Xception, EfficientNetV2L, and ConvNeXtXLarge) were trained, validated, and tested using transfer learning.

RESULTS
A total of 540 normal, 298 SD, and 314 ID were used in this study. In the differentiation between normal and abnormal images, the highest accuracy rates were achieved with ResNet50 (93.3%) and InceptionResNetV2 (90.6%) CNN models. Then, after using automated cropping preprocessing, the highest accuracy rates were achieved with ConvNeXtXLarge (96.9%), ResNet50 (95.5%), and InceptionResNetV2 (95.5%). The highest accuracy in the differentiation between SD and ID was achieved with EfficientNetV2L (94.6%).

CONCLUSION
Deep learning models can be integrated into radiographs located in the emergency departments as a decision support system with high accuracy rates in pediatric GI obstructions by immediately alerting the physicians about abnormal radiographs and possible etiologies.

CLINICAL SIGNIFICANCE
This paper describes a novel area of utilization of well-known deep learning algorithm models. Although some studies in the literature have shown the efficiency of CNN models in identifying small bowel obstruction with high accuracy for the adult population or some specific diseases, our study is unique for the pediatric population and for evaluating the requirement of surgical versus medical treatment.

KEYWORDS
Abdominal X-ray, ileus, pediatric radiology, convolutional neural networks, deep learning

You may cite this article as: Ayaz E, Güçlü H, Oktay AB. Diagnostic accuracy of convolutional neural network algorithms to distinguish gastrointestinal obstruction on conventional radiographs in a pediatric population. *Diagn Interv Radiol*. 2026;32(2):233-241.

The imaging of the gastrointestinal (GI) system is challenging in children, and often, the initial modality of choice is either an abdominal radiograph or ultrasound, both in the emergency and outpatient settings. Abdominal radiography is cheap, widely available, exposes less radiation compared with computed tomography (CT), and provides specific appearances for some pediatric conditions such as duodenal atresia and necrotizing enterocolitis (NEC).[1] The common causes of GI obstructions in pediatric patients are more varied and different than in adults and often require dedicated radiological evaluation to recognize peculiar imaging features.[2] The bowel can be obstructed or dilated by a wide range of diseases classified as congenital, developmental, inflammatory, infectious, and neoplastic lesions.[3] Delay in the diagnosis and surgical management of such pediatric acute bowel obstruction increases the risk of bowel necrosis, perforation, and death. Therefore, accurate diagnostic management is crucial to improve patient outcomes.[4] Previous studies in adult populations have revealed that the 3 most sensitive radiographic signs for bowel obstruction are air-fluid levels in loops of the bowel wider than 2.5 cm, 2 or more air-fluid levels, and multiple air-fluid levels within 1 loop of the bowel differing 5 mm.[2]

In recent years, there has been a growing number of studies on integrating artificial intelligence (AI) as a diagnostic support model into image-based medical fields such as radiology and pathology. Artificial neural networks have become the most preferred models for image classification among the subfields of AI due to their high accuracy rates.[5]

Convolutional neural network (CNN), a deep artificial neural network, possesses the ability to distinguish and classify images by extracting and comparing specific features from them. However, the main limitation of CNNs is their need for large datasets for training. The capacity of a CNN trained on a large dataset can be transferred to differentiate similar images.[6] With the proliferation of digital radiography and Picture Archiving and Communication Systems (PACS), significant advancements have been made in acquiring radiographic data in recent years. Although radiography involves single-section and two-dimensional imaging, CT and magnetic resonance imaging provide multi-sectional and three-dimensional imaging. Therefore, radiographs can be processed with simpler deep-learning models.

Main points

- Pre-trained convolutional neural network models can be accurately used in abdominal radiographs with the transfer learning method.
- Fine-tuning should be performed to improve the performance of the model and to decrease the validation and training loss.
- The automated cropping process significantly improves the performance of all models, probably due to factors such as the non-standard nature of the radiographs taken under emergency and outpatient conditions, improper positioning, and inappropriate adjustment of the imaging area.

In daily practice, many abdominal radiographs are performed on children in emergency rooms and outpatient clinics. In Turkey, most of these are not evaluated by radiologists but by emergency or outpatient physicians under time constraints. According to a report prepared by the Turkish Society of Radiology instead of Radiology Association in 2018, the number of radiologists per 100,000 people in Türkiye was 5, whereas this number was 2–3 times higher in Organization for Economic Co-operation and Development countries.[7] Due to the lack of sufficient time for evaluating radiographs or the inexperience of the evaluating physician, additional tests may be unnecessarily requested for patients with false-negative evaluations, or patients with a condition may be incorrectly deemed normal and sent home. Conversely, unnecessary treatments or surgical interventions may be performed on patients with false-positive evaluations. Since children often cannot accurately express their complaints and because laboratory findings can change rapidly, radiological examinations hold even greater importance.[4]

Therefore, if the radiographs taken in the emergency room are classified by a CNN model integrated into the PACS system and presented to the relevant physician, it can enable more careful evaluation by the physician.

This study aims to retrain current CNN models on abdominal radiographs and assess which models are more successful in classifying normal and pathological radiographs. It also proposes differentiating between pathological radiographs that resolve with medical treatment (infectious) and those requiring surgical intervention.

Methods

Institutional review board approval was obtained from the Diyarbakır Gazi Yaşargil Training and Research Hospital Non-Interventional Clinical Research Ethics Committee (decision no: 2022/108, decision date: 10.06.2022) for this study's retrospective data collection and analysis. Informed consent was waived because of the retrospective nature of the study.

Image acquisition

After obtaining the approval of the ethical committee, abdominal radiographs taken in the outpatient clinic and emergency department between January 1, 2019, and June 1, 2022, were reviewed using the radiology PACS archive of our institution. They were included if the patients had multiple images within the same disease course and before the surgery. X-ray devices used in the outpatient clinic and emergency department were single-tube Jumong model digital X-ray imaging systems (SG Healthcare Co, Gyeonggi-Do, South Korea). Automatic exposure control (AEC) sensors were used during imaging, and dose parameters for each imaging were adjusted accordingly. Shielding was not used to avoid overexposure due to AEC measurements. Peak tube voltage (kVp), tube current (mA), exposure time (msec), and dose area product (DAP) were recorded for each examination. Due to vast body size variations in the study cohort (0–18 years), peak tube voltage was changed between 80,100 and 120 kVp according to tissue thickness, requiring more photon penetration. For the routine posteroanterior erect abdominal radiograph performed in the outpatient clinic and emergency department, the patient-tube distance was 110 cm.

The images were retrieved for the study using JPEG compression. For comparison, a control group was formed from patients with normal GI findings on abdominal radiographs, who were imaged for other reasons, such as kidney stones, with a balanced age distribution from 0 to 18 years. The dataset consisted of three main groups: (1) patients with GI obstruction requiring surgical intervention [surgically-corrected dilatation (SD)], (2) patients with bowel dilatation/ileus treated without surgery [inflammatory/infectious dilatation (ID)], and (3) a normal control group. The age and sex characteristics of the patients and the diagnoses of the diseases for groups with pathological findings were recorded. The first group requiring surgery was diagnosed surgically. While labeling the second group, if examinations remained indeterminate, the cases were discussed by an experienced pediatric radiologist (with 7 years of experience) and the referring pediatrician. Six cases that remained indeterminate after enhanced clinical-radiological review were excluded, as no meaningful label could be assigned.

Consequently, a total of 612 radiographs with the findings of bowel dilatation or obstruction were included. For the first group (patients who underwent surgery), 298 images from 107 patients were obtained from the archive, and for the second group (patients who did not require surgery), 314 images from 189 patients were obtained. For comparison, a control group of 540 normal abdominal radiographs, 1 for each case, was created, considering a balanced age distribution between 0 and 18 years. The flowchart of the study is presented in Figure 1.

Training, testing dataset, and preprocessing

Images were retrieved from the PACS station with a resolution of 1,040 × 624 pixels and down-sampled by bicubic interpolation automatically in the CNN to match the input layers. Afterward, 32 batches, each including 36 images, were composed of 1,152 images. Each batch was split into training, validation, and test sets using a ratio of 28:3:5, respectively. This ratio was designed to maintain a sufficient training dataset while providing adequate statistical power for the testing dataset. A test set sample size of 160 enabled a statistical power of 0.8 for detecting an area under the curve (AUC) of 0.65 with a type 1 error of 0.025.[8] Data augmentation was performed on the training dataset with horizontal flipping and rotation by Keras library. The images formed with data augmentation would be similar to those not taken in the correct position due to patient rotation during the shooting or sent incorrectly to the PACS system. This approach aims to provide flexibility for the model to evaluate images that are not properly positioned (Supplementary Figure 1).

To determine the impact of standardizing the imaging area on the model's performance, an additional dataset was created by applying an automated cropping process to the data using a cropping code set to remove rows or columns from all edges until a white-toned pixel was found. During the automated cropping process, some images had data labels on the edges of the image, causing the cropping to stop before the model reached the image (Figures 2a-d). This situation represented a limitation of the model compared with manual cropping. Since this study aimed to provide the classification result directly to the physician via automated preprocessing and model analysis of the image obtained from X-ray imaging, manual cropping was not preferred.

Neural networks training and testing

All CNN training, testing, and other processes were performed using the Keras 2.1.5 library with TensorFlow 1.7 as the backend framework in Python (version 3.7.3), and Google Colab was used as a notebook service provider with its integrated graphics processing units.[9-11]

Five CNNs used in this study were publicly available and pre-trained on the ImageNet data set: ResNet50, InceptionResNetV2, Xception, EfficientNet, and ConvNeXt.[12-17]

The architecture of the models is briefly described in Supplementary Figure 2. The background of the networks was developed to detect everyday objects such as vehicles, flowers, or animals, but the top layers were completely new and acquired their parameters based on the radiographs used in the training database, called the transfer learning method. Since the datasets on which CNN models were pre-trained contained a large amount of data, during training with our smaller dataset, the process of determining the filter weights for feature extraction was limited to the first few convolutional layers (usually the first three), and the training of the last layers did not occur.[18] To overcome this, a particular process called fine-tuning was applied, and the layers close to the input

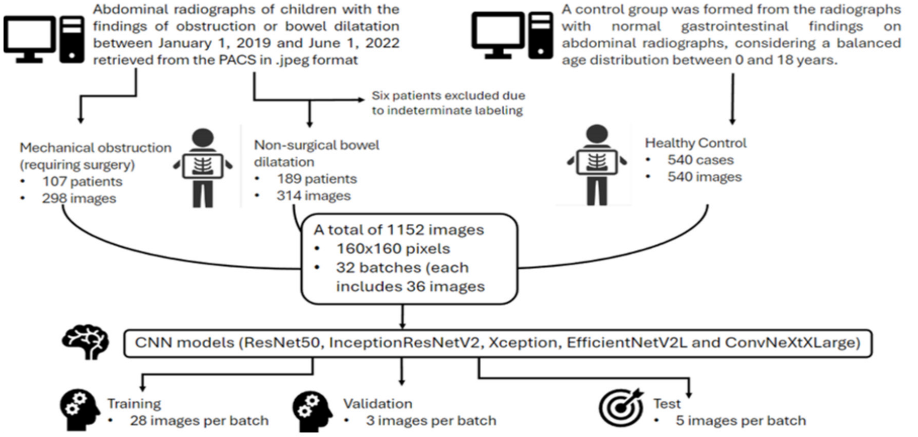

Abdominal radiographs of children with the findings of obstruction or bowel dilatation between January 1, 2019 and June 1, 2022 retrieved from the PACS in .jpeg format
A control group was formed from the radiographs with normal gastrointestinal findings on abdominal radiographs, considering a balanced age distribution between 0 and 18 years.
Six patients excluded due to indeterminate labeling
Mechanical obstruction (requiring surgery)
• 107 patients
• 298 images
Non-surgical bowel dilatation
• 189 patients
• 314 images
Healthy Control
• 540 cases
• 540 images
A total of 1152 images
• 160x160 pixels
• 32 batches (each includes 36 images
CNN models (ResNet50, InceptionResNetV2, Xception, EfficientNetV2L and ConvNeXtXLarge)
Training
• 28 images per batch
Validation
• 3 images per batch
Test
• 5 images per batch

Figure 1. Flow diagram of the study. CNN, convolutional neural network; PACS: Picture Archiving and Communication Systems.

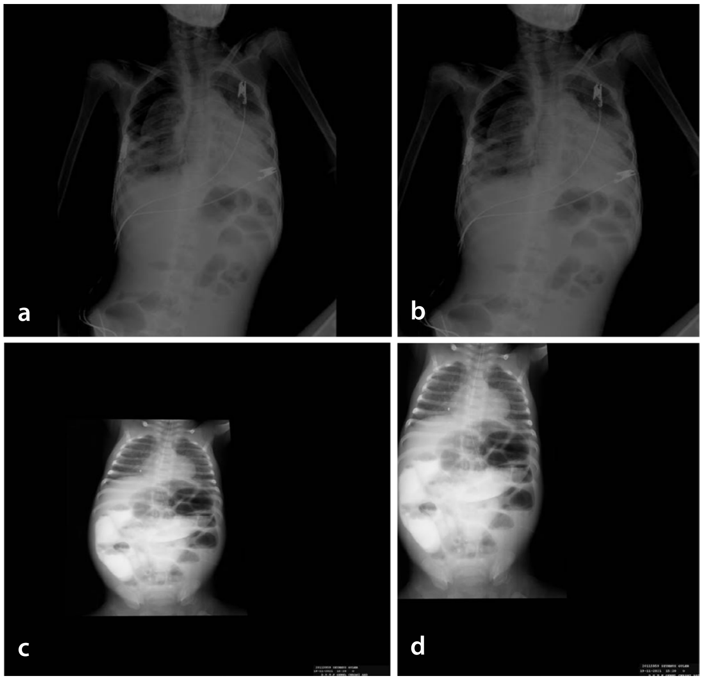


Figure 2. Example of before **(a)** and after **(b)** successful automated cropping to remove unnecessary parts and suboptimal cropping **(c, d)** due to white pixels of a label at the right lower corner of the image.

were frozen to ensure that the weights for feature extraction in the subsequent layers were determined. Fine-tuning was routinely applied, especially in transfer learning methods used in image analysis, and it improved the model's performance. Although epoch durations were longer compared with the standard training process, a significant decrease in training and validation errors was achieved with fewer epochs (Supplementary Figure 3).

The five models used in the experiments were trained for 100 epochs during the training phase. To enhance performance during transfer learning and to allow the models trained on another dataset to adapt to the features of our data, an additional 20 epochs were run for fine-tuning. In this study, the models were first presented with original data and then cropped data from the normal control and abnormal patient groups. All models were tested on 224 images after training, and their success was evaluated using performance metrics. Finally, to determine which diseases, ages, and sexes the misdiagnosed cases (false positives or false negatives) belonged to, the dataset was analyzed using the most successful model.

Statistical analysis

The mean and standard deviation (SD) values for age and the median and quartiles were presented. The descriptive statistics of the pathological groups were calculated. Pearson's chi-squared test was used to compare gender data. Kolmogorov–Smirnov test showed that the age data did not follow a normal distribution ($P < 0.001$ for all groups). The median and interquartile ranges (IQRs) were presented for non-normally distributed dose parameters. A non-parametric Mann–Whitney U test was applied since the pathological groups did not show a normal distribution. Statistical analyses were performed using the IBM SPSS version 23.0 software package (IBM Corporation, Armonk, NY, USA). A single receiver operating characteristic (ROC) curve and cut-off analysis were used for the internal test, whereas two ROC curves with independent groups were designed to compare the external and internal validation tests. Two-tailed P values < 0.05 were considered statistically significant. After completing the training phase, the models were tested using the dataset created for testing. The performance of the models was measured by metrics such as accuracy, precision, sensitivity, specificity, F1 score, and the AUC.

Results

The age and sex distribution of the three groups within the dataset are presented in Table 1. No significant difference was found between the two patient groups regarding age ($P = 0.928$). However, there were more boys in the SD group than in the ID group ($P < 0.001$). Regarding dose parameters, 725 examinations were performed with 80 kVp, 338 with 100 kVp, and 89 with 120 kVp. The median tube current was 320 mA (IQR: 80). The mean (± SD) exposure time was 37.22 ± 7.51 milliseconds, and the median DAP was 165 mGy·cm^2 (IQR: 349). In the SD group, a total of 16 different causes of obstruction were identified. The most prevalent cause, ileus due to postoperative adhesions, was observed in 83 radiographs of 27 patients (27.9%). This was followed by complicated appendicitis, seen in 67 radiographs of 30 patients (22.5%), and NEC, found in 35 radiographs of 11 patients (11.7%). It is worth noting that some cases of ileus due to postoperative adhesions were observed during follow-up after surgeries of patients with other etiologies, which is why the total number of cases appears higher than the total number of patients in this group when the cases from both groups are combined. The age and sex distribution according to the types of diseases is presented in Table 2.

NEC, hypertrophic pyloric stenosis, meconium ileus, Hirschsprung's disease, duodenal atresia/stenosis, and inguinal hernia cases are observed in the neonatal and infant periods, whereas abscess/peritonitis secondary to intraperitoneal catheter and intussusception cases occur in early childhood. Complicated appendicitis and Crohn's disease are predominantly seen in the group aged over 10 years. The disease groups with the broadest age distribution are also the two most common diseases: ileus due to postoperative adhesions and complicated appendicitis. Among the common diseases, groups with similar ages were compared statistically using the Student's t-test. The ages of patients with complicated appendicitis were found to

Table 1. Age and sex distribution of the study groups and the control group

	Healthy control group	SD group	ID group
Sex [male (%)/female (%)]	262 (48.5)/278 (51.5)	232 (77.5)/66 (22.1)	180 (57.3)/134 (42.7)
Age (mean ± standard deviation), years	7.29 ± 5.05	5.47 ± 5.82	4.22 ± 4.44
Age [median (interquartile ranges)], years	6.5 (3.1–11.3)	3 (0.3–10.0)	2.1 (1.3–6.0)

SD, surgically-corrected dilatation; ID, inflammatory/infectious dilatation.

Table 2. Number, age, and sex features of patients within the surgically corrected obstruction group

Diagnosis	Number of cases/ number of images	Sex: male (%)/ female (%)	Age: mean ± standard deviation in years
Ileus due to postoperative adhesion	27/83	22 (81.5)/5 (18.5)	6.48 ± 5.32
Complicated acute appendicitis	30/67	22 (73.3)/8 (26.7)	11.33 ± 4.75
Necrotizing enterocolitis	11/35	5(45.5)/6 (54.5)	0.39 ± 0.53
Hypertrophic pyloric stenosis	12/19	12 (100)/0	0.11 ± 0.06
Hirschsprung's disease	7/15	6 (85.7)/1 (14.3)	1.02 ± 1.38
Abscess/peritonitis secondary to intraabdominal catheter	5/15	4 (80)/1 (20)	5.13 ± 6.08
Meconium ileus or meconium plug syndrome	5/12	4 (80)/1 (20)	0.34 ± 0.27
Duodenal atresia or stenosis	2/11	1 (50)/1 (50)	1.15 ± 0.10
Intussusception	9/9	5 (55.6)/4 (44.4)	3.18 ± 4.41
Complicated inguinal hernia	4/8	4 (100)/0	0.95 ± 0.78
Complicated Crohn's disease	2/8	2 (100)/0	11.89 ± 1.48
Midgut volvulus	3/6	3 (100)/0	1.80 ± 2.26
Other	4/9	4 (100)/0	6.87 ± 4.79

be significantly higher than those with ileus due to postoperative adhesions (*P* < 0.001), and the ages of patients with NEC were significantly higher than those with hypertrophic pyloric stenosis (*P* = 0.003). No significant difference was found between cases of postoperative adhesions and catheter infections (*P* = 0.379) or between Hirschsprung's disease and duodenal atresia/stenosis (*P* = 0.719).

In the third group, which included cases of non-ileus, no infectious agent was detected in 142 patients, from whom 231 (73.6%) radiographs were obtained. In 41 patients (68 radiographs, 21.7%), rotavirus was detected in 2 patients (3 radiographs, 1%), adenovirus antigen in 2 patients (6 radiographs, 1.9%), and amoeba in the stool of 2 patients (6 radiographs, 1.9%). In 2 patients with 6 radiographs (1.9%), GI involvement due to multisystem inflammatory syndrome secondary to coronavirus disease-2019 was diagnosed. When comparing the ages of the rotavirus cases and other cases, it was found that rotavirus cases were significantly higher in the younger age groups (*P* < 0.001).

All models were tested separately on 224 images using both the original and cropped datasets after training. The confusion matrices of the models are presented in Table 3, and the performance metrics are presented in Table 4. Although the highest performance metrics of the models were observed in different rankings across different datasets, they were generally achieved with ResNet50 on raw data, EfficientNetV2L between pathological groups, and ConvNeXtXLarge on cropped data and overall. The pixels with the most weight in classification, as indicated by the gradient-weighted class activation map applied to some data in the analysis of the Xception model, are shown in Figure 3.

Finally, to determine which diseases and ages the misclassified cases (false positives or false negatives) belonged to, our dataset was analyzed using the ConvNeXtXLarge model, which had the highest F1 score. The model was run 3 times using 224 images randomly distributed across 7 packages in each analysis. Four images with SD and six with ID were labeled as false negatives in the three-model analyses. In the normal control group, 33 images were classified as false positives across the 3 analyses. Examples of patients who were classified as abnormal but were healthy, according to the model analysis, are presented with their ages and sexes in Supplementary Figure 4. The cases labeled as normal despite being in the SD group are presented in Figure 4. Since the false negative cases occurred in three different disease groups and involved common diseases in the dataset, we could not conclude that a specific disease group was undetectable.

Discussion

Very few studies utilize deep learning applications on abdominal radiographs, and there is even less literature regarding the pediatric population.[4,19-21] Studies on X-rays in the literature primarily focus on chest radiography, mainly due to the large volume of accessible data.[22-25] Our study demonstrated that in classifying normal and abnormal radiographs, an accuracy above 90% was achieved with the ResNet50 (93.3%) and InceptionResNetV2 (90.6%) CNN models. After applying the cropping preprocessing step to the same data groups, an accuracy above 90% was achieved with EfficientNetV2L (94.6%), and an accuracy above 95% was reached with ResNet50 (95.5%), InceptionResNetV2 (95.5%), and ConvNeXtXLarge (96.9%). In the analysis conducted on cropped images to distinguish surgically corrected GI obstruction from other GI dilations, an accuracy above 90% was achieved with InceptionResNetV2 (90.2%), EfficientNetV2L (94.6%), and ConvNeXtXLarge (91.1%). It is evident that the cropping preprocessing step significantly impacts the performance of all models. This

Table 3. Confusion matrices of the convolutional neural networks' test results used in the study

CNN model	Data type	Labels		Actual	
				Normal (or SD group)	Abnormal (or ID group)
Classification results with **ResNet50** CNN model	Raw images	**Predicted**	Normal	109	15
			Abnormal	0	100
	Cropped images		Normal	109	7
			Abnormal	3	105
	Cropped images		SD group	117	4
			ID group	21	82
Classification results with **InceptionResNetV2** CNN model	Raw images	**Predicted**	Normal	103	3
			Abnormal	18	100
	Cropped images		Normal	119	1
			Abnormal	9	95
	Cropped images		SD group	106	6
			ID group	16	96
Classification results with **Xception** CNN model	Raw images	**Predicted**	Normal	120	10
			Abnormal	28	66
	Cropped images		Normal	100	17
			Abnormal	13	94
	Cropped images		SD group	104	21
			ID group	7	92
Classification results with **EfficientNetV2L** CNN model	Raw images	**Predicted**	Normal	84	28
			Abnormal	0	112
	Cropped images		Normal	118	0
			Abnormal	12	94
	Cropped images		SD group	108	5
			ID group	7	104
Classification results with **ConvNeXtXLarge** CNN model	Raw images	**Predicted**	Normal	102	6
			Abnormal	17	99
	Cropped images		Normal	121	3
			Abnormal	4	96
	Cropped images		SD group	107	7
			ID group	13	97

CNN, convolutional neural network; SD, surgically-corrected dilatation; ID, inflammatory/infectious dilatation.

improvement is likely due to factors such as the non-standard nature of radiographs taken under emergency and outpatient conditions, improper positioning, inappropriate adjustment of the imaging area, and the failure to remove contrast-inducing items from patients during imaging.

Abdominal radiographs are generally the first preferred method for GI diseases due to their affordability, widespread availability, rapid application and interpretation (especially with digital radiographs), and ability to comprehensively show intestinal gas distribution. Radiography is superior to ultrasound, particularly for diagnosing GI obstructions.[26] Typical imaging findings are observed in diseases such as NEC and duodenal atresia, which are seen in the neonatal and infant periods. Additionally, in patients with acute severe clinical symptoms where bowel perforation (rupture) is suspected, radiographs can reveal free air in the abdominal cavity. However, the sensitivity of abdominal radiographs in children with abdominal pain is relatively low, with the rate of pathological findings reported between 2% and 20%.[26] Abdominal radiographs in newborns and young children are usually taken while the patient is lying down. In older children, an upright abdominal radiograph may better display air-fluid levels and bowel loop distention, especially in conditions where peristalsis is impaired. In some cases, lateral decubitus radiographs are taken by positioning the patient on their side to show air-fluid levels, free fluid, or free air in the abdomen.

The following studies stood out when reviewing previous deep-learning research in the literature on diagnosing GI diseases using abdominal radiographs. In the study by Kwon et al.[21], 11,384 abdominal radiographs (1,449 with intussusception and 9,935 without) were retrieved from three hospitals to detect intussusception. Diagnosing intussusception from abdominal radiographs is challenging and requires expertise. Therefore, the diagnosis is typically made by ultrasound. The interobserver agreement among radiologists with limited experience in abdominal radiographs is less than 50%.[27] In the study by Kwon et al.[21], for binary classification, the CNN model used was ResNet. The average sensitivity achieved was 81.6%, with a specificity of 92.5%. The highest accuracy reported was 76%, the lowest was 73%, and the average was 74%. In our study, an analysis of the SD cases classified as false negatives revealed that two of the four cases were complicated appendicitis, one was bowel obstruction (ileus due to postoperative adhesions), and one was Hirschsprung's disease. Notably, no misclassification was detected in intussusception cases. Additionally, an accuracy rate of 93.3% was achieved with the ResNet50 model in our study, making it the model with the highest accuracy on raw data.

In another study on small bowel obstruction, a total of 3,663 upright abdominal ra-

Table 4. Performance metrics of the convolutional neural network models according to datasets

CNN model	Dataset	Accuracy	Specificity	Sensitivity	F1 score
ResNet50	Normal vs. abnormal (raw data)	0.933	1.000	0.869	0.930
	Normal vs. abnormal (cropped data)	0.955	0.973	0.938	0.955
	SD vs. ID group	0.889	0.848	0.953	0.868
InceptionResNetV2	Normal vs. abnormal (raw data)	0.906	0.851	0.970	0.905
	Normal vs. abnormal (cropped data)	0.955	0.930	0.990	0.950
	SD vs. ID group	0.902	0.869	0.941	0.897
Xception	Normal vs. abnormal (raw data)	0.839	0.811	0.868	0.776
	Normal vs. abnormal (cropped data)	0.866	0.885	0.847	0.862
	SD vs. ID group	0.875	0.937	0.814	0.868
EfficientNetV2L	Normal vs. abnormal (raw data)	0.875	1.000	0.800	0.889
	Normal vs. abnormal (cropped data)	0.946	0.908	1.000	0.940
	SD vs. ID group	0.946	0.939	0.954	0.945
ConvNeXtXLarge	Normal vs. abnormal (raw data)	0.897	0.857	0.943	0.896
	Normal vs. abnormal (cropped data)	0.969	0.968	0.970	0.965
	SD vs. ID group	0.911	0.892	0.933	0.907

CNN, convolutional neural network; SD, surgically-corrected dilatation; ID, inflammatory/infectious dilatation.

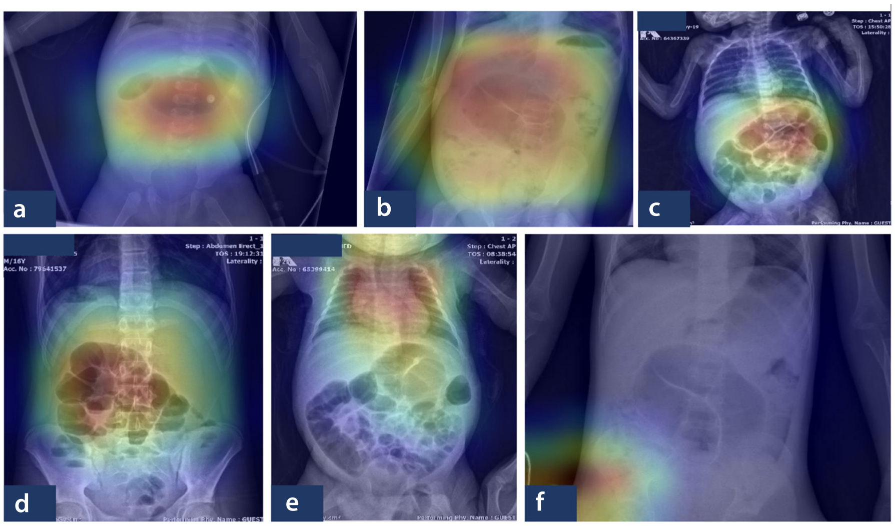


Figure 3. In the gradient-weighted class activation map (Grad-CAM) heatmap, the location of the findings was correctly identified for patients with gastrointestinal dilatation requiring surgery with diagnoses of duodenal atresia (**a**), midgut volvulus (**b**), meconium ileus (**c**), and perforated appendicitis (**d**). However, in two patients diagnosed with intestinal malrotation/midgut volvulus (**e, f**), the weight of the Grad-CAM heatmap was incorrectly identified.

diographs (2,210 for training and 1,453 for testing) were used, with 74 showing signs of obstruction.[19] In this study, the pre-trained InceptionV3 CNN model was fine-tuned using the transfer learning method, trained with their dataset, and then tested. The AUC was calculated as 0.84, the sensitivity as 83.8%, and the specificity as 68.1%.

In a subsequent study conducted by the same team, a new dataset consisting of 5,558 radiographs was created using images obtained from their hospital and a second hospital.[20] The average age of the patients in this dataset was 59.1 and 59.9 years, which differed significantly from the causes of obstruction in our patient group. Again, using InceptionV3, the researchers trained and tested the model with the second dataset.

For comparison, 1,453 test images were independently evaluated by three radiologists. The sensitivity of the radiologists ranged from 28.5% to 65.5% (average 44%), whereas the CNN model achieved 82.9%. The specificity of the radiologists ranged from 96.4% to 99.6% (average 98.4%), whereas the CNN model achieved 92.5%. The radiologists' positive predictive value (PPV) ranged from 43% to 78% (average 62%), whereas the CNN model's PPV was 28%. The low PPV in the CNN model was due to a high number of false positives.

Upon examining these false positives, it was found that while the intestinal segments were within physiological limits and considered normal clinically and radiologically, the CNN model identified them as positive. Increasing the number of similar images in the training set could potentially improve the model's performance and address this issue.

In another UK-based study on the same subject, a dataset of abdominal radiographs (445 normal and 445 with GI obstructions) from 990 adult patients was classified using transfer learning and ensemble modeling with five pre-trained CNN models: VGG16, DenseNet121, NasNetLarge, InceptionV3, and Xception.[4] Of the dataset, 800 images were used for training, 80 for validation, and 110 for testing. Among the 110 test images, there were 5 false negatives and 4 false positives. Among the models, DenseNet121 was trained using CheXNet, which consisted of chest radiograph images, whereas the other models were trained with ImageNet. The validation loss rate of the DenseNet121 model was significantly lower than that of the other models, at 43%. In previous studies where CNN models were applied to abdominal radiographs, the highest accuracy rate achieved was 92%. Although similar or slightly better performance metrics were achieved in our study, ours is the first to reach these levels in a pediatric patient group. Additionally, upon examining the image samples from the aforementioned study, it is evident that the images were standardized in size and cropped to include only the abdomen. In our study, automatic cropping was applied, but the cropping process only sometimes achieved the desired level in every dataset. This may have caused a decrease in performance metrics. The performance metrics of our study and the aforementioned studies are presented in the Supplementary Table 1.

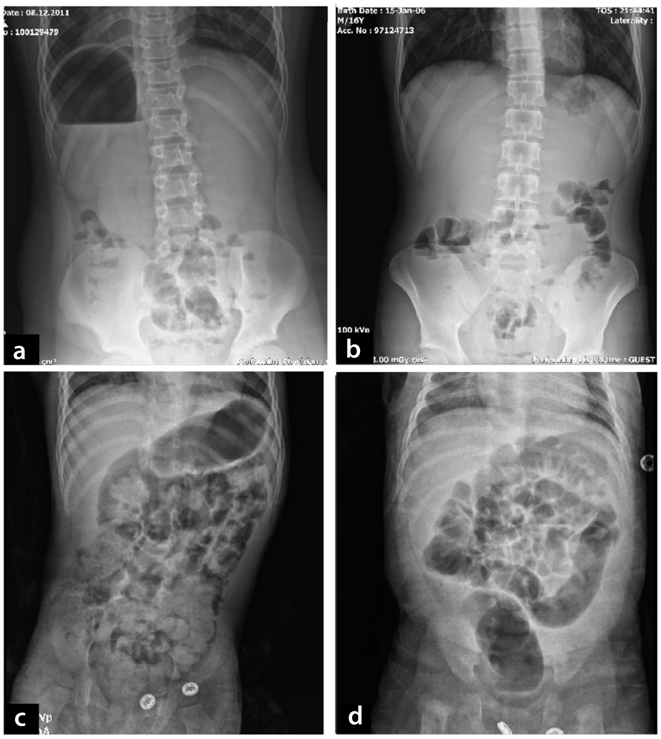


Figure 4. Surgical diagnosis, age, sex, and radiographs of abnormal cases classified as normal (false negatives) when tested with ConvNeXtXLarge are shown. The name labels on images were manually cropped before presenting in the figure. **(a)** An 11-year-old boy with situs inversus and perforated appendicitis; **(b)** a 16-year-old-boy with perforated appendicitis; **(c)** a 2-year-old boy with postoperative adhesions and Ladd band excision; **(d)** a 2-month-old boy with Hirschsprung's disease.

In all three test runs of the model on our dataset, false-positive results were more frequent than false negatives. At first glance, this could potentially lead to unnecessary surgical or medical treatment. However, since patients with positive results will also be evaluated through laboratory data, clinical examinations, and symptoms, the likelihood of unnecessary surgery due to false positives is very low. It could, however, result in a loss of time and resources due to additional tests and examinations. However, false-negative cases are more dangerous, as they could lead to the oversight of positive cases in the busy working environment of emergency rooms or outpatient clinics. In the model analysis, false negatives were about one-third as frequent as false positives, with 60% of these being patients within the ID group. The false-negative rate was relatively low for more critical SD cases. When examining sensitivity, the performance metric most affected by false-negative data, the sensitivity in the InceptionResNetV2, EfficientNetV2L, and ConvNeXtXLarge models was above 95%.

The main limitation of the study is the small sample size. In CNN models, the amount of data is one of the most important factors for performance improvement. For radiographic studies, there are open-access chest radiograph datasets provided by different institutions, with the number of images approaching 225,000.[28] However, to our knowledge, no such dataset currently exists for abdominal radiographs. In children, radiographs are used far less frequently than in adults due to the potential harm of ionizing radiation. Therefore, multicenter studies are needed to reach sufficient sample sizes. To mitigate this limitation, data augmentation was applied during the training phase. However, data augmentation could result in higher performance metrics than what might be achieved in practical applications.

The SD group in the study included 16 different etiologies, and since the number of cases for each disease was too small when evaluated individually, performance metrics for specific disease groups could not be assessed separately. Another limitation of our study is that some patients had multiple radiographs taken on different days during their illness, and radiographs taken during follow-up after a diagnosis was made were also included in the study to increase the sample size. As the diagnostic process progresses, signs of GI obstruction become more pronounced in radiographs taken later. Therefore, if only radiographs from the initial visit had been used, performance metrics might have been lower.

When creating the control dataset, the aim was to include images representing all age groups between 0 and 18 years to ensure balanced representation during model training. However, patients with abnormal findings were mostly infants and young children. As a result, the average age of the control group (7.29 ± 5.05 years) was higher than that of the patient groups (SD: 5.47 ± 5.82 and ID: 4.22 ± 4.44 years). It is generally expected that there should be no significant difference in the age and sex distribution between the study and control groups, which may have introduced bias in our study. However, we intentionally chose to create a balanced control group for ages 0–18, as we believe our model can be applied across all stages of childhood. In the future, if large open-access datasets are made available, it would be beneficial to use age filters when selecting data for such studies.

In conclusion, this study has verified that training with transfer learning can be used in deep learning to identify GI obstruction in children with high accuracy. The appropriate preprocessing steps and fine-tuning significantly improve the performance of all models. Although there are inconsistent features in the heat map of some correctly labeled cases, these models can also be useful for depicting the location of obstruction requiring surgery and for monitoring dilatation requiring medical treatment.

Acknowledgments

Sample codes used in the study for binary classification can be accessed at https://github.com/RadDrEAyaz/Convolutional-Neural-Netwok-Usage-on-Abdominal-Radiographs-for-the-Diagnosis-of-Gastrointestinal-Obs.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary: https://d2v96fxpocvxx.cloudfront.net/c1dc3a38-51db-436b-af33-1bc7522029b3/content-images/75e9cdf5-3e32-45f1-b92c-d412498063fc.pdf

References

1. Kandasamy D, Sharma R, Gupta AK. Bowel imaging in children: part 1. *Indian J Pediatr*. 2019;86(9):805-816. [Crossref]
2. Hryhorczuk AL, Lee EY. Imaging evaluation of bowel obstruction in children: updates in imaging techniques and review of imaging findings. *Semin Roentgenol*. 2012;47(2):159-170. [Crossref]
3. Kandasamy D, Sharma R, Gupta AK. Bowel imaging in children: part 2. *Indian J Pediatr*. 2019;86:817-829. [Crossref]
4. Kim DH, Wit H, Thurston M, et al. An artificial intelligence deep learning model for identification of small bowel obstruction on plain abdominal radiographs. *Br J Radiol*. 2021;94(1122):20201407. [Crossref]
5. Atlan F, Pençe İ. An overview of artificial intelligence and medical imaging technologies. *ACIN*. 2021;5(1):207-230. [Crossref]
6. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Arvix*. 2014;27:3320-3328. [Crossref]
7. Turk Radyoloji Derneği. Radyolojik tetkik yoğunlugu. Published in January 2018. Accessed in 08.09.2024. [Crossref]
8. Goksuluk D, Korkmaz S, Zararsiz G, Karaagaoglu AE. easyROC: an interactive web-tool for ROC curve analysis using R language environment. *The R Journal*. 2016;8:213-230. [Crossref]
9. F. Chollet, Keras: the python deep learning library. 2015. [Crossref]
10. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. 2015. [Crossref]
11. Bisong, E. Google colaboratory. In: building machine learning and deep learning models on Google Cloud Platform. Apress, Berkeley, CA. 2019;59-64. [Crossref]
12. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: 2016:770-778. [Crossref]
13. Szegedy C, Liu W, Jia Y, et al. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: 2015:1-9. [Crossref]
14. Chollet F. Xception: deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017:1251-1258. [Crossref]
15. Tan M, Le QV. Efficientnet: rethinking model scaling for convolutional neural networks. In: *International conference on machine learning*. 2019;6105-6114. [Crossref]
16. Liu Z, Mao H, Wu CY, et al. A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022;11976-11986. [Crossref]
17. Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. *CVPR09*. 2009. [Crossref]
18. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging*. 2016;35(5):1299-1312. [Crossref]
19. Cheng PM, Tejura TK, Tran KN, Whang G. Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. *Abdom Radiol (NY)*. 2018;43(5):1120-1127. [Crossref]
20. Cheng PM, Tran KN, Whang G, Tejura TK. Refining convolutional neural network detection of small-bowel obstruction in conventional radiography. *AJR Am J Roentgenol*. 2019;212(2):342-350. [Crossref]
21. Kwon G, Ryu J, Oh J, et al. Deep learning algorithms for detecting and visualising intussusception on plain abdominal radiography in children: a retrospective multicenter study. *Sci Rep*. 2020;10(1):17582. [Crossref]
22. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal Appl*. 2021;24(3):1207-1220. [Crossref]
23. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: a retrospective study. *PLoS Med*. 2018;15(11):e1002697. [Crossref]
24. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Med Hypotheses*. 2020;140:109761. [Crossref]
25. Almutairi TM, Ismail MMB, Bchir O. X-ray based COVID-19 classification using lightweight EfficientNet. *J Artif Intell*. 2022;4(3):167-187. [Crossref]
26. Rothrock SG, Green SM, Hummel CB. Plain abdominal radiography in the detection of major disease in children: a prospective analysis. *Ann Emerg Med*. 1992;21(12):1423-1429. [Crossref]
27. Carroll AG, Kavanagh RG, Ni Leidhin C, Cullinan NM, Lavelle LP, Malone DE. Comparative effectiveness of imaging modalities for the diagnosis and treatment of intussusception:

a critically appraised topic. *Acad Radiol*. 2017;24(5):521-529. [Crossref]

28. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019;33:590-597. [Crossref]

29. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* 2017;31(1):4278-4284. [Crossref]

30. Mehmood A. Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks. *IEEE Access*. 2021;9:138283-138295. [Crossref]