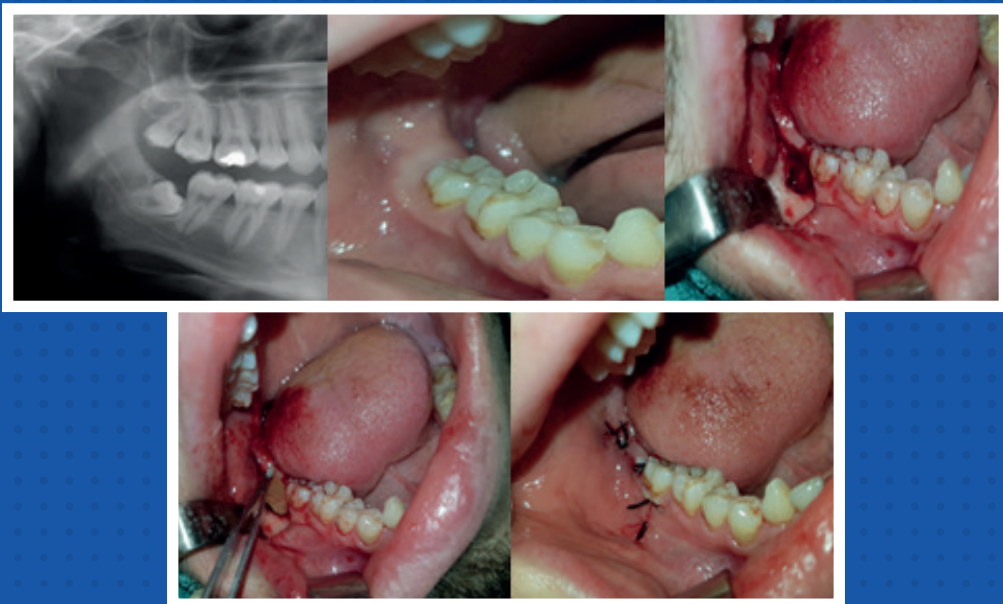


ISSN: 2979-9538

# **NORTHWESTERN MEDICAL JOURNAL**

Year: 2026 ▪ Volume: 6 ▪ Issue: 4



**Official Journal of Izzet Baysal  
Training and Research Hospital**

# **NORTHWESTERN MEDICAL JOURNAL**

Year: 2026 ▪ Volume: 6 ▪ Issue: 4

ISSN: 2979-9538

[www.nwmedj.org](http://www.nwmedj.org)

Official Journal of Izzet Baysal  
Training and Research Hospital

# Northwestern Medical Journal

**ISSN:** 2979-9538 (Online)

## **Owner**

Owner on behalf of the Bolu İzzet Baysal Training and Research Hospital  
Assoc. Prof. Süleyman Özsarı, MD  
Director, Bolu İzzet Baysal Training and Research Hospital, Türkiye

## **Publication Type**

International peer-reviewed journal

## **Publication Frequency and Language**

Quarterly (January, April, July, October), English

## **Abstracting and Indexing**

Northwestern Medical Journal indexed in TR Index, DOAJ, GALE, J-Gate, and TürkMedline.

## **Editor in Chief**

Prof. Ahmet Ural, MD  
Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Otorhinolaryngology,  
Head & Neck Surgery, Türkiye

## **Publisher**

İzzet Baysal Training and Research Hospital

## **Publisher Address**

İzzet Baysal Training and Research Hospital, Gölköy 14280 Bolu, Türkiye  
Phone: +90 374 253 46 18 / +90 374 253 46 56  
E-mail: [nwmedj@gmail.com](mailto:nwmedj@gmail.com)  
Web: [izzetbaysaleah.saglik.gov.tr](http://izzetbaysaleah.saglik.gov.tr)

## **Publishing Services**

Akdema Informatics and Publishing  
Address: Kızılay Mah. Gazi Mustafa Kemal Bulvarı No: 23/8 06420 Çankaya/Ankara  
Certificate number: 52576  
E-mail: [bilgi@akdema.com](mailto:bilgi@akdema.com)  
Tel: +90 533 166 80 80  
Web: [www.akdema.com](http://www.akdema.com)

Northwestern Medical Journal is an open access journal. All articles are published under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

You can reach all publication policies and author guidelines from [www.nwmedj.org](http://www.nwmedj.org)

## Editorial Board

### Editor in Chief

Prof. Ahmet Ural, MD

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Otorhinolaryngology,  
Head & Neck Surgery, Türkiye

E-mail: ahmetural@ibu.edu.tr

ORCID: 0000-0002-6088-1415

### Associate Editors

Prof. Muhammed Nur ÖGÜN, MD

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Neurology, Türkiye

E-mail: dr.mogun@gmail.com

ORCID: 0000-0001-5524-5767

Prof. Ümit Doğan, MD

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Ophthalmology, Türkiye

E-mail: u\_dogan@hotmail.com

ORCID: 0000-0002-8249-2621

Prof. Tülin Fırat, MD

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Histology and Embryology, Türkiye

E-mail: tulins2000@gmail.com

ORCID: 0000-0002-8317-7092

Assist. Prof. Murat Taşçı, MD

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Internal Medicine, Division of  
Rheumatology, Türkiye

E-mail: dr\_murat9113@hotmail.com

ORCID: 0000-0002-4635-5179

Assoc. Prof. Başar Erdivanlı, MD

Recep Tayyip Erdogan University, Faculty of Medicine, Department of Anesthesiology and Reanimation,  
Rize, Türkiye

E-mail: berdivanli@gmail.com

ORCID: 0000-0002-3955-8242

Prof. Amir Minovi, MD

St. Elisabeth-Krankenhaus Köln-Hohenlind, Department of Otorhinolaryngology - Head & Neck Surgery,  
Cologne, Germany

E-mail: amir.minovi@hohenlind.de

ORCID: 0000-0002-9560-0249

Assoc. Prof. Norsuhana Omar  
Sains Malaysia University, Department of Physiology, Penang, Malaysia  
E-mail: suhanakk@usm.my  
ORCID: 0000-0002-1975-6487

**Biostatistics Editor**

Assoc. Prof. Mehmet Karadağ  
Hatay Mustafa Kemal University, Faculty of Medicine Department of Biostatistics, Hatay, Türkiye  
E-mail: mkarad@gmail.com  
ORCID: 0000-0001-9539-4193

**Language Editor**

Müge Bakioğlu  
Akdema Informatics and Publishing, Ankara, Türkiye

**Publishing Assistant**

Social Worker Ayfer Say  
Bolu İzzet Baysal Training and Research Hospital, Türkiye  
E-mail: nwmedj@gmail.com  
ORCID: 0000-0002-7591-1007

# Northwestern Medical Journal

## Advisory Board

### **Prof. Niyazi Acer, PhD**

İstanbul Arel University, Faculty of Medicine, Department of Anatomy, Türkiye

### **Elif Açar, Consultant, MD**

Yozgat Sorgun Public Hospital, Department of Chest Diseases, Türkiye

### **Assist. Prof. Nurcan Akbaş Güneş, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Family Medicine, Türkiye

### **Assoc. Prof. Barış Akcan, MD**

Aydın Adnan Menderes University, Faculty of Medicine, Department of Pediatrics, Division of Neonatology, Türkiye

### **Assoc. Prof. Remzi Adnan Akdoğan, MD**

Recep Tayyip Erdoğan University, Faculty of Medicine, Department of Internal Medicine, Division of Gastroenterology, Türkiye

### **Şengül Aksakal, MD**

Health Sciences University, Samsun Training and Research Hospital, Faculty of Medicine, Department of Internal Medicine, Division of Immunology, Türkiye

### **Assoc. Prof. Gülali Aktaş, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Internal Medicine, Türkiye

### **Assoc. Prof. Ferit Akıl, MD**

Dicle University, Faculty of Medicine, Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

### **Assoc. Prof. İbrahim Faruk Aktürk, MD**

Bakırköy Training and Research Hospital, Department of Cardiology, Türkiye

### **Assist. Prof. Murat Alışık, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Biochemistry, Türkiye

### **Assoc. Prof. Turhan Aran, MD**

Karadeniz Technical University, Faculty of Medicine, Department of Gynecology and Obstetrics, Türkiye

### **Assist. Prof. Ali Osman Arslan, PhD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of Medicine, Department of Medical Biology, Türkiye

### **Assoc. Prof. Hesna Müzeyyen Astarıcı, MD**

Lokman Hekim University, Faculty of Medicine, Department of Pathology, Türkiye

### **Burçin Meryem Atak, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of Medicine, Department of Internal Medicine, Türkiye

### **Assoc. Prof. Elif Ateş, MD**

Karadeniz Technical University, Faculty of Medicine, Department of Family Medicine, Türkiye

### **Assist. Prof. Asma Ali Awadh, MD**

Public Health, Ministry of Health, Nairobi, Kenya

### **Prof. Dr. Erol Ayaz, PhD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Parasitology, Türkiye

### **Assoc. Prof. Pınar Çağlar Aytaç, MD**

Başkent University Hospital Adana, Faculty of Medicine, Department of Obstetrics and Gynecology, Türkiye

### **Assist. Prof. Zümrüt Bahat, MD**

Karadeniz Technical University, Faculty of Medicine, Department of Radiation Oncology, Türkiye

### **Assist. Prof. Abdallah Bajaber, MD**

University of Nairobi, Kenyatta National Hospital, Department of Pediatrics, Kenya

### **Assoc. Prof. Seval Bayrak, DDS, PhD**

Bolu Abant İzzet Baysal University, Faculty of Dentistry, Department of Dentomaxillofacial Radiology, Türkiye

### **Assist. Prof. Mustafa Behçet, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Microbiology, Türkiye

### **Prof. Devrim Bektaş, MD**

Giresun University, Faculty of Medicine, Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

### **Cansu Benli, Consultant, PhD**

Health Sciences University, Samsun Training and Research Hospital, Faculty of Medicine, Department of Pathology, Türkiye

### **Assoc. Prof. Güray Can, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Internal Medicine, Division of Gastroenterology, Türkiye

### **Assist. Prof. Hatice Sonay Yalçın Cömert, MD**

Karadeniz Technical University, Faculty of Medicine, Department of Pediatric Surgery, Türkiye

### **Prof. Meryem Çam, MD**

İstanbul Arel University, Faculty of Medicine, Department of Histology and Embryology, Türkiye

### **Assist. Prof. Oğuz Çatal, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of General Surgery, Türkiye

### **Assist. Prof. Kaan Çelik, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Emergency Medicine, Türkiye

### **Assoc. Prof. Ayhan Çetinkaya, PhD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Physiology, Türkiye

**Assoc. Prof. Hatice Bengü Çobanoğlu, MD**

Karadeniz Technical University, Faculty of Medicine, Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

**Assoc. Prof. Emine Dağıstan, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Radiology, Türkiye

**Assoc. Prof. Cem Dane, MD**

İstinye University, Faculty of Medicine, Department of Gynecology and Obstetrics, İstanbul, Türkiye

**Ayşegül Danış, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of Medicine, Department of Pediatrics, Division of Pediatric Neurology, Türkiye

**Prof. Abdullah Demirhan, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Anesthesiology and Reanimation, Türkiye

**Assist. Prof. Murat Dıramalı, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of Medicine, Department of Anatomy, Türkiye

**Assoc. Prof. Mustafa Dilek, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Pediatrics, Division of Neonatology, Türkiye

**Assist. Prof. İbrahim Dönmez, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Cardiology, Türkiye

**Assoc. Prof. Görkem Dülger, MD**

Düzce University, Faculty of Medicine, Department of Medical Biology, Türkiye

**Assoc. Prof. Esra Ercan, DDS, PhD**

Çanakkale Onsekiz Mart University, Faculty of Dentistry, Department of Periodontology, Türkiye

**Emrah Erdal, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of Medicine, Department of Cardiology, Türkiye

**Assoc. Prof. Güliz Erdem, MD**

İstanbul Kent University, Department of Cardiology, Faculty of Health Sciences, Türkiye

**Assist. Prof. Selma Erdoğan Düzcü, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Pathology, Türkiye

**Assoc. Prof. Hashairi Fauzi, MD**

Universiti Kebangsaan Malaysia, Faculty of Medicine, Department of Emergency, Kubang Kerian, Kelantan, Malaysia

**Assist. Prof. Ali Gökkaaya, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Plastic - Reconstructive and Aesthetic Surgery, Türkiye

**Assoc. Prof. Pınar Göksedef, MD**

Yeniüyüzyıl University, Faculty of Medicine, Department of Gynecology and Obstetrics, Türkiye

**Assoc. Prof. Hasan Tahsin Gözdaş, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Infectious Diseases, Türkiye

**Prof. Adnan Gücük, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Urology, Türkiye

**Prof. Ender Güçlü, MD**

Düzce University, Faculty of Medicine, Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

**Assoc. Prof. Mehmet Yalçın Günal, MD**

Alanya Alaaddin Keykubat University, Faculty of Medicine, Department of Physiology, Türkiye

**Assoc. Prof. Akif Güneş, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

**Assoc. Prof. Orçun Gürbüz, MD**

Bursa International Ceylan Hospital, Department of Cardiovascular Surgery, Türkiye

**Assist. Prof. Müjgan Gürler, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Internal Medicine, Türkiye

**Assist. Prof. Ali Güvey, MD**

Kütahya Health Sciences University, Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

**Prof. Ali İrfan Güzel, PhD**

Recep Tayyip Erdoğan University, Faculty of Medicine, Department of Medical Biology, Rize, Türkiye

**Filiz Halıcı, Consultant, MD**

Dışkapı Yıldırım Beyazıt Training and Research Hospital, Department of Public Health, Türkiye

**Assist. Prof. Abdullahi Mohammed Hasan, MD**

Abrar University, Faculty of Medicine, Department of Cardiology, Mogadishu, Somalia

**Assist. Prof. Tuan Salwani Tuan Ismail**

Universiti Sains Malaysia, Faculty of Medicine, Department of Pathology, Kubang Kerian, Kelantan, Malaysia

**Prof. Cengiz Işık, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Orthopedics and Traumatology, Türkiye

**Prof. Mustafa İmamoğlu, MD**

Karadeniz Technical University, Faculty of Medicine, Department of Pediatric Surgery, Division of Pediatric Urology, Türkiye

**Assist. Prof. Dono İndarto**

Universitas Sebelas Maret, Faculty of Medicine, Department of Physiology and Biomedical Laboratory, Surakarta, Indonesia

**Prof. Yıldız İyidoğan, PhD**

İstanbul Arel University, Faculty of Medicine, Department of Biochemistry, Türkiye

**Assoc. Prof. İbrahim Karagöz, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine, Department of Anesthesiology and Reanimation, Türkiye

**Assist. Prof. Elif Karalı, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Otorhinolaryngology - Head & Neck Surgery, Türkiye

**Assist. Prof. Yasin Emre Kaya, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Orthopedics and Traumatology, Türkiye

**Prof. Arzu Kılıç, MD**

Balıkesir University, Faculty of Medicine, Department of  
Dermatology, Türkiye

**Assist. Prof. Aysu Kıyan, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Public Health, Türkiye

**Gencehan Kumtepe, Consultant, MD**

Isparta Meddem Hospital, Department of Cardiovascular Surgery,  
Türkiye

**Assoc. Prof. Şerife Mehlika Kuşkonmaz, MD**

Ankara Training and Research Hospital, Department of Internal  
Medicine, Endocrinology And Metabolic Diseases, Türkiye

**Assoc. Prof. Mehmet Zahid Koçak, MD**

Necmettin Erbakan University, Faculty of Medicine, Department of  
Internal Medicine, Türkiye

**İsmail Mall, PhD**

Durban University of Technology, Faculty of Health Science,  
Department of Pharmacology, South Africa

**Aslı Kurtar Mansiroğlu, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of  
Medicine, Department of Cardiology, Türkiye.

**Ozayr Haroon Mahomed, MD**

University of Kwazulu Natal, Faculty of Medicine, Department  
Public Health, Durban South Africa

**Assist. Prof. Rohimah Mohamud**

Universiti Sains Malaysia, Faculty of Medicine, Department of  
Immunology, Kubang Kerian, Kelantan, Malaysia

**Assoc. Prof. Salim Neşelioğlu, MD**

Yıldırım Beyazıt University, Faculty of Medicine, Department of  
Biochemistry, Türkiye

**Assist. Prof. Mohd Zarawi Mat Nor**

Universiti Sains Malaysia, Faculty of Medicine, Department of  
Medical Education, Kubang Kerian, Kelantan, Malaysia

**Assoc. Prof. Julia Omar**

Universiti Sains Malaysia, Faculty of Medicine, Department of  
Pathology, Kubang Kerian, Kelantan, Malaysia

**Ali Can Önal, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of  
Medicine, Department of Pathology, Türkiye

**Assoc. Prof. Zerrin Özergin Coşkun, MD**

Recep Tayyip Erdoğan University, Faculty of Medicine, Department  
of Otorhinolaryngology - Head & Neck Surgery, Rize, Türkiye

**Assoc. Prof. Zeynep Güneş Özünal, MD**

İstanbul Maltepe University, Faculty of Medicine, Department of  
Pharmacology, Türkiye

**Hande Özkalaycı, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty of  
Medicine, Department of Medical Genetics, Türkiye

**Assist. Prof. Can Özlü, MD**

Kütahya Health Sciences University, Faculty of Medicine,  
Department of Internal Medicine, Division of Hematology, Türkiye

**Assoc. Prof. Yusuf Öztürk, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Child & Adolescent Psychiatry, Türkiye

**Prof. Mualla Polat, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Dermatology, Türkiye

**Assist. Prof. Ömer Polat, MD**

Düzce University, Faculty of Medicine, Department of  
Neurosurgery, Türkiye

**Kop Roly, MD**

Hospital Beyond Boundaries, Phnom Penh, Cambodia

**Prof. Muslim Uddin Sabuj**

Bangladesh Chattogram International Medical College and Hospital,  
Department of Pediatrics, Chittagong, Bangladesh

**Assist. Prof. Parisa Sharafi, PhD**

TOBB ETU University, Faculty of Medicine, Department of Medical  
Biology, Türkiye

**Assist. Prof. Emrullah Söğütülen, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Urology, Türkiye

**Assist. Prof. Mürsel Şahin, MD**

Karadeniz Technical University, Faculty of Medicine, Department of  
Cardiology, Türkiye

**Assist. Prof. Aslıhan Şaylan, PhD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Histology and Embryology, Türkiye

**Assoc. Prof. Tuba Taslamacıoğlu Duman, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Internal Medicine, Türkiye

**Duygu Tecer, Consultant, MD**

Health Sciences University, Gülhane Training and Research  
Hospital, Faculty of Medicine, Department of Internal Medicine,  
Division of Rheumatology, Türkiye

**Ferhat Tek, Consultant, MD**

Asklepios Klinikum Harburg, Otorhinolaryngology - Head & Neck  
Surgery, St George Hamburg, Germany

**Assist. Prof. Beliz Öztok Tekten, MD**

Bolu Abant İzzet Baysal University Faculty of Medicine Department  
of Emergency Medicine, Türkiye

**Assoc. Prof. Hakan Temiz, MD**

Dicle University, Faculty of Medicine, Department of Microbiology,  
Diyarbakır, Türkiye

**Prof. Mustafa Tunalı, DDS, PhD**

Çanakkale Onsekiz Mart University, Faculty of Dentistry, Departmet  
of Periodontology, Türkiye

**Assist. Prof. Sevim Türay, MD**

Düzce University, Faculty of Medicine, Department of Pediatrics,  
Division of Pediatric Neurology, Türkiye

**Assist. Prof. Tengku Ahmad Damitri Al-Astani Tengku Din, PhD**

Universiti Sains Malaysia, Department of Pathology, Kubang Kerian,  
Kelantan, Malaysia

**Assist. Prof. Serdar Ferit Toprak, MD**

Dicle University, Faculty of Medicine, Department of  
Otorhinolaryngology - Head & Neck Surgery, Türkiye

**Assist. Prof. Sinan Tüfekci, MD**

Tekirdağ Namık Kemal University, Faculty of Medicine, Department  
of Pediatrics, Division of Neonatology, Türkiye

**Meral Uluköylü Mengüç, Consultant, MD**

Bolu İzzet Baysal Training and Research Hospital, Faculty  
of Medicine, Department of Internal Medicine, Division of  
Hematology, Türkiye

**Assoc. Prof. Gülbahar Ustaoglu, DDS, PhD**

Bolu Abant İzzet Baysal University, Faculty of Dentistry,  
Department of Periodontology, Türkiye

**Prof. Özge Uzun, PhD**

İstanbul Arel University, Faculty of Medicine, Department of  
Pharmacology, Türkiye

**Hüseyin Übeyli, Consultant, MD**

Asklepios Klinikum Harburg, Orthopedics and Traumatology, St  
George Hamburg, Germany

**Assist. Prof. Brian Wasita**

Universitas Sebelas Maret, Faculty of Medicine, Department of  
Anatomical Pathology, Surakarta, Indonesia

**Assist. Prof. Osman Yakşı, MD**

Bolu Abant İzzet Baysal University Faculty of Medicine Department  
of Thoracic Surgery, Türkiye

**Assist. Prof. Kerem Yaman, PhD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Parasitology, Türkiye

**Assist. Prof. Mohamad Nurman Yaman, MD**

University of Kwazulu Natal, Faculty of Medicine, Department  
Public Health, Durban South Africa

**Assist. Prof. Mustafa Fatih Yaşar, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Physical Medicine and Rehabilitation, Türkiye

**Assoc. Prof. İsa Yıldız, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Anesthesiology and Reanimation, Türkiye

**Banu Yılmaz, Consultant, MD**

Bursa Provincial Health Directorate, Department of Public Health,  
Türkiye

**Assist. Prof. Murat Yılmaz, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Neurology, Türkiye

**Assoc. Prof. Hamit Yoldaş, MD**

Bolu Abant İzzet Baysal University, Faculty of Medicine,  
Department of Anesthesiology and Reanimation, Türkiye

**Assoc. Prof. Sadık Yurttutan, MD**

Kahramanmaraş Sütçü İmam University, Faculty of Medicine,  
Department of Pediatrics, Division of Neonatology, Türkiye

**Assoc. Prof. Ahmet Yüksel, MD**

Bursa City Hospital, Department of Cardiovascular Surgery, Türkiye

**Assist. Prof. Susilorini, MD, MSi, Med, SpPA**

Universitas Islam Sultan Agung, Faculty of Medicine, Department of  
Pathology Anatomy, Semarang Indonesia

# Contents

## Editorial

|                  |   |
|------------------|---|
| Ahmet Ural ..... | x |
|------------------|---|

## Research Articles

### **Laser Doppler evaluation of gingival and pulpal microcirculation after impacted mandibular third molar surgery: A retrospective comparison of L-PRF and T-PRF**

|                                |     |
|--------------------------------|-----|
| Büşra Meşeci, Neşet Akay ..... | 275 |
|--------------------------------|-----|

### **Applicability and validation of a Clavien-Dindo classification for postoperative complications in lung cancer surgery**

|                                                                                                     |     |
|-----------------------------------------------------------------------------------------------------|-----|
| Osman Yakşı, Hacı Ali Kılıçgün, Ezel Erşen, Muhammed Yavuz Haktanır, Akif Turna, Kamil Kaynak ..... | 289 |
|-----------------------------------------------------------------------------------------------------|-----|

### **Analysis of maxillary aesthetic region morphology for immediate implant planning: A CBCT study**

|                                  |     |
|----------------------------------|-----|
| Berfin Canpolat, Raif Alan ..... | 296 |
|----------------------------------|-----|

### **Readability assessment of ChatGPT-4 and ChatGPT-5 responses to ear care questions**

|                                            |     |
|--------------------------------------------|-----|
| Cemal Özyılmaz, Mehmet Fatih Karakuş ..... | 308 |
|--------------------------------------------|-----|

### **Comparative evaluation of four large language models for postoperative patient education after septorhinoplasty: A bilingual cross-sectional study**

|                                                                        |     |
|------------------------------------------------------------------------|-----|
| Zülküf Küçüktağ, Aykut Özdoğan, Gökçe Saygı Uysal, Yağmur Uluhan ..... | 317 |
|------------------------------------------------------------------------|-----|

### **AI-based diagnosis of hallux valgus: evaluation of measurements derived from foot X-ray images**

|                                                                                    |     |
|------------------------------------------------------------------------------------|-----|
| Mahmut Kurtboğan, Gülçin Ray, Abdullah Ray, İbrahim Kürtül, Mihriban Karatop ..... | 330 |
|------------------------------------------------------------------------------------|-----|

### **Comparison of thoracolumbar fascia flexibility and its relationship with trunk proprioception and spinal posture in individuals with nonspecific low back pain and healthy peers**

|                                         |     |
|-----------------------------------------|-----|
| Hüsna Nur Çördükçü, Ramazan Kurul ..... | 343 |
|-----------------------------------------|-----|

### **Peripheral blood inflammation indices versus age in predicting gastritis histopathology: A cross-sectional study**

|                                         |     |
|-----------------------------------------|-----|
| Muhammet Fatih Keyif, Ferdi Bolat ..... | 354 |
|-----------------------------------------|-----|

## Case Report

### **Catamenial bilateral conjunctival hemorrhage associated with endometriosis: A rare case report**

|                        |     |
|------------------------|-----|
| Elif Betül Esmer ..... | 364 |
|------------------------|-----|

# Editorial

## Reflecting on Growth, Charting the Future: Northwestern Medical Journal in 2026

As we approach the final quarter of 2026, it is an opportune moment to reflect on the trajectory of Northwestern Medical Journal (NWMJ) and the evolving landscape of open-access medical publishing. Since its inaugural issue in 2021, NWMJ has remained steadfast in its mission to serve as a platform for contemporary data across surgery, clinical sciences, and basic medical research. What began as the official scientific publication of Izzet Baysal Training and Research Hospital has steadily matured into an international, peer-reviewed journal indexed in DOAJ, Crossref, and other prominent databases.

The October 2026 issue reflects the diversity that has come to define our publication. The articles featured in this issue span a wide spectrum of medical disciplines. This breadth is not accidental; it is a deliberate reflection of our commitment to fostering interdisciplinary dialogue across the health sciences. We remain particularly proud that NWMJ does not charge authors for submission or processing, ensuring that scholarly dissemination remains equitable and accessible to researchers regardless of institutional funding. Yet, growth brings responsibility. The coming years will demand that we remain vigilant about the rigor of our peer-review processes and the ethical standards of the research we publish. As the volume of submissions increases, so too must our collective commitment to methodological transparency and reproducibility.

Looking ahead to 2027 and beyond, NWMJ is poised to expand its reach further. Our editorial board, representing diverse specialties from Türkiye and abroad, is actively working to strengthen international collaborations and attract submissions from a broader geographic base. We encourage our readers—clinicians, researchers, and students alike—to consider NWMJ not only as a source of information but as a partner in the dissemination of their work. The future of medical knowledge lies in collective effort, and we are honored to play our part.

Sincerely,  
**Prof. Ahmet Ural, M.D.**  
Editor-in-chief

# Laser Doppler evaluation of gingival and pulpal microcirculation after impacted mandibular third molar surgery: A retrospective comparison of L-PRF and T-PRF

Büşra Meşeci<sup>1</sup>, Neşet Akay<sup>2</sup>

<sup>1</sup>Department of Oral and Maxillofacial Surgery, Okmeydanı Oral & Dental Health Center, İstanbul, Türkiye

<sup>2</sup>Department of Oral and Maxillofacial Surgery, Specialist Clinic, Bolu, Türkiye

**Cite as:** Meşeci B, Akay N. Laser Doppler evaluation of gingival and pulpal microcirculation after impacted mandibular third molar surgery: A retrospective comparison of L-PRF and T-PRF. Northwestern Med J. 2026;6(4):275-288.

## Abstract

**Objective:** The aim of this study was to compare the effects of leukocyte- and platelet-rich fibrin (L-PRF) and titanium-prepared platelet-rich fibrin (T-PRF), placed into extraction sockets after impacted mandibular third molar surgery, on the gingival and pulpal microcirculation of the adjacent mandibular second molar.

**Methods:** This retrospective comparative clinical study included 60 patients aged 20–30 years with fully impacted mesioangular mandibular third molars. Patients were divided into two groups according to the platelet concentrate used during surgery: L-PRF (n=30) and T-PRF (n=30). Gingival blood flow in the mucoperiosteal flap and pulpal blood flow in the adjacent mandibular second molar were measured with laser Doppler flowmetry, a well-established noninvasive method. Measurements were performed preoperatively and on postoperative day 7. Perfusion values were recorded as perfusion units (PU) and analyzed using non-parametric statistical methods with Bonferroni correction applied for multiple comparisons.

**Results:** Both the L-PRF and T-PRF groups demonstrated significant increases in pulpal and gingival perfusion values on postoperative day 7 compared with baseline measurements (all Bonferroni-adjusted  $p < 0.001$ ). In the T-PRF group, pulpal perfusion increased by 8.10 PU and gingival perfusion increased by 11.09 PU, whereas the corresponding increases in the L-PRF group were 6.52 PU and 6.81 PU, respectively. Direct intergroup comparison of change scores (postoperative – preoperative) revealed that the increase in both pulpal blood flow ( $U = 759.0$ ;  $Z = 4.568$ ;  $p < 0.001$ ;  $r = 0.590$ ) and gingival blood flow ( $U = 814.0$ ;  $Z = 5.382$ ;  $p < 0.001$ ;  $r = 0.695$ ) was significantly greater in the T-PRF group. Intergroup comparisons further demonstrated higher postoperative perfusion values in the T-PRF group. The largest between-group difference was observed for gingival perfusion ( $U = 900.0$ ;  $Z = 6.653$ ;  $p < 0.001$ ;  $r = 0.859$ ), representing a very large effect size. A moderate effect size was observed for postoperative pulpal perfusion ( $r = 0.332$ ).

**Conclusion:** Both platelet-rich fibrin protocols were associated with increased gingival and pulpal perfusion during the early postoperative healing period. T-PRF was associated with higher postoperative perfusion measurements,

**Corresponding author:** Neşet Akay **E-mail:** nesetakay@gmail.com

**Received:** 09.03.2026 **Accepted:** 24.08.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

particularly in gingival tissues; this finding may indicate a more pronounced early microcirculatory response. However, because clinical healing outcomes were not directly evaluated and no untreated control group was included, the clinical significance of the observed perfusion differences remains uncertain and should be confirmed by prospective randomized controlled studies.

**Keywords:** impacted third molar, laser doppler flowmetry, platelet-rich fibrin

## INTRODUCTION

The extraction of impacted mandibular third molars remains one of the most frequently performed surgical procedures in oral and maxillofacial surgery. Despite standardized surgical techniques, postoperative complications such as pain, swelling, trismus, bleeding, alveolar osteitis, and delayed soft and hard tissue healing are still commonly encountered (1–3). These outcomes are closely related to the severity of surgical trauma and the inflammatory and vascular responses that develop at the extraction site (2,3). Therefore, strategies aimed at enhancing local microcirculation and promoting wound healing have attracted increasing clinical interest (4–6).

Platelet concentrates have been extensively investigated as autologous biomaterials capable of enhancing tissue regeneration by delivering growth factors and cytokines directly to the surgical site (7–9). Among these, platelet-rich fibrin (PRF) is a second-generation platelet concentrate characterized by a fibrin matrix rich in platelets and leukocytes and obtained without the use of anticoagulants (7,10). Leukocyte- and platelet-rich fibrin (L-PRF) has been shown to promote the sustained release of angiogenic and regenerative mediators, including platelet-derived growth factor, vascular endothelial growth factor (VEGF), and transforming growth factor- $\beta$ , thereby promoting angiogenesis, modulation of inflammation, and early wound healing (5,6,8).

Titanium-prepared platelet-rich fibrin (T-PRF) was subsequently developed as a modification of conventional PRF protocols in which titanium tubes are used instead of glass tubes for platelet activation (11,12). Experimental and clinical studies have demonstrated that T-PRF forms a denser and more organized fibrin network and exhibits a longer resorption time, which may allow a more prolonged and controlled release of growth factors (11,13,14).

These structural characteristics suggest that T-PRF may provide a more sustained biological environment for tissue repair than conventional L-PRF (13,15,16).

Microcirculation plays a critical role in wound healing because adequate blood flow is essential for oxygen delivery, nutrient transport, immune cell migration, and angiogenesis (17–19). Laser Doppler flowmetry (LDF) is an established and non-invasive method for the objective assessment of tissue microcirculation and has been used to evaluate microvascular blood flow in oral tissues, including the dental pulp and gingiva (20–22). Because LDF enables quantitative assessment of perfusion changes over time, it has become a valuable tool for monitoring healing dynamics following oral surgical procedures (21,23).

However, evidence comparing the effects of L-PRF and T-PRF on gingival and pulpal microcirculation following impacted mandibular third molar surgery remains limited. This retrospective comparative clinical study aimed to compare the effects of L-PRF and T-PRF, placed into extraction sockets after the surgical removal of impacted mandibular third molars, on gingival blood flow of the mucoperiosteal flap and pulpal blood flow of the adjacent mandibular second molar, using LDF. It was hypothesized that both platelet concentrates would be associated with increased local microcirculation during early wound healing, with a potentially greater response in the T-PRF group (6,13).

## MATERIALS AND METHODS

### Study design and review of records

In this retrospective comparative clinical study, the records of patients who underwent extraction of impacted mandibular third molars with placement of PRF into the extraction socket and who underwent routine LDF measurements preoperatively and on postoperative day 7 were reviewed.

Patients were grouped according to the PRF protocol routinely used during the respective treatment period. During the early phase of the study, L-PRF served as the standard clinical protocol. Following the introduction and routine availability of titanium tubes in our clinic, T-PRF became the preferred protocol. Therefore, allocation of patients to the L-PRF and T-PRF groups was based on chronological treatment periods rather than on patient-related clinical characteristics, surgeon preference, or patient selection. No patient-related clinical criteria were used to determine which platelet concentrate to apply.

Of the eligible records, 60 patients aged 20–30 years with fully impacted mesioangular mandibular third molars who received either L-PRF or T-PRF were included in the study. The baseline demographic and clinical characteristics of the study groups are presented in Table 1. All included teeth were classified as mesioangular impactions and exhibited a similar level of surgical difficulty, according to the Pell and Gregory and the Winter classification systems (24). To minimize operator-related variability, all surgical procedures were performed by the same experienced oral and maxillofacial surgeon using a standardized surgical protocol. Changes in gingival and pulpal blood flow were analyzed retrospectively.

## Ethical approval and informed consent

Ethical approval was obtained from the institutional Clinical Research Ethics Committee before the study (date: 27.04.2021; decision no: 2021/62). Patient records containing signed informed consent forms were evaluated. The study was conducted in accordance with the principles of the Declaration of Helsinki.

## Inclusion criteria

Patients meeting the following criteria were included in the study:

- Being between 20 and 30 years of age,
- Having a fully impacted mesioangular mandibular third molar,
- Having been treated with L-PRF or T-PRF,
- Having complete baseline and postoperative day 7 LDF recordings.

## Exclusion criteria

Patients were excluded if any of the following conditions were present:

- Presence of a systemic disease,
- Use of systemic medications,

**Table 1.** Baseline characteristics of the study groups.

| Variable                                        | L-PRF<br>(n=30)  | T-PRF<br>(n=30)  | Exact<br>p value | Bonferroni-adjusted<br>p value |
|-------------------------------------------------|------------------|------------------|------------------|--------------------------------|
| Age (years), mean $\pm$ SD                      | 22.97 $\pm$ 2.47 | 22.33 $\pm$ 2.67 | 0.160            | —                              |
| Age (years), median (min–max)                   | 22.0 (20–30)     | 22 (20–30)       | —                | —                              |
| Female, n (%)                                   | 18 (60.0)        | 22 (73.3)        | 0.273            | —                              |
| Male, n (%)                                     | 12 (40.0)        | 8 (26.7)         | 0.273            | —                              |
| Left side (tooth #38), n (%)                    | 11 (36.7)        | 18 (60.0)        | 0.071            | —                              |
| Right side (tooth #48), n (%)                   | 19 (63.3)        | 12 (40.0)        | 0.071            | —                              |
| Tooth impaction status                          | Fully impacted   | Fully impacted   | —                | —                              |
| Pell-Gregory Classification                     | Class II-B       | Class II-B       | 1.000            | —                              |
| Winter Classification                           | Mesioangular     | Mesioangular     | 1.000            | —                              |
| Baseline pulpal perfusion (PU), mean $\pm$ SD   | 15.10 $\pm$ 3.47 | 14.17 $\pm$ 2.99 | 0.021            | 0.086                          |
| Baseline gingival perfusion (PU), mean $\pm$ SD | 14.97 $\pm$ 2.86 | 15.77 $\pm$ 3.28 | 0.017            | 0.071                          |

Data are presented as mean  $\pm$  SD, median (min–max), or n (%), where applicable. All included teeth were fully impacted mesioangular mandibular third molars. Patients were selected from cases with comparable surgical difficulty according to the Pell and Gregory and Winter classification systems. L-PRF: leukocyte- and platelet-rich fibrin; T-PRF: titanium-prepared platelet-rich fibrin; SD: standard deviation; min–max: minimum–maximum; n: number of patients; #38 and #48: mandibular left and right third molars (FDI notation); PU: perfusion unit.

- Use of local or topical oral medications,
- Smoking,
- Pregnancy,
- History of radiotherapy or chemotherapy to the head and neck region,
- Bisphosphonate use by any route of administration (oral, intravenous, or subcutaneous),
- Acute infection,
- Pain or swelling associated with the impacted tooth,
- Presence of a lesion within the flap design,
- Previous surgical intervention in the same region.

In addition, patients were excluded if any of the following conditions were present in the adjacent mandibular second molar:

- Missing tooth,
- Presence of dental caries,
- Extensive restoration or prosthetic restoration,
- Periapical radiolucency,
- Previous root canal treatment,
- Deep periodontal pocket,
- Periodontal destruction.

### Study groups

Group 1 (L-PRF group) comprised patients in whom L-PRF was placed in the extraction socket (n=30).

Group 2(T-PRF group) consisted of patients in whom T-PRF was placed into the extraction socket (n=30).

### Blinding of LDF measurements

Preoperative and postoperative LDF measurements were performed by a clinician blinded to the platelet concentrate used in each socket. The measurement procedure was non-invasive and did not affect the postoperative healing process. This blinded assessment strategy reduced observer-related bias, thereby enhancing the internal validity of the perfusion assessment.

### Laser Doppler flowmetry protocol

The LDF device (Perimed Periflux 5000, Perimed, Sweden) was calibrated according to the manufacturer's instructions. The same dental unit

was used for all measurements and was fixed at a 90° angle to the ground. Patients were seated in an upright position, with the Frankfurt horizontal plane positioned parallel to the floor, and were instructed to remain motionless during recording.

All LDF measurements were performed under standardized clinical conditions using the same measurement setup and probe stabilization protocol. Patients were instructed to avoid caffeine-containing beverages, vigorous physical activity, and analgesic intake for at least 2 hours before the measurements. All recordings were obtained in the same clinical examination room under standardized environmental conditions (25). To avoid potential alterations in local perfusion, measurements on postoperative day 7 were performed before suture removal.

To standardize probe placement, alginate impressions were taken before surgery and plaster models were prepared. A 2-mm-thick transparent acrylic splint(Essix-type) was fabricated for each patient and was extended distally to include the flap boundaries. In accordance with previous studies emphasizing the importance of probe stabilization and position standardization in LDF measurements (26,27), guide slots were created to ensure that the probe could be positioned at the same location during each measurement.

The pulpal measurement point (A) was defined as the midpoint of the mesiodistal and occlusocervical axes of the adjacent second molar. Gingival measurements were obtained from three points located within the flap boundaries:

Point B: The intersection of a point 5 mm apical to the incision line and a point 5 mm distal to the distal surface of the adjacent second molar,

Point C: A point 5 mm distal to the cervical corner of the adjacent second molar,

Point D: A point 5 mm apical to the distocervical corner of the adjacent second molar.

Measurements were recorded for 30 seconds at each point using PeriSoft for Windows®(Perimed) software.

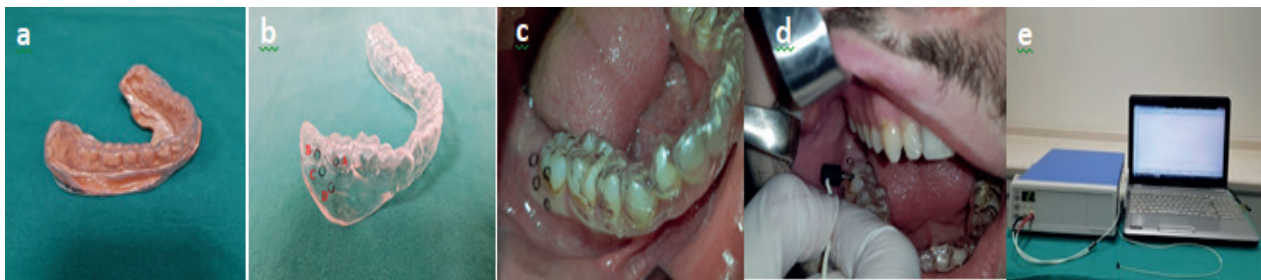


**Figure 1.** Standardization of laser Doppler flowmetry measurements.

- (a) Acrylic splint prepared on the plaster model.
- (b) Measurement points for pulpal (A) and gingival blood flow (B–D).
- (c) Intraoral placement of the acrylic splint.
- (d) Positioning of the laser Doppler flowmetry probe during measurement.
- (e) Laser Doppler flowmetry system (PeriFlux 5000, Perimed, Sweden) connected to a computer for real-time monitoring and recording of tissue perfusion measurements.

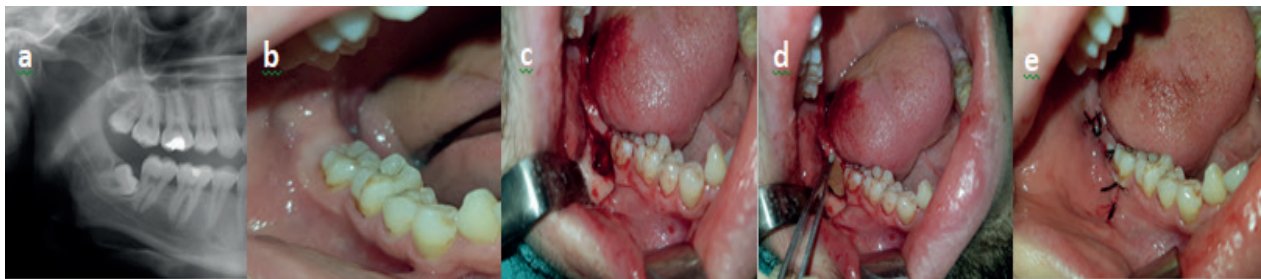


**Figure 2.** Surgical procedure and application of platelet-rich fibrin.

- (a, b) Impacted mandibular third molar teeth included in the study.
- (c) Extraction socket after tooth removal.
- (d) Placement of L-PRF or T-PRF into the extraction socket.
- (e) Primary wound closure with sutures.

The use of individualized acrylic splints, predefined reference points, and a standardized measurement environment was preferred to reduce variability and improve the reproducibility of LDF recordings (26,27) (Figure 1).

### Surgical procedure

Following the preoperative LDF measurements, a local anesthetic solution containing articaine hydrochloride and epinephrine (Maxicaine Fort®, Vem Medicine, Türkiye) was administered for inferior alveolar, lingual, and buccal nerve blocks. A full-thickness triangular mucoperiosteal flap was elevated. Bone removal was performed using a round bur under sterile saline irrigation. Tooth sectioning was performed with a tungsten-carbide fissure bur, and the tooth fragments were removed. The socket was curetted, the sharp bony edges were smoothed, and the area was irrigated

with 20 mL of saline solution. Hemostasis was achieved before PRF application (Figure 2).

### Preparation and application of L-PRF

Twenty milliliters of venous blood were drawn into two sterile glass tubes (10 mL each) without anticoagulant and were immediately centrifuged using the same centrifugation system (IntraSpin®, Intra-Lock, USA) throughout the study period. Centrifugation was performed for 12 minutes at 2700 rpm according to the standard Choukroun protocol; this corresponds to an approximate relative centrifugal force (RCF) of  $400 \times g$ . The fibrin clot was obtained and separated from the erythrocyte layer and was placed into the extraction socket in two layers before wound closure (7,10). All PRF preparation procedures were performed by the same clinician according to a standardized protocol (Figure 3).

### Preparation and application of T-PRF

Twenty milliliters of venous blood were transferred into two sterile Grade IV titanium tubes (10 mL each) and immediately centrifuged for 12 minutes at 2700 rpm using the centrifugation system (IntraSpin®, Intra-Lock, USA), according to the standard T-PRF preparation protocol. This procedure corresponded to an approximate RCF of  $400 \times g$ . The fibrin clot was obtained and separated from the erythrocyte fraction and was placed into the extraction socket in two layers before wound closure (11,12) (Figure 4).

### Postoperative period and postoperative measurements

The flaps were primarily closed using 3/0 silk sutures. In accordance with the routine protocol, patients were prescribed postoperative antibiotic therapy (amoxicillin/clavulanic acid), a nonsteroidal anti-inflammatory drug (naproxen sodium), and a 0.12% chlorhexidine mouthwash. Sutures were removed on postoperative day 7, and LDF measurements were repeated using the same acrylic splints and measurement points.



**Figure 3.** Preparation of leukocyte- and platelet-rich fibrin (L-PRF).

- (a, b) Blood collection into glass-coated plastic tubes.
- (c) Centrifugation procedure.
- (d, e) Separation of the fibrin clot.
- (f) Final L-PRF membranes ready for application.



**Figure 4.** Preparation of titanium-prepared platelet-rich fibrin (T-PRF).

- (a, b) Blood collection into titanium tubes.
- (c) Centrifugation procedure.
- (d) Separation of the fibrin clot.
- (e) Final T-PRF membranes ready for application.

## Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows (version 30.0; IBM Corp., Armonk, NY, USA). The distribution of continuous variables was assessed using the Shapiro–Wilk test, Kolmogorov–Smirnov test, and evaluation of skewness and kurtosis coefficients. Since the assumptions of normality were not satisfied for most variables, non-parametric statistical methods were used. Descriptive data were presented as mean  $\pm$  standard deviation (SD) and median (minimum–maximum).

Intergroup comparisons were performed using the Mann–Whitney U test, whereas within-group preoperative–postoperative comparisons were analyzed using the Wilcoxon signed-rank test. Two-tailed exact p values, standardized Z statistics, rank-biserial correlation effect sizes (r), and their corresponding 95% confidence intervals (CIs) were calculated and reported. Effect sizes were interpreted according to Cohen’s criteria as small (r=0.10), medium (r=0.30), and large (r=0.50).

Bonferroni correction was applied to control the risk of Type I error arising from multiple comparisons. Because four primary comparisons were considered, the adjusted significance level was set at  $p < 0.0125$ . Both unadjusted (exact) p values and Bonferroni-adjusted p values were reported.

To evaluate the adequacy of the sample size, an a priori power analysis was performed using G\*Power software (version 3.1.9.7; Heinrich Heine University, Düsseldorf, Germany), assuming a significance level of  $\alpha = 0.05$  and 80% statistical power. The planned sample size of 30 participants per group exceeded the calculated minimum sample size required to detect a large effect size (Cohen’s  $d \geq 0.80$ ) in independent-group comparisons. The post hoc power value based on the observed effect size was calculated for descriptive purposes only. Unless otherwise specified, for analyses with Bonferroni correction, the level of statistical significance was accepted as  $p < 0.05$ .

## RESULTS

The baseline demographic and clinical characteristics of the study groups are summarized in Table 1.

Pulpal blood flow of the adjacent mandibular second molar and gingival blood flow of the mucoperiosteal flap were expressed in perfusion units (PU). Gingival perfusion was calculated as the mean of the measurements at the three predefined measurement sites.

Assessment of normality, performed using the Shapiro–Wilk and Kolmogorov–Smirnov tests (Table 2), revealed that most variables were not normally distributed. Therefore, non-parametric statistical methods were used for all analyses. Bonferroni correction was applied to reduce the risk of error arising from multiple comparisons.

### Within-Group Comparisons

In both groups, pulpal perfusion values on postoperative day 7 were significantly increased compared with baseline values. In the T-PRF group, pulpal perfusion increased from  $14.17 \pm 2.99$  PU to  $22.28 \pm 4.81$  PU ( $\Delta = +8.10$  PU;  $Z = 4.535$ ; Bonferroni-adjusted  $p < 0.001$ ;  $r = 0.586$ , 95% CI: 0.389–0.731). In the L-PRF group, pulpal perfusion increased from  $15.10 \pm 3.47$  PU to  $21.63 \pm 3.49$  PU ( $\Delta = +6.52$  PU;  $Z = 4.782$ ; Bonferroni-adjusted  $p < 0.001$ ;  $r = 0.617$ , 95% CI: 0.431–0.753). Large effect sizes were observed for both comparisons. Within-group comparisons are presented in Table 3.

Similarly, gingival perfusion values increased significantly in both groups. In the T-PRF group, gingival perfusion increased from  $15.77 \pm 3.28$  PU to  $26.87 \pm 1.03$  PU ( $\Delta = +11.09$  PU;  $Z = 4.782$ ; Bonferroni-adjusted  $p < 0.001$ ;  $r = 0.617$ , 95% CI: 0.431–0.753). In the L-PRF group, gingival perfusion increased from  $14.97 \pm 2.86$  PU to  $21.79 \pm 0.91$  PU ( $\Delta = +6.81$  PU;  $Z = 4.782$ ; Bonferroni-adjusted  $p < 0.001$ ;  $r = 0.617$ , 95% CI: 0.431–0.753). Comparable effect sizes were obtained for gingival perfusion. The mean postoperative increases in pulpal and gingival perfusion values are shown in Figure 5.

**Table 2.** Normal distribution tests of measurement variables (Shapiro–Wilk, Kolmogorov–Smirnov, and Skewness–Kurtosis Coefficients).

| Variable                            | Group | K-S Statistic | K-S p  | S-W Statistic | S-W p  | Skewness | Kurtosis | Decision   |
|-------------------------------------|-------|---------------|--------|---------------|--------|----------|----------|------------|
| Pulpal Blood Flow (Preoperative)    | T-PRF | 0.330         | 0.002  | 0.716         | <0.001 | -1.966   | 3.012    | Non-Normal |
|                                     | L-PRF | 0.230         | 0.071  | 0.739         | <0.001 | -1.920   | 2.881    | Non-Normal |
| Pulpal Blood Flow (Postoperative)   | T-PRF | 0.372         | <0.001 | 0.590         | <0.001 | -2.291   | 3.955    | Non-Normal |
|                                     | L-PRF | 0.207         | 0.132  | 0.753         | <0.001 | -2.058   | 4.168    | Non-Normal |
| Gingival Blood Flow (Preoperative)  | T-PRF | 0.267         | 0.022  | 0.742         | <0.001 | -1.740   | 2.029    | Non-Normal |
|                                     | L-PRF | 0.264         | 0.025  | 0.694         | <0.001 | -2.128   | 3.663    | Non-Normal |
| Gingival Blood Flow (Postoperative) | T-PRF | 0.121         | 0.729  | 0.961         | 0.322  | 0.129    | -0.659   | Normal     |
|                                     | L-PRF | 0.083         | 0.976  | 0.976         | 0.715  | 0.133    | -0.612   | Normal     |

L-PRF, leukocyte- and platelet-rich fibrin; T-PRF, titanium-prepared platelet-rich fibrin; K-S, Kolmogorov–Smirnov test; S-W, Shapiro–Wilk test. Normality was assessed using the Shapiro–Wilk test (n = 30 per group). Because normality assumptions were not consistently satisfied, non-parametric tests were used for all analyses.

**Table 3.** Preoperative–postoperative comparison of measurements within each PRF group separately.

| Group | Measurement         | Time   | Mean ± SD    | Postop-Preop Difference | T    | Z      | p (Bonf.) | r (95% CI)          |
|-------|---------------------|--------|--------------|-------------------------|------|--------|-----------|---------------------|
| T-PRF | Pulpal Blood Flow   | Preop  | 14.17 ± 2.99 | +8.10                   | 12.0 | +4.535 | <0.001    | 0.586 (0.389–0.731) |
|       |                     | Postop | 22.28 ± 4.81 |                         |      |        |           |                     |
|       | Gingival Blood Flow | Preop  | 15.77 ± 3.28 | +11.09                  | 0.0  | +4.782 | <0.001    | 0.617 (0.431–0.753) |
|       |                     | Postop | 26.87 ± 1.03 |                         |      |        |           |                     |
| L-PRF | Pulpal Blood Flow   | Preop  | 15.10 ± 3.47 | +6.52                   | 0.0  | +4.782 | <0.001    | 0.617 (0.431–0.753) |
|       |                     | Postop | 21.63 ± 3.49 |                         |      |        |           |                     |
|       | Gingival Blood Flow | Preop  | 14.97 ± 2.86 | +6.81                   | 0.0  | +4.782 | <0.001    | 0.617 (0.431–0.753) |
|       |                     | Postop | 21.79 ± 0.91 |                         |      |        |           |                     |

PRF, platelet-rich fibrin; L-PRF, leukocyte- and platelet-rich fibrin; T-PRF, titanium-prepared platelet-rich fibrin; SD, standard deviation; Preop, preoperative; Postop, postoperative; T, Wilcoxon signed-rank test statistic; Z, standardized test statistic; p (Bonf.), Bonferroni-adjusted p value; r, effect size (rank-biserial correlation); 95% CI, 95% confidence interval. Postop – Preop differences were calculated as postoperative minus preoperative values; positive values indicate an increase. Significant p values are shown in bold italics.

### Intergroup comparisons

Although preoperative pulpal and gingival perfusion values differed in the unadjusted analyses, these differences were not statistically significant after Bonferroni correction (adjusted p = 0.086 and 0.071, respectively). Therefore, after correction for multiple comparisons, no statistically significant baseline imbalance was detected between the groups. Intergroup comparisons are summarized in Table 4.

Postoperative pulpal perfusion was higher in the T-PRF group than in the L-PRF group (U = 624.0; Z = 2.572;

exact p = 0.010; r = 0.332; 95% CI: 0.085–0.540); however, the difference did not remain significant after Bonferroni correction. However, a direct comparison of change scores demonstrated a greater increase in the T-PRF group (U = 759.00; Z = 4.568; p < 0.001; r = 0.590). This result corresponds to a moderate effect size. Similarly, postoperative gingival perfusion was higher in the T-PRF group (U = 900.0; Z = 6.653; exact p < 0.001; r = 0.859; 95% CI: 0.774–0.914). This finding represents a very large effect size. Effect size estimates and their corresponding 95% CIs are presented in Figure 6.

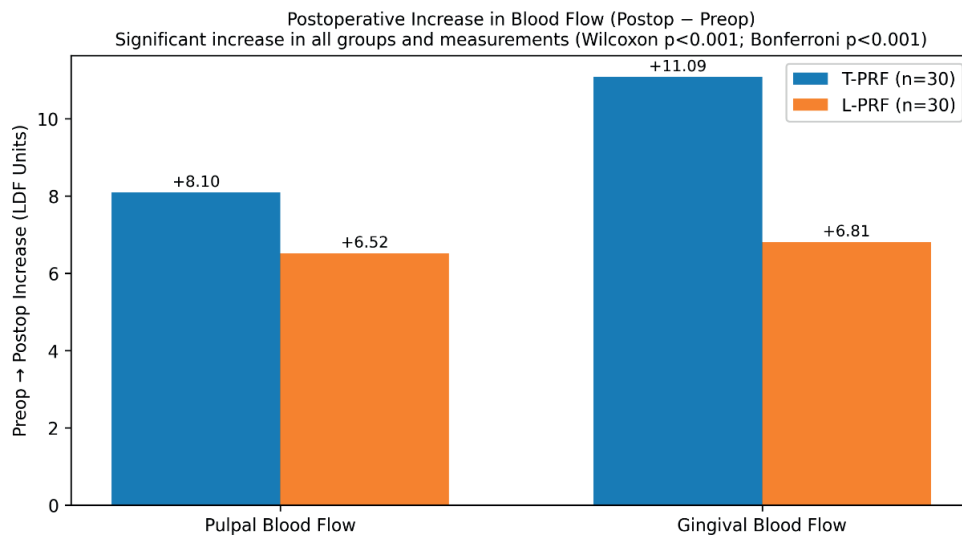


**Figure 5.** Mean postoperative increases in pulpal and gingival blood flow (postoperative minus preoperative values) in the T-PRF and L-PRF groups. Both platelet concentrates were associated with significant within-group increases in blood flow following surgery (Wilcoxon signed-rank test,  $p < 0.001$  for all comparisons). Direct comparison of change scores (postoperative – preoperative) showed larger increases in the T-PRF group for both pulpal and gingival blood flow measurements, with the most pronounced difference observed for gingival perfusion.

L-PRF, leukocyte- and platelet-rich fibrin; T-PRF, titanium-prepared platelet-rich fibrin; LDF, laser Doppler flowmetry.

**Table 4.** Comparison of measurement values according to PRF groups.

| Variable                            | Group | Mean $\pm$ SD    | Median (Min–Max)    | U      | Z      | P (Exact) | P (Bonf.) | r (95% CI)          |
|-------------------------------------|-------|------------------|---------------------|--------|--------|-----------|-----------|---------------------|
| Pulpal Blood Flow (Preoperative)    | T-PRF | 14.17 $\pm$ 2.99 | 14.92 (5.15–17.45)  | 294.00 | –2.306 | 0.021     | 0.086     | 0.298 (0.047–0.513) |
|                                     | L-PRF | 15.10 $\pm$ 3.47 | 16.20 (4.67–18.41)  |        |        |           |           |                     |
| Pulpal Blood Flow (Postoperative)   | T-PRF | 22.28 $\pm$ 4.81 | 23.81 (6.45–25.67)  | 624.00 | +2.572 | 0.010     | 0.041     | 0.332 (0.085–0.540) |
|                                     | L-PRF | 21.63 $\pm$ 3.49 | 22.54 (9.18–24.91)  |        |        |           |           |                     |
| Gingival Blood Flow (Preoperative)  | T-PRF | 15.77 $\pm$ 3.28 | 16.52 (6.34–19.80)  | 611.00 | +2.380 | 0.017     | 0.071     | 0.307 (0.058–0.521) |
|                                     | L-PRF | 14.97 $\pm$ 2.86 | 15.85 (6.23–17.70)  |        |        |           |           |                     |
| Gingival Blood Flow (Postoperative) | T-PRF | 26.87 $\pm$ 1.03 | 26.63 (25.19–29.23) | 900.00 | +6.653 | <0.001    | <0.001    | 0.859 (0.774–0.914) |
|                                     | L-PRF | 21.79 $\pm$ 0.91 | 21.87 (20.23–23.70) |        |        |           |           |                     |

PRF, platelet-rich fibrin; L-PRF, leukocyte- and platelet-rich fibrin; T-PRF, titanium-prepared platelet-rich fibrin; SD, standard deviation; Min–Max, minimum–maximum; U, Mann–Whitney U statistic; Z, standardized test statistic; p(Exact), exact two-tailed p value; p (Bonf.), Bonferroni-adjusted p value; r, effect size (rank-biserial correlation); 95% CI, 95% confidence interval. Significant p values are shown in bold italics. Significant between-group differences were observed for postoperative pulpal and gingival blood flow, with higher values in the T-PRF group.

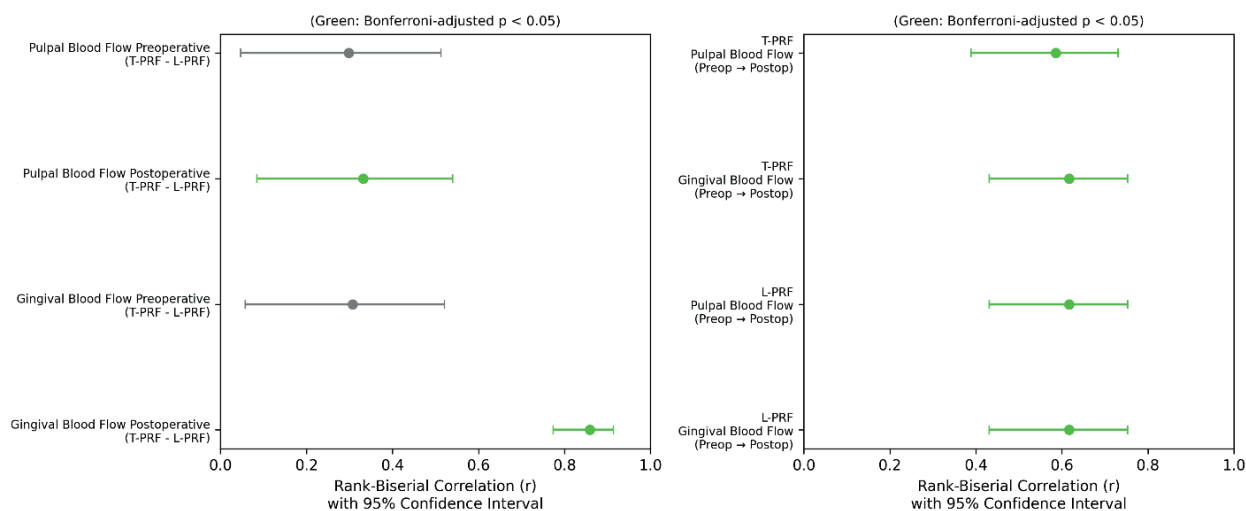


**Figure 6.** Effect Sizes (Rank-Biserial Correlation) and 95% Confidence Intervals for Intergroup (Left Panel) and Within-Group (Right Panel) Comparisons. Green markers indicate statistically significant comparisons after Bonferroni correction. Error bars represent 95% confidence intervals.

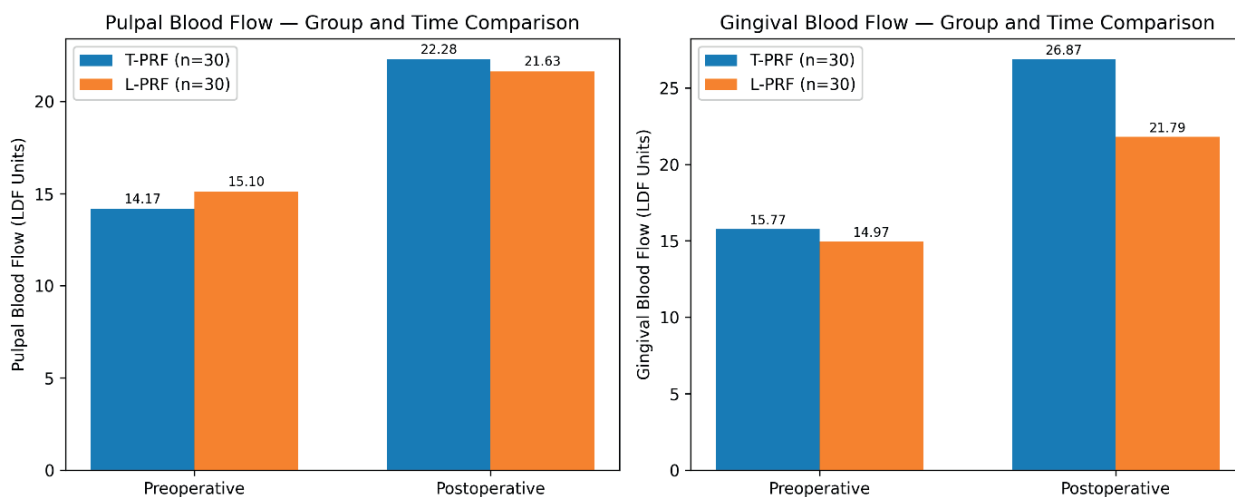


**Figure 7.** Comparison of mean preoperative and postoperative pulpal and gingival blood flow values between the T-PRF and L-PRF groups. Bars represent mean perfusion values measured by laser Doppler flowmetry (LDF). Higher postoperative perfusion values were observed in the T-PRF group; however, because of the retrospective design, these findings should be interpreted as an association rather than evidence of a causal effect.

Overall, both PRF preparations were associated with significant increases in pulpal and gingival perfusion following surgery. However, postoperative perfusion values were higher in the T-PRF group (Figure 7). To further account for potential baseline differences, change scores (postoperative – preoperative) were directly compared between the two groups using the Mann–Whitney U test. The magnitude of change

was greater in the T-PRF group for both pulpal blood flow (T-PRF: +8.10 PU vs. L-PRF: +6.52 PU;  $U = 759.00$ ;  $Z = 4.568$ ;  $p < 0.001$ ;  $r = 0.590$  [95% CI: 0.378–0.768]) and gingival blood flow (T-PRF: +11.09 PU vs. L-PRF: +6.81 PU;  $U = 814.00$ ;  $Z = 5.382$ ;  $p < 0.001$ ;  $r = 0.695$  [95% CI: 0.512–0.842]), with the largest difference observed in gingival perfusion (Table 5).

**Table 5.** Between-group comparison of change scores ( $\Delta$  = postoperative – preoperative) in pulpal and gingival blood flow (Mann–Whitney U test; n = 30 per group).

| Variable                                | Group | Mean $\pm$ SD    | Median (Min–Max)      | U      | Z      | p (Exact) | p (Bonf.) | r (95% CI)             |
|-----------------------------------------|-------|------------------|-----------------------|--------|--------|-----------|-----------|------------------------|
| Pulpal Blood Flow Change ( $\Delta$ )   | T-PRF | 8.10 $\pm$ 4.88  | 9.04<br>(–8.10–16.61) | 759.00 | +4.568 | <0.001    | <0.001    | 0.590<br>(0.378–0.768) |
|                                         | L-PRF | 6.52 $\pm$ 1.09  | 6.27<br>(3.93–8.74)   |        |        |           |           |                        |
| Gingival Blood Flow Change ( $\Delta$ ) | T-PRF | 11.09 $\pm$ 3.47 | 10.21<br>(7.36–20.09) | 814.00 | +5.382 | <0.001    | <0.001    | 0.695<br>(0.512–0.842) |
|                                         | L-PRF | 6.81 $\pm$ 2.87  | 5.98<br>(3.61–17.22)  |        |        |           |           |                        |

L-PRF, leukocyte- and platelet-rich fibrin; T-PRF, titanium-prepared platelet-rich fibrin; SD, standard deviation; Min–Max, minimum–maximum; U, Mann–Whitney U test statistic; Z, standardized test statistic; p (Bonf.), Bonferroni-adjusted p value; r, effect size (rank-biserial correlation); CI, confidence interval;  $\Delta$ , postoperative – preoperative change; PU, perfusion unit.

The T-PRF group demonstrated significantly greater increases in pulpal and gingival blood flow than the L-PRF group, with the largest effect observed for gingival perfusion.

## DISCUSSION

This study demonstrated that both L-PRF and T-PRF were associated with significant increases in gingival and pulpal perfusion on postoperative day 7 compared with baseline values. Intergroup comparisons revealed higher postoperative perfusion values in the T-PRF group; the most pronounced difference was observed in gingival perfusion ( $r=0.859$ ), whereas a moderate effect size was observed for postoperative pulpal perfusion ( $r=0.332$ ).

Wound healing is a complex and multistage process consisting of the inflammatory, proliferative, and remodeling phases (17,18). The proliferative phase, which begins within the first few days after injury and generally continues until the second or third postoperative week, is characterized by angiogenesis, granulation tissue formation, and collagen deposition (17,18). Angiogenesis is a key event during this period because the formation of new capillary vessels supports tissue regeneration by increasing local blood flow (18,19). Therefore, evaluation on postoperative day 7 is particularly important for assessing the biological activities of platelet concentrates during the early healing period (8,17).

PRF-based biomaterials are known to provide the sustained release of angiogenic growth factors, particularly VEGF, during the early healing period

(5,6,8). Previous studies have shown that L-PRF releases VEGF and other cytokines predominantly within the first 7 days, corresponding to the period of peak angiogenic activity (8). The significant increase in gingival and pulpal perfusion observed in the L-PRF group in the present study is consistent with these findings and supports a potential role for L-PRF in enhancing early vascular responses (4,5).

One of the most important findings of this study is the magnitude of the observed effect sizes. Although both PRF protocols were associated with significant increases in perfusion values, the difference between the groups was particularly pronounced for gingival perfusion, showing a very large effect size ( $r=0.859$ ). In contrast, the effect size for postoperative pulpal perfusion was moderate ( $r=0.332$ ). These findings suggest that the effect of T-PRF on early postoperative microcirculation may be more pronounced in gingival tissues than in pulpal tissues. However, the biological mechanisms underlying this difference remain speculative and require further investigation.

The significantly greater increase in perfusion observed in the T-PRF group may be related to the unique fibrin architecture of this material. Experimental studies have revealed that T-PRF forms a thicker and more compact fibrin network than L-PRF, which may be resorbed more slowly and therefore maintain its scaffold function for a longer period (11,13,14). Its prolonged persistence

within the tissues may allow the sustained release of growth factors and cytokines, thereby contributing to angiogenic signaling and enhanced microvascular perfusion (13,15). The present findings are consistent with previous histological and biochemical studies reporting greater cellular entrapment and more prolonged growth factor activity in T-PRF compared with conventional PRF formulations (13,14). However, fibrin architecture, growth factor release, cytokine profiles, angiogenesis, and histological healing were not directly evaluated in the present study. Therefore, these mechanisms should be regarded as possible biological explanations rather than confirmed causes of the observed perfusion differences.

LDF was selected as the assessment method because tissue perfusion is a sensitive indicator of wound-healing quality (20,21,26). Increased PU values reflect enhanced microcirculation and oxygenation, which are essential for effective healing and resistance to infection (17,18). In the present study, standardized acrylic splints were used to ensure stabilization of the LDF probe, thereby improving measurement reproducibility and minimizing artifacts related to probe movement or positioning (26,27). This methodological approach strengthens the reliability of the observed perfusion changes.

Another clinically noteworthy finding was a significant increase in pulpal blood flow in the adjacent mandibular second molar, particularly in the T-PRF group. Increased pulpal perfusion may reflect changes in the vascular response of the pulp tissue during the early postoperative period. However, pulpal vitality, sensibility responses, clinical symptoms, and long-term pulp status were not evaluated in this study. Therefore, the increase observed in LDF values should not be interpreted as evidence of pulp preservation or maintenance of vitality, but rather as an indicator of changes in pulpal microcirculation. Furthermore, the increased perfusion values may reflect transient postoperative vascular or inflammatory responses. Prospective studies incorporating vitality tests, sensibility assessments, and long-term clinical follow-up are required to clarify the biological and clinical significance of these findings and to determine whether the observed perfusion changes translate into clinically meaningful pulpal benefits.

This study has several limitations. First, its retrospective, non-randomized design carries an inherent risk of selection bias. In addition, allocation to the treatment groups was not randomized; L-PRF was used during the early phase of the study, whereas T-PRF was used after titanium tubes were introduced into routine clinical practice. This assignment of groups according to chronological treatment periods (L-PRF applied during the earlier period and T-PRF during the later) may represent an important source of bias, as surgical experience, clinical routines, patient selection, measurement practices, and postoperative care protocols may have changed over time and thereby intergroup differences. Therefore, despite the use of standardized surgical and measurement protocols and the similar surgical characteristics of the groups, residual confounding effects related to time-dependent factors and unmeasured variables cannot be completely excluded.

After applying Bonferroni correction for multiple comparisons, no statistically significant difference was detected in baseline pulpal or gingival perfusion values. Although the unadjusted analyses suggested possible differences between the groups, these findings did not retain statistical significance after correction. Therefore, the available data do not demonstrate a statistically significant imbalance in baseline perfusion measurements. Nevertheless, because the study is retrospective and non-randomized, residual confounding from unmeasured factors cannot be completely excluded, and the findings should be interpreted with caution.

Another important limitation is the absence of a control group that did not receive PRF. Although increases in gingival and pulpal perfusion compared with baseline values were observed in both the L-PRF and T-PRF groups, the current study design does not allow the determination of whether these changes were directly attributable to the platelet concentrates or merely reflected the normal physiological healing response following third molar surgery. Therefore, the findings should primarily be interpreted as a comparison between two PRF protocols and should not be considered definitive evidence that either material alone enhances postoperative healing or microcirculation.

A strength of the present study is the inclusion of normality testing, effect size calculations, CIs, and Bonferroni-adjusted analyses, which improve the robustness of the statistical evaluation. Nevertheless, the relatively limited sample size and the retrospective design remain important factors that should be considered when interpreting the findings.

In addition, the follow-up period was limited to the early healing phase. Long-term clinical outcomes, such as bone regeneration, periodontal status of the adjacent tooth, postoperative complications, and patient-reported outcome measures, were not evaluated. However, the primary aim of the present study was to assess early microcirculatory changes occurring during the proliferative phase of wound healing, and postoperative day 7 represents a biologically meaningful and widely accepted time point for this purpose (8,17). Therefore, the present findings should be regarded as indicators of early microcirculatory responses rather than evidence of long-term clinical superiority.

Prospective randomized controlled studies including an untreated control group, larger sample sizes, formal corrections for multiple comparisons, and longer follow-up periods are needed to determine whether the observed perfusion differences translate into meaningful long-term clinical benefits.

Accordingly, these findings should be interpreted as physiological observations rather than evidence of superior clinical effectiveness. However, the clinical significance of these perfusion differences, and whether they translate into improved long-term healing outcomes, have not yet been clearly established.

## CONCLUSIONS

Within the limitations of this retrospective study, both L-PRF and T-PRF were associated with significant increases in gingival and pulpal perfusion on postoperative day 7, and T-PRF was associated with greater increases, most notably in gingival perfusion. However, because no non-PRF control

group was included and the design was retrospective, these differences cannot be causally attributed to the platelet concentrates and may partly reflect the normal postoperative healing response. Prospective randomized controlled studies with a non-PRF control group, larger samples, and longer follow-up are needed to determine whether these microcirculatory differences translate into clinically meaningful healing outcomes.

## Ethical approval

Approved by the Bolu Abant İzzet Baysal University Clinical Research Ethics Committee (27.04.2021; No. 2021/62). Written informed consent was obtained from all participants prior to inclusion in the study.

## Author contribution

Concept and design: NA; Data acquisition: NA, BM; Data analysis and interpretation: NA; Drafting of the manuscript: NA, BM; Critical revision of the manuscript: NA; Final approval of the version to be published: NA, BM. All authors meet the ICMJE authorship criteria and approve the final version of the manuscript.

## Source of funding

The authors declare the study received no funding.

## Conflict of interest

The authors declare that there is no conflict of interest.

## Use of artificial intelligence

During the preparation of this manuscript, the authors used ChatGPT (GPT5.5, OpenAI, San Francisco, CA, USA) exclusively for language editing and grammatical refinement. The authors have critically reviewed, revised, and approved all generated content and take full responsibility for the final manuscript.

## Acknowledgement

The authors thank Prof. Dr. Gülbahar Ustaoglu, Department of Periodontology, Faculty of Dentistry, Bolu Abant İzzet Baysal University, for her valuable contribution to the study.

## REFERENCES

1. Menziletoglu D, Guler AY, Basturk F, Isik BK, Erdur EA. Comparison of two different flap designs for bilateral impacted mandibular third molar surgery. *J Stomatol Oral Maxillofac Surg.* 2020; 121(4): 368-72. [\[Crossref\]](#)
2. Torul D, Omezli MM, Kahveci K. Evaluation of the effects of concentrated growth factors or advanced platelet rich-fibrin on postoperative pain, edema, and trismus following lower third molar removal: a randomized controlled clinical trial. *J Stomatol Oral Maxillofac Surg.* 2020; 121(6): 646-51. [\[Crossref\]](#)
3. Kim JW, Seong TW, Cho S, Kim SJ. Randomized controlled trial on the effectiveness of absorbable collagen sponge after extraction of impacted mandibular third molar: split-mouth design. *BMC Oral Health.* 2020; 20(1): 77. [\[Crossref\]](#)
4. He Y, Chen J, Huang Y, Pan Q, Nie M. Local application of platelet-rich fibrin during lower third molar extraction improves treatment outcomes. *J Oral Maxillofac Surg.* 2017; 75(12): 2497-506. [\[Crossref\]](#)
5. Afat IM, Akdoğan ET, Gönül O. Effects of leukocyte- and platelet-rich fibrin alone and combined with hyaluronic acid on early soft tissue healing after surgical extraction of impacted mandibular third molars: a prospective clinical study. *J Craniomaxillofac Surg.* 2019; 47(2): 280-6. [\[Crossref\]](#)
6. Ratajczak J, Vangansewinkel T, Gervois P, et al. Angiogenic properties of 'leukocyte- and platelet-rich fibrin'. *Sci Rep.* 2018; 8(1): 14632. [\[Crossref\]](#)
7. Choukroun J, Diss A, Simonpieri A, et al. Platelet-rich fibrin (PRF): a second-generation platelet concentrate. Part IV: clinical effects on tissue healing. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod.* 2006; 101(3): e56-60. [\[Crossref\]](#)
8. Dohan Ehrenfest DM, de Peppo GM, Doglioli P, Sammartino G. Slow release of growth factors and thrombospondin-1 in Choukroun's platelet-rich fibrin (PRF): a gold standard to achieve for all surgical platelet concentrates technologies. *Growth Factors.* 2009; 27(1): 63-9. [\[Crossref\]](#)
9. Dohan Ehrenfest DM, Rasmusson L, Albrektsson T. Classification of platelet concentrates: from pure platelet-rich plasma (P-PRP) to leukocyte- and platelet-rich fibrin (L-PRF). *Trends Biotechnol.* 2009; 27(3): 158-67. [\[Crossref\]](#)
10. Dohan DM, Choukroun J, Diss A, et al. Platelet-rich fibrin (PRF): a second-generation platelet concentrate. Part I: technological concepts and evolution. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod.* 2006; 101(3): e37-44. [\[Crossref\]](#)
11. Tunalı M, Özdemir H, Küçükodacı Z, Akman S, Fıratlı E. In vivo evaluation of titanium-prepared platelet-rich fibrin (T-PRF): a new platelet concentrate. *Br J Oral Maxillofac Surg.* 2013; 51(5): 438-43. [\[Crossref\]](#)
12. Tunalı M, Özdemir H, Küçükodacı Z, et al. A novel platelet concentrate for guided bone regeneration: titanium prepared platelet-rich fibrin (T-PRF). *Gulhane Med J.* 2015; 57(2): 102-6. [\[Crossref\]](#)
13. Ravi S, Santhanakrishnan M. Mechanical, chemical, structural analysis and comparative release of PDGF-AA from L-PRF, A-PRF and T-PRF - an in vitro study. *Biomater Res.* 2020; 24: 16. [\[Crossref\]](#)
14. Bhattacharya HS, Gummaluri SS, Astekar M, Gummaluri RK. Novel method of determining the periodontal regenerative capacity of T-PRF and L-PRF: an immunohistochemical study. *Dent Med Probl.* 2020; 57(2): 137-44. [\[Crossref\]](#)
15. Ustaoglu G, Ercan E, Tunalı M. The role of titanium-prepared platelet-rich fibrin in palatal mucosal wound healing and histoconduction. *Acta Odontol Scand.* 2016; 74(7): 558-64. [\[Crossref\]](#)
16. Mitra DK, Potdar PN, Prithyani SS, Rodrigues SV, Shetty GP, Talati MA. Comparative study using autologous platelet-rich fibrin and titanium prepared platelet-rich fibrin in the treatment of infrabony defects: an in vitro and in vivo study. *J Indian Soc Periodontol.* 2019; 23(6): 554-61. [\[Crossref\]](#)
17. Broughton G, Janis JE, Attinger CE. Wound healing: an overview. *Plast Reconstr Surg.* 2006; 117(7 Suppl): 1e-S-32e-S. [\[Crossref\]](#)
18. Velnar T, Bailey T, Smrkolj V. The wound healing process: an overview of the cellular and molecular mechanisms. *J Int Med Res.* 2009; 37(5): 1528-42. [\[Crossref\]](#)
19. Kondo T, Ishida Y. Molecular pathology of wound healing. *Forensic Sci Int.* 2010; 203(1-3): 93-8. [\[Crossref\]](#)
20. Gazelius B, Olgart L, Edwall B, Edwall L. Non-invasive recording of blood flow in human dental pulp. *Endod Dent Traumatol.* 1986; 2(5): 219-21. [\[Crossref\]](#)
21. Retzepi M, Tonetti M, Donos N. Gingival blood flow changes following periodontal access flap surgery using laser Doppler flowmetry. *J Clin Periodontol.* 2007; 34(5): 437-43. [\[Crossref\]](#)
22. Barry O, Wang Y, Wahl G. Determination of baseline alveolar mucosa perfusion parameters using laser Doppler flowmetry and tissue spectrophotometry in healthy adults. *Acta Odontol Scand.* 2020; 78(1): 31-7. [\[Crossref\]](#)
23. Naoum S, Goonewardene M, Abbott PV, Karunanayake K, Budgeon C. Changes in pulp blood flow and pulp sensibility resulting from surgically assisted rapid maxillary expansion: a clinical study. *Am J Orthod Dentofacial Orthop.* 2019; 155(5): 632-41. [\[Crossref\]](#)
24. Juodzbalys G, Daugela P. Mandibular third molar impaction: review of literature and a proposal of a classification. *J Oral Maxillofac Res.* 2013; 4(2): e1. [\[Crossref\]](#)
25. Hartmann A, Azérad J, Boucher Y. Environmental effects on laser Doppler pulpal blood-flow measurements in man. *Arch Oral Biol.* 1996; 41(4): 333-9. [\[Crossref\]](#)
26. Ramsay DS, Artun J, Martinen SS. Reliability of pulpal blood-flow measurements utilizing laser Doppler flowmetry. *J Dent Res.* 1991; 70(11): 1427-30. [\[Crossref\]](#)
27. Roebuck EM, Evans DJ, Stirrups D, Strang R. The effect of wavelength, bandwidth, and probe design and position on assessing the vitality of anterior teeth with laser Doppler flowmetry. *Int J Paediatr Dent.* 2000; 10(3): 213-20. [\[Crossref\]](#)

# Applicability and validation of a Clavien-Dindo classification for postoperative complications in lung cancer surgery

Osman Yakşı<sup>1</sup>, Hacı Ali Kılıçgün<sup>1</sup>, Ezel Erşen<sup>2</sup>, Muhammed Yavuz Haktanır<sup>1</sup>, Akif Turna<sup>2</sup>, Kamil Kaynak<sup>2</sup>

<sup>1</sup>Department of Thoracic Surgery, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

<sup>2</sup>Department of Thoracic Surgery, Cerrahpaşa Faculty of Medicine, İstanbul University-Cerrahpaşa, İstanbul, Türkiye

**Cite as:** Yakşı O, Kılıçgün HA, Erşen E, Haktanır MY, Turna A, Kaynak K. Applicability and validation of a Clavien-Dindo classification for postoperative complications in lung cancer surgery. Northwestern Med J. 2026;6(4):289-295.

## Abstract:

**Background:** There is no standard scale to evaluate complications after lung cancer surgery. Many surgical fields use the Clavien-Dindo classification for postoperative complications. We adapted the Clavien-Dindo classification by modifying it to reflect the complications that developed in patients who underwent surgery for lung cancer at our clinic. We aimed to demonstrate the applicability and accuracy of this classification.

**Methods:** A retrospective study involving 802 patients who underwent lung cancer surgery at two sites was conducted between January 2002 and December 2019. Age, gender, surgical method, procedure, complications, and survival times were documented. One hundred eighty patients who experienced complications were assessed using the Clavien-Dindo classification. The correlation between this grading and survival was examined.

**Results:** Of 802 patients who underwent surgery for lung cancer, 180 patients developed 224 complications. The distribution of complications according to the Clavien-Dindo classification was as follows: Grade I: 46 patients (25.6%); Grade II: 64 patients (35.6%); Grade IIIa: 22 patients (12.2%); Grade IIIb: 12 patients (6.7%); Grade IVa: 10 patients (5.6%); Grade IVb: 10 patients (5.6%); Grade V: 16 patients (8.9%). In the survival analysis, survival decreased significantly with increasing Clavien-Dindo complication grade ( $p < 0.001$ ).

**Conclusion:** There is no standardized classification of complications after lung cancer resection. We found the Clavien-Dindo classification to be applicable and verifiable for analyzing complications after lung cancer surgery. The effect of the Clavien-Dindo classification on survival underscores the importance of classifying complications after lung cancer surgery. To clarify this important issue, a database should be created, and large-scale multicenter studies should be conducted.

**Keywords:** Clavien-Dindo classification, lung cancer, postoperative complications, surgery

**Corresponding author:** Osman Yakşı **E-mail:** osmanyakshi@gmail.com

**Received:** 10.04.2026 **Accepted:** 09.06.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

## INTRODUCTION

Lung cancer is the most prevalent disease responsible for mortality globally (1,2). Tobacco use accounts for 80–90% of all lung cancer cases; however, other hereditary and environmental risk factors also contribute. TNM classification is employed for the staging of lung cancer (3). Surgery is the preferred treatment for early-stage and locally progressed lung cancer when clinically indicated (4,5).

Postoperative complications following lung cancer surgery elevate morbidity and mortality rates (6). Approximately 30–40% of people have problems (6). Classifying postoperative problems and reaching consensus are crucial for their management and prevention. The Clavien-Dindo classification is the predominant system utilized to categorize postoperative problems arising from surgical interventions (7,8). This categorization technique categorizes problems by severity (7,8) (Table 1).

However, there is no standard classification method for classifying complications after lung cancer surgery (6). The European Society of Thoracic Surgeons (ESTS) has developed a classification system, Thoracic Morbidity and Mortality (TM&M), utilizing the ESTS database to categorize complications following lung surgery (6). The European Society of Thoracic Surgeons (ESTS) classified the complications after lung cancer surgery into two groups: minor and major (Table 2). They

classified Grade I-II as minor complications, Grade III-IV as major complications, and grade V as mortality (9).

Our study sought to elucidate the complications and mortality, and to assess the impact of these complications on survival in patients who underwent lung cancer surgery. Our objective was to assess complications following lung cancer surgery according to the Clavien-Dindo classification and demonstrate the use of this classification in thoracic surgery.

## PATIENTS AND METHODS

**Study Design:** The study included patients who underwent lung resection for primary lung cancer at two clinics between January 2002 and December 2019. The study included 802 patients, all of whom had complete control and follow-up data, with no data loss. The study recorded whether the patients developed postoperative complications. We retrospectively reviewed data from 180 patients in this series who developed complications. We recorded the age, gender, type of complication, surgical technique, type of resection applied, and survival times. The Clavien-Dindo classification graded the complications as Grade 1, Grade 2, Grade 3A, Grade 3B, Grade 4A, Grade 4B, and Grade 5. Because 180 patients developed a total of 224 complications, we recorded data on the highest-grade complication for patients who experienced more than one.

**Table 1.** Clavien-Dindo classification of surgical complications.

|                                         |                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Grade I                                 | Any deviation from the normal postoperative course without the need for pharmacological treatment or surgical, endoscopic, and radiological interventions Allowed therapeutic regimens are: drugs as antiemetics, antipyretics, analgesics, diuretics, electrolytes, and physiotherapy. This grade also includes wound infections opened at the bedside |
| Grade II                                | Requiring pharmacological treatment with drugs other than such allowed for grade I complications Blood transfusions and total parenteral nutrition are also included                                                                                                                                                                                    |
| Grade III<br>-Grade IIIa<br>-Grade IIIb | Requiring surgical, endoscopic or radiological intervention -Intervention not under general anesthesia<br>-Intervention under general anesthesia                                                                                                                                                                                                        |
| Grade IV<br>-Grade IVa<br>-Grade IVb    | Life-threatening complication (including central nervous system complications) requiring intermediate care/intensive care unit management -Single organ dysfunction (including dialysis) -Multiorgan dysfunction                                                                                                                                        |
| Grade V                                 | Death of a patient                                                                                                                                                                                                                                                                                                                                      |

**Table 2.** Grading of complication following TM & M classification system.

| Grading             |            | Definition/Treatment Performed                                                                      |
|---------------------|------------|-----------------------------------------------------------------------------------------------------|
| Minor Complications | Grade I    | Adverse event which alters the standard postoperative course without requiring a specific treatment |
|                     | Grade II   | Pharmacologic treatment or minor intervention required                                              |
| Major Complications | Grade IIIa | Surgical, radiologic, endoscopic treatment, or multitherapy required without general anesthesia     |
|                     | Grade IIIb | Surgical, radiologic, endoscopic treatment, or multitherapy required with general anesthesia        |
|                     | Grade IVa  | Intensive care unit treatment for single organ dysfunction required                                 |
|                     | Grade IVb  | Intensive care unit treatment for multiple organ dysfunction required                               |
| Mortality           | Grade V    | Adverse event which leads to death                                                                  |

**Preoperative Evaluation:** All patients underwent conventional oncological and surgical assessments. Radiological examinations included chest X-ray, computed tomography, and positron emission tomography (PET/CT). Comprehensive blood analyses and biochemical assessments were conducted. Respiratory function assessments were conducted. Advanced respiratory and cardiac evaluations were conducted when deemed essential during the patient's clinical assessment. Patients exhibiting peripheral tumor localization with a tumor diameter less than 3 cm and no mediastinal lymph node involvement as assessed by PET/CT did not receive preoperative mediastinal staging. Nonetheless, preoperative mediastinal staging was conducted using mediastinoscopy in all patients not meeting these criteria. Preoperative mediastinal staging guided management of patients with N2 disease toward systemic therapy, including neoadjuvant chemotherapy and/or definitive chemoradiation.

**Postoperative follow-up:** Patients underwent outpatient clinic evaluations at three-month intervals post-surgery. All outpatient clinic examinations included chest X-rays and biochemical analyses. Surveillance and management were conducted using thoracic tomography every six months during the first two years and annually thereafter. All patients were directed to the oncology outpatient clinic following surgery. Systemic adjuvant therapy, including chemotherapy and/or radiotherapy, was administered based on appropriate indications such as N1 or N2

disease and the presence of T2 tumors. Patients undergoing adjuvant oncological treatment were evaluated for treatment responses using PET/CT at designated intervals. All patient data were documented during these follow-ups.

**Exclusion Criteria:** The study excluded patients with benign pathology results, patients with missing data, and patients with incomplete survival information during follow-up.

**Statistical Evaluation:** Data analysis was performed using SPSS version 20.0 (SPSS Inc., Chicago, USA). Quantitative parametric data were expressed as mean  $\pm$  standard deviation (SD), and quantitative non-parametric data were expressed as median (minimum–maximum). Kaplan-Meier and Cox regression analyses were performed. A p-value  $<0.05$  was considered significant.

**Ethics:** Bolu Abant Izzet Baysal University Clinical Research Ethics Committee approval was obtained (22.06.2021/172).

## RESULTS

Of 802 patients who underwent surgery with a diagnosis of lung cancer, 180 developed 224 complications. Complications developed in 19 female (10.6%) and 161 male (89.4%) patients. The mean age was  $62.2 \pm 9.4$  (29–83 years) (Table 3).

**Table 3.** Demographic data.

|        | Patients (n=180, 22.4%) |
|--------|-------------------------|
| Age    | Mean:62.2± 9.4 (29–83)  |
| Gender |                         |
| Female | 19 (10.6%)              |
| Male   | 161 (89.4%)             |

**Table 4.** Frequencies of complications.

| Type of Complications             | N   | %     |
|-----------------------------------|-----|-------|
| Prolonged Air Leak                | 54  | 6.73  |
| Pneumonia                         | 22  | 2.74  |
| Arrhythmia                        | 19  | 2.36  |
| Hemorrhage                        | 17  | 2.11  |
| Respiratory Failure               | 15  | 1.87  |
| Atelectasis                       | 10  | 1.24  |
| Pneumothorax/Re-expansion Failure | 7   | 0.87  |
| Cardiac Arrest                    | 7   | 0.87  |
| Bronchopleural Fistula            | 7   | 0.87  |
| Subileus/ileus                    | 6   | 0.74  |
| Other                             | 60  | 7.48  |
| Total                             | 224 | 27.88 |

Out of 180 patients, 22.4% experienced 224 complications. The most common complication was a prolonged air leak, occurring in 54 patients (6.73%). Other common complications included pneumonia (22 patients, 2.74%), arrhythmia (19 patients, 2.36%), respiratory failure (15 patients, 1.87%), hemorrhage (17 patients, 2.11%), atelectasis (10 patients, 1.24%), pneumothorax/expansion defect (7 patients, 0.87%), cardiac arrest (7 patients, 0.87%), bronchopleural fistula (7 patients, 0.87%), and subileus/ileus (6 patients, 0.74%). Table 4 provides the types and rates of complications.

Patients with complications most frequently underwent right upper lobectomies (55 patients, 30.6%). The second most common resection type was a left upper lobectomy (30 patients, 16.7%). The least common resection type in patients with complications was right middle lobectomy (3 patients, 1.7%) (Table 5).

**Table 5.** Distribution of surgery types.

| Resection Types | N   | %    |
|-----------------|-----|------|
| Lobectomy       |     |      |
| RUL             | 55  | 30.6 |
| RML             | 3   | 1.7  |
| RLL             | 19  | 10.6 |
| LUL             | 30  | 16.7 |
| LLL             | 24  | 13.3 |
| Pneumonectomy   |     |      |
| RP              | 9   | 5.0  |
| LP              | 17  | 9.4  |
| Bilobectomy     |     |      |
| RUBL            | 5   | 2.8  |
| RLBL            | 8   | 4.4  |
| SR              | 10  | 5.6  |
| Total           | 180 | 100  |

RUL: Right upper lobectomy, RML: Right middle lobectomy, RLL: Right lower lobectomy, LUL: Left upper lobectomy, LLL: Left lower lobectomy, RP: Right pneumonectomy, LP: Left pneumonectomy, RUBL: Right upper bilobectomy, RLBL: Right lower bilobectomy, SR: Sublobar resection (wedge and segmentectomy).

**Table 6.** Complication rates according to the Clavien-Dindo classification.

| Grade               |            | N  | %    |
|---------------------|------------|----|------|
| Minor Complications | Grade I    | 46 | 25.6 |
|                     | Grade II   | 64 | 35.6 |
| Major Complications | Grade IIIa | 22 | 12.2 |
|                     | Grade IIIb | 12 | 6.7  |
|                     | Grade IVa  | 10 | 5.6  |
|                     | Grade IVb  | 10 | 5.6  |
| Mortality           | Grade V    | 16 | 8.9  |

The distribution of complications according to the Clavien-Dindo classification was as follows: Grade I: 46 patients (25.6%); Grade II: 64 patients (35.6%); Grade IIIa: 22 patients (12.2%); Grade IIIb: 12 patients (6.7%); Grade IVa: 10 patients (5.6%); Grade IVb: 10 patients (5.6%); Grade V: 16 patients (8.9%) (Table 6).

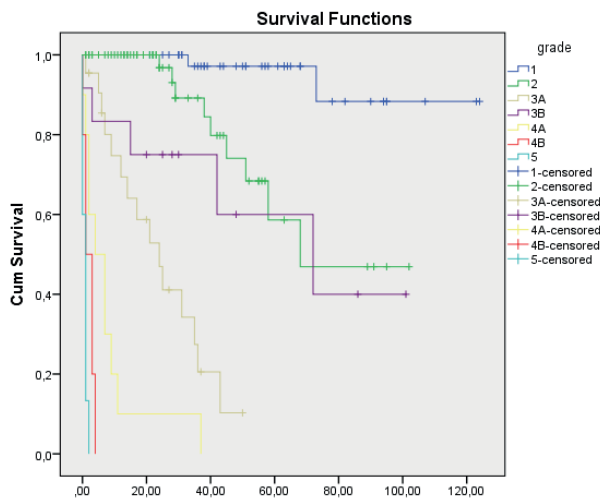


**Figure 1.** Survival analysis according to Clavien-Dindo classification. As the complication stage increases, survival is observed to be statistically significantly worse ( $p < 0.001$ ).

In the survival analysis, it was seen that survival decreased statistically significantly as the Clavien-Dindo complication grade increased ( $p < 0.001$ ) (Figure 1).

The 5-year survival of all patients was 62.3% ( $179 \pm 28.7$ ) (Figure 2). In the regression analysis performed on all patients, cancer stage, development of complications, age, and tumor size were significantly associated with survival.

## DISCUSSION

Our study demonstrated a correlation between the Clavien-Dindo classification of complications and long-term survival in patients undergoing lung resection for lung cancer. We demonstrated the usability and suitability of the Clavien-Dindo classification following thoracic surgical procedures.

Although the morbidity and mortality of lung cancer surgery have decreased in recent years, they are still high (6,10,11). There are various risk factors for complications after lung cancer surgery (12). There

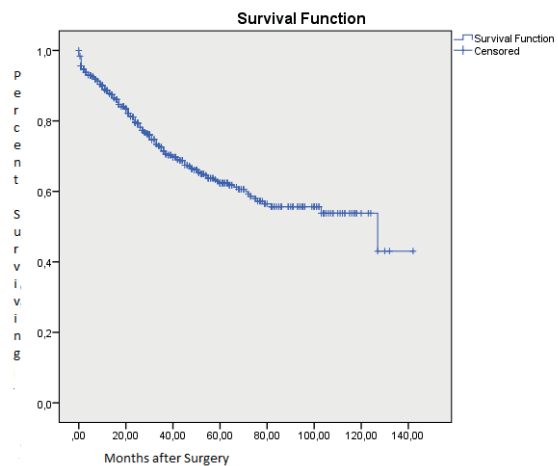


**Figure 2.** Survival analysis.

are differences in both the definition and classification of postoperative complications (6,9,13,14). There is a need for a common complication definition and classification used worldwide (6,9,13,14).

Dindo et al. emphasized that there is no consensus yet on how to define postoperative complications and that this prevents standardization (7,8). Therefore, they conducted an international survey of general surgeons and evaluated 6336 patients. They modified their classification recommendations, which they had issued since 1992, and attempted to standardize them in this study. In their 2009 study, the same researchers conducted a 5-year evaluation, revealing the validity and use of this classification across various surgical branches and many countries worldwide (7,8).

The most common complication after thoracic surgery is arrhythmia, followed by prolonged air leakage (15). Postoperative pain and prolonged air leakage are the most common causes of delayed discharge (15). Prolonged air leakage is also among the most important causes of postoperative morbidity (15,16). The literature reports that air leakage lasting more than five days causes 9.6% of morbidity after lobectomy, and 5.9% after segmentectomy (15,16). Varela et al. showed that prolonged air leakage is associated with important pulmonary complications such as atelectasis, pneumonia, and empyema (17).

Numerous studies have evaluated the complications that arise from thoracic surgery operations, using the Clavien-Dindo classification (18,19). The Clavien-Dindo classification was used by Esteva et al. to look at the problems that came up after 31 Nuss surgeries for pectus excavatum (18). They found that the Clavien-Dindo classification can be used and can be repeated after pectus excavatum surgery (18).

Watanabe et al. used the Clavien-Dindo classification in their study evaluating 43 patients receiving hemodialysis who required lung resection (19). They showed that increased smoking and prolonged surgery time caused the development of Grade 2 and higher complications (19).

The Clavien-Dindo classification has been modified and used to classify complications in thoracic surgery (6,9). The European Society of Thoracic Surgeons (ESTS) arranged the Clavien-Dindo classification using the thoracic surgery database, with Grade I-II being minor complications, Grade III-IV being major complications, and Grade V being mortality (6,9). When we revised our Clavien-Dindo classification according to the TM & M classification, 110 (61.2%) of the patients were classified as having minor complications, 54 (30.1%) as having major complications, and 16 (8.9%) died.

Due to data loss, we evaluated the limitations of our retrospective study, which excluded some patients.

In our study, we modified the Clavien-Dindo classification for complications after lung cancer surgery and showed that Clavien-Dindo grading is an applicable method for analyzing these complications.

### Ethical approval

This study has been approved by the Bolu Abant İzzet Baysal University Clinical Research Ethics Committee (approval date 22/06/2021, number 172).

### Author contribution

Concept: OY, HAK; Design: EE, MYH, AT, KK; Data Collection or Processing: MYH, EE, HAK; Analysis or Interpretation: KK, OY, EE, HAK; Literature Search: OY, MYH, EE; Writing: OY, HAK, MYH, EE. All authors

reviewed the results and approved the final version of the article.

### Source of funding

The authors declare the study received no funding.

### Conflict of interest

The authors declare that there is no conflict of interest.

## REFERENCES

1. Schabath MB, Cote ML. Cancer progress and priorities: lung cancer. *Cancer Epidemiol Biomarkers Prev.* 2019; 28(10): 1563-79. [\[Crossref\]](#)
2. Thandra KC, Barsouk A, Saginala K, Aluru JS, Barsouk A. Epidemiology of lung cancer. *Contemp Oncol (Pozn).* 2021; 25(1): 45-52. [\[Crossref\]](#)
3. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: proposals for revision of the TNM stage groupings in the forthcoming (Eighth) edition of the TNM classification for lung cancer. *J Thorac Oncol.* 2016; 11(1): 39-51. [\[Crossref\]](#)
4. Chaft JE, Rimner A, Weder W, Azzoli CG, Kris MG, Cascone T. Evolution of systemic therapy for stages I-III non-metastatic non-small-cell lung cancer. *Nat Rev Clin Oncol.* 2021; 18(9): 547-57. [\[Crossref\]](#)
5. Grunewald DH. Surgery for advanced stage lung cancer. *Semin Surg Oncol.* 2000; 18(2): 137-42. [\[Crossref\]](#)
6. Allen MS, Darling GE, Pechet TT, et al. Morbidity and mortality of major pulmonary resections in patients with early-stage lung cancer: initial results of the randomized, prospective ACOSOG Z0030 trial. *Ann Thorac Surg.* 2006; 81(3): 1013-9; discussion 1019-20. [\[Crossref\]](#)
7. Clavien PA, Barkun J, de Oliveira ML, et al. The Clavien-Dindo classification of surgical complications: five-year experience. *Ann Surg.* 2009; 250(2): 187-96. [\[Crossref\]](#)
8. Bolliger M, Kroehnert JA, Molineus F, Kandioler D, Schindl M, Riss P. Experiences with the standardized classification of surgical complications (Clavien-Dindo) in general surgery patients. *Eur Surg.* 2018; 50(6): 256-61. [\[Crossref\]](#)
9. Salati M, Refai M, Pompili C, Xiumè F, Sabbatini A, Brunelli A. Major morbidity after lung resection: a comparison between the European Society of Thoracic Surgeons Database system and the Thoracic Morbidity and Mortality system. *J Thorac Dis.* 2013; 5(3): 217-22. [\[Crossref\]](#)
10. Jacobson FL, Dezube AR, Bravo-Iñiguez C, et al. Preserving NLST mortality benefits and acceptable morbidity for lung cancer surgery in a community hospital. *J Surg Oncol.* 2021; 124(1): 124-34. [\[Crossref\]](#)

11. Ginsberg RJ. Lung cancer surgery: acceptable morbidity and mortality, expected results and quality control. *Surg Oncol.* 2002; 11(4): 263-6. [\[Crossref\]](#)
12. Motono N, Ishikawa M, Iwai S, Iijima Y, Usuda K, Uramoto H. Individualization of risk factors for postoperative complication after lung cancer surgery: a retrospective study. *BMC Surg.* 2021; 21(1): 311. [\[Crossref\]](#)
13. Rotman JA, Plodkowski AJ, Hayes SA, et al. Postoperative complications after thoracic surgery for lung cancer. *Clin Imaging.* 2015; 39(5): 735-49. [\[Crossref\]](#)
14. He T, Guo J, Chen N, et al. MediMLP: using Grad-CAM to extract crucial variables for lung cancer postoperative complication prediction. *IEEE J Biomed Health Inform.* 2020; 24(6): 1762-71. [\[Crossref\]](#)
15. Shaw JP, Locicero III J . Postoperative management of the general thoracic surgical patient. In: Shields TW, Locicero III J, Reed CE, Feins RH, editors. *General Thoracic Surgery.* 7th ed. Philadelphia: Lippincott Williams and Wilkins; 2009: 571-579.
16. Haithcock BE, Feins RH. Complications of pulmonary resection. In: Shields TW, Locicero III J, Reed CE, Feins RH, editors. *General Thoracic Surgery.* 7th ed. Philadelphia: Lippincott Williams and Wilkins; 2009: 551-559.
17. Varela G, Jiménez MF, Novoa N, Aranda JL. Estimating hospital costs attributable to prolonged air leak in pulmonary lobectomy. *Eur J Cardiothorac Surg.* 2005; 27(2): 329-33. [\[Crossref\]](#)
18. Esteva Miró C, Núñez García B, Pérez-Gaspar M, et al. Applicability of Clavien-Dindo classification in Nuss procedure surgical complications. *Cir Pediatr.* 2020; 33(4): 154-9.
19. Watanabe Y, Hattori A, Fukui M, et al. Postoperative complications and perioperative management in patients on hemodialysis undergoing lung resection. *Gen Thorac Cardiovasc Surg.* 2021; 69(3): 478-86. [\[Crossref\]](#)

# Analysis of maxillary aesthetic region morphology for immediate implant planning: A CBCT study

Berfin Canpolat<sup>1</sup>, Raif Alan<sup>1</sup>

<sup>1</sup>Department of Periodontology, Faculty of Dentistry, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye

**Cite as:** Canpolat B, Alan R. Analysis of maxillary aesthetic region morphology for immediate implant planning: A CBCT study. Northwestern Med J. 2026;6(4):296-307.

## ABSTRACT

**Aim:** To evaluate morphometric characteristics relevant to immediate implant planning in the maxillary anterior region using cone-beam computed tomography (CBCT).

**Materials and Methods:** In this retrospective study, CBCT images of 597 maxillary anterior teeth, including central incisors, lateral incisors, and canines, from 119 individuals were analyzed. Labial bone thickness, sagittal root position, apical bone height, labial concavity, undercut depth, and alveolar width were assessed. For each tooth, two virtual implant placement strategies were simulated: prosthetically driven and bone-driven.

**Results:** Labial bone thickness was <1 mm in most cases, and sagittal root position predominantly exhibited a Class III distribution (77.8%). Apical bone height was generally adequate for primary stability. More pronounced undercuts and concavities were observed in the lateral incisors. Prosthetically driven implant placement was associated with an increased risk of facial bone compromise, whereas the bone-driven approach resulted in safer implant positioning.

**Conclusion:** Immediate implant planning in the maxillary anterior region frequently requires a bone-driven approach, primarily because of thin labial bone and root-position characteristics. CBCT-based three-dimensional evaluation plays a critical role in achieving safe, predictable, and individualized implant planning.

**Keywords:** Cone-beam computed tomography, dental implants, regional anatomy

## INTRODUCTION

The three-dimensional positioning of a dental implant within the alveolar bone constitutes one of the principal factors determining the success of dental implant therapy (1,2). This requirement is particularly

critical in the maxillary anterior region, where aesthetic expectations are high (3,4). In this region, the thin facial bone structure, limited bone volume, and variable root position within the alveolus make immediate implant placement technically challenging and increase the risk of complications (5,6).

**Corresponding author:** Berfin Canpolat **E-mail:** drberfincanpolat@gmail.com

**Received:** 27.04.2026 **Accepted:** 18.09.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

Immediate implant placement is increasingly preferred because it can shorten treatment time and preserve hard- and soft-tissue contours (1,5,7,8). However, the success of this approach depends not only on surgical technique but also on the extent to which implant positioning is compatible with the existing bone morphology (1,2,5,9). In cases with insufficient facial bone thickness, implant placement along the prosthetic axis may result in facial bone perforation and soft-tissue complications (5,6,10,11).

Post-extraction alveolar bone resorption further complicates implant procedures (12,13). Pronounced early volumetric loss of the facial bone limits the availability of adequate bone support for implant placement and often necessitates a more palatal implant position (12–14). Therefore, immediate implant planning should consider prosthetic requirements and the existing bone morphology.

Cone-beam computed tomography (CBCT) has become an important imaging modality in implant planning because it enables three-dimensional evaluation of alveolar bone structure (15,16). CBCT data allow detailed analysis of preoperative parameters and are critical for safe implant positioning (6,10,11,15,17).

However, only a few studies have comprehensively evaluated critical morphological parameters, such as labial bone thickness, alveolar ridge dimension, sagittal root position, and labial concavity within the same sample (10,18–20). Moreover, studies directly comparing implant placement based on the prosthetic axis with placement optimized according to existing bone morphology remain limited (20). This gap may introduce uncertainty in clinical decision-making about which placement approach is morphologically safer and more predictable for immediate implant planning.

A comprehensive assessment of these interrelated anatomical parameters within the same population may provide a more integrated understanding of the morphological constraints affecting immediate implant placement.

Therefore, this study aimed to analyze alveolar bone morphology in the maxillary anterior region using CBCT and to evaluate the effect of these morphological

characteristics on immediate implant placement strategies.

## MATERIALS AND METHODS

### Ethical approval

This retrospective cross-sectional study was approved by the Non-Interventional Clinical Research Ethics Committee of Canakkale Onsekiz Mart University (decision no. 2025-07/07-19; April 17, 2025) and was conducted in accordance with the principles of the Declaration of Helsinki.

### Sample selection

CBCT images obtained for various clinical indications between 2022 and 2025 at the Faculty of Dentistry of Canakkale Onsekiz Mart University were retrospectively evaluated. The inclusion criteria were as follows:

- Individuals aged  $\geq 20$  years
- Presence of at least one fully erupted maxillary anterior tooth
- Teeth with complete root development
- CBCT records of sufficient image quality
- The exclusion criteria were as follows:
- Presence of periapical pathology
- Active orthodontic treatment
- History of surgical procedures in the relevant region
- Records showing pronounced alveolar bone loss or imaging artifacts
- Presence of restorations, crowns, posts, orthodontic appliances, or implants in the relevant region

A total of 597 maxillary anterior teeth, including central incisors, lateral incisors, and canines, from 119 individuals who met the study criteria were included in the analysis (Figure 1).

### Image analysis

CBCT images obtained using a NewTom 5G XL device (QR Systems, Verona, Italy) with a field of view (FOV) of  $10 \times 5$  cm, tube voltage of 110 kV, voxel size of 0.200 mm, and exposure time of 3.6 s were evaluated using NNT® Viewer software, and

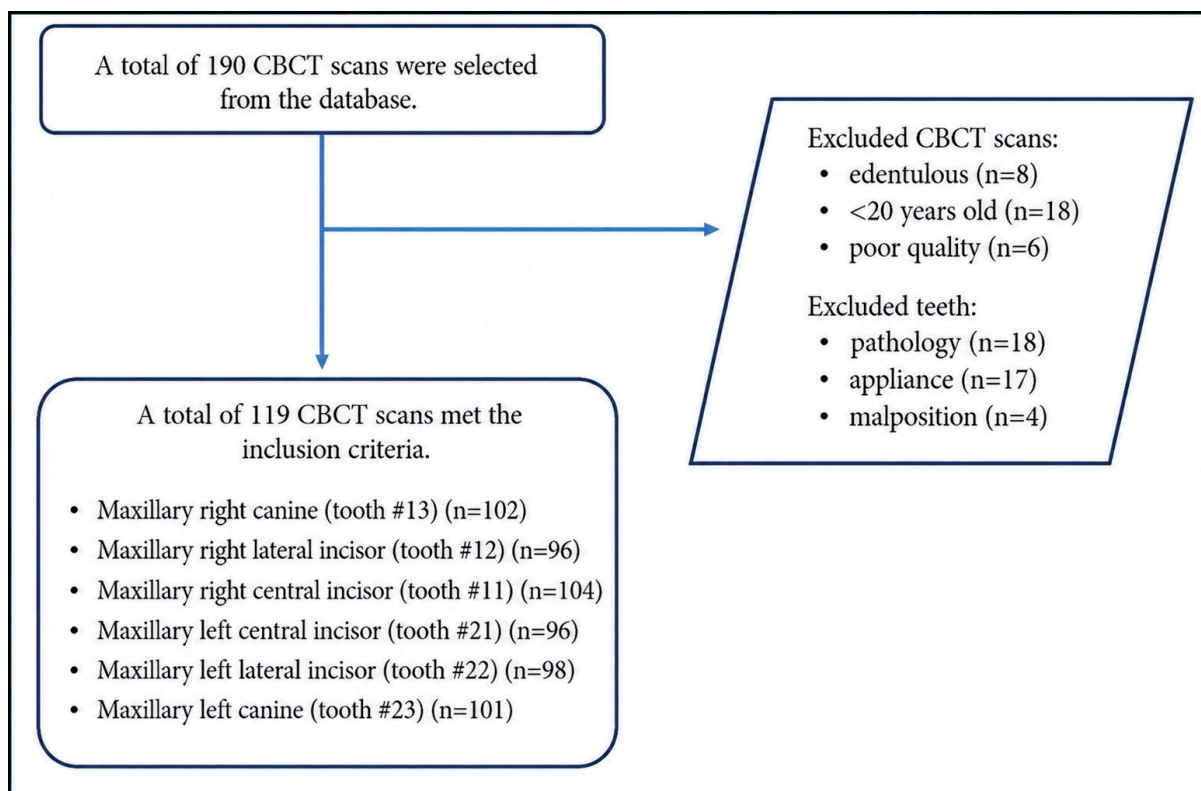


**Figure 1.** Selection of cone-beam computed tomography (CBCT) scans and distribution of the included teeth according to tooth type and exclusion criteria.

all measurements were conducted on sagittal, axial, and coronal reconstructions to ensure accuracy and standardization.

### Morphometric evaluation

All morphometric measurements were performed on sagittal sections using a reference plane that passed through the mesiodistal midpoint of the tooth and was aligned with its long axis. Labial bone thickness (LBT) was measured perpendicular to the long axis at 0, 1, 2, and 4 mm apical to the alveolar crest (LBT0-4) (21). The faciopalatal position of the root within the alveolus was assessed as the sagittal root position (SRP) according to the classification described by Gluckman et al. (11).

Apical bone support was determined by measuring the linear distance between the root apex and the relevant anatomical landmark (RABH-t) and comparing it with the distance between the virtual implant apex and the

same landmark (RABH-b) (22). Alveolar concavity was analyzed using undercut depth (LUD) and concavity angle (LCA), based on a reference line tangent to the labial cortical plate and parallel to the alveolar crest axis (9,23,24). Additionally, alveolar width was evaluated by measuring the mesiodistal distance (MD) between adjacent root surfaces and the faciopalatal dimension (FP) on sagittal sections (25).

In virtual implant placement (VIP), standard implant dimensions were selected according to tooth type: 4.1 × 10 mm for central incisors and canines and 3.3 × 10 mm for lateral incisors (20). The implant platform was positioned approximately 1 mm apical to the alveolar crest, in accordance with previous literature (26). For each tooth, two placement strategies were simulated: prosthetically-driven placement, parallel to the long axis of the tooth (VIP-t), and bone-driven placement, optimized according to the existing bone morphology (VIP-b) (Figure 2).

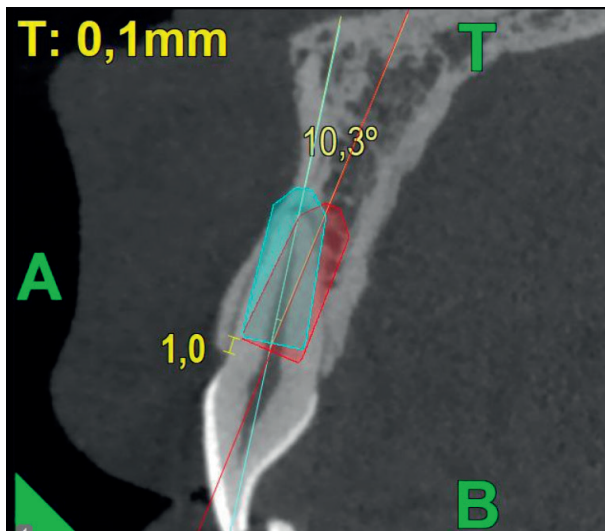


**Figure 2.** Virtual implant placement on cone-beam computed tomography (CBCT) images. Blue: virtual implant positioned parallel to the long axis of the tooth; Red: virtual implant positioned parallel to the long axis of the alveolar crest; Yellow: angle between the two implant positions.

In both approaches, the relationship between the implant and the facial bone was evaluated, and the distance between the implant and the labial bone was classified according to clinical applicability as exposed, <1 mm, or  $\geq 1$  mm. Additionally, the angular difference between the two placement strategies (VIP-angle) was recorded (20).

### Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics 22.0 (IBM Corp., Armonk, NY, USA). The distributional characteristics of continuous variables were assessed using the Kolmogorov-Smirnov test and graphical methods, including histograms and Q-Q plots. Descriptive statistics are presented as mean  $\pm$  standard error for continuous variables and as frequency (n) and percentage for categorical variables.

For comparisons between groups, the *independent-samples t-test* was used for two-group comparisons, and one-way analysis of variance (ANOVA) was used when the data were normally distributed. When ANOVA detected a significant difference, multiple

comparisons were performed using Tukey's post hoc test. Pearson correlation analysis was used to assess the relationships between continuous variables and age.

To assess intra-examiner reproducibility, the intraclass correlation coefficient (ICC) was calculated based on repeated measurements performed at a 2-week interval on selected images, and an ICC  $\geq 0.90$  was considered to indicate "high reliability." For all tests,  $p < 0.05$  was considered statistically significant.

## RESULTS

A total of 597 maxillary anterior teeth from 119 individuals were evaluated. In the sample, 57.6% were female patients ( $n = 344$ ) and 42.4% were male patients ( $n = 253$ ), with a balanced distribution between the right (50.6%) and left (49.4%) sides. The largest age group was 20–29 years (30.7%) (Table 1).

**Table 1.** Demographic and descriptive characteristics of the study sample.

| General Characteristics | n   | %    |
|-------------------------|-----|------|
| Gender                  |     |      |
| Female                  | 344 | 57.6 |
| Male                    | 253 | 42.4 |
| Location                |     |      |
| Right                   | 302 | 50.6 |
| Left                    | 295 | 49.4 |
| Age                     |     |      |
| 20–29                   | 183 | 30.7 |
| 30–39                   | 125 | 20.9 |
| 40–49                   | 160 | 26.8 |
| $\geq 50$               | 129 | 21.6 |
| Tooth                   |     |      |
| 11                      | 104 | 17.4 |
| 12                      | 96  | 16.1 |
| 13                      | 102 | 17.1 |
| 21                      | 96  | 16.1 |
| 22                      | 98  | 16.4 |
| 23                      | 101 | 16.9 |

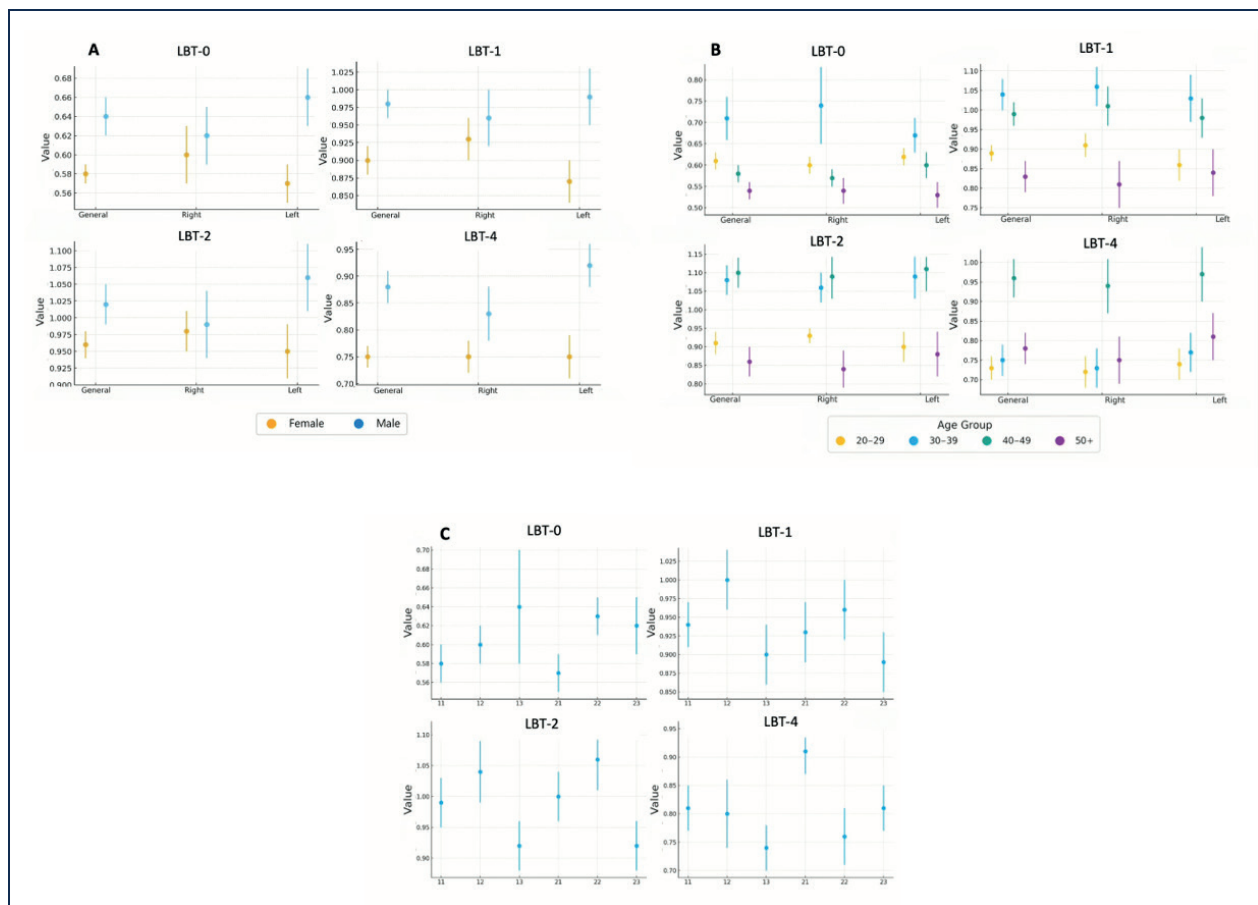


**Figure 3.** A. Distribution of LBT measurements by gender B. Association between LBT measurements and age groups C. Distribution of LBT measurements across teeth in the maxillary anterior region.

LBT: labial bone thickness

LBT values were  $<1$  mm at all levels in most cases. Significantly higher LBT values were observed in male patients at the LBT-1 and LBT-4 levels ( $p = 0.044$  and  $p = 0.002$ , respectively) (Figure 3A). Significant differences were detected among age groups at all LBT levels ( $p < 0.05$ ), with the highest values generally observed in the 30-39- and 40-49-year age groups. Correlation analyses revealed weak negative correlations between age and LBT-0 ( $r = -0.110$ ,  $p = 0.007$ ) and LBT-1 ( $r = -0.082$ ,  $p = 0.046$ ) (Figure 3B). Although no significant differences were observed among tooth types, LBT-0 exhibited the lowest values, whereas LBT-1 and LBT-2 exhibited the highest values (Figure 3C).

In the SRP distribution, Class III was markedly predominant (77.8%), whereas the remaining classes were observed at substantially lower frequencies. SRP distribution was independent of sex, age, and location (Table 2).

No significant sex-related difference was observed in RABH-t values ( $p > 0.05$ ), whereas RABH-b values, calculated based on bone-driven placement, were significantly higher in male patients ( $p < 0.001$ ) (Figure 4A). While RABH-t values remained comparable across age groups, RABH-b values differed significantly ( $p < 0.001$ ), with the highest values observed in the 40-49-

**Table 2.** Associations of SRP categories with gender, age group and location.

|       |       |                       | Gender |        | Total  |                        | Age Group |        |        |        | Total               |            | Location |        |        |
|-------|-------|-----------------------|--------|--------|--------|------------------------|-----------|--------|--------|--------|---------------------|------------|----------|--------|--------|
|       |       |                       | Female | Male   |        |                        | 20-29     | 30-39  | 40-49  | ≥50    |                     |            | Right    | Left   | Total  |
| SRP   | CI-2A | Count                 | 5      | 9      | 14     | Count                  | 4         | 2      | 5      | 3      | 14                  | Count      | 7        | 7      | 14     |
|       |       | % Gender              | 1.5%   | 3.6%   | 2.4%   | % Age Group            | 2.2%      | 1.6%   | 3.2%   | 2.3%   | 2.4%                | % Location | 2.3%     | 2.4%   | 2.4%   |
|       | CI-2B | Count                 | 34     | 39     | 73     | Count                  | 24        | 21     | 12     | 16     | 73                  | Count      | 43       | 30     | 73     |
|       |       | % Gender              | 10.0%  | 15.4%  | 12.3%  | % Age Group            | 13.1%     | 16.8%  | 7.6%   | 12.4%  | 12.3%               | % Location | 14.3%    | 10.2%  | 12.3%  |
|       | CI-3  | Count                 | 276    | 186    | 462    | Count                  | 149       | 94     | 125    | 94     | 462                 | Count      | 230      | 232    | 462    |
|       |       | % Gender              | 80.9%  | 73.5%  | 77.8%  | % Age Group            | 81.4%     | 75.2%  | 79.6%  | 72.9%  | 77.8%               | % Location | 76.7%    | 78.9%  | 77.8%  |
|       | CI-4  | Count                 | 6      | 4      | 10     | Count                  | 1         | 0      | 4      | 5      | 10                  | Count      | 4        | 6      | 10     |
|       |       | % Gender              | 1.8%   | 1.6%   | 1.7%   | % Age Group            | 0.5%      | 0.0%   | 2.5%   | 3.9%   | 1.7%                | % Location | 1.3%     | 2.0%   | 1.7%   |
|       | CI-5  | Count                 | 20     | 15     | 35     | Count                  | 5         | 8      | 11     | 11     | 35                  | Count      | 16       | 19     | 35     |
|       |       | % Gender              | 5.9%   | 5.9%   | 5.9%   | % Age Group            | 2.7%      | 6.4%   | 7.0%   | 8.5%   | 5.9%                | % Location | 5.3%     | 6.5%   | 5.9%   |
| Total |       | Count                 | 341    | 253    | 594    | Count                  | 183       | 125    | 157    | 129    | 594                 | Count      | 300      | 294    | 594    |
|       |       | % Gender              | 100.0% | 100.0% | 100.0% | % Age Group            | 100.0%    | 100.0% | 100.0% | 100.0% | 100.0%              | % Location | 100.0%   | 100.0% | 100.0% |
|       |       | χ² = 7.254; p = 0.123 |        |        |        | χ² = 19.497; p = 0.077 |           |        |        |        | χ² = 2.921; p=0.571 |            |          |        |        |

SRP: Sagittal root position, CI: Class; CI 2–5: SRP classes according to Gluckman et al.

year age group. A weak positive correlation between RABH-b and age was identified ( $r = 0.095$ ,  $p = 0.021$ ; Figure 4B). Among tooth types, RABH-t differed significantly ( $p < 0.001$ ), whereas RABH-b did not differ significantly ( $p = 0.364$ ) (Figure 4C).

LUD values did not differ by sex ( $p > 0.05$ ; Figure 5A). However, LUD increased significantly across age groups ( $p < 0.001$ ), with the highest values observed in the 40-49-year-old age group (Figure 5B). A significant positive correlation was found between age and LUD ( $r = 0.172$ ,  $p < 0.001$ ). Significant differences were also detected among tooth types ( $p < 0.001$ ), with the highest values observed in lateral incisors.

LCA values were significantly higher in male patients ( $p < 0.001$ ; Figure 5C). Significant differences were observed among age groups ( $p < 0.001$ ), with the highest values in the 30-39-year age group. No significant correlation was found between age and LCA ( $p > 0.05$ ; Figure 5D).

Mesiodistal measurements were higher in male patients ( $p < 0.05$ ) (Figure 6A) and differed significantly

among age groups ( $p = 0.001$ ) (Figure 6B). Significant differences were also observed among tooth types ( $p < 0.001$ ) (Figure 6C). Faciopalatal width was significantly greater in male patients across all analyses ( $p < 0.001$ ) (Figure 7A). A weak negative correlation was found between age and FP ( $r = -0.253$ ,  $p < 0.001$ ) (Figure 7B). FP measurements differed significantly among maxillary anterior teeth ( $p < 0.001$ ; Figure 7C).

In VIP-t, a significant sex-related difference was observed ( $p < 0.001$ ), with a higher proportion of Class A among female patients and of Class C among male patients. In VIP-b, most cases were classified as Class C (95.3%). Significant differences were observed among age groups for both VIP-t ( $p < 0.001$ ) and VIP-b ( $p < 0.05$ ), with more favorable placement patterns observed in older age groups. No location-related effect was observed ( $p > 0.05$ ).

The mean VIP-angle was approximately  $10^\circ$ – $11^\circ$  and was positively correlated with age ( $r = 0.183$ ,  $p < 0.001$ ) (Figure 8B). No significant differences were observed by sex or tooth type (Figure 8A and Figure 8C).

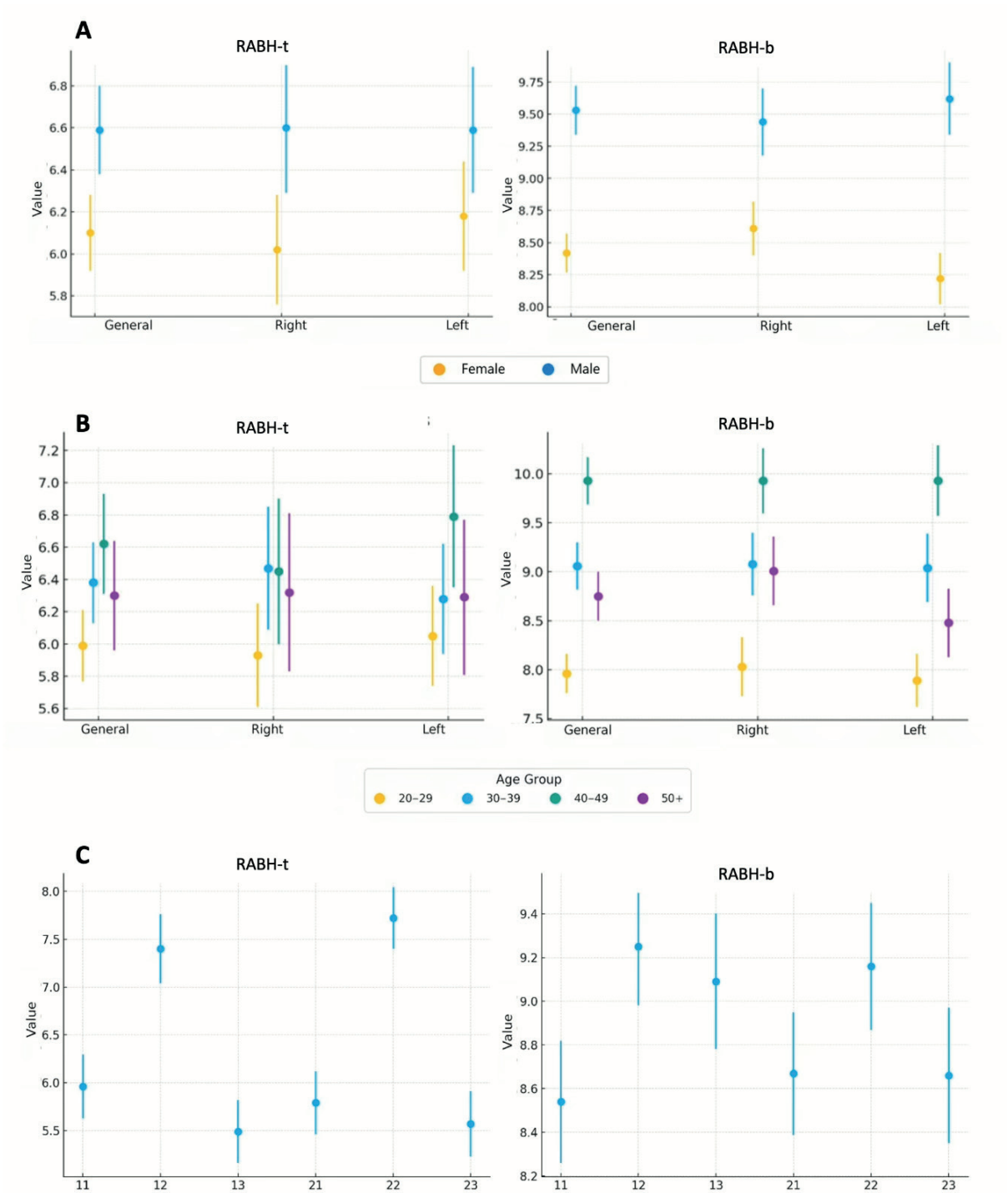


**Figure 4.** A. Distribution of RABH-t/RABH-b measurements by gender B. Association between RABH-t/RABH-b measurements and age groups C. Distribution of RABH-t/RABH-b measurements across teeth in the maxillary anterior region.

RABH-t: linear distance between the root apex and the relevant anatomical landmark, RABH-b: linear distance between the virtual implant apex and the same landmark.

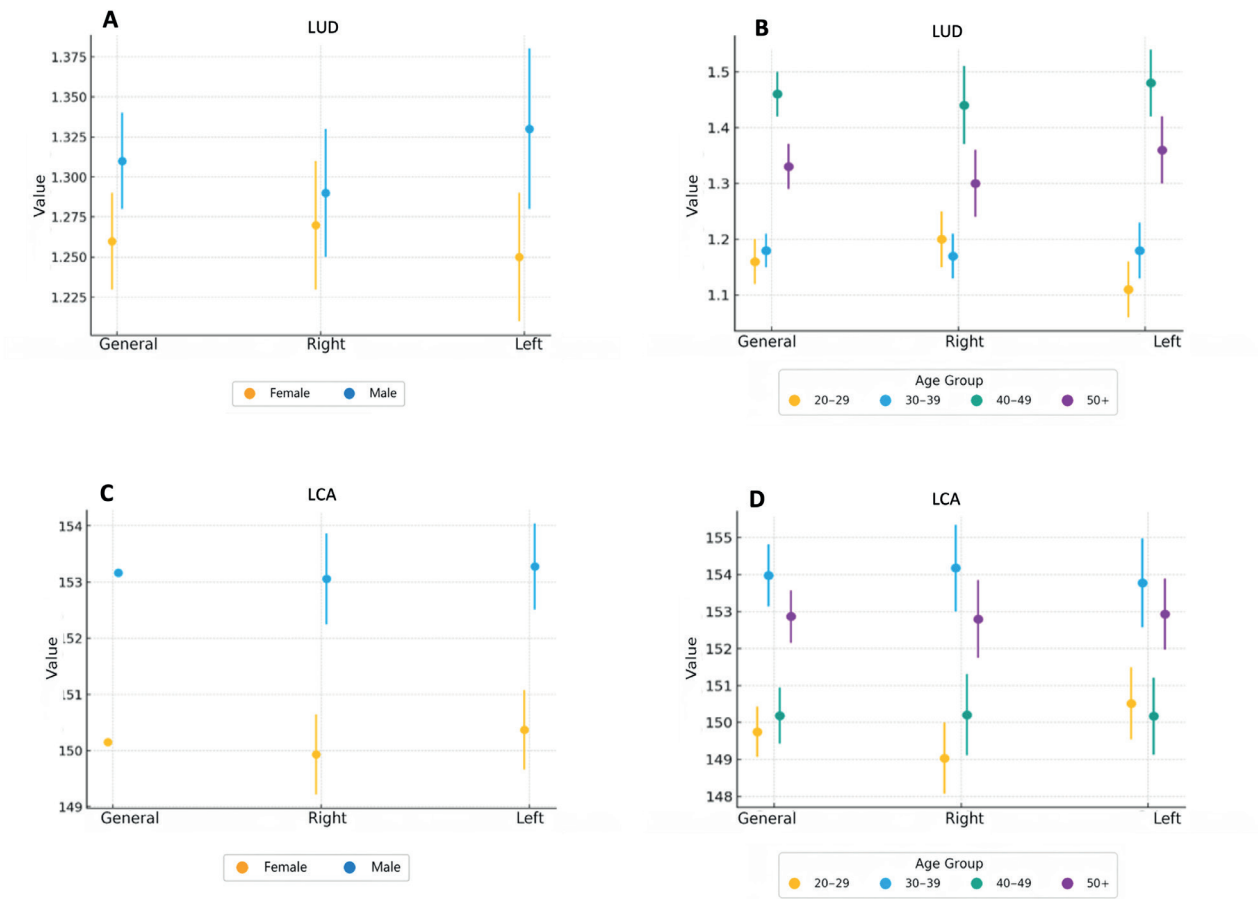


**Figure 5.** A. Distribution of LUD measurements by gender B. Association between LUD measurements and age groups. C. Distribution of LCA measurements by gender D. Association between LCA measurements and age groups. LUD: Labial undercut depth, LCA: Labial concavity angle.

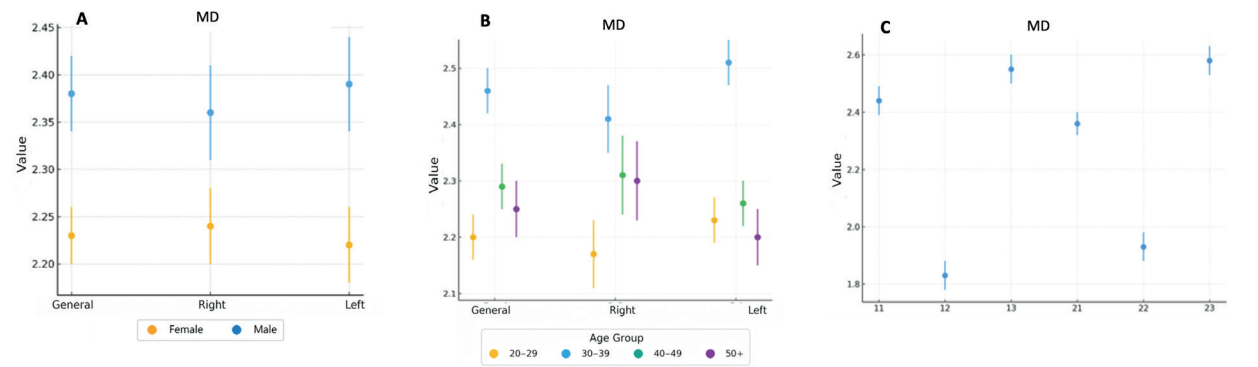


**Figure 6.** A. Distribution of MD measurements by gender B. Association between MD measurements and age groups C. Distribution of MD measurements across teeth in the maxillary anterior region MD: mesiodistal distance between adjacent root surfaces.

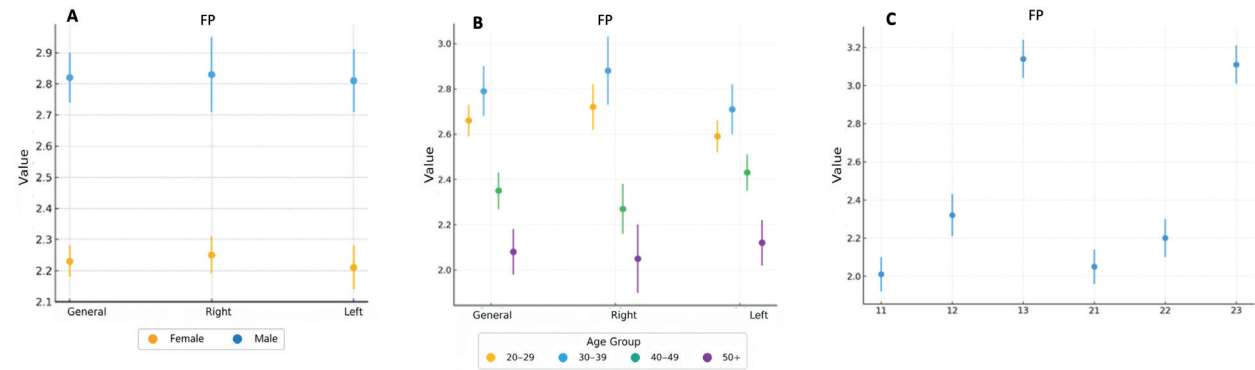


**Figure 7.** A. Distribution of FP measurements by gender B. Association between FP measurements and age groups C. Distribution of FP measurements across teeth in the maxillary anterior region  
FP: faciopalatal dimension on sagittal sections.

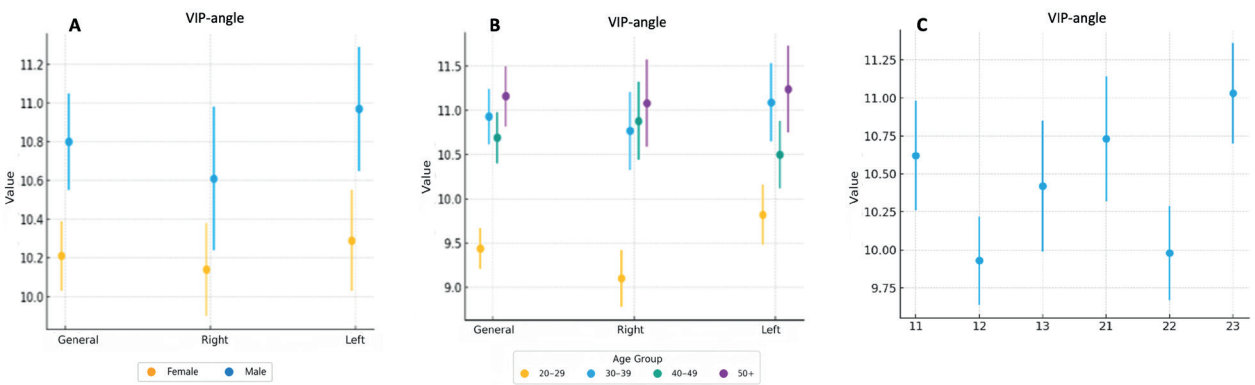


**Figure 8.** A. Distribution of VIP-angle measurements by gender B. Association between VIP-angle measurements and age groups C. Distribution of VIP-angle measurements across teeth in the maxillary anterior region.  
VIP-angle: angular difference between the two placement strategies.

## DISCUSSION

The maxillary anterior region is an anatomically demanding area where aesthetic and functional requirements must be precisely balanced, and the success of immediate implant placement in this region is largely governed by the three-dimensional morphological characteristics of the alveolar bone (3,6,27). The present findings demonstrate that implant placement in the anterior maxilla cannot be dictated solely by prosthetic requirements; rather, existing bone morphology acts as a decisive and often limiting factor.

A notable finding of this study is that labial bone thickness was predominantly  $<1$  mm. This observation is consistent with previous reports describing the inherently thin nature of the facial bone in the anterior maxilla (6,10,11,17). Such limited thickness predisposes the facial plate to early post-extraction resorption and consequently increases the risk of aesthetic complications following immediate implant placement (28,29). In this context, the frequently cited threshold of  $\geq 2$  mm facial bone thickness (30,31) appeared to be clinically unattainable in most cases within the sampled population. Accordingly, these findings support the frequent need for augmentation procedures in immediate implant therapy.

The predominance of Class III sagittal root position carries important clinical implications for implant planning. This distribution reflects a labially inclined root within the alveolar housing, thereby restricting prosthetically-driven implant placement in most cases. The consistency of similar findings across different populations (17,24,32) suggests that this is a reproducible anatomical pattern rather than a population-specific variation. A proclined root position increases the likelihood that the root will be close to the labial cortex and elevates the risk of cortical perforation; consequently, palatal implant placement often becomes necessary.

The present findings highlight the discrepancy between prosthetically-driven and bone-driven approaches in immediate implant planning. In most cases, LBT was <1 mm at all evaluated levels, indicating limited facial bone availability. Accordingly, implant placement along the prosthetic axis may increase the risk of facial bone perforation when it is incongruent with the existing bone morphology (20). Furthermore, the mean angular difference of approximately 10°–11° between prosthetically driven and bone-driven virtual implant positions demonstrates that the prosthetically determined implant position may not always be compatible with the available alveolar bone morphology. These findings emphasize the importance of considering both prosthetic requirements and individual bone morphology during immediate implant planning.

The observation that apical bone height was sufficient in most cases represents a favorable condition for immediate implant placement. Fulfillment of minimum apical bone requirements for primary stability, as emphasized in the literature (23,33,34), is a critical factor for successful osseointegration. This finding suggests that, in the anterior maxilla, anatomical limitations are primarily concentrated in the coronal and facial aspects, whereas the apical region generally provides adequate support.

Alveolar concavity and the presence of a labial undercut are key morphological determinants that directly influence implant angulation. A more pronounced concavity was observed in lateral teeth in the present study, suggesting increased technical complexity in this region and a potentially higher risk of perforation,

as also reported in the literature (19). These findings reinforce the importance of CBCT-based preoperative assessment and highlight the need for individualized surgical planning.

Mesiodistal distances exceeding the thresholds recommended in the literature ( $\geq 1.5$ –2 mm) (35–37) may be advantageous for papilla stability in the aesthetic zone. However, the predominance of faciopalatal width in favor of the palatal bone suggests that palatal implant positioning is not only necessary but also biomechanically advantageous (6,20,38). This supports the concept that implant placement strategies in the anterior maxilla should be guided by the distribution of available bone rather than prosthetic alignment alone.

A key strength of this study is the integrated assessment of multiple morphological parameters within the same population, together with the direct comparison of prosthetically-driven and bone-driven virtual implant positioning. CBCT-based three-dimensional planning may improve implant placement accuracy and help reduce potential bone-related risks by enabling detailed assessment of the available alveolar bone morphology (39). The present findings support this concept and indicate that bone-driven planning may provide greater morphological compatibility with the available alveolar bone in the anterior maxilla. Analysis of implant angulation revealed that implant positioning deviates from the prosthetic axis in most cases. This finding emphasizes the need to balance restorative objectives with anatomical constraints. The variability in angulation values reported across different populations (23) indicates that morphological variability plays a decisive role in clinical decision-making.

This study has certain limitations. The single-center dataset may limit the generalizability of the findings. Additionally, patient-related factors that may influence alveolar bone morphology were not evaluated. Furthermore, the absence of a direct correlation between morphological parameters and clinical outcomes necessitates cautious interpretation of their impact on treatment success. Prospective longitudinal studies are required to further elucidate the clinical relevance of these morphological factors.

Within the limitations of this study, immediate implant planning in the maxillary anterior region appears to be predominantly governed by alveolar bone morphology. The presence of a thin labial bone plate and the high prevalence of Class III sagittal root positions restrict prosthetically-driven implant placement in most cases and necessitate a bone-driven, palatally-oriented approach. These findings indicate that implant planning in the anterior region should be based on individualized morphological assessment. CBCT-based three-dimensional analysis, combined with a bone-driven placement strategy, should be considered a fundamental clinical approach for preserving facial bone integrity and achieving long-term aesthetic success.

### Ethical approval

The study was approved by the Non-Interventional Clinical Research Ethics Committee of Çanakkale Onsekiz Mart University (approval no: 2025-07/07-19; April 17, 2025) and conducted in accordance with the ethical principles of the Declaration of Helsinki.

### Author contribution

Concept: RA; Design: RA, BC; Data Collection or Processing: BC; Analysis or Interpretation: BC; Literature Search: BC; Writing: BC, RA; Critical Review: RA. All authors reviewed the results and approved the final version of the manuscript.

### Source of funding

The authors declare the study received no funding.

### Conflict of interest

The authors declare that there is no conflict of interest.

### Use of artificial intelligence

OpenAI ChatGPT (OpenAI, San Francisco, CA, USA) was used to improve the language and readability of the text. All AI-assisted outputs were critically reviewed, verified, and edited by the authors, who assume full responsibility for the presented work.

## REFERENCES

1. Buser D, Chappuis V, Belser UC, Chen S. Implant placement post extraction in esthetic single tooth sites: when immediate, when early, when late? *Periodontol* 2000. 2017; 73(1): 84-102. [\[Crossref\]](#)
2. Cosyn J, Eghbali A, De Bruyn H, Collys K, Cleymaet R, De Rouck T. Immediate single-tooth implants in the anterior maxilla: 3-year results of a case series on hard and soft tissue response and aesthetics. *J Clin Periodontol*. 2011; 38(8): 746-53. [\[Crossref\]](#)
3. Liu F, Wang YH, Xiong YJ, Ni YL. Modified oral implant restoration improves dental aesthetics, function, and patient satisfaction in maxillary anterior tooth loss. *Am J Transl Res*. 2025; 17(2): 983-91. [\[Crossref\]](#)
4. Bahar AD, Sagi MS, Mohd Zuhairi FA, Wan Hassan WN. Dental aesthetics and self-esteem of patients seeking orthodontic treatment. *Healthcare (Basel)*. 2024; 12(16): 1576. [\[Crossref\]](#)
5. Tonetti MS, Cortellini P, Graziani F, et al. Immediate versus delayed implant placement after anterior single tooth extraction: the timing randomized controlled clinical trial. *J Clin Periodontol*. 2017; 44(2): 215-24. [\[Crossref\]](#)
6. Todorovic VS, Postma TC, Hoffman J, van Zyl AW. Buccal and palatal alveolar bone dimensions in the anterior maxilla: a micro-CT study. *Clin Implant Dent Relat Res*. 2023; 25(2): 261-70. [\[Crossref\]](#)
7. Patel R, Ucer C, Wright S, Khan RS. Differences in dental implant survival between immediate vs. delayed placement: a systematic review and meta-analysis. *Dent J (Basel)*. 2023; 11(9): 218. [\[Crossref\]](#)
8. Donos N, Asche NV, Akbar AN, et al. Impact of timing of dental implant placement and loading: summary and consensus statements of group 1-The 6th EAO Consensus Conference 2021. *Clin Oral Implants Res*. 2021; 32(Suppl 21): 85-92. [\[Crossref\]](#)
9. Botermans A, Lidén A, de Carvalho Machado V, Chrcanovic BR. Immediate implant placement in the maxillary aesthetic zone: a cone beam computed tomography study. *J Clin Med*. 2021; 10(24): 5853. [\[Crossref\]](#)
10. Heimes D, Schiegnitz E, Kuchen R, Kämmerer PW, Al-Nawas B. Buccal bone thickness in anterior and posterior teeth-a systematic review. *Healthcare (Basel)*. 2021; 9(12): 1663. [\[Crossref\]](#)
11. Gluckman H, Pontes CC, Du Toit J. Radial plane tooth position and bone wall dimensions in the anterior maxilla: a CBCT classification for immediate implant placement. *J Prosthet Dent*. 2018; 120(1): 50-6. [\[Crossref\]](#)
12. Kim JH, Susin C, Min JH, et al. Extraction sockets: erratic healing impeding factors. *J Clin Periodontol*. 2014; 41(1): 80-5. [\[Crossref\]](#)
13. Couso-Queiruga E, Stuhr S, Tattan M, Chambrone L, Avila-Ortiz G. Post-extraction dimensional changes: a systematic review and meta-analysis. *J Clin Periodontol*. 2021; 48(1): 126-44. [\[Crossref\]](#)

14. Araújo MG, Silva CO, Misawa M, Sukekava F. Alveolar socket healing: what can we learn? *Periodontol* 2000. 2015; 68(1): 122-34. [\[Crossref\]](#)
15. Fokas G, Vaughn VM, Scarfe WC, Bornstein MM. Accuracy of linear measurements on CBCT images related to presurgical implant treatment planning: a systematic review. *Clin Oral Implants Res*. 2018; 29(Suppl 16): 393-415. [\[Crossref\]](#)
16. Rios HF, Borgnakke WS, Benavides E. The use of cone-beam computed tomography in management of patients requiring dental implants: an American Academy of Periodontology best evidence review. *J Periodontol*. 2017; 88(10): 946-59. [\[Crossref\]](#)
17. Kan JYK, Roe P, Rungcharassaeng K, et al. Classification of sagittal root position in relation to the anterior maxillary osseous housing for immediate implant placement: a cone beam computed tomography study. *Int J Oral Maxillofac Implants*. 2011; 26(4): 873-6.
18. El Nahass H, N Naiem S. Analysis of the dimensions of the labial bone wall in the anterior maxilla: a cone-beam computed tomography study. *Clin Oral Implants Res*. 2015; 26(4): e57-61. [\[Crossref\]](#)
19. Chung MP, Wang IC, Chan HL, Wang HL. Evaluation of buccal bone concavity in the esthetic zone: a cadaver study. *Implant Dent*. 2017; 26(5): 751-5. [\[Crossref\]](#)
20. Lim HC, Kang DU, Baek H, et al. Cone-beam computed tomographic analysis of the alveolar ridge profile and virtual implant placement for the anterior maxilla. *J Periodontal Implant Sci*. 2019; 49(5): 299-309. [\[Crossref\]](#)
21. Shi M, Wang X, Zeng P, et al. Analysis of the sagittal root angle and its correlation with hard and soft tissue indices in anterior teeth for immediate implant evaluation: a retrospective study. *BMC Oral Health*. 2021; 21(1): 494. [\[Crossref\]](#)
22. Parra-Moreno FJ, Schiavo-Di Flaviano V, Egido-Moreno S, Saka-Herranz C, Estrugo-Devesa A, López-López J. Pre-surgical socket analysis for immediate implant placement. *Med Oral Patol Oral Cir Bucal*. 2024; 29(3): e305-17. [\[Crossref\]](#)
23. Zhang W, Skrypczak A, Weltman R. Anterior maxilla alveolar ridge dimension and morphology measurement by cone beam computerized tomography (CBCT) for immediate implant treatment planning. *BMC Oral Health*. 2015; 15: 65. [\[Crossref\]](#)
24. Ikbal M, Shen YW, Ruslin M, Fuh LJ, Hsu JT. Evaluation of sagittal root position and labial alveolar bone concavity in the maxillary anterior tooth area for immediate implant placement. *J Chin Med Assoc*. 2023; 86(6): 565-70. [\[Crossref\]](#)
25. Wilson JP, Johnson TM. Frequency of adequate mesiodistal space and faciolingual alveolar width for implant placement at anterior tooth positions. *J Am Dent Assoc*. 2019; 150(9): 779-87. [\[Crossref\]](#)
26. Monje A, Chan HL, Fu JH, Suarez F, Galindo-Moreno P, Wang HL. Are short dental implants (<10 mm) effective? a meta-analysis on prospective clinical trials. *J Periodontol*. 2013; 84(7): 895-904. [\[Crossref\]](#)
27. AlAli F, Atieh MA, Hannawi H, et al. Anterior maxillary labial bone thickness on cone beam computed tomography. *Int Dent J*. 2023; 73(2): 219-27. [\[Crossref\]](#)
28. Nowzari H, Molayem S, Chiu CHK, Rich SK. Cone beam computed tomographic measurement of maxillary central incisors to determine prevalence of facial alveolar bone width  $\geq 2$  mm. *Clin Implant Dent Relat Res*. 2012; 14(4): 595-602. [\[Crossref\]](#)
29. Tan WL, Wong TLT, Wong MCM, Lang NP. A systematic review of post-extraction alveolar hard and soft tissue dimensional changes in humans. *Clin Oral Implants Res*. 2012; 23(Suppl 5): 1-21. [\[Crossref\]](#)
30. Monje A, Rocuzzo A, Buser D, Wang HL. Influence of buccal bone wall thickness on the peri-implant hard and soft tissue dimensional changes: a systematic review. *Clin Oral Implants Res*. 2023; 34(3): 157-76. [\[Crossref\]](#)
31. Spray JR, Black CG, Morris HF, Ochi S. The influence of bone thickness on facial marginal bone response: stage 1 placement through stage 2 uncovering. *Ann Periodontol*. 2000; 5(1): 119-28. [\[Crossref\]](#)
32. Xu D, Wang Z, Sun L, et al. Classification of the root position of the maxillary central incisors and its clinical significance in immediate implant placement. *Implant Dent*. 2016; 25(4): 520-4. [\[Crossref\]](#)
33. Lebon A, Cortasse B, Emonet-Denand V, Campan P, Galibourg A. Digital evaluation of the apical 3 mm portion of various implant designs on the surface area in contact with bone. *Digital Dentistry Journal*. 2025; 2(2): 100021. [\[Crossref\]](#)
34. Fenner M, Vairaktaris E, Fischer K, Schlegel KA, Neukam FW, Nkenke E. Influence of residual alveolar bone height on osseointegration of implants in the maxilla: a pilot study. *Clin Oral Implants Res*. 2009; 20(6): 555-9. [\[Crossref\]](#)
35. Chen ST, Buser D, Sculean A, Belser UC. Complications and treatment errors in implant positioning in the aesthetic zone: diagnosis and possible solutions. *Periodontol* 2000. 2023; 92(1): 220-34. [\[Crossref\]](#)
36. Han HS, Kim PJ, Kim KT, Kim S, Ku Y, Cho YD. Dental implant proximity to adjacent teeth: a retrospective study. *Clin Implant Dent Relat Res*. 2022; 24(6): 733-9. [\[Crossref\]](#)
37. Vela X, Méndez V, Rodríguez X, Segalá M, Tarnow DP. Crestal bone changes on platform-switched implants and adjacent teeth when the tooth-implant distance is less than 1.5 mm. *Int J Periodontics Restorative Dent*. 2012; 32(2): 149-55.
38. López-Jarana P, Díaz-Castro CM, Falcão A, Falcão C, Ríos-Santos JV, Herrero-Climent M. Thickness of the buccal bone wall and root angulation in the maxilla and mandible: an approach to cone beam computed tomography. *BMC Oral Health*. 2018; 18(1): 194. [\[Crossref\]](#)
39. Lascala CA, Panella J, Marques MM. Analysis of the accuracy of linear measurements obtained by cone beam computed tomography (CBCT-NewTom). *Dentomaxillofac Radiol*. 2004; 33(5): 291-4. [\[Crossref\]](#)

# Readability assessment of ChatGPT-4 and ChatGPT-5 responses to ear care questions

Cemal Özyılmaz<sup>1</sup>, Mehmet Fatih Karakuş<sup>1</sup>

<sup>1</sup>Department of Otorhinolaryngology, Faculty of Medicine, Ordu University, Ordu, Türkiye

**Cite as:** Özyılmaz C, Karakuş MF. Readability assessment of ChatGPT-4 and ChatGPT-5 responses to ear care questions. Northwestern Med J. 2026;6(4):308-316.

## ABSTRACT

**Aim:** To assess the readability of ChatGPT-4o and ChatGPT-5 responses to questions about ear care.

**Materials and Methods:** Ten questions about ear care from the World Health Organization's 'Questions and Answers' section were each posed seven times to both ChatGPT-4o and ChatGPT-5. A total of 70 responses were obtained and analyzed for readability.

**Results:** For ChatGPT-4o, the mean readability scores were: Flesch Reading Ease  $62.84 \pm 8.32$ , Flesch–Kincaid Grade Level  $8.13 \pm 1.59$ , Simple Measure of Gobbledygook  $7.35 \pm 1.42$ , Gunning Fog Index  $8.44 \pm 1.79$ , Coleman–Liau Index  $9.62 \pm 1.71$ , and Automated Readability Index  $8.70 \pm 2.02$ . For ChatGPT-5, the corresponding scores were: Flesch Reading Ease  $72.93 \pm 7.69$ , Flesch–Kincaid Grade Level  $5.75 \pm 1.17$ , Simple Measure of Gobbledygook  $6.35 \pm 0.79$ , Gunning Fog Index  $7.37 \pm 1.22$ , Coleman–Liau Index  $7.32 \pm 1.52$ , and Automated Readability Index  $5.78 \pm 1.42$ . ChatGPT-5 demonstrated better readability than ChatGPT-4o across all readability indices ( $p < 0.001$ ).

**Conclusion:** The readability of the answers produced by ChatGPT-4o is "standard", and that of ChatGPT-5 is "fairly easy". ChatGPT-5 responses showed improved readability over ChatGPT-4o, aligning more closely with patient education readability standards and underscoring its potential as an accessible source of health information.

**Keywords:** artificial intelligence, ChatGPT health information, ear care, readability

**Corresponding author:** Mehmet Fatih Karakuş **E-mail:** mfkarakus@yahoo.com

**Received:** 09.05.2026 **Accepted:** 24.08.2026 **Published:** 06.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

## INTRODUCTION

Ear care is an important part of personal hygiene and plays a fundamental role in maintaining ear function. Behaviours such as inserting foreign objects into the ear, prolonged exposure to loud noise, and swimming in unclean water can adversely affect ear health. If proper ear care is not carried out, various infections and related complications may arise. However, public knowledge and awareness of ear care are insufficient. Impacted cerumen and the associated hearing loss that develop as a result of incorrect ear care practices have become a health issue that requires attention (1-3).

Cerumen, formed by secretions from the ceruminous and sebaceous glands in the external auditory canal, plays an important role in maintaining ear health. A certain amount of cerumen moistens the ear canal, forms a physical barrier against foreign objects, and has antimicrobial properties. However, attempts by individuals to clean their ears with cotton buds or unsuitable tools can disrupt this natural mechanism, leading to trauma to the external auditory canal, damage to the tympanic membrane, cerumen impaction, and infection. Societal awareness of the potential harms associated with the use of cotton buds is generally low. Studies show that individuals have insufficient knowledge about ear cleaning habits and the complications associated with these habits (3,4).

Currently, individuals access health information from various sources, including healthcare professionals, printed materials, online search engines, social media, and health portals. In recent years, with the proliferation of the internet, online search engines, social media, and health portals have become among the most frequently used sources of information for the public (5). However, rapid advances in artificial intelligence have ushered in a new era in the ways individuals access healthcare information. Chat Generative Pre-trained Transformer (ChatGPT) is an artificial intelligence (AI) system developed by OpenAI and based on large language models. It can understand and generate coherent and accurate responses to text-based queries in natural language. By learning the structures of language, ChatGPT can provide information on a wide range of topics and engage in human-like conversations. It is trained on a vast corpus

of text data using deep learning techniques to acquire general knowledge. The ability of AI-powered language models such as ChatGPT to provide users with quick, comprehensive and accessible answers on health matters increases the potential for these platforms to become one of the primary sources for seeking health-related information (6,7).

ChatGPT has been shown to perform well in medical examinations and provide accurate information across various medical specialities (8-10). The literature reports that ChatGPT achieves high success rates in answering questions in the ear, nose, and throat (ENT) field. Its success in the audiology competency exam suggests it could be a potential tool in hearing care services (11); it has been reported to provide highly accurate answers to ENT questions from the public (12) and that its diagnostic and treatment approaches in easy to moderately difficult ENT scenarios largely align with those of physicians (13).

Artificial intelligence-based platforms can not only facilitate access to information but also influence individuals' health-related decision-making processes. Making informed health decisions depends not only on the accuracy of the information provided but also on the target audience's ability to understand this information (14). At this point, the concept of 'readability' comes to the fore. Readability is defined as the characteristics that facilitate or hinder comprehension of reading material. Readability is generally expressed as the minimum number of years of formal education required to read and understand a text (15). Today, individuals of all educational backgrounds can make health decisions based on information obtained from AI-based platforms such as ChatGPT. Therefore, in addition to the accuracy of ChatGPT's responses, their readability is also critical parameter. This study aimed to evaluate the readability of responses generated by ChatGPT-4o and ChatGPT-5 to questions about ear care, to compare readability between the two models, and to assess their readability relative to the World Health Organization's (WHO) materials on the same topic. The findings of this study may provide insights into whether newer AI models offer more patient-friendly health information, thereby supporting the informed use of large language models in healthcare settings.

## MATERIALS AND METHODS

This study does not involve any research conducted with human or animal subjects, and ChatGPT, used in this study, is a publicly available application. Therefore, ethical approval is not required.

Ten questions related to ear care from the Question and Answer (Q&A) section on the World Health Organisation's (WHO) official website were selected and asked to ChatGPT (16).

1. How should I clean my (or my child's) ears?
2. What is ear wax? Should it be removed?
3. When does wax need to be removed from the ear?
4. Can I use Q-tips or cotton buds to remove ear wax?
5. What causes ear pain?
6. What should I do for ear pain?
7. Can I take any antibiotic for ear pain and ear discharge?
8. What causes fluid or pus from ears? Can it be dangerous?
9. What should I do if my ears feel itchy or if an insect enters my ear?
10. What should I do when water goes into my ear?

These questions were posed in separate new chat sessions using ChatGPT's GPT-4o model on 12 June 2025 and the GPT-5 model on 15 August 2025. Default settings were used, and no custom instructions, web search functions, external tools, or additional prompts were employed. After all ten questions had been posed, the session was terminated and a new chat session was initiated. The complete set of questions was repeated seven times in separate chat sessions for each model.

The readability of ChatGPT answers was assessed using six validated, commonly used readability assessment tools: Flesch Reading Ease (FRE) score, Flesch–Kincaid grade level (FKGL), Gunning Fog Index (GFI), Coleman–Liau Index (CLI), Simple Measure of Gobbledygook (SMOG) grade, and Automated Readability Index (ARI). In 1948, Rudolph Flesch developed the Flesch

Reading Ease (FRE) formula. Later, in 1975, Kincaid and colleagues built upon the FRE to create the Flesch–Kincaid Grade Level (FKGL) formula for the U.S. Navy, providing a grade-level indicator for written material (17). FRE is calculated from the total number of words, syllables, and sentences in a written passage using the following formula:  $206.835 - 1.015 (\text{total words/total sentences}) - 84.6 (\text{total syllables/total words})$ . The score obtained from FRE ranges from 0 to 100, and the higher the score, the easier the text is to read (18). FKGL, on the other hand, provides a score corresponding to the US grade level (17). The Gunning Fog Index is a formula used to estimate the number of years of formal education required to understand a text on first reading. It is based on the average number of words per sentence and the percentage of complex words in the text. Higher scores indicate greater reading difficulty (19). The SMOG grade is calculated by comparing the number of polysyllabic words (three or more syllables) to the number of sentences. This provides an estimate of the number of years of education required for the average person to understand a given piece of writing (20). The Coleman-Liau Index is a measure that takes into account the average number of characters per word and the average number of sentences per 100 words in a given text sample. This index determines the grade level at which a text is written using a specific formula (21). The Automated Readability Index (ARI) estimates the grade level needed to understand a text on first reading, using the average number of characters per word and words per sentence (19).

Readability scores were calculated by copying ChatGPT's responses to each question into a free online readability calculator (22). Statistical analyses were conducted using the Statistical Package for the Social Sciences (SPSS) version 15.0. The normality of the data was assessed using the Kolmogorov–Smirnov and Shapiro–Wilk tests. Descriptive statistics are presented as mean  $\pm$  standard deviation and median (minimum–maximum). The independent samples t-test was used for comparisons of parametric data, while the Mann–Whitney U test was applied for non-parametric data. Effect sizes were calculated as  $r (Z/\sqrt{N})$  for the Mann–Whitney U test and as Cohen's  $d$  for independent-samples t-tests. A  $p\text{-value} < 0.05$  was considered statistically significant.

## RESULTS

The mean FRE score was  $72.93 \pm 7.69$  for ChatGPT-5, corresponding to the 'fairly easy' category, while it was  $62.84 \pm 8.32$  for ChatGPT-4, which corresponds to the 'standard' level ( $p < 0.001$ ). The mean FKGL score was  $5.75 \pm 1.17$  for ChatGPT-5, which is equivalent to an elementary reading level, whereas the mean score for ChatGPT-4o was  $8.13 \pm 1.59$ , corresponding to a middle school level. Across indices, the reading levels of ChatGPT-5 responses ranged from approximately 5th to 7th grade (FKGL:  $5.75 \pm 1.17$ , SMOG:  $6.35 \pm 0.79$ , GFI:  $7.37 \pm 1.22$ , CLI:  $7.32 \pm 1.52$ , ARI:  $5.78 \pm 1.42$ ), while those of ChatGPT-4 responses ranged from approximately 7th to 9th grade (FKGL:  $8.13 \pm 1.59$ , SMOG:  $7.35 \pm 1.42$ , GFI:  $8.44 \pm 1.79$ , CLI:  $9.62 \pm 1.71$ , ARI:  $8.70 \pm 2.02$ ). ChatGPT-5 showed better readability than ChatGPT-4 across all readability indices ( $p < 0.001$ ). The effect sizes ranged from 0.31 to 1.68, indicating

moderate to large differences between ChatGPT-4o and ChatGPT-5. Readability scores for ChatGPT-4o and ChatGPT-5 responses are shown in Table 1.

The mean FRE score for the questions answered by the World Health Organization was  $76.9 \pm 8.44$ , with the other readability indices indicating levels ranging from 5th to 7th grade (FKGL  $6.36 \pm 1.58$ , SMOG  $5.6 \pm 2.39$ , GFI  $7.02 \pm 1.67$ , CLI  $6.9 \pm 1.71$ , ARI  $6.56 \pm 1.64$ ). There was no statistically significant in readability between ChatGPT-5 and WHO responses; however, ChatGPT-4o responses differed significantly from WHO responses on all readability indices and were more difficult to read. A comparison of readability indices between ChatGPT models and WHO responses is shown in Table 2.

Table 3 presents ChatGPT responses with the highest and lowest FRE scores to the questions.

**Table 1.** Readability scores of ChatGPT-4o and ChatGPT-5 responses.

| Readability Indices | ChatGPT-4o       |                   | ChatGPT-5        |                   | p       | Effect sizes |
|---------------------|------------------|-------------------|------------------|-------------------|---------|--------------|
|                     | Mean $\pm$ SD    | Median(min-max)   | Mean $\pm$ SD    | Median(min-max)   |         |              |
| FRE                 | 62.84 $\pm$ 8.32 | 65.0 (40.0-78.0)  | 72.93 $\pm$ 7.69 | 74.0 (52.0-86.0)  | <0.001  | 0.55         |
| FKGL                | 8.13 $\pm$ 1.59  | 7.78 (5.39-11.91) | 5.75 $\pm$ 1.17  | 5.54 (3.57-8.97)  | <0.001  | 0.67         |
| ARI                 | 8.7 $\pm$ 2.02   | 8.4 (5.12-13.01)  | 5.78 $\pm$ 1.42  | 5.58 (2.86-9.61)  | <0.001* | 1.68         |
| GFI                 | 8.44 $\pm$ 1.79  | 8.0 (6.0-13.1)    | 7.37 $\pm$ 1.22  | 7.15 (5.6-10.6)   | <0.001  | 0.31         |
| CLI                 | 9.62 $\pm$ 1.71  | 9.13 (7.0-13.59)  | 7.32 $\pm$ 1.52  | 7.12 (4.61-11.65) | <0.001  | 0.38         |
| SMOG                | 7.35 $\pm$ 1.42  | 6.95 (5.02-10.85) | 6.35 $\pm$ 0.79  | 6.24 (5.05-8.67)  | <0.001  | 0.61         |

\*Independent-samples t-test. All other comparisons were performed using the Mann-Whitney U test.

**Table 2.** Comparison of readability indices between ChatGPT models and WHO responses.

| Readability Indices | ChatGPT-5 vs. WHO responses |              | ChatGPT-4o vs. WHO responses |              |
|---------------------|-----------------------------|--------------|------------------------------|--------------|
|                     | p                           | Effect sizes | p                            | Effect sizes |
| FRE                 | 0.104                       | 0.18         | <0.001                       | 0.44         |
| FKGL                | 0.150*                      | 0.49         | 0.002                        | 0.35         |
| ARI                 | 0.115*                      | 0.54         | 0.002*                       | 1.08         |
| GFI                 | 0.288                       | 0.12         | 0.013                        | 0.28         |
| CLI                 | 0.428*                      | 0.27         | <0.001                       | 0.41         |
| SMOG                | 0.171                       | 0.15         | 0.011                        | 0.28         |

\*Independent-samples t-test. All other comparisons were performed using the Mann-Whitney U test.

**Table 3.** Examples of questions with the lowest and highest scores according to FRE.

| Model      | The question with the easiest readability based on FRE score                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | The question with the most difficult readability based on FRE score                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ChatGPT-4o | <p><b>Can I use Q-tips or cotton buds to remove ear wax?</b><br/>(FRE:78.0)</p> <p><b>Answer:</b> No, it's best not to use Q-tips or cotton buds inside your ear canal. Here's why:</p> <ul style="list-style-type: none"> <li>• They can push the wax deeper into the ear, causing blockages or even damage.</li> <li>• They might injure the sensitive skin inside the ear canal or the eardrum.</li> <li>• Using them too often can cause irritation or infections.</li> </ul> <p>If you want to clean your ears, just gently wipe the outer ear with a soft, damp cloth. For wax buildup deeper inside, see a healthcare professional for safe removal.</p> <p>Would you like tips on safer ways to manage ear wax at home?</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | <p><b>Can I take any antibiotic for ear pain and ear discharge?</b><br/>(FRE:40.0)</p> <p><b>Answer:</b> No, you should not take any antibiotic for ear pain or ear discharge without a doctor's advice. Here's why:</p> <ul style="list-style-type: none"> <li>• Ear pain and discharge can have different causes, and not all of them need antibiotics.</li> <li>• Many ear infections are viral, which don't respond to antibiotics.</li> <li>• Using antibiotics unnecessarily can cause side effects and contribute to antibiotic resistance.</li> <li>• A healthcare provider needs to examine your ear to determine if you have a bacterial infection that requires antibiotics, and which antibiotic is appropriate.</li> </ul> <p>If you have ear pain with discharge, fever, or worsening symptoms, see a doctor promptly for proper diagnosis and treatment.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| ChatGPT-5  | <p><b>What should I do when water goes into my ear?</b><br/>(FRE:86.0)</p> <p><b>Answer:</b> Getting water trapped in the ear is usually harmless but can sometimes cause discomfort or infection. Here's what to do:</p> <ol style="list-style-type: none"> <li>1. Remove the water safely <ul style="list-style-type: none"> <li>• Tilt your head to the side and gently pull on the earlobe to help the water drain.</li> <li>• Jump or shake your head gently while tilted to encourage drainage.</li> <li>• Lie on your side with the affected ear facing down for a few minutes.</li> </ul> </li> <li>2. Dry the ear <ul style="list-style-type: none"> <li>• Use a soft towel to gently dry the outer ear.</li> <li>• Do not insert cotton swabs or fingers into the ear canal.</li> </ul> </li> <li>3. Optional techniques <ul style="list-style-type: none"> <li>• Gravity + jaw movement: Chewing or yawning can help water escape.</li> <li>• Alcohol-vinegar drops (for adults, not children under 1 year): A few drops can help dry the ear and prevent infection—but only if the eardrum is intact.</li> </ul> </li> <li>4. Watch for signs of infection ("swimmer's ear") <ul style="list-style-type: none"> <li>• Pain, redness, swelling</li> <li>• Discharge or foul odor</li> <li>• Itching or fever</li> </ul> </li> </ol> <p>If any of these appear, see a doctor promptly.</p> <p>I can also make a quick home routine to prevent water-related ear problems, especially for kids who swim or bathe often. Do you want me to do that?</p> | <p><b>What causes fluid or pus from ears? Can it be dangerous?</b><br/>(FRE:52.0)</p> <p><b>Answer:</b> Fluid or pus from the ear (otorrhea) can happen for several reasons — some mild, some potentially serious.</p> <ol style="list-style-type: none"> <li>1. Common causes <ul style="list-style-type: none"> <li>• Outer ear infection (otitis externa / swimmer's ear) – infection of the ear canal, often after swimming or scratching.</li> <li>• Middle ear infection (otitis media) with eardrum perforation – pus drains out when pressure builds up behind the eardrum and it ruptures.</li> <li>• Chronic suppurative otitis media – long-term ear infection with ongoing discharge.</li> <li>• Fungal infection – discharge may look white, black, or fluffy.</li> <li>• Ear injury – from foreign objects, trauma, or sudden pressure changes.</li> </ul> </li> <li>2. Less common but serious causes <ul style="list-style-type: none"> <li>• Cholesteatoma – abnormal skin growth in the middle ear that can erode bone.</li> <li>• Mastoiditis – infection of the bone behind the ear (often after untreated ear infection).</li> <li>• Cerebrospinal fluid (CSF) leak – rare, usually after head injury or surgery; fluid may be clear and watery.</li> </ul> </li> <li>3. When it can be dangerous <p>Seek urgent medical care if discharge is:</p> <ul style="list-style-type: none"> <li>• Accompanied by fever, severe pain, swelling, or redness around the ear</li> <li>• Mixed with blood</li> <li>• Clear and watery after head trauma (possible CSF leak)</li> <li>• Persistent for more than a few days despite basic care</li> <li>• Associated with hearing loss, dizziness, or facial weakness</li> </ul> </li> <li>4. Key points <ul style="list-style-type: none"> <li>• Never put anything in the ear unless prescribed by a doctor.</li> <li>• Keep the ear dry until evaluated.</li> <li>• The correct treatment (antibiotics, antifungals, surgery) depends on the exact cause — guessing can delay proper care.</li> </ul> </li> </ol> <p>If you want, I can make you a visual chart showing ear discharge types, colors, and their likely causes — it's very useful for quick recognition, especially for parents.</p> |

## DISCUSSION

The internet has become a major platform for individuals seeking health-related information, with more than half of the US population relying on it as a primary source (23,24). Recent advances in artificial intelligence have further expanded these opportunities, providing patients with access to a new and rapidly growing source of medical information (25). Among these tools, ChatGPT—an AI-based language model that has gained substantial attention—has the potential to become a widely used resource for health information. Therefore, ensuring the reliability and readability of its responses is of critical importance. To date, numerous studies have evaluated the reliability of ChatGPT responses. It has generally been reported to provide accurate and trustworthy information on topics such as asthma, spinal cord injuries, thyroid cancer, and hypothyroidism in pregnancy (26-29). In addition, ChatGPT has demonstrated success in medical exams (8,30). While these findings highlight its potential as a reliable source of medical knowledge, the readability of its responses is equally important, as information that is difficult to comprehend may undermine its practical value and contribute to negative health outcomes. Therefore, our study focused on assessing the readability of ChatGPT's responses regarding ear care.

Readability is a key component of effective health communication. When health information is presented at a level that is too complex, patients' ability to understand, retain, and apply the information may be limited, ultimately reducing the effectiveness of health education efforts. Therefore, assessing the readability of responses provided by large language models such as ChatGPT is essential to ensure that AI-supported health information remains accessible and understandable to a wide audience. Patients who are unable to fully comprehend medical information may be disadvantaged as active participants in their own healthcare decisions, which can jeopardize their access to the highest quality of care (24). The American Medical Association recommends that patient education materials be written at a sixth-grade reading level, while the National Institutes of Health recommends an eighth-grade level (31).

Among readability measures, the Flesch Reading Ease (FRE) is the most commonly used to evaluate the readability of written health information materials (17). Based on the FRE scores, the most understandable response was at the 'fairly easy' level for ChatGPT-4o and at the 'easy' level for ChatGPT-5 (78.0 and 86.0, respectively). The mean FRE score in our study was  $62.84 \pm 8.32$  for ChatGPT-4o and  $72.93 \pm 7.69$  for ChatGPT-5; indicating that responses generated by ChatGPT-4o are "standard" and ChatGPT-5 are 'fairly easy'. Similarly, the mean FKGL score was  $8.13 \pm 1.59$  for ChatGPT-4o (corresponding to an eighth-grade level) and  $5.75 \pm 1.17$  for ChatGPT-5 (corresponding to an elementary level). Studies suggest that patient comprehension may be compromised if content exceeds a 7th-grade reading level, which is the average American reading level identified by the United States Department of Health and Human Services (32). Taken together, our findings suggest that ChatGPT-5 provides health-related information in a more accessible and comprehensible manner compared with ChatGPT-4o, which may, in turn, facilitate better understanding and application of medical information by readers.

Our findings indicate that, when comparing ChatGPT with the WHO, ChatGPT-5 responses were closer to the readability levels of WHO materials than responses from ChatGPT-4o. WHO and ChatGPT-5 responses corresponded to a fifth- to seventh-grade reading level, which aligns more closely with recommended standards. In contrast, ChatGPT-4o responses were consistently higher, ranging from 7th- to 9th-grade levels, indicating a more complex readability profile. This comparison highlights that responses from ChatGPT-5 closely match the readability of WHO materials, indicating a notable improvement over ChatGPT-4o in producing health information accessible to the general public. However, these findings should be interpreted with caution, as WHO materials and ChatGPT-generated responses differ fundamentally in their development processes and intended purposes. WHO patient education materials are developed through expert review and editorial oversight, whereas ChatGPT responses are dynamically generated and may vary across interactions. Therefore, the present comparison reflects similarities in readability rather than equivalence in content quality, accuracy, reliability, or suitability for patient education.

In our study, the SMOG index was 7.35 for ChatGPT-4o and 6.35 for ChatGPT-5. In contrast, a previous study on the surgical treatment of retinal diseases reported a SMOG index above 13 for ChatGPT-4o responses, corresponding to college-level readability. This may be explained by the more complex medical content of that study and the fact that the questions were prepared for clinicians (33). Similarly, a study evaluating the readability of ChatGPT-3.5-turbo model responses on lung cancer reported considerably higher scores (GFI  $13.63 \pm 1.54$ , CLI  $14.24 \pm 1.90$ , SMOG  $14.61 \pm 1.45$ , ARI  $15.04 \pm 1.97$ ) compared to those observed in our study (19). The technical nature of the question set used in that study may be associated with these high scores. The readability of ChatGPT-4's responses regarding hypothyroidism during pregnancy was also found to be difficult, often requiring a college-level education to comprehend the text; however, the questions in that study targeted both patients and clinicians, which may account for the higher readability levels (29). Evidence on ChatGPT-5 readability is still limited, but one recent study on polytrauma reported higher values (SMOG:  $7.98 \pm 1.12$ , CLI:  $14.22 \pm 3.24$ , ARI:  $11.23 \pm 1.86$ , FRE:  $42.26 \pm 6.08$ , FKGL:  $11.24 \pm 1.46$ ) compared to ours (34). Overall, higher readability scores reported in studies of various diseases, compared with our findings, may be explained by our study's focus on ear care-related questions, which contain less complex medical content and are designed for the general public. This suggests that readability levels may depend not only on the AI model used but also on the complexity of medical topics and the intended audience for the questions.

This study has some limitations. First, it compared only ChatGPT-4o and ChatGPT-5, excluding other large language models or previous model versions. Second, the analysis was limited to English-language responses; therefore, the results may not be generalizable to responses in other languages. Although each response was generated in a separate chat session, repeated outputs from the same prompt may not be completely independent. Therefore, the effective sample size may be smaller than the nominal number of responses analysed; the findings should be interpreted accordingly. Future studies may benefit from applying repeated-measures or mixed-effects modelling approaches to account for potential within-question correlations. Another limitation is that responses

from GPT-4o and GPT-5 were collected on different dates. Changes in the underlying model infrastructure, system-level settings, or platform updates occurring between the two data collection periods may have influenced the generated responses. Therefore, the observed differences cannot be attributed exclusively to the model versions.

This study demonstrated that ChatGPT-5 provides health-related responses with greater readability and closer alignment to the recommended readability standards for patient education materials than ChatGPT-4o. These findings suggest that ChatGPT-5 may enhance the general public's understanding of medical information and potentially support health education efforts. This study addresses a clear gap in the literature by evaluating the readability of ChatGPT-5 responses for which there is currently limited evidence. A direct comparison of ChatGPT-4o and ChatGPT-5 provides insight into the progress of large language models and their potential to generate more understandable health-related content. That ChatGPT-5 produces more readable outputs than ChatGPT-4o highlights ongoing advancements in artificial intelligence and suggests its growing potential as an accessible source of health information. However, the readability of ChatGPT outputs may vary depending on the complexity of the medical topic and the target audience. Since such responses may directly influence individuals' health-related decisions, presenting the information clearly and comprehensibly is essential. In our study, the use of questions designed for the general public increases the relevance of the findings to real-world health communication and patient education. Importantly, readability reflects only the ease with which information can be read and understood and does not necessarily indicate accuracy, completeness, reliability, or clinical appropriateness. Therefore, the superior readability observed for ChatGPT-5 should not be interpreted as evidence of greater medical accuracy or overall information quality. Future studies should evaluate these additional dimensions to provide a more comprehensive assessment of large language models in health communication. Future development of ChatGPT should focus on tailoring responses to users' educational backgrounds to maximize comprehension and ensure safe, equitable access to health information.

## CONCLUSION

In conclusion, ChatGPT-5 generated significantly more readable responses to ear care questions than ChatGPT-4o, with readability levels closer to those of WHO patient education materials. These findings suggest that newer large language models may have greater potential for providing accessible health information to the general public. However, readability alone does not reflect the accuracy, completeness, or clinical appropriateness of AI-generated information, and further studies evaluating these dimensions are warranted.

## Ethical approval

In this study, since no patient or experimental animal data were used, ethical committee approval was not required.

## Author contribution

Concept: CÖ; Design: CÖ, MFK; Data Collection or Processing: CÖ, MFK; Analysis or Interpretation: CÖ, MFK; Literature Search: CÖ, MFK; Writing: CÖ. All authors reviewed the results and approved the final version of the article.

## Source of funding

The authors declare the study received no funding.

## Conflict of interest

The authors declare that there is no conflict of interest.

## REFERENCES

1. Koirala B, Manandhar S, Shah RK. Knowledge, attitude and practice of ear care among pre clinical medical students at Birat Medical College Teaching Hospital. *Birat Journal of Health Sciences*. 2020; 5(3): 1236-1240. [\[Crossref\]](#)
2. Sağlam Ö, Saat B, Dursun E, et al. The prevalence of cerumen in young male population. *KBB-Forum*. 2013; 12(2): 32-5.
3. Alateeq OM. Knowledge Attitude and practice of ear care in Saudi community. *Baqai J Health Sci*. 2017; 20(2): 11-4.
4. Alrajhi MS, Alim BM, Aldokhayel SD, Zeitouni LM, Al Tawil LK, Alzaharani FA. Knowledge, attitudes, and practices pertaining to cotton-bud usages and the complications related to their misuse among outpatients in an ear, nose, and throat clinic. *Journal of Nature and Science of Medicine*. 2019; 2(4): 220-5.
5. Tonsaker T, Bartlett G, Trpkov C. Health information on the Internet: gold mine or minefield? *Can Fam Physician*. 2014; 60(5): 407-8.
6. Beheshti M, Toubal IE, Alaboud K, et al. Evaluating the reliability of chatgpt for health-related questions: a systematic review. *Informatics*. 2025; 12(1): 9. [\[Crossref\]](#)
7. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. *Research Square [Preprint]*. 2023. [\[Crossref\]](#)
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023; 2(2): e0000198. [\[Crossref\]](#)
9. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023; 3(4): 100324. [\[Crossref\]](#)
10. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023; 20: 1. [\[Crossref\]](#)
11. Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X. Exploring the performance of ChatGPT-4 in the Taiwan audiologist qualification examination: Preliminary observational study highlighting the potential of AI chatbots in hearing care. *JMIR Med Educ*. 2024; 10: e55595. [\[Crossref\]](#)
12. Zalzal HG, Abraham A, Cheng J, Shah RK. Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investig Otolaryngol*. 2024; 9(1): e1193. [\[Crossref\]](#)
13. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. *OTO Open*. 2023; 7(3): e67. [\[Crossref\]](#)
14. Soon S, Perry B. Paging Dr. ChatGPT: safety, accuracy and readability of ChatGPT in ENT emergencies. *Australian Journal of Otolaryngology*. 2025; 8: 8. [\[Crossref\]](#)
15. Adams RC, Smith TP, Metts JK, Ross JW. Readability-its applicability to education of patients by pharmacy. *Hosp Pharm*. 1979; 14(11): 654-6.
16. World Health Organization (WHO). Deafness and hearing loss: ear care. Available at: <https://www.who.int/news-room/questions-and-answers/item/deafness-and-hearing-loss-ear-care> (Accessed on June 4, 2025).
17. Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of flesch formula. *Educ Health (Abingdon)*. 2017; 30(1): 84-8. [\[Crossref\]](#)

18. Flesch R. A new readability yardstick. *J Appl Psychol.* 1948; 32(3): 221-33. [\[Crossref\]](#)
19. Gencer A. Readability analysis of ChatGPT's responses on lung cancer. *Sci Rep.* 2024; 14(1): 17234. [\[Crossref\]](#)
20. Hedman AS. Using the SMOG formula to revise a health-related document. *Am J Health Educ.* 2008; 39(1): 61-4. [\[Crossref\]](#)
21. Azer SA, AlOlayan TI, AlGhamdi MA, AlSanea MA. Inflammatory bowel disease: An evaluation of health information on the internet. *World J Gastroenterol.* 2017; 23(9): 1676-96. [\[Crossref\]](#)
22. Readability Scoring System. Available at: <https://readabilityformulas.com/readability-scoring-system.php> (Accessed on August 20, 2025).
23. Baumann E, Czerwinski F, Rosset M, Seelig M, Suhr R. How do people in Germany seek health information? Insights from the first wave of HINTS Germany. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz.* 2020; 63(9): 1151-60. [\[Crossref\]](#)
24. Kianian R, Carter M, Finkelshtein I, Eleswarapu SV, Kachroo N. Application of artificial intelligence to patient-targeted health information on kidney stone disease. *J Ren Nutr.* 2024; 34(2): 170-6. [\[Crossref\]](#)
25. Fritsch SJ, Blankenheim A, Wahl A, et al. Attitudes and perception of artificial intelligence in healthcare: A cross-sectional survey among patients. *Digit Health.* 2022; 8: 20552076221116772. [\[Crossref\]](#)
26. Høj S, Thomsen SF, Ulrik CS, Meteran H, Sigsgaard T, Meteran H. Evaluating the scientific reliability of ChatGPT as a source of information on asthma. *J Allergy Clin Immunol Glob.* 2024; 3(4): 100330. [\[Crossref\]](#)
27. Özcan F, Örüçü Atar M, Köroğlu Ö, Yılmaz B. Assessment of the reliability and usability of ChatGPT in response to spinal cord injury questions. *J Spinal Cord Med.* 2025; 48(5): 852-7. [\[Crossref\]](#)
28. Cavnar Helvacı B, Hepsen S, Candemir B, et al. Assessing the accuracy and reliability of ChatGPT's medical responses about thyroid cancer. *Int J Med Inform.* 2024; 191: 105593. [\[Crossref\]](#)
29. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep.* 2024; 14(1): 243. [\[Crossref\]](#)
30. Long C, Lowe K, Santos A, et al. Evaluating ChatGPT-4 in otolaryngology-head and neck surgery board examination using the CVSA model. *medRxiv [Preprint].* 2023. [\[Crossref\]](#)
31. Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *J Patient Exp.* 2021; 8: 2374373521998847. [\[Crossref\]](#)
32. Walsh TM, Volsko TA. Readability assessment of internet-based consumer health information. *Respir Care.* 2008; 53(10): 1310-15. [\[Crossref\]](#)
33. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina.* 2023; 7(10): 862-8. [\[Crossref\]](#)
34. Şimşek SD, Şensöğüt C. The use of readability formulas in polytrauma. *Acta Medica Young Doctors.* 2025; 1(1): 2-9. [\[Crossref\]](#)

# Comparative evaluation of four large language models for postoperative patient education after septorhinoplasty: A bilingual cross-sectional study

Zülküf Küçüktağ<sup>1</sup>, Aykut Özdoğan<sup>1</sup>, Gökçe Saygı Uysal<sup>1</sup>, Yağmur Uluhan<sup>1</sup>

<sup>1</sup>Department of Otorhinolaryngology-Head and Neck Surgery, Etlik City Hospital, Ankara, Türkiye

**Cite as:** Küçüktağ Z, Özdoğan A, Saygı Uysal G, Uluhan Y. Comparative evaluation of four large language models for postoperative patient education after septorhinoplasty: A bilingual cross-sectional study. Northwestern Med J. 2026;6(4):317-329.

## Abstract

**Aim:** To compare the quality, readability, and emergency referral performance of four major language models in answering post-septorhinoplasty patient questions in Turkish and English.

**Materials and Methods:** A cross-sectional evaluation study comprised 31 questions: 20 routine postoperative and 11 emergency/red-flag scenarios. Each question was submitted once to ChatGPT, Google Gemini, DeepSeek, and Claude in Turkish and English within independent conversation windows. Two otolaryngologists independently evaluated the responses using the Quality Analysis of Medical Artificial Intelligence tool. English responses were additionally assessed using Flesch Reading Ease, Flesch-Kincaid Grade Level, and Simple Measure of Gobbledygook indices. Emergency referral accuracy was evaluated separately.

**Results:** Total quality scores differed significantly among models for Turkish routine postoperative questions ( $p<0.001$ , Kendall's  $W=0.775$ ) and for Turkish emergency/red-flag questions ( $p=0.009$ ,  $W=0.350$ ), whereas the corresponding overall differences for English questions were not statistically significant. Within-model language comparisons showed significant Turkish-English differences for routine questions across all four models, whereas no language-related differences were observed for emergency/red-flag questions. ChatGPT, DeepSeek, and Claude provided appropriate emergency referrals in all 22 responses, whereas Gemini did so in 16 of 22 responses (72.7%). All models exceeded recommended readability thresholds for patient education materials.

**Conclusion:** The performance of large language models in post-septorhinoplasty patient education varied by language context and question category. Although emergency referral performance was generally high, model-specific omissions occurred, and all models exceeded recommended readability thresholds. These findings support language-specific evaluation, simplification to improve readability, and expert review before clinical use.

**Keywords:** artificial intelligence, large language models, patient education, rhinoplasty, septorhinoplasty, postoperative care

**Corresponding author:** Zülküf Küçüktağ **E-mail:** zkucuk79@yahoo.com

**Received:** 11.06.2026 **Accepted:** 23.09.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

## INTRODUCTION

Artificial intelligence (AI)-based large language models (LLMs) show increasing potential to support patient communication and education in healthcare. Postoperative information provision is one area where this potential is particularly evident; in recent years, studies evaluating the responses provided by LLMs to postoperative patient questions have begun to be published, primarily in the field of otolaryngology but also across various other surgical specialties (1,2). These studies address various aspects, including the alignment of AI tools with clinical guidelines, their potential contributions to patient accessibility, and the challenges in integrating them into healthcare systems (3,4).

In response to this need, studies investigating the use of LLM models in postoperative patient counseling for rhinoplasty and related facial plastic surgery are on the rise (5,6). These studies examine the capacity of AI models to provide reliable and patient-centered information on common concerns related to rhinoplasty and postoperative recovery (7,8). The postoperative course following septorhinoplasty includes a range of expected symptoms, such as swelling, bruising, nasal obstruction, and discomfort that patients may find difficult to distinguish from findings requiring urgent medical assessment. Patients may therefore seek rapid guidance from easily accessible AI-based tools when deciding whether a symptom represents a normal part of recovery or a potential complication. In this context, inaccurate reassurance, incomplete advice, or failure to recommend timely medical evaluation could have direct patient-safety implications. Evaluating not only the overall quality of information generated by LLMs but also their ability to distinguish routine postoperative concerns from emergency and red-flag scenarios is therefore clinically relevant.

The performance of LLMs in different languages and their reliability for patients constitute another important aspect of research in this field. Multilingual comparative studies have shown that LLMs can generate competitive responses in languages other than English (9); however, there are concerns regarding the readability and comprehensibility of AI responses for patients (10). From the patient's perspective,

trust in and acceptance of AI tools have been shown to vary according to demographic characteristics and technology use; however, whether patient trust aligns with clinical reliability remains a subject of debate (11).

However, the vast majority of existing studies evaluate a single model, are conducted in English only, and do not systematically address the distinction between routine postoperative questions and emergency clinical situations. Furthermore, the performance of AI tools in different language groups, such as Turkish-speaking patient populations, remains largely unknown. Although recent studies have compared multiple LLMs in rhinoplasty-related patient education, including the evaluation of postoperative red-flag scenarios (12,13), evidence regarding their comparative performance across different languages remains limited.

This study aimed to evaluate and compare responses from ChatGPT, Gemini, DeepSeek, and Claude to post-septorhinoplasty patient questions in Turkish and English using the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool; assess total QAMAI scores and effect sizes; determine English response readability via Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), and Simple Measure of Gobbledygook (SMOG) indices; and compare emergency referral accuracy across languages. To our knowledge, this is the first study to evaluate questions from septorhinoplasty patients using a bilingual (Turkish-English) design and a systematic distinction between routine and emergency cases. The primary research question was whether the quality of LLM-generated responses to questions from post-septorhinoplasty patients differs across AI models, languages (Turkish vs English), and question types (routine postoperative vs emergency/red-flag). We hypothesized that response quality and emergency referral performance would vary across models and language contexts.

## MATERIALS AND METHODS

### Study design and ethics

This study is a cross-sectional evaluation comparing the responses of four AI models to patient questions posed following septorhinoplasty. Only text generated by AI systems was analyzed; no patient data were used.

Therefore, approval from an ethics committee was not required. The study was reported with reference to the Chatbot Assessment Reporting Tool (CHART) (14), a reporting framework specifically developed for studies evaluating chatbot-generated health advice. Given the cross-sectional comparative design of the study, relevant principles of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) (15) statement were also considered.

### Artificial intelligence models

Four AI models were evaluated: ChatGPT (GPT-5.3 Instant, OpenAI, San Francisco, USA), Google Gemini 2.0 Flash (Google LLC, Mountain View, USA), DeepSeek V3 (DeepSeek AI, Hangzhou, China), and Claude Sonnet 4.6 (Anthropic, San Francisco, USA). These models were selected to represent widely used, general-purpose LLMs developed by different organizations that are capable of generating multilingual responses relevant to patient-facing health information. Their inclusion also enabled comparison across multiple contemporary LLM platforms, rather than limiting the evaluation to a single model or developer. ChatGPT was accessed through a ChatGPT Plus subscription (GPT-5.3 Instant), while Claude was accessed through a paid subscription. Gemini and DeepSeek were accessed through their web-based interfaces; their subscription status was not recorded during data collection. All queries were performed through the models' web-based user interfaces without API access. Data were collected between April 1 and April 30, 2026, using the following LLMs: ChatGPT (GPT-5.3 Instant), Google Gemini 2.0 Flash, DeepSeek V3, and Claude Sonnet 4.6. To implement a blind evaluation procedure, the models were coded as M1 (ChatGPT), M2 (Gemini 2.0 Flash), M3 (DeepSeek V3), and M4 (Claude). Each Turkish and English question was submitted once to each model in a new and independent conversation window to minimize contextual bias. A standard prompt structure was used for all queries: "You are answering a question from a patient who has undergone septorhinoplasty surgery. Provide an answer that is accurate, clear, and patient-friendly according to current medical standards." The methodological framework was designed to align with the multi-model AI comparison study in the field of otolaryngology published by Alsudays et al. (1).

### Question groups

The question set was developed by the research team based on common postoperative patient concerns and clinically relevant warning signs encountered after septorhinoplasty, and was not derived from a single formal clinical guideline. The final set of 31 questions was selected to provide broad coverage of predefined routine and emergency/red-flag domains. Because this study evaluated AI-generated responses using a predefined question set rather than a patient- or population-based sample, no formal a priori sample size or power calculation was performed. The questions were divided into two main categories. Group A — Routine Postoperative Questions (n=20): questions covering expected postoperative findings, activity restrictions, and common patient concerns. These span a wide range, from managing swelling and bruising to interpreting asymmetrical swelling. Group B — Emergency and Red Flag Questions (n=11): This group includes clinical scenarios that require urgent medical evaluation, such as active nasal bleeding, signs of surgical site infection, possible vascular compromise, and orbital complications. All questions were prepared in both Turkish and English, and their translation accuracy and clinical content validity were verified by a bilingual ENT specialist. Questions were reviewed by two senior ENT specialists for clinical relevance and comprehensiveness prior to data collection, providing face validity for the question set. Each question was submitted separately in Turkish and English within independent conversation windows to prevent cross-language contextual bias.

### Assessment tool

The responses were analyzed using the QAMAI tool, which was developed and validated by Vaira et al. (16) to assess the quality of health information provided by AI platforms. QAMAI measures five core criteria on a Likert scale ranging from 1 (inadequate) to 5 (excellent): Accuracy, Clarity, Relevance, Completeness, and Usefulness. The maximum total QAMAI score is 25; the total QAMAI score is defined as the sum of these five criteria (range: 5–25). Two independent ENT specialists (R1 and R2) with ≥10 years of clinical experience evaluated all responses independently, with the models' identities blinded. The study comprised 31 questions × 4 models × 2

languages, yielding 248 model-generated responses. Each response was evaluated by two raters across five QAMAI criteria, corresponding to 2,480 individual criterion-level ratings (248 responses  $\times$  5 criteria  $\times$  2 raters). For inter-rater agreement analyses, ratings from the two evaluators were paired for each criterion, yielding 1,240 expected criterion-level rating pairs (248 responses  $\times$  5 criteria). Because one criterion-level rating was missing, 1,239 complete criterion-level rating pairs were available for analysis.

### Readability analysis

Three readability metrics were applied to the English responses: FRE, FKGL, and the SMOG index (17,18). The American Medical Association (AMA) and the U.S. National Institutes of Health (NIH) recommend the SMOG index as the primary measure of readability for patient education materials; they set the target level at SMOG  $\leq 8$  and FKGL  $\leq 6$  (6th-grade reading level). Since no psychometrically validated readability tool is available for Turkish, responses in Turkish were not included in this analysis.

### Accuracy of urgent referral recommendations

For Group B responses, both raters independently assessed whether the AI model (a) correctly identified the clinical urgency of the scenario, and (b) included a clear recommendation for the patient to visit the emergency department or to contact their physician immediately. Responses were classified into two categories: correct (appropriate and sufficient emergency referrals) and incorrect (inadequate communication of urgency or a lack of referral).

### Statistical Analysis

The distribution of continuous variables was formally assessed using the Shapiro–Wilk test. Response length was quantified as the number of whitespace-delimited words in each model-generated response, excluding the question text and model labels. Because the same questions were answered by all four models and the response-length distributions did not consistently satisfy parametric assumptions, between-model differences in response length were analyzed using the Friedman test, with the question as the matched unit. Because the same LLM responses were

independently evaluated by both raters, the two sets of QAMAI ratings were treated as paired rather than independent observations. For the primary QAMAI analyses, scores assigned by the two raters were averaged at the individual response level, yielding a single combined score for each question–model combination. Total QAMAI scores were calculated only when all five criterion scores were available from both raters; missing values were not imputed. Because QAMAI scores are bounded ordinal ratings and several distributions showed ceiling effects and departures from normality, non-parametric methods were used. Model comparisons within each language and question group were performed using the Friedman test, with question as the matched unit. Significant omnibus tests were followed by pairwise Wilcoxon signed-rank tests with Holm correction. Kendall’s W was reported as the effect-size measure for Friedman tests. For the mean total QAMAI scores, 95% confidence intervals were estimated using percentile bootstrap resampling of response-level scores (20,000 resamples).

Turkish–English comparisons of total QAMAI scores within each model were performed using paired Wilcoxon signed-rank tests, with the same question serving as the matching unit. Holm correction was applied within each question group for the four prespecified model-wise language comparisons, and matched-pairs rank-biserial correlation ( $r_{rb}$ ) was reported as the effect-size measure.

Inter-rater agreement analyses were performed using available complete rating pairs. Linearly weighted Cohen’s kappa was calculated for individual QAMAI criteria, and intraclass correlation coefficients [ICC (2,1) and ICC (2,k)] were calculated for total QAMAI scores. For total QAMAI ICC estimates, 95% confidence intervals were obtained using nonparametric bootstrap resampling of complete response-level pairs (20,000 resamples).

Emergency referral accuracy was compared across the four models within each language using Cochran’s Q test, accounting for the matched structure of the 11 emergency/red-flag questions. When the omnibus test was significant, pairwise comparisons were performed using exact McNemar tests with Holm correction. Turkish–English differences in emergency referral

accuracy within each model were assessed using exact McNemar tests for paired binary data. For emergency-referral accuracy proportions, 95% confidence intervals were calculated using Wilson score intervals. Overall, Turkish–English emergency referral rates were reported only descriptively.

Because the Turkish and English datasets comprised paired versions of the same questions, inferential analyses were performed separately for each language; no inferential analysis was conducted on pooled Turkish–English data. All analyses were performed using Python/SciPy v1.11. Statistical significance was set at  $p < 0.05$ .

## RESULTS

### Inter-rater agreement

The inter-rater reliability analysis yielded an overall linear-weighted  $\kappa$  of 0.284 at the criterion level. For total QAMAI scores, the overall ICC (2,1) was 0.364 (95% CI, 0.260–0.457) and ICC (2,k) was 0.534 (95% CI, 0.412–0.627) (Table 1). Across language–question group datasets, total-score ICC (2,1) values ranged from 0.415 to 0.553 in Turkish and from 0.179 to 0.208 in English. Overall, 1,218 of 1,239 complete paired criterion-level ratings (98.3%) differed by no more than 1 scale point, and no discrepancy exceeded 2 points. One criterion-level rating was inadvertently

**Table 1.** Inter-rater agreement across all response categories.

| Dataset                                                                      | Simple $\kappa$ | Weighted $\kappa$ (linear) | Interpretation | Exact agreement          | $\pm 1$ agreement          | ICC (2,1) [95% CI]         | ICC (2,k) [95% CI]         |
|------------------------------------------------------------------------------|-----------------|----------------------------|----------------|--------------------------|----------------------------|----------------------------|----------------------------|
| <b>Criterion-level scores (5 criteria <math>\times</math> all responses)</b> |                 |                            |                |                          |                            |                            |                            |
| Turkish — Group A (n=400)                                                    | 0.252           | 0.338                      | Fair           | 255/400 (63.7%)          | 393/400 (98.2%)            | 0.432                      | 0.603                      |
| Turkish — Group B (n=220)                                                    | 0.323           | 0.363                      | Fair           | 149/220 (67.7%)          | 216/220 (98.2%)            | 0.409                      | 0.581                      |
| English — Group A (n=399)                                                    | 0.188           | 0.190                      | Slight         | 261/399 (65.4%)          | 392/399 (98.2%)            | 0.161                      | 0.277                      |
| English — Group B (n=220)                                                    | 0.219           | 0.234                      | Fair           | 145/220 (65.9%)          | 217/220 (98.6%)            | 0.253                      | 0.404                      |
| <b>Overall (n=1,239)</b>                                                     | <b>0.239</b>    | <b>0.284</b>               | <b>Fair</b>    | <b>810/1,239 (65.4%)</b> | <b>1,218/1,239 (98.3%)</b> | <b>0.332</b>               | <b>0.499</b>               |
| <b>Total QAMAI score (range 5–25)</b>                                        |                 |                            |                |                          |                            |                            |                            |
| Turkish — Group A (n=80)                                                     | —               | —                          | —              | —                        | —                          | 0.415 [0.248–0.555]        | 0.587 [0.397–0.713]        |
| Turkish — Group B (n=44)                                                     | —               | —                          | —              | —                        | —                          | 0.553 [0.321–0.718]        | 0.713 [0.486–0.836]        |
| English — Group A (n=79)                                                     | —               | —                          | —              | —                        | —                          | 0.208 [–0.004–0.394]       | 0.344 [–0.008–0.565]       |
| English — Group B (n=44)                                                     | —               | —                          | —              | —                        | —                          | 0.179 [–0.120–0.432]       | 0.304 [–0.273–0.603]       |
| <b>Overall (n=247)</b>                                                       | <b>—</b>        | <b>—</b>                   | <b>—</b>       | <b>—</b>                 | <b>—</b>                   | <b>0.364 [0.260–0.457]</b> | <b>0.534 [0.412–0.627]</b> |

Weighted  $\kappa$  (linear weights) = primary metric for ordinal Likert data. ICC(2,1) = single-rater; ICC(2,k) = average-measures intraclass correlation coefficient (two-way random effects, absolute agreement). Values in brackets represent 95% confidence intervals. The 95% confidence intervals for total QAMAI ICC estimates were obtained using nonparametric bootstrap resampling of complete response-level pairs (20,000 resamples).  $\kappa$  interpretation (Landis & Koch, 1977): 0.00–0.20 = Slight; 0.21–0.40 = Fair; 0.41–0.60 = Moderate. Overall, 98.3% of complete paired criterion-level ratings differed by no more than 1 scale point. One criterion-level rating (Evaluator 2, Accuracy, English Group A) was missing, resulting in 399/400 complete criterion-level rating pairs and 79/80 complete total QAMAI rating pairs; no imputation was performed. QAMAI: Quality Analysis of Medical Artificial Intelligence.

left blank by Evaluator 2 in the English routine dataset and was therefore excluded from the corresponding complete-pair analyses; no imputation was performed.

### Total QAMAI score

As shown in Table 2, total QAMAI scores differed significantly across models for Turkish routine postoperative questions ( $\chi^2=46.52$ ,  $p<0.001$ , Kendall's  $W=0.775$ ). Pairwise Wilcoxon signed-rank tests with Holm correction showed significant differences between ChatGPT and Gemini (adjusted  $p=0.001$ ), between ChatGPT and DeepSeek (adjusted  $p<0.001$ ), between ChatGPT and Claude (adjusted  $p=0.002$ ), between Gemini and DeepSeek (adjusted  $p<0.001$ ), and between DeepSeek and Claude (adjusted  $p<0.001$ ); however, the difference between Gemini and Claude was not significant (adjusted  $p=0.207$ ). No significant overall model differences were observed for English routine postoperative questions ( $\chi^2=7.60$ ,  $p=0.055$ , Kendall's  $W=0.133$ ). For emergency/red-flag questions, a significant overall model difference was observed in Turkish ( $\chi^2=11.56$ ,  $p=0.009$ , Kendall's  $W=0.350$ ), although, after the Holm correction, no pairwise comparison remained significant. The

corresponding overall model difference for English was not statistically significant ( $\chi^2=7.63$ ,  $p=0.054$ , Kendall's  $W=0.231$ ).

### QAMAI criteria scores-group A: Routine postoperative questions

As shown in Table 3, significant differences among models were observed for the Turkish routine questions in Accuracy ( $\chi^2=30.86$ ,  $p<0.001$ , Kendall's  $W=0.514$ ), Clarity ( $\chi^2=25.90$ ,  $p<0.001$ , Kendall's  $W=0.432$ ), Completeness ( $\chi^2=50.83$ ,  $p<0.001$ , Kendall's  $W=0.847$ ), and Usefulness ( $\chi^2=37.79$ ,  $p<0.001$ , Kendall's  $W=0.630$ ); Relevance did not differ significantly ( $p=0.392$ ). Holm-adjusted pairwise comparisons showed significant differences between DeepSeek and each of the other three models for Accuracy. For clarity, significant differences were observed between Gemini and each of ChatGPT, DeepSeek, and Claude. All six pairwise comparisons were significant for both Completeness and Usefulness. For English routine questions, significant model differences were observed for Clarity ( $\chi^2=22.96$ ,  $p<0.001$ ,  $W=0.383$ ) and Completeness ( $\chi^2=9.12$ ,  $p=0.028$ ,  $W=0.152$ ), whereas Accuracy, Relevance, and Usefulness did not

**Table 2.** Total QAMAI score by model, language, and question group.

| Group                            | Language | ChatGPT<br>Mean±SD<br>[95% CI] | Gemini<br>Mean±SD<br>[95% CI] | DeepSeek<br>Mean±SD<br>[95% CI] | Claude<br>Mean±SD<br>[95% CI] | $\chi^2$ | p       | Kendall's<br>W |
|----------------------------------|----------|--------------------------------|-------------------------------|---------------------------------|-------------------------------|----------|---------|----------------|
| Routine Postoperative Questions  | TR       | 23.20±0.70<br>[22.88–23.45]    | 24.08±0.80<br>[23.73–24.40]   | 20.88±1.18<br>[20.38–21.38]     | 24.35±0.69<br>[24.02–24.60]   | 46.52    | <0.001* | 0.775          |
| Routine Postoperative Questions† | EN       | 23.85±0.92<br>[23.45–24.23]    | 23.00±1.30<br>[22.40–23.50]   | 23.68±0.77<br>[23.35–24.00]     | 23.66±1.25<br>[23.11–24.18]   | 7.60     | 0.055   | 0.133          |
| Emergency and Red Flag Questions | TR       | 23.82±1.10<br>[23.14–24.36]    | 22.45±1.59<br>[21.55–23.32]   | 22.41±1.28<br>[21.64–23.09]     | 24.00±1.95<br>[22.77–24.82]   | 11.56    | 0.009*  | 0.350          |
| Emergency and Red Flag Questions | EN       | 23.95±0.69<br>[23.55–24.32]    | 22.68±1.12<br>[22.05–23.32]   | 23.00±1.18<br>[22.32–23.64]     | 23.91±1.04<br>[23.32–24.45]   | 7.63     | 0.054   | 0.231          |

Values are presented as mean ± SD [95% CI] based on response-level scores obtained by averaging the two raters. The 95% confidence intervals for mean total QAMAI scores were estimated using percentile bootstrap resampling (20,000 resamples). Model comparisons within each language and question group were performed using the Friedman test, with question as the matched unit. Kendall's  $W$  is reported as the effect-size measure. When the omnibus Friedman test was significant, pairwise Wilcoxon signed-rank tests with Holm correction were performed. †For English routine postoperative questions, one Claude response with an incomplete rater assessment was excluded; Claude  $n=19$ , while the other model-specific descriptive estimates used  $n=20$ . The complete-case Friedman analysis used  $n=19$  matched questions. TR = Turkish; EN = English. Total QAMAI score range: 5–25.  $p < 0.05$ .

differ significantly. Pairwise comparisons for Clarity were significant for ChatGPT versus Gemini, ChatGPT versus Claude, Gemini versus DeepSeek, and DeepSeek versus Claude after Holm correction. Although the omnibus test for Completeness was significant, no pairwise comparison remained significant after Holm correction.

### QAMAI criteria scores-group B: Emergency and red flag questions

As shown in Table 3, significant model differences were observed for Turkish emergency/red-flag questions in Clarity ( $\chi^2=16.50$ ,  $p<0.001$ , Kendall's  $W=0.500$ ) and Completeness ( $\chi^2=16.42$ ,  $p<0.001$ , Kendall's  $W=0.498$ ); whereas Accuracy, Relevance, and

**Table 3.** QAMAI criterion scores by question group and language.

| Group              | Lang. | Criterion    | n  | ChatGPT Mean±SD | Gemini Mean±SD | DeepSeek Mean±SD | Claude Mean±SD | $\chi^2$ | p                 | Kendall's W |
|--------------------|-------|--------------|----|-----------------|----------------|------------------|----------------|----------|-------------------|-------------|
| Routine            | TR    | Accuracy     | 20 | 4.90±0.26       | 4.90±0.21      | 4.38±0.46        | 4.95±0.22      | 30.86    | <b>&lt;0.001*</b> | 0.514       |
| Routine            | TR    | Clarity      | 20 | 4.47±0.11       | 4.90±0.21      | 4.40±0.42        | 4.58±0.24      | 25.90    | <b>&lt;0.001*</b> | 0.432       |
| Routine            | TR    | Relevance    | 20 | 4.97±0.11       | 5.00±0.00      | 5.00±0.00        | 5.00±0.00      | 3.00     | 0.392             | 0.050       |
| Routine            | TR    | Completeness | 20 | 4.40±0.21       | 4.65±0.33      | 3.17±0.44        | 4.92±0.18      | 50.83    | <b>&lt;0.001*</b> | 0.847       |
| Routine            | TR    | Usefulness   | 20 | 4.45±0.15       | 4.62±0.32      | 3.92±0.44        | 4.90±0.26      | 37.79    | <b>&lt;0.001*</b> | 0.630       |
| Routine            | EN    | Accuracy     | 19 | 4.82±0.34       | 4.71±0.51      | 4.84±0.24        | 4.76±0.31      | 1.19     | 0.754             | 0.021       |
| Routine            | EN    | Clarity      | 20 | 4.97±0.11       | 4.53±0.38      | 4.85±0.24        | 4.50±0.43      | 22.96    | <b>&lt;0.001*</b> | 0.383       |
| Routine            | EN    | Relevance    | 20 | 5.00±0.00       | 5.00±0.00      | 5.00±0.00        | 5.00±0.00      | 0.00     | 1.000             | 0.000       |
| Routine            | EN    | Completeness | 20 | 4.45±0.32       | 4.40±0.31      | 4.47±0.38        | 4.72±0.41      | 9.12     | <b>0.028*</b>     | 0.152       |
| Routine            | EN    | Usefulness   | 20 | 4.60±0.31       | 4.38±0.36      | 4.50±0.28        | 4.65±0.43      | 5.82     | 0.121             | 0.097       |
| Emergency/Red flag | TR    | Accuracy     | 11 | 4.77±0.34       | 4.55±0.47      | 4.64±0.45        | 4.77±0.61      | 4.62     | 0.202             | 0.140       |
| Emergency/Red flag | TR    | Clarity      | 11 | 4.86±0.23       | 4.45±0.15      | 4.41±0.20        | 4.73±0.34      | 16.50    | <b>&lt;0.001*</b> | 0.500       |
| Emergency/Red flag | TR    | Relevance    | 11 | 5.00±0.00       | 5.00±0.00      | 5.00±0.00        | 5.00±0.00      | 0.00     | 1.000             | 0.000       |
| Emergency/Red flag | TR    | Completeness | 11 | 4.55±0.35       | 4.32±0.51      | 3.95±0.27        | 4.82±0.46      | 16.42    | <b>&lt;0.001*</b> | 0.498       |
| Emergency/Red flag | TR    | Usefulness   | 11 | 4.64±0.45       | 4.14±0.60      | 4.41±0.58        | 4.68±0.64      | 6.22     | 0.101             | 0.188       |
| Emergency/Red flag | EN    | Accuracy     | 11 | 4.91±0.20       | 4.73±0.34      | 4.73±0.26        | 4.73±0.26      | 3.69     | 0.297             | 0.112       |
| Emergency/Red flag | EN    | Clarity      | 11 | 4.91±0.20       | 4.45±0.42      | 4.91±0.20        | 4.64±0.45      | 10.66    | <b>0.014*</b>     | 0.323       |
| Emergency/Red flag | EN    | Relevance    | 11 | 5.00±0.00       | 5.00±0.00      | 4.91±0.20        | 4.82±0.25      | 8.25     | <b>0.041*</b>     | 0.250       |
| Emergency/Red flag | EN    | Completeness | 11 | 4.55±0.27       | 4.27±0.61      | 4.18±0.40        | 5.00±0.00      | 17.07    | <b>&lt;0.001*</b> | 0.517       |
| Emergency/Red flag | EN    | Usefulness   | 11 | 4.59±0.30       | 4.23±0.41      | 4.27±0.41        | 4.73±0.26      | 11.04    | <b>0.012*</b>     | 0.335       |

Values are mean±SD based on criterion-level scores averaged across the two raters. Model comparisons were performed using the Friedman test, with Kendall's W as the effect size; significant omnibus tests were followed by pairwise Wilcoxon signed-rank tests with Holm correction. †For English routine Accuracy, one incomplete rater assessment resulted in n=19; all other routine analyses included n=20 and emergency/red-flag analyses n=11. TR=Turkish; EN=English.  $p<0.05$ .

Usefulness did not differ significantly. Holm-adjusted pairwise comparisons showed significant differences in Clarity between ChatGPT and Gemini and between ChatGPT and DeepSeek; for Completeness, only the comparison between DeepSeek and Claude remained significant. For English emergency/red-flag questions, significant overall differences were observed for Clarity ( $\chi^2=10.66$ ,  $p=0.014$ ,  $W=0.323$ ), Relevance ( $\chi^2=8.25$ ,  $p=0.041$ ,  $W=0.250$ ), Completeness ( $\chi^2=17.07$ ,  $p<0.001$ ,  $W=0.517$ ), and Usefulness ( $\chi^2=11.04$ ,  $p=0.012$ ,  $W=0.335$ ); whereas Accuracy did not differ significantly. After the Holm correction, no individual pairwise comparisons remained significant for Clarity, Relevance, or Usefulness. For completeness, significant differences remained between Claude and each of ChatGPT, Gemini, and DeepSeek.

### Readability Analysis and Response Length

As shown in Table 4, the mean readability scores of all models exceeded the recommended AMA/NIH targets for patient education materials (SMOG  $\leq 8$  and FKGL  $\leq 6$ ). SMOG scores ranged from  $11.0\pm 0.2$  (DeepSeek, Group A) to  $12.4\pm 0.2$  (Claude, Group B), corresponding to a reading level of grades 11–12, while FRE scores ranged from  $36.2\pm 0.5$  to  $48.6\pm 0.7$  across models and question groups. Regarding response length, DeepSeek generated the shortest responses in Turkish Group A ( $41.8\pm 8.4$  words), whereas Claude generated the longest responses in English Group B ( $153.0\pm 43.2$  words). Between-model differences in response length were statistically significant in all four language–question group combinations (Turkish routine:  $p<0.001$ ; English routine:  $p=0.034$ ; Turkish emergency/red-flag:  $p=0.003$ ; English emergency/red-flag:  $p<0.001$ ).

**Table 4.** Response characteristics and readability metrics.

| Metric                                                               | ChatGPT<br>Mean $\pm$ SD | Gemini<br>Mean $\pm$ SD | DeepSeek<br>Mean $\pm$ SD | Claude<br>Mean $\pm$ SD | p                                         |
|----------------------------------------------------------------------|--------------------------|-------------------------|---------------------------|-------------------------|-------------------------------------------|
| <b>Group A: Routine Postoperative Questions (n=20)</b>               |                          |                         |                           |                         |                                           |
| FRE score†                                                           | 39.6 $\pm$ 0.7           | 44.5 $\pm$ 0.6          | 48.6 $\pm$ 0.6            | 36.7 $\pm$ 0.5          | <0.001*                                   |
| FKGL score†                                                          | 11.0 $\pm$ 0.3           | 10.6 $\pm$ 0.3          | 9.8 $\pm$ 0.2             | 11.5 $\pm$ 0.2          | <0.001*                                   |
| SMOG index†                                                          | 11.9 $\pm$ 0.2           | 11.7 $\pm$ 0.2          | 11.0 $\pm$ 0.2            | 12.2 $\pm$ 0.2          | <0.001*                                   |
| Word count (TR)                                                      | 76.1 $\pm$ 22.7          | 96.6 $\pm$ 32.5         | 41.8 $\pm$ 8.4            | 86.3 $\pm$ 27.5         | <0.001*                                   |
| Word count (EN)                                                      | 110.7 $\pm$ 40.8         | 116.5 $\pm$ 45.0        | 87.3 $\pm$ 21.7           | 115.8 $\pm$ 44.8        | 0.034*                                    |
| <b>Group B: Emergency and Red Flag Questions (n=11)</b>              |                          |                         |                           |                         |                                           |
| FRE score†                                                           | 39.3 $\pm$ 0.8           | 45.6 $\pm$ 0.5          | 48.6 $\pm$ 0.7            | 36.2 $\pm$ 0.5          | <0.001*                                   |
| FKGL score†                                                          | 11.1 $\pm$ 0.4           | 10.1 $\pm$ 0.3          | 9.8 $\pm$ 0.3             | 11.7 $\pm$ 0.2          | <0.001*                                   |
| SMOG index†                                                          | 12.1 $\pm$ 0.2           | 11.3 $\pm$ 0.2          | 11.1 $\pm$ 0.2            | 12.4 $\pm$ 0.2          | <0.001*                                   |
| Word count (TR)                                                      | 90.2 $\pm$ 30.8          | 106.9 $\pm$ 57.5        | 50.0 $\pm$ 29.6           | 96.5 $\pm$ 45.5         | 0.003*                                    |
| Word count (EN)                                                      | 138.9 $\pm$ 20.3         | 148.9 $\pm$ 54.3        | 91.4 $\pm$ 22.5           | 153.0 $\pm$ 43.2        | <0.001*                                   |
| <b>AMA/NIH Recommended Threshold for Patient Education Materials</b> |                          |                         |                           |                         |                                           |
| Target reading level                                                 | FRE $\geq 61$            | FKGL $\leq 6$           | SMOG $\leq 8$             | —                       | All models exceeded recommended threshold |

Mean $\pm$ SD. †Applied to English responses only (no validated readability instruments for Turkish). FRE=Flesch Reading Ease (higher=easier to read). FKGL=Flesch-Kincaid Grade Level. SMOG=Simple Measure of Gobbledygook (preferred for medical patient education per AMA/NIH). AMA/NIH patient education target: SMOG  $\leq 8$ , FKGL  $\leq 6$  (sixth-grade reading level). Word-count comparisons across models were performed using the Friedman test, with question as the matched unit.  $p<0.05$ .

## Accuracy of Emergency Referral

As shown in Table 5, ChatGPT, DeepSeek, and Claude provided appropriate emergency referrals in all 22 responses (100%). Gemini provided an appropriate emergency referral in 16 of 22 responses (72.7%). When emergency referral accuracy was compared across the four models using Cochran's Q test, a significant overall difference was observed for the Turkish responses ( $Q=12.00$ ,  $p=0.007$ ), whereas no significant difference was found for the English responses ( $Q=6.00$ ,  $p=0.112$ ). Pairwise exact McNemar tests with Holm correction showed no statistically significant pairwise differences between models in Turkish (all adjusted  $p \geq 0.750$ ). In paired Turkish-English comparisons within each model, no significant language-related difference in

emergency referral accuracy was observed (Gemini: exact McNemar  $p=0.500$ ; ChatGPT, DeepSeek, and Claude:  $p=1.000$ ). In Turkish, Gemini failed to provide an appropriate emergency referral for Q22 (fever and foul-smelling discharge), Q23 (post-traumatic deformity), Q25 (marked respiratory distress), and Q28 (post-tamponade bleeding).

## Comparison of Response Quality Between Turkish and English

For routine postoperative questions, paired Turkish-English comparisons showed significant differences in total QAMAI scores for all four models after Holm correction (Table 6). English scores were higher than Turkish scores for ChatGPT ( $p=0.004$ ,  $r_{rb}=-0.813$ ) and

**Table 5.** Emergency referral accuracy-group B: emergency and red flag questions.

| Language        | n         | ChatGPT                     | Gemini                     | DeepSeek                    | Claude                      | Cochran's Q | p      |
|-----------------|-----------|-----------------------------|----------------------------|-----------------------------|-----------------------------|-------------|--------|
| Turkish         | 11        | 11/11 (100.0%)              | 7/11 (63.6%)               | 11/11 (100.0%)              | 11/11 (100.0%)              | 12.00       | 0.007* |
|                 |           | [95% CI 74.1–100.0%]        | [95% CI 35.4–84.8%]        | [95% CI 74.1–100.0%]        | [95% CI 74.1–100.0%]        |             |        |
| English         | 11        | 11/11 (100.0%)              | 9/11 (81.8%)               | 11/11 (100.0%)              | 11/11 (100.0%)              | 6.00        | 0.112  |
|                 |           | [95% CI 74.1–100.0%]        | [95% CI 52.3–94.9%]        | [95% CI 74.1–100.0%]        | [95% CI 74.1–100.0%]        |             |        |
| <b>Overall†</b> | <b>22</b> | <b>22/22 (100.0%)</b>       | <b>16/22 (72.7%)</b>       | <b>22/22 (100.0%)</b>       | <b>22/22 (100.0%)</b>       | —           | —      |
|                 |           | <b>[95% CI 85.1–100.0%]</b> | <b>[95% CI 51.8–86.8%]</b> | <b>[95% CI 85.1–100.0%]</b> | <b>[95% CI 85.1–100.0%]</b> |             |        |

Values are presented as n/N (%) [95% CI]. The 95% confidence intervals for referral accuracy proportions were calculated using Wilson score intervals. Within-language differences among the four models were evaluated using Cochran's Q test because the same emergency questions were assessed across models. When the omnibus test was significant, pairwise exact McNemar tests with Holm correction were used. †Overall values are descriptive only and combine the Turkish and English responses; no pooled inferential test was performed.  $p < 0.05$ .

**Table 6.** Language comparison of total QAMAI scores (Turkish vs English).

| Group                                            | Model    | n   | Turkish Mean±SD | English Mean±SD | p       | r <sub>rb</sub> | Direction |
|--------------------------------------------------|----------|-----|-----------------|-----------------|---------|-----------------|-----------|
| <b>Group A: Routine Postoperative Questions</b>  |          |     |                 |                 |         |                 |           |
| Routine                                          | ChatGPT  | 20  | 23.20±0.70      | 23.85±0.92      | 0.004*  | -0.813          | EN>TR     |
| Routine                                          | Gemini   | 20  | 24.08±0.80      | 23.00±1.30      | 0.001*  | +1.000          | TR>EN     |
| Routine                                          | DeepSeek | 20  | 20.88±1.18      | 23.68±0.77      | <0.001* | -0.990          | EN>TR     |
| Routine                                          | Claude   | 19† | 24.47±0.42      | 23.66±1.25      | 0.014*  | +0.691          | TR>EN     |
| <b>Group B: Emergency and Red Flag Questions</b> |          |     |                 |                 |         |                 |           |
| Emergency                                        | ChatGPT  | 11  | 23.82±1.10      | 23.95±0.69      | 1.000   | -0.028          | ns        |
| Emergency                                        | Gemini   | 11  | 22.45±1.59      | 22.68±1.12      | 1.000   | 0.000           | ns        |
| Emergency                                        | DeepSeek | 11  | 22.41±1.28      | 23.00±1.18      | 1.000   | -0.400          | ns        |
| Emergency                                        | Claude   | 11  | 24.00±1.95      | 23.91±1.04      | 1.000   | +0.356          | ns        |

Values are mean±SD based on response-level total QAMAI scores averaged across the two raters. Turkish-English comparisons were performed using paired Wilcoxon signed-rank tests with Holm correction;  $r_{rb}$  indicates matched-pairs rank-biserial correlation (positive=TR>EN; negative=EN>TR). †Claude routine analysis included n=19 because of one incomplete English assessment; all other routine analyses included n=20 and emergency/red-flag analyses n=11. TR=Turkish; EN=English; ns=not significant.  $p < 0.05$ .

DeepSeek ( $p < 0.001$ ,  $r_{rb} = -0.990$ ), whereas Turkish scores were higher for Gemini ( $p = 0.001$ ,  $r_{rb} = +1.000$ ) and Claude ( $p = 0.014$ ,  $r_{rb} = +0.691$ ). For emergency or red-flag questions, no significant difference between Turkish and English was observed in any model after Holm correction (all adjusted  $p = 1.000$ ).

## DISCUSSION

This study provides a comparative evaluation of four major language models for post-septorhinoplasty patient education, using a bilingual Turkish-English design and systematically distinguishing between routine and emergency/red-flag questions. The findings of this study highlight model differences that have direct implications for patient safety, health equity, and the clinical integration of AI tools.

A clinically relevant finding was model-specific omissions in emergency referral recommendations, including infection, post-tamponade bleeding, and marked respiratory distress. These findings align with Alsudays et al. (1), who demonstrated inter-model differences in FESS patient education, with Gemini generally receiving lower ratings than ChatGPT-4 and DeepSeek. Related otolaryngology studies evaluating postoperative guidance and emergency management have also identified limitations in AI-generated medical advice, supporting the need for continued clinical oversight (19,20). Although Gemini showed lower descriptive emergency-referral accuracy in Turkish than in English, the paired-language comparison was not statistically significant, indicating that this apparent difference should be interpreted with caution. Our findings should also be considered in the context of recent comparative studies evaluating LLMs for rhinoplasty-related patient communication. Meyer et al. (12) compared ChatGPT, Gemini, and Claude in answering rhinoplasty consultation questions and identified differences in response quality among models, while also highlighting limitations in completeness and readability. Similarly, Palmer et al. (13) evaluated ChatGPT, Claude, Perplexity, and Gemini using patient-initiated questions following functional rhinoplasty, including red-flag scenarios, and reported generally favorable performance, particularly for

ChatGPT, although readability remained an important limitation. These findings are broadly consistent with our results, which demonstrate generally high response quality but also meaningful model-specific differences. The present study extends this literature by evaluating both Turkish and English responses within the same framework, thereby allowing potential language-related differences in response quality and emergency referral performance to be examined.

Model performance varied across language contexts, with differences between models more evident in Turkish routine postoperative responses than in the corresponding English responses. This finding highlights the importance of evaluating LLMs in the language patients are expected to use them, rather than extrapolating performance from English-only assessments. Sallam et al. (9) reported no significant language-related differences in the performance of DeepSeek-R1 and ChatGPT-4o across Arabic and English ophthalmology queries, whereas in our study DeepSeek showed lower response quality in Turkish than in English for routine postoperative questions. Differences across clinical domains and languages may therefore affect the generalizability of multilingual LLM evaluations, supporting language- and context-specific assessment before clinical implementation.

Sari et al. (10) reported mean FRE scores of 28.77 for DeepSeek-V3 and 30.4 for ChatGPT-4, with responses from both models classified as very difficult to read; they further noted that such complex language may reduce accessibility for patients with limited health literacy. The readability findings in Table 4 show that all models exceeded the recommended thresholds for patient education materials (SMOG  $\leq 8$ , FKGL  $\leq 6$ ). This result is consistent with the study by Ayoub et al. (3), which reported lower understandability and actionability scores for ChatGPT-generated postoperative instructions compared with institutional materials. Our findings also suggest a need for a balance between content coverage and readability. Indeed, Claude, which generated the most comprehensive responses, also provided the responses with the lowest readability levels (Group B SMOG:  $12.4 \pm 0.2$ ). This suggests that more detailed and comprehensive information is often presented using a more complex

language structure. Therefore, it is important that AI-generated patient education materials be reviewed for readability and be simplified as needed before clinical use.

Notably, Claude's Turkish response to the persistent bright-red nasal bleeding scenario appropriately recognized the urgency of the situation and recommended immediate contact with the surgeon or immediate evaluation in the emergency department. However, it advised against applying pressure to the soft part of the nose, which represents a contradictory first-aid instruction. Although this response was classified as correct in terms of emergency referral accuracy, it illustrates that appropriate recognition of an emergency and a referral recommendation do not necessarily ensure that all accompanying first-aid instructions provided by an LLM are clinically appropriate. These findings are consistent with Alghamdi et al. (5), who reported that ChatGPT-4 provided reliable responses to common post-rhinoplasty questions, with performance comparable to that of a senior consultant facial plastic surgeon, and with Ibas et al. (6), who found that ChatGPT-4o provided clinically accurate and safe responses to common postoperative rhinoplasty questions, with no critical errors identified by expert evaluators. Additionally, Capelleras et al. (7) highlighted the potential of ChatGPT to support patient education during rhinoplasty recovery by providing general information and reassurance. Taken together, these findings indicate that an appropriate emergency referral should not be interpreted as evidence that all accompanying clinical advice is accurate; they highlight the need for multidimensional assessment and expert oversight of LLM-generated postoperative guidance.

In Group B, no significant between-model differences in accuracy were observed for either Turkish ( $p=0.202$ ) or English ( $p=0.297$ ). However, significant between-model differences were observed in other quality dimensions, particularly Clarity and Completeness, suggesting that the evaluation of responses to emergency/red-flag questions should not rely solely on the accuracy criterion. These findings indicate that clinically significant differences may stem not only from the accuracy of medical information but also from whether this information is conveyed in a

complete and understandable manner. Therefore, when evaluating the clinical reliability of AI tools, it appears more appropriate to use multidimensional quality assessment tools rather than focusing solely on accuracy metrics.

Patient trust in AI tools does not necessarily reflect clinical reliability. Mayo-Yáñez et al. (11) demonstrated that patient trust in and acceptance of AI may vary according to demographic characteristics and technology use; however, the model-specific variation in emergency referral performance observed in our study suggests that such trust should not be assumed to indicate consistent performance in safety-critical scenarios. Aliyeva et al. (2) demonstrated the potential utility of ChatGPT-4 as a supplementary information resource for postoperative patients, particularly in settings with limited access to healthcare professionals. More broadly, Crowson and Nwosu (4) emphasized the importance of careful clinical integration, ethical considerations, and appropriate oversight when implementing AI tools in otolaryngology practice.

Inter-rater agreement was stronger in the Turkish datasets than in the English datasets, although agreement estimates were not uniformly high, supporting cautious interpretation of between-model differences, particularly when effect sizes were small.

Our findings have several clinical implications. First, AI models should be evaluated within the target patient population's language context, as English-only assessments may not reflect multilingual performance. Second, accuracy alone is insufficient for emergency scenarios; completeness and emergency referral accuracy must also be considered. Third, since all models exceeded recommended readability thresholds, AI-generated materials should be reviewed and simplified before clinical use. Finally, the observed variation in emergency referral performance across models highlights the importance of model- and language-specific safety evaluation before the use of LLMs in safety-critical patient education.

This study has several limitations. First, the question set was developed by the research team and reviewed for face validity by two senior ENT specialists; however, it was not derived directly from patient-reported

questions or subjected to formal psychometric validation. Second, each question was submitted only once to each model. Therefore, the potential variability of LLM-generated responses across repeated queries was not assessed. Third, the findings reflect the specific model versions evaluated during the study period, and subsequent model updates may limit the temporal generalizability of the results. Fourth, the analysis was limited to Turkish and English, and the findings cannot necessarily be generalized to other languages or linguistic contexts. Fifth, no validated Turkish readability instrument was available, precluding a direct comparison of readability between the Turkish and English responses. Sixth, evaluation by only two ENT specialists may have limited the precision of inter-rater reliability estimates and introduced some degree of subjectivity into QAMAI scoring. Finally, this study evaluated AI-generated responses under standardized experimental conditions rather than their use by actual patients; therefore, the effects of these responses on patient comprehension, decision-making, satisfaction, adherence, and clinical outcomes were not assessed. Prospective studies incorporating patient-generated questions, repeated model outputs, additional languages, larger expert panels, and patient-centered outcome measures are warranted.

## CONCLUSION

This study compares four major language models for post-septorhinoplasty patient education across two languages and question categories. Significant between-model differences were observed for Turkish routine postoperative questions (Kendall's  $W=0.775$ ), whereas the corresponding overall comparison in English did not reach statistical significance. Emergency referral accuracy was high for several models, although model-specific omissions were observed, and differences in individual QAMAI dimensions varied by language and question category. All models exceeded recommended readability thresholds, underscoring the need for simplification before clinical use. AI chatbots may support postoperative patient education but cannot replace professional medical evaluation in safety-critical scenarios.

Future studies should extend comparisons to other otolaryngology procedures and prospectively

assess the impact on clinical outcomes and patient satisfaction.

## Ethical approval

In this study, since no patient or experimental animal data were used, ethical committee approval was not required.

## Author contribution

Concept: ZK; Design: ZK, GSU; Data Collection or Processing: GSU, AÖ, ZK, YU; Analysis or Interpretation: GSU, ZK, YU; Literature Search: ZK; Writing: ZK, AÖ. All authors reviewed the results and approved the final version of the article.

## Source of funding

The authors declare the study received no funding.

## Conflict of interest

The authors declare that there is no conflict of interest.

## Use of artificial intelligence

During the preparation and revision of this manuscript between June and September 2026, ChatGPT (OpenAI, San Francisco, CA, USA) was used solely for English language editing to improve grammar, clarity, and readability. All scientific content, data interpretation, and conclusions were determined by the authors. The authors reviewed and approved the final manuscript and take full responsibility for its content.

## REFERENCES

1. Alsudays AM, Almanea KA, Alhajlah AA, Alroqi A. Artificial intelligence in postoperative sinus care: a comparative study of ChatGPT-4, Google Gemini, and DeepSeek in patient education and support. *J Craniofac Surg.* 2026; 37(3-4): 490-3. [\[Crossref\]](#)
2. Aliyeva A, Alaskarov E, Sari E. Postoperative management of tympanoplasty with ChatGPT-4.0. *J Int Adv Otol.* 2025; 21(1): 1-6. [\[Crossref\]](#)
3. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and Google Search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg.* 2023; 149(6): 556-8. [\[Crossref\]](#)

4. Crowson MG, Nwosu OI. The integration and impact of artificial intelligence in otolaryngology-head and neck surgery: navigating the last mile. *Otolaryngol Clin North Am*. 2024; 57(5): 887-95. [\[Crossref\]](#)
5. Alghamdi AM, Alzahrani RS, Alghamdi FS, Albilasi TM, Doubi AO. The role of artificial intelligence in assisting patient education post-rhinoplasty. *International Journal of Medicine in Developing Countries*. 2024; 8(10): 2794-8. [\[Crossref\]](#)
6. Ibas M, Dursun S, Paksoy M, Ocal R, Karatas E. Accuracy and safety of ChatGPT-4o responses in rhinoplasty postoperative counseling: a panel-based study. *Acta Otolaryngol*. 2025; 145(9): 851-6. [\[Crossref\]](#)
7. Capelleras M, Soto-Galindo GA, Cruellas M, Apaydin F. ChatGPT and rhinoplasty recovery: an exploration of AI's role in postoperative guidance. *Facial Plast Surg*. 2024; 40(5): 628-31. [\[Crossref\]](#)
8. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthetic Plast Surg*. 2023; 47(5): 1985-93. [\[Crossref\]](#)
9. Sallam M, Alasfoor IM, Khalid SW, et al. Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English. *Narra J*. 2025; 5(1): e2371. [\[Crossref\]](#)
10. Sari F, Çelik Z, Mirza Y. ChatGPT-4 vs. DeepSeek-V3: a comparative study of response quality, reliability, usefulness, and readability for exercise and rehabilitation strategies in patients with ankylosing spondylitis. *Clin Rheumatol*. 2026; 45(1): 187-95. [\[Crossref\]](#)
11. Mayo-Yáñez M, Rameau A, Vaira LA, et al. Patient perceptions of artificial intelligence in otolaryngology-head and neck surgery: an international study. *Ear Nose Throat J*. 2025; undefined: 1455613251351774. [\[Crossref\]](#)
12. Meyer MKR, Kandathil CK, Davis SJ, et al. Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. *Aesthetic Plast Surg*. 2025; 49(7): 1868-73. [\[Crossref\]](#)
13. Palmer WJ, Michlin D, Estephan L, et al. Utility and safety of artificial intelligence for patient-initiated contact after functional rhinoplasty. *OTO Open*. 2025; 9(4): e70170. [\[Crossref\]](#)
14. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). *BMJ*. 2025; 390: e083305. [\[Crossref\]](#)
15. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007; 370(9596): 1453-7. [\[Crossref\]](#)
16. Vaira LA, Lechien JR, Abbate V, et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur Arch Otorhinolaryngol*. 2024; 281(11): 6123-31. [\[Crossref\]](#)
17. McLaughlin GH. SMOG grading - a new readability formula. *The Journal of Reading*. 1969; 12(8): 639-46.
18. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Educ Behav*. 2006; 33(3): 352-73. [\[Crossref\]](#)
19. Dhar S, Kothari D, Vasquez M, et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. *Int J Pediatr Otorhinolaryngol*. 2024; 179: 111901. [\[Crossref\]](#)
20. Chen FJ, Nightingale J, You WS, Anderson D, Morrissey D. Assessment of ChatGPT vs. Bard vs. guidelines in the artificial intelligence (AI) preclinical management of otorhinolaryngological (ENT) emergencies. *Aust J Otolaryngol*. 2024; 7: 19. [\[Crossref\]](#)

# AI-based diagnosis of hallux valgus: evaluation of measurements derived from foot X-ray images

Mahmut Kurtboğan<sup>1</sup>, Gülçin Ray<sup>2</sup>, Abdullah Ray<sup>2</sup>, İbrahim Kürtül<sup>2</sup>, Mihriban Karatop<sup>3</sup>

<sup>1</sup>Department of Orthopedics and Traumatology, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

<sup>2</sup>Department of Anatomy, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

<sup>3</sup>Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

**Cite as:** Kurtboğan M, Ray G, Ray A, Kürtül İ, Karatop M. AI-based diagnosis of hallux valgus: evaluation of measurements derived from foot X-ray images. Northwestern Med J. 2026;6(4):330-342.

## ABSTRACT

**Objective:** This study aimed to diagnose hallux valgus (HV) using machine learning algorithms (MLAs) applied to angular measurements obtained from foot X-ray images.

**Materials and Methods:** The study was a retrospective observational investigation. The sample comprised 50 patients diagnosed with HV and 50 controls. The intermetatarsal angle (IMA), interphalangeal angle (IPA), talometatarsal angle (TMTA), metatarsal tangent angle (MTTA), metatarsus adductus angle (MTAA), and metatarsal parabola angle (MTPA) were measured on the anteroposterior X-ray image, and calcaneal inclination angle (CIA) and the lateral talocalcaneal angle (LTCA) were measured on the lateral X-ray image. The following algorithms were used as machine learning algorithms (MLAs): Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), LightGBM (LGBM), AdaBoost (ADA), Extra Trees (ET), K-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA). Statistical analyses were performed using the open-source RStudio software.

**Results:** The diagnostic accuracy rates of the algorithms (CV) were 90% for XGB, ET, RF, LGBM, and LDA; 85% for ADA, LR, SVM, and GB; and 80% for KNN. The mean  $\pm$  SD values of the variables for the control and patient groups, respectively, were calculated as  $21.71 \pm 1.83$  and  $22.52 \pm 3.32$  for IMA;  $17.86 \pm 1.79$  and  $18.35 \pm 2.28$  for MTAA;  $143.43 \pm 9.20$  and  $143.25 \pm 9.14$  for MTPA; and  $50.96 \pm 5.27$  and  $49.42 \pm 6.98$  for LTCA. The median (min-max) values were: for IPA, 36.48 (25.73–45.82) and 45.41 (29.60–76.90); for TMTA, 7.60 (5.91–11.45) and 12.54 (6.62–19.50); for MTTA, 7.93 (1.07–13.20) and 7.13 (2.60–12.07); and for CIA, 28.52 (22.31–37.65) and 27.19 (10.85–47.46).

**Conclusions:** Angular measurements derived from X-ray images enabled MLAs to diagnose HV with an accuracy of up to 90%. This finding demonstrates the applicability of artificial intelligence for diagnostic purposes in imaging-based applications.

**Keywords:** hallux valgus deformity, machine learning algorithms, foot anatomy, x-ray images, intermetatarsal angle

**Corresponding author:** Mahmut Kurtboğan **E-mail:** kurtbogann@gmail.com

**Received:** 25.06.2026 **Accepted:** 27.07.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

## INTRODUCTION

Hallux valgus (HV), one of the most common foot deformities, develops in general with the progressive subluxation of the metatarsophalangeal joint at the pollex, medial deviation of the metatarsal bone, and medial displacement of the pollex (1). A bony prominence developing on the medial aspect of the first metatarsal is the characteristic appearance of the deformity (1,2). A study has reported that 23% of the adult population and 36% of the geriatric population are diagnosed with HV (2). The incidence of the disease is higher in women and increases with age (3). The biomechanical cause of the pathology can be described as the deviation of the bones and joints in the anteromedial region of the foot from their normal axes, leading to impaired muscular stabilization and ligamentous laxity (3,4). This biomechanical disturbance is accompanied by pain, impairment in physical and daily-life activities, and an increased risk of falls due to disturbed balance (4). Because the severity of deformity progresses over time, early diagnosis and intervention are essential (5). An increase in deformity severity corresponds then to an increase in the HV (5,6). Although several methods such as ultrasound, footprint analysis, and gait and plantar pressure analysis are used to diagnose the deformity, the most accurate and widely used method is radiographic imaging (7). For the diagnosis, orthopedic surgeons rely on the HV and intermetatarsal angle IMA, which can be measured on weight-bearing anteroposterior radiographs of the foot (5–7). On the basis of these angles, the deformity is classified as mild, moderate, or severe (8). The HV is measured between the first metatarsal and the proximal phalanx of the pollex, and the IMA is measured between the parallel axes of the first and second metatarsals (7,8). As a surgical intervention in particular is planned, the severity of the deformity, degenerative changes in the joint, and the size and shape of the metatarsal bone are also evaluated (9). After surgical correction, the recurrence rate of HV is considerably high (10). Risk factors such as a high preoperative HV, the presence and degree of subluxation of the sesamoid bone, hereditary predisposition, body mass index, and the patient's preference for narrow, high-heeled footwear, contribute to recurrence (9–11). The rapid progression of disease severity, high recurrence rate despite

surgical correction, and resulting deterioration in quality of life require an easily applicable tool for early diagnosis and follow-up.

A review of the literature reveals that machine learning algorithms (MLAs)—a class of artificial intelligence (AI) applications providing description, classification, and diagnostic functions—have been used on radiographic images for this purpose (12–16). Recently popular AI applications in healthcare perform analysis and classification on medical images, develop predictive models, and predict prognosis with high accuracy (12,13). Algorithms allow large datasets to be processed efficiently, capture correlations among variables, and consequently provide highly accurate predictions (17,18). Nevertheless, further research is needed to enable AI applications to differentiate HV from concomitant deformities such as osteoarthritis, bunionette deformity, or pes planus. With those in mind, the study aimed to demonstrate the accuracy of HV diagnosis by MLAs using angular variables derived from foot X-ray images.

## MATERIALS AND METHODS

This retrospective study was initiated following decision No. 2026/217 of the Bolu Abant İzzet Baysal University Non-Interventional Clinical Research Ethics Committee. The study included 50 patients (39 women, 11 men) who presented to the Bolu Abant İzzet Baysal University Department of Orthopedics and Traumatology between January 2020 and February 2026 and were diagnosed with hallux valgus, and 50 controls (29 women, 21 men) who had no history of disease involving the foot. For both groups, individuals with a history of surgical intervention, fracture, trauma, or congenital anomaly in the foot or ankle region, as well as those with unclear or inadequate radiographic images, were excluded. The images of the sample group were retrieved retrospectively from the hospital's PACS archive, transferred to the Radiant DICOM Viewer software, and used to obtain the predetermined angular measurements. An expert orthopedic surgeon classified participants as patients or controls based on HV measurements performed on the images. The intermetatarsal angle (IMA), interphalangeal angle (IPA), talometatarsal angle (TMTA), metatarsal tangent angle (MTTA), metatarsus adductus angle (MTAA),

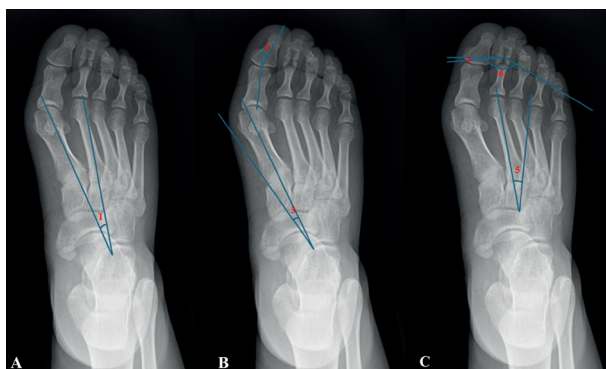


**Figure 1.** Reference points for angular measurements on the anteroposterior foot radiograph of a 32-year-old male participant. A. intermetatarsal angle; B. interphalangeal angle, talometatarsal angle; C. metatarsal tangent angle, metatarsus adductus angle, metatarsal parabola angle.

and metatarsal parabola angle (MTPA) variables (Figure 1) were measured on the anteroposterior foot radiographs, and the inclination angle (CIA) and the lateral talocalcaneal angle (LTCA) variables were measured on the lateral foot radiographs (Figure 2). All radiographs, both anteroposterior and lateral, were obtained under weight-bearing conditions, with patients standing on the affected foot at the time of image acquisition. To eliminate inter-observer error, each measurement was performed twice by the same investigator, and the mean values were recorded in Excel format.

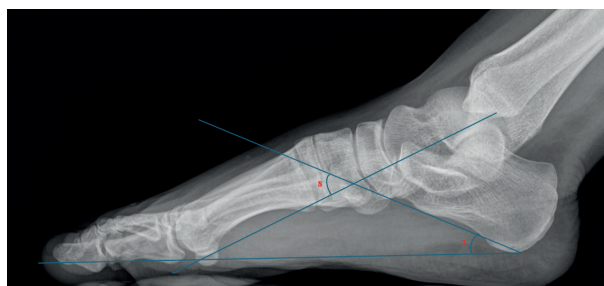


**Figure 2.** Reference points for angular measurements on the lateral foot radiograph of a 32-year-old male participant. Calcaneal inclination angle; lateral talocalcaneal angle.

Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), LightGBM (LGBM), AdaBoost (ADA), Extra Trees (ET), K-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA) were used as MLAs, and their hyperparameters are shown in Table 1.

### Training/test strategy

The dataset was split into training (80%) and test (20%) sets. To preserve class proportions in both subsets, stratified splitting (random state = 42) was applied. For evaluation, 10-fold cross-validation (CV) was used to partition the dataset and measure classification performance. All variables were standardized prior to model training using z-score normalization (StandardScaler:  $\mu = 0$ ,  $\sigma = 1$ ).

**Table 1.** Best hyperparameters for the MLAs.

| Algorithm | Best Hyperparameters (after GridSearchCV)                                                                                       |
|-----------|---------------------------------------------------------------------------------------------------------------------------------|
| LR        | {'clf__C': 0.01, 'clf__penalty': 'l2', 'clf__solver': 'lbfgs'}                                                                  |
| SVM       | {'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel': 'rbf'}                                                                    |
| RF        | {'clf__max_depth': None, 'clf__max_features': 'sqrt', 'clf__min_samples_split': 5, 'clf__n_estimators': 100}                    |
| GB        | {'clf__learning_rate': 0.05, 'clf__max_depth': 3, 'clf__n_estimators': 100, 'clf__subsample': 1.0}                              |
| XGB       | {'clf__colsample_bytree': 1.0, 'clf__learning_rate': 0.2, 'clf__max_depth': 2, 'clf__n_estimators': 100, 'clf__subsample': 1.0} |
| LGBM      | {'clf__learning_rate': 0.05, 'clf__max_depth': -1, 'clf__n_estimators': 100, 'clf__num_leaves': 15}                             |
| ADA       | {'clf__learning_rate': 0.5, 'clf__n_estimators': 50}                                                                            |
| ET        | {'clf__max_depth': None, 'clf__min_samples_split': 2, 'clf__n_estimators': 100}                                                 |
| KNN       | {'clf__metric': 'euclidean', 'clf__n_neighbors': 3, 'clf__weights': 'uniform'}                                                  |
| LDA       | {'clf__solver': 'svd'}                                                                                                          |

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

## Statistical analyses

Analyses were performed using the open-source RStudio software. A One-way analysis of variance was applied to all variables, and residual normality was assessed using the Anderson–Darling test. Logarithmic and square-root transformations were applied to variables that did not conform to a normal distribution. For pairwise comparisons, parametric variables were analyzed with the Independent Samples t-test, and nonparametric variables were analyzed with the Mann–Whitney U test. Descriptive statistics were calculated for all variables.

## RESULTS

The performance of the ten MLAs in distinguishing the HV and control groups is presented in an integrated manner in Table 2 using cross-validation accuracy (CV Accuracy, mean  $\pm$  SD), test accuracy, area under the curve (AUC), F1 score, sensitivity (recall), specificity, precision, balanced accuracy, and the Matthews correlation coefficient (MCC). In the CV analysis, mean accuracy values ranged from  $0.8500 \pm 0.1561$  to  $0.9125 \pm 0.0800$ . The highest CV accuracy was obtained for GB at  $0.9125 \pm 0.0800$ , followed by ADA at  $0.9000 \pm 0.0935$  and SVM at  $0.9000 \pm 0.1225$ . ET, RF, and LGBM yielded a mean CV accuracy of 0.8875, with standard deviations of  $\pm 0.1038$ ,  $\pm 0.1179$ , and  $\pm 0.0875$ , respectively. Logistic Regression (LR) was  $0.8875 \pm 0.0875$ , XGB was  $0.8750 \pm 0.1118$ , LDA was  $0.8750 \pm 0.1250$ , and KNN was  $0.8500 \pm 0.1561$ . Based on the examined standard deviation values, the lowest variation was observed for GB (0.0800) and the highest for KNN (0.1561), indicating that GB exhibited the most stable performance across CV folds, whereas KNN produced the most variable results.

With respect to test set performance, the models clustered into two distinct groups. XGB, ET, RF, LGBM, and LDA formed the top tier with a test accuracy of 0.9000, while ADA, LR, SVM, and GB achieved 0.8500, and KNN reached 0.8000. In terms of AUC, the highest discriminative capacity was obtained for XGB and ET at 0.9400, followed by RF, LGBM, and LDA at 0.9300; ADA

at 0.9050; LR at 0.9000; SVM at 0.8900; GB at 0.8500; and KNN at 0.8150. In the F1 score analysis, XGB, ET, RF, LGBM, and LDA shared the highest value of 0.9000, whereas ADA, LR, SVM, and GB scored 0.8421, while KNN scored 0.7778.

When sensitivity and specificity were considered together, XGB, ET, RF, LGBM, and LDA displayed a balanced profile, with a value of 0.9000 for both metrics. ADA, LR, SVM, and GB yielded sensitivity of 0.8000 and specificity of 0.9000; KNN yielded sensitivity of 0.7000 and specificity of 0.9000. Precision values were: 0.9000 for XGB, ET, RF, LGBM, and LDA; 0.8889 for the four mid-tier models (ADA, LR, SVM, and GB); and 0.8750 for KNN. Balanced accuracy, reflecting equally weighted performance across classes, was 0.9000 for XGB, ET, RF, LGBM, and LDA; 0.8500 for ADA, LR, SVM, and GB; and 0.8000 for KNN. With respect to the Matthews correlation coefficient (MCC) which evaluates classification quality in a manner sensitive to class imbalance, the highest MCC of 0.8000 was observed for XGB, ET, RF, LGBM, and LDA; ADA, LR, and SVM yielded MCCs of 0.7035; GB yielded 0.6124; and KNN yielded 0.6118. Likewise, the XGBoost algorithm (test accuracy 0.9000; AUC 0.9400; F1 score 0.9000; sensitivity 0.9000; specificity 0.9000; precision 0.9000; balanced accuracy 0.9000; MCC 0.8000) emerged as the highest-performing model in Table 2, providing the most consistent diagnostic performance across all metrics for HV diagnosis. By contrast, the KNN algorithm with a test accuracy of 0.8000, an AUC of 0.8150, an F1 score of 0.7778, and an MCC of 0.6118 delivered the lowest performance across all evaluated parameters and was identified as the model with the most limited diagnostic utility for clinical decision support.

To evaluate models' sample-independent generalization capacity and stability across folds, the ten machine learning algorithms were subjected to 10-fold CV. The accuracy values obtained for each fold (Folds 1–10) are presented in Table 2. For the LR algorithm, fold accuracy values ranged from 0.7500 to 1.0000 (Fold 1: 0.8750; Fold 2: 1.0000; Fold 3: 1.0000; Fold 4: 0.8750; Fold 5: 0.8750; Fold 6: 0.8750;

**Table 2.** MLA classification results for Hallux Valgus vs Control groups.

| Algorithm | CV Accuracy (Mean $\pm$ SD) | Test Accuracy | AUC (ROC) | F1-Score | Sensitivity (Recall) | Specificity | Precision | Balanced Accuracy | MCC    |
|-----------|-----------------------------|---------------|-----------|----------|----------------------|-------------|-----------|-------------------|--------|
| XGB       | 0.8750 $\pm$ 0.1118         | 0.9000        | 0.9400    | 0.9000   | 0.9000               | 0.9000      | 0.9000    | 0.9000            | 0.8000 |
| ET        | 0.8875 $\pm$ 0.1038         | 0.9000        | 0.9400    | 0.9000   | 0.9000               | 0.9000      | 0.9000    | 0.9000            | 0.8000 |
| RF        | 0.8875 $\pm$ 0.1179         | 0.9000        | 0.9300    | 0.9000   | 0.9000               | 0.9000      | 0.9000    | 0.9000            | 0.8000 |
| LGBM      | 0.8875 $\pm$ 0.0875         | 0.9000        | 0.9300    | 0.9000   | 0.9000               | 0.9000      | 0.9000    | 0.9000            | 0.8000 |
| LDA       | 0.8750 $\pm$ 0.1250         | 0.9000        | 0.9300    | 0.9000   | 0.9000               | 0.9000      | 0.9000    | 0.9000            | 0.8000 |
| ADA       | 0.9000 $\pm$ 0.0935         | 0.8500        | 0.9050    | 0.8421   | 0.8000               | 0.9000      | 0.8889    | 0.8500            | 0.7035 |
| LR        | 0.8875 $\pm$ 0.0875         | 0.8500        | 0.9000    | 0.8421   | 0.8000               | 0.9000      | 0.8889    | 0.8500            | 0.7035 |
| SVM       | 0.9000 $\pm$ 0.1225         | 0.8500        | 0.8900    | 0.8421   | 0.8000               | 0.9000      | 0.8889    | 0.8500            | 0.7035 |
| GB        | 0.9125 $\pm$ 0.0800         | 0.8500        | 0.8500    | 0.8421   | 0.8000               | 0.9000      | 0.8889    | 0.8500            | 0.7035 |
| KNN       | 0.8500 $\pm$ 0.1561         | 0.8000        | 0.8150    | 0.7778   | 0.7000               | 0.9000      | 0.8750    | 0.8000            | 0.6124 |

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

Fold 7: 0.7500; Fold 8: 0.7500; Fold 9: 1.0000; Fold 10: 0.8750), and the overall mean was calculated as  $0.8875 \pm 0.0875$ . For the SVM, fold values ranged from 0.6250 to 1.0000 (Fold 1: 1.0000; Fold 2: 1.0000; Fold 3: 1.0000; Fold 4: 0.8750; Fold 5: 1.0000; Fold 6: 0.8750; Fold 7: 0.6250; Fold 8: 0.7500; Fold 9: 1.0000; Fold 10: 0.8750), yielding a mean accuracy of  $0.9000 \pm 0.1225$ . The RF algorithm exhibited its lowest starting value in the first fold (0.6250). The subsequent folds were recorded as Fold 2: 1.0000; Fold 3: 1.0000; Fold 4: 1.0000; Fold 5: 0.8750; Fold 6: 0.8750; Fold 7: 0.8750; Fold 8: 0.7500; Fold 9: 1.0000; Fold 10: 0.8750, producing a mean of  $0.8875 \pm 0.1179$ . The GB model—with fold values of 0.8750, 1.0000, 1.0000, 0.8750, 1.0000, 0.8750, 0.8750, 0.8750, and 0.8750—exhibited the most stable performance and reached a mean accuracy of  $0.9125 \pm 0.0800$ . The fold distribution of XGB, highlighted in yellow in Table 2, contained the values 0.6250 (Fold 1), 1.0000 (Fold 2), 1.0000 (Fold 3), 1.0000 (Fold 4), 0.8750 (Fold 5), 0.8750 (Fold 6), 0.7500 (Fold 7), 0.8750 (Fold 8), 0.8750 (Fold 9), and 0.8750 (Fold 10). The mean CV accuracy of the model was calculated as  $0.8750 \pm 0.1118$ . For the LGBM, fold values ranged from 0.7500 to 1.0000 (0.7500; 1.0000; 1.0000; 0.8750; 0.8750; 0.8750; 0.8750; 0.8750; 1.0000; 0.8750), with a mean accuracy of  $0.8875 \pm 0.0875$  (Figure 3).

For the ADA, fold accuracies of 0.7500, 1.0000, 1.0000, 0.8750, 1.0000, 0.8750, 0.8750, 0.8750, 1.0000, and 0.8750 were obtained, with a mean of  $0.9000 \pm 0.0935$ . The ET produced fold values of 0.8750, 1.0000, 1.0000, 0.8750, 0.8750, 0.8750, 0.8750, 0.8750, 1.0000, and 0.7500, yielding a mean of  $0.8875 \pm 0.1038$ . Likewise, for the KNN, fold values were distributed over a notably wide range, from 0.5000 to 1.0000 (1.0000; 1.0000; 1.0000; 0.7500; 0.7500; 0.7500; 0.7500; 0.5000; 0.8750); the model reached a mean accuracy of  $0.8500 \pm 0.1561$  and exhibited the highest standard deviation among all evaluated algorithms. The LDA algorithm produced fold values of 0.6250, 1.0000, 1.0000, 0.8750, 0.8750, 0.8750, 0.7500, 1.0000, 1.0000, and 0.7500, with a mean performance of  $0.8750 \pm 0.1250$ . Examination of the inter-fold distribution pattern revealed that all ten algorithms achieved an accuracy of 1.0000 in at least one fold; in contrast, the lowest fold accuracy, 0.5000, was observed for KNN in Fold 9. When the standard deviations were ranked from lowest to highest, the order was as follows: GB (0.0800), LGBM (0.0875), LR (0.0875), ADA (0.0935), ET (0.1038), XGB (0.1118), RF (0.1179), SVM (0.1225), LDA (0.1250), and KNN (0.1561) (Table 3 and Figure 4).

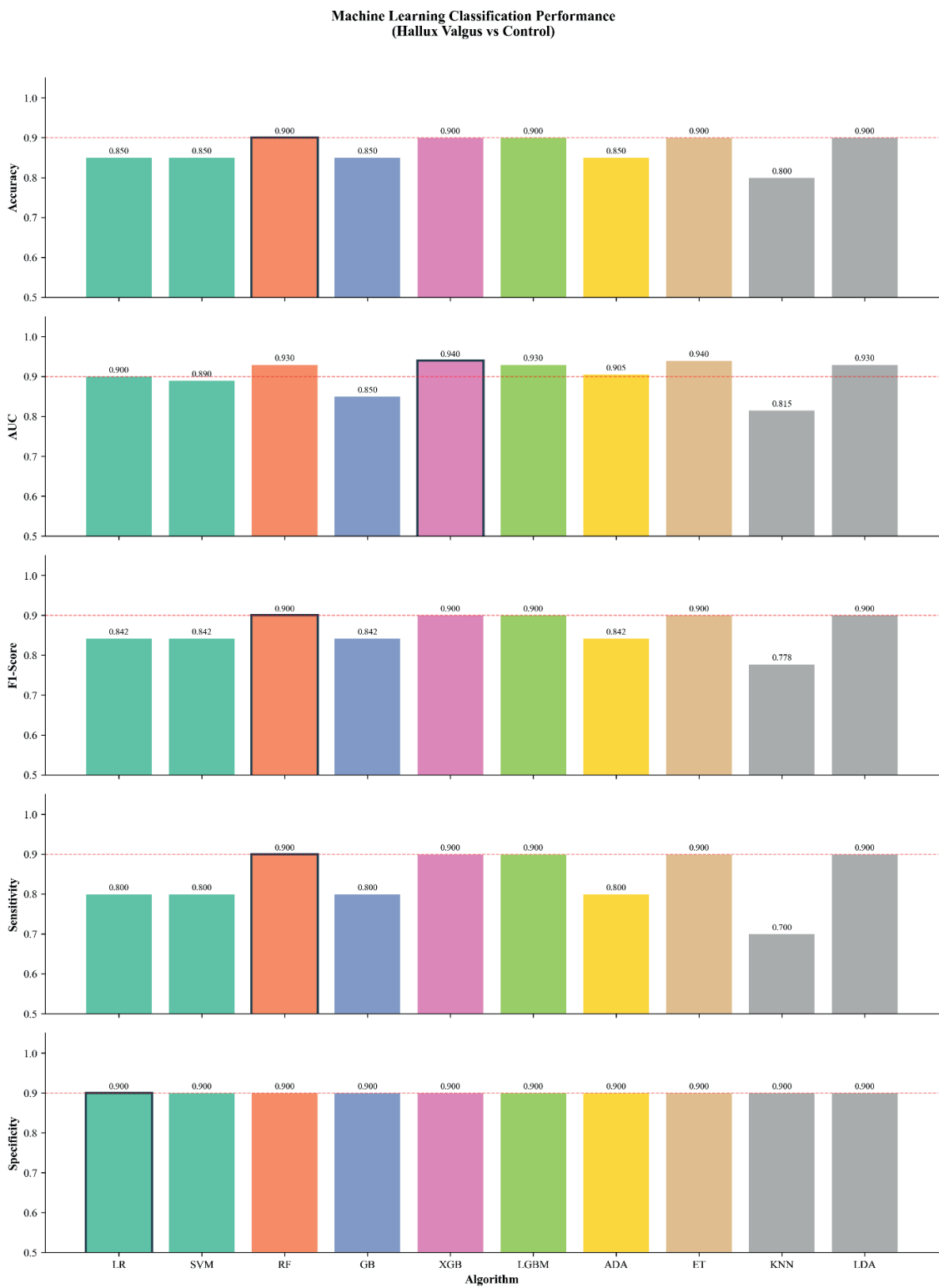
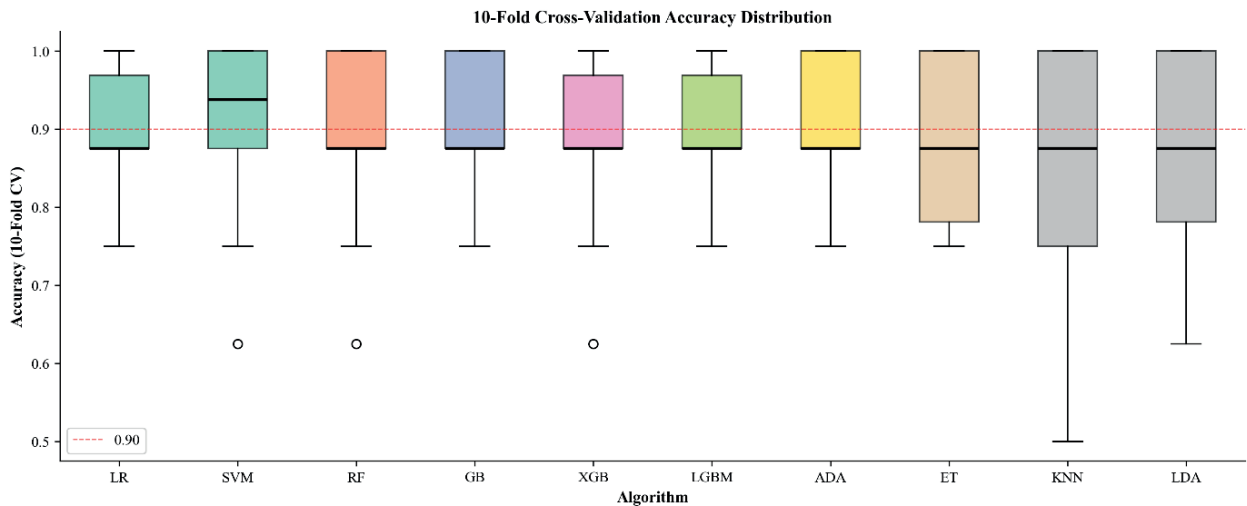


**Figure 3.** Machine learning algorithms classification performance.

**Table 3.** 10-fold CV results.

| Algorithm | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Mean $\pm$ SD       |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|---------------------|
| LR        | 0.8750 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.8750 | 0.7500 | 0.7500 | 1.0000 | 0.8750  | 0.8875 $\pm$ 0.0875 |
| SVM       | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 1.0000 | 0.8750 | 0.6250 | 0.7500 | 1.0000 | 0.8750  | 0.9000 $\pm$ 0.1225 |
| RF        | 0.6250 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.8750 | 0.7500 | 1.0000 | 0.8750  | 0.8875 $\pm$ 0.1179 |
| GB        | 0.8750 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 1.0000 | 0.7500 | 0.8750 | 0.8750 | 0.8750  | 0.9125 $\pm$ 0.0800 |
| XGB       | 0.6250 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.7500 | 0.8750 | 0.8750 | 0.8750  | 0.8750 $\pm$ 0.1118 |
| LGBM      | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.8750 | 0.7500 | 0.8750 | 0.8750  | 0.8875 $\pm$ 0.0875 |
| ADA       | 0.7500 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.8750 | 0.7500 | 1.0000 | 0.8750  | 0.9000 $\pm$ 0.0935 |
| ET        | 0.8750 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.7500 | 0.7500 | 1.0000 | 0.7500  | 0.8875 $\pm$ 0.1038 |
| KNN       | 0.8750 | 1.0000 | 1.0000 | 1.0000 | 0.7500 | 0.7500 | 0.7500 | 0.5000 | 1.0000 | 0.8750  | 0.8500 $\pm$ 0.1561 |
| LDA       | 0.6250 | 1.0000 | 1.0000 | 1.0000 | 0.8750 | 0.8750 | 0.8750 | 0.7500 | 1.0000 | 0.7500  | 0.8750 $\pm$ 0.1250 |

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

**Figure 4.** Accuracy distribution of the 10-fold cross-validation.

The relationship between sensitivity (true positive rate) and 1 – specificity (false positive rate) for the ten classifiers evaluated to discriminate between the HV and control groups is visualized in Figure 2 using Receiver Operating Characteristic (ROC) curves. The area under the curve AUC values indicated that the models had discriminative capacity significantly greater than the random-guess line (AUC = 0.50). According to the ROC analysis, the highest discriminative power was obtained by XGB and ET, with an AUC of 0.940; their curves were closest to the upper-left corner of the coordinate system, achieving an optimal balance between sensitivity and specificity. They were

followed by RF, LGBM, and LDA, each with an AUC of 0.930; these three algorithms exhibited closely aligned performance. The ADA (0.905), LR (0.900), and SVM (0.890) were ranked among the algorithms with high discriminative capacity. By contrast, GB displayed lower performance, with an AUC of 0.850, while KNN provided the weakest discriminative power among all models, with an AUC of 0.815. The ROC curve for KNN was noticeably closer to the random-guess line than the ROC curves of the other models. Overall, the AUC values ranged from 0.815 to 0.940, and nine models exceeded the 0.850 threshold, indicating clinically acceptable diagnostic performance (Figure 5).

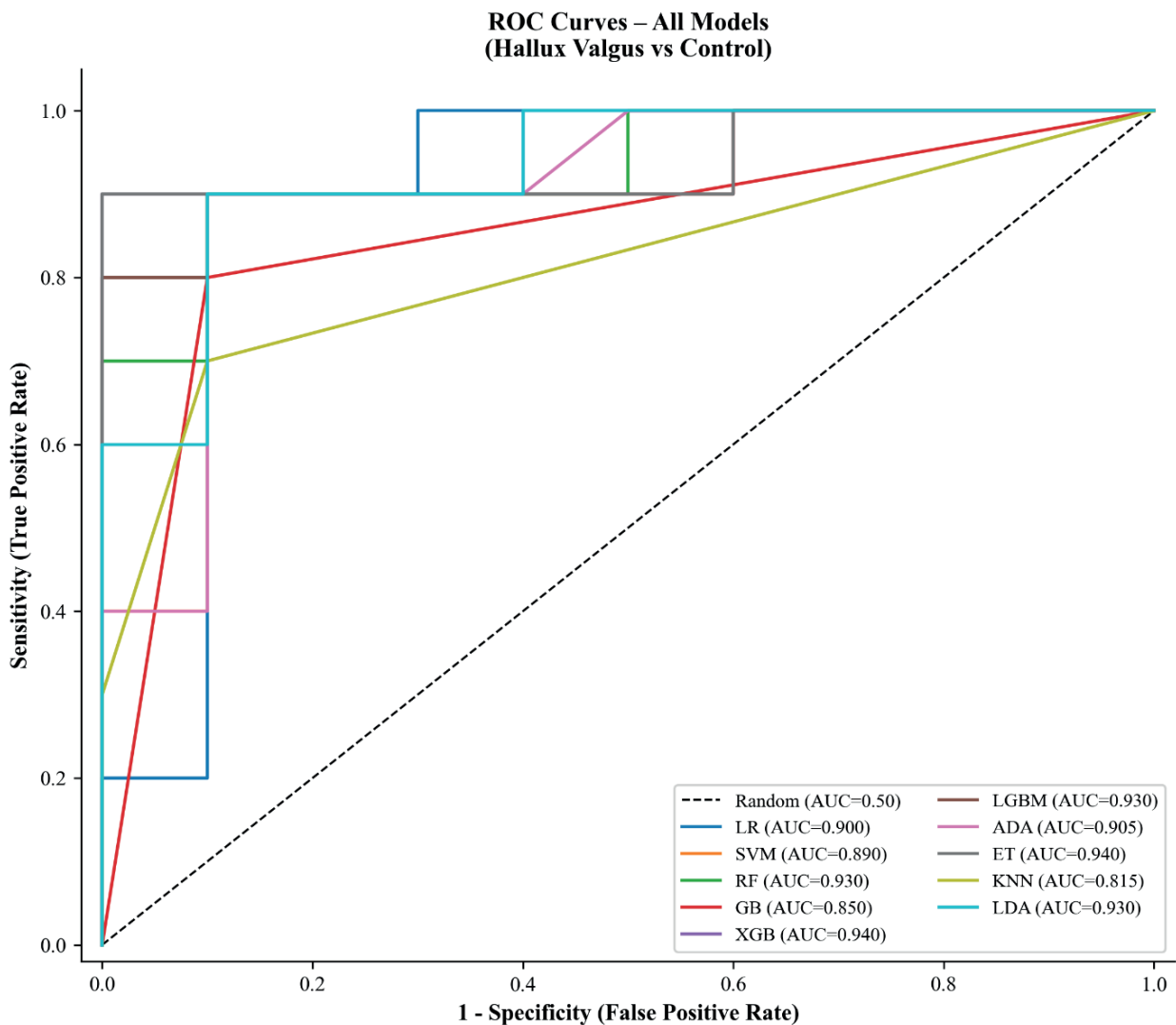


**Figure 5.** The ROC curve graphic of all the models.

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

The relationship between the actual and predicted labels for the HV and control (CN) classes was presented in Figure 6 using normalized confusion matrices for each algorithm. The test set comprised ten cases for each class ( $n = 10$  HV;  $n = 10$  CN; total  $n = 20$ ), and the proportions in each cell were calculated relative to the number of cases in the corresponding class. Regarding recognition of the control class, all algorithms exhibited a uniform performance pattern: across all ten models, 90% ( $n = 9$ ) of the actual control cases were correctly classified as CN, while the

remaining 10% ( $n = 1$ ) were misclassified as HV (false positive). However, marked differences among the models were observed in recognizing the HV class. RF, XGB, LGBM, ET, and LDA algorithms correctly classified 90% ( $n = 9$ ) of the actual HV cases as HV and yielded false negatives in only 10% ( $n = 1$ ); the accuracy of these five models was 90.0%. The LR, SVM, GB, and ADA algorithms achieved a correct identification rate of 80% ( $n = 8$ ) for the HV class; the remaining 20% ( $n = 2$ ) were misclassified as controls (false negatives). The overall accuracy of these four models was 85.0%. The

KNN algorithm exhibited the highest false-negative rate in the HV class: only 70% ( $n = 7$ ) of the actual HV cases were correctly classified, while 30% ( $n = 3$ ) were misclassified as controls. Consequently, KNN's overall accuracy was 80.0%, the lowest among all models.

The learning curve, obtained to evaluate the generalization behavior of the XGB algorithm as a function of training-set size, is shown in Figure 7; it compares the training score with the CV score. The analysis was conducted in five consecutive stages, during which the training-set size was increased from approximately 48 to 80 cases. The 95% confidence interval for the CV score was shown as a shaded band around the curve. The training score remained constant at 1.000 across all training-set sizes examined, indicating that the model achieved a perfect fit on the training data. The CV score, by contrast, displayed a marked upward trend as training-set size increased. Beginning at 0.73 for the smallest training-set size ( $n \sim 48$ ), the CV score increased to 0.82 when the number of training samples reached 56, and to 0.89 when it reached 64. Expanding the training set to 72 cases brought the CV score to its peak of 0.93, at which point the gap between training and validation

scores was also minimized. When the training set was further enlarged to 80 cases, the CV score declined slightly to 0.90, while the confidence interval widened. The difference between the training and CV scores narrowed markedly as training set size increased: an initial separation of approximately 0.27 units decreased to 0.11 at 64 cases and to 0.07 at 72 cases.

Figure 8 shows the relative contributions of the radiographic angular variables used by the XGB algorithm to distinguish the HV and CN groups, presented in descending order as feature importance scores. Eight radiographic angular parameters were evaluated in the analysis, and a normalized importance score was calculated for each parameter. The variable with the highest contribution to the classification decision of the model was TMTA, with an importance score of 0.586; this value alone accounted for more than half of the total relative contribution of the eight variables. IPA was second to TMTA, with an importance score of 0.126, and made a marked contribution to the model's prediction process. The LTCA ranked third with an importance score of 0.106 and was highlighted in red in the figure. Variables providing moderate contributions to the model's prediction were

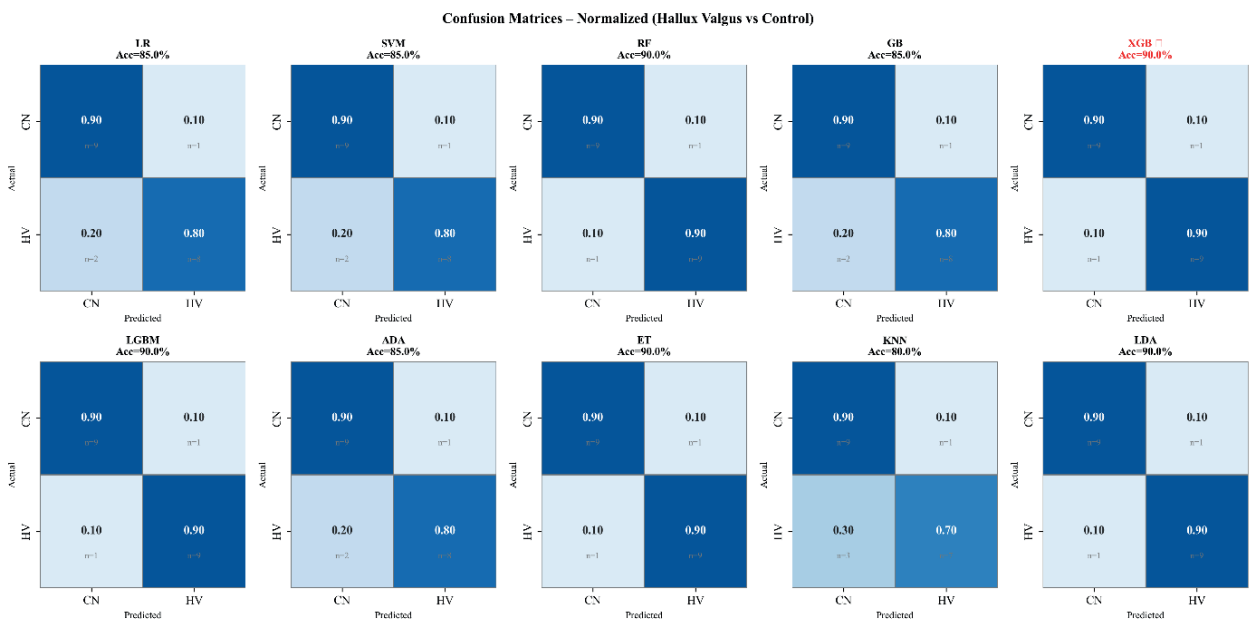


**Figure 6.** Confusion matrix tables of all the models.

ADA: AdaBoost; ET: Extra Trees; GB: Gradient Boosting; KNN: K-Nearest Neighbor; LDA: Linear Discriminant Analysis; LGBM: LightGBM; LR: Logistic Regression; RF: Random Forest; SVM: Support Vector Machine; XGB: XGBoost.

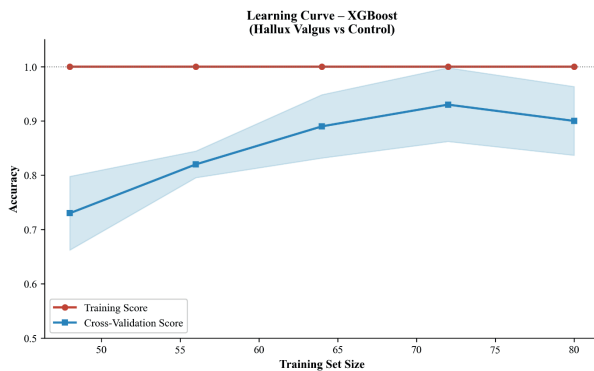


**Figure 7.** Training curve of the XGB algorithm.

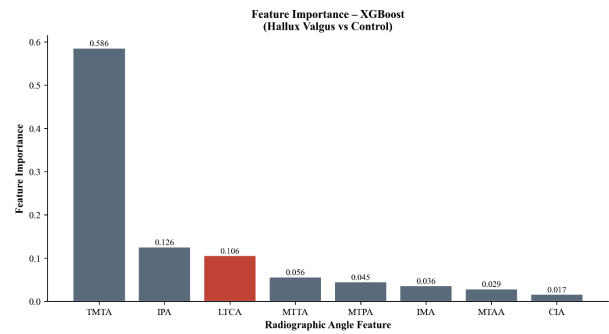


**Figure 8.** Feature importance ranking of the XGBoost algorithm.

MTTA (0.056) and MTPA (0.045). The parameters with relatively low contributions were IMA (0.036), MTAA (0.029), and CIA (0.017); these variables contributed least to the model's classification decision. Examination of the overall distribution showed that the cumulative importance of the top three variables (TMTA, IPA, LTCA) reached 0.818, indicating that the model's predictive power depended largely on these

three radiographic angles. By contrast, the cumulative contribution of the remaining five variables (MTTA, MTPA, IMA, MTAA, CIA) remained at 0.183.

The statistical results and descriptive values for all variables are shown in Table 4.

**Table 4.** Descriptive statistics and analytical results for the variables (n = 50 per group; CN vs HV).

| Variable | CN                           |                 |                   | HV                           |                 |                   | Test Score       |            |
|----------|------------------------------|-----------------|-------------------|------------------------------|-----------------|-------------------|------------------|------------|
|          | Mean ± SD / Median (min-max) | (Q1-Q3)         | (95% CI Low-High) | Mean ± SD / Median (min-max) | (Q1-Q3)         | (95% CI Low-High) | p-value          | Sig.       |
| IMA      | 21.71 ± 1.83                 | (20.50–23.07)   | (21.19–22.23)     | 22.52 ± 3.32                 | (20.53–24.48)   | (21.57–23.46)     | 0.1344           | ns         |
| IPA      | 36.48<br>(25.73–45.82)       | (33.40–39.63)   | (34.57–38.63)     | 45.41<br>(29.60–76.90)       | (40.09–50.09)   | (42.47–47.82)     | <b>&lt;0.001</b> | <b>***</b> |
| TMTA     | 7.60<br>(5.91–11.45)         | (7.23–8.57)     | (7.42–7.88)       | 12.54<br>(6.62–19.50)        | (10.72–14.49)   | (11.50–13.42)     | <b>&lt;0.001</b> | <b>***</b> |
| MTTA     | 7.93<br>(1.07–13.20)         | (7.46–8.89)     | (7.65–8.23)       | 7.13<br>(2.60–12.07)         | (5.84–8.30)     | (6.48–7.38)       | <b>0.0266</b>    | <b>*</b>   |
| MTAA     | 17.86 ± 1.79                 | (16.72–19.01)   | (17.35–18.37)     | 18.35 ± 2.28                 | (16.95–19.82)   | (17.70–19.00)     | 0.2351           | ns         |
| MTPA     | 143.43 ± 9.20                | (135.37–150.58) | (140.82–146.05)   | 143.25 ± 9.14                | (137.44–147.63) | (140.66–145.85)   | 0.9230           | ns         |
| CIA      | 28.52<br>(22.31–37.65)       | (26.97–30.33)   | (27.54–29.31)     | 27.19<br>(10.85–47.46)       | (25.21–29.23)   | (25.82–28.40)     | 0.2632           | ns         |
| LTCA     | 50.96 ± 5.27                 | (47.30–54.27)   | (49.46–52.46)     | 49.42 ± 6.98                 | (45.62–53.82)   | (47.43–51.40)     | 0.2153           | ns         |

Groups: CN = Control group; HV = Patient/Study group (n = 50 per group). Mean ± SD: mean ± standard deviation — measure of central tendency and dispersion of the data (parametric summary). Median (min-max): median value with the smallest and largest observed values — central tendency and range unaffected by extreme values. (Q1-Q3): interquartile range — boundaries of the central 50% of the data. (95% CI Low-High): 95% confidence interval for the mean — the interval within which the population mean lies with 95% probability. Test type: statistical approach selected according to the result of the normality test. Parametric → t-test (normal distribution); Non-parametric → Mann-Whitney U (non-normal distribution). All numerical values are rounded to two decimal places. p-value: statistical significance of the between-group difference based on one-way ANOVA. p < 0.05 was considered statistically significant. Significance codes: \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05; ns = not significant (p ≥ 0.05). Rows highlighted in red indicate statistically significant differences.

## DISCUSSION

HV is a common foot deformity that significantly impairs activities of daily living, particularly among older adults. Studies employing artificial intelligence tools to diagnose or classify the disease by extracting reference points or morphometric measurements from foot radiographs, the conventional diagnostic method, are steadily increasing. Similarly, ten MLAs were used in this study to classify patients and controls based on angular measurements obtained from radiographic images. Overall, the study yielded accuracy rates ranging from 80% to 90%.

In a study aimed at automatically determining HV and IMA measurements from radiographic images, a deep neural network model was developed. By means of this model, the angles, consistent with the present study, were measured with an accuracy comparable to that of expert observers (12). Another study aimed at performing automatic IPA measurement on radiographs via bone segmentation employed a U-Net neural network. The AI-assisted measurements were found to coincide with the manual measurements, and the measurement results were consistent with the patient group of the present study (1). In a further study aimed at developing a deep learning model that automatically annotates reference points and computes the HV and IMA values, labels were placed on the radiographic images. The inter-rater agreement performance in HV and IMA measurements was found to be satisfactory, and the IMA values were comparable to those of the present study (19). In a separate study designed to measure HV angular values automatically using deep learning models and to compare these measurements with those made by surgeons, a UNet++-based DL model was developed. The HV and IMA measurements were found to be highly accurate, and the IMA values were consistent with those of the control and patient groups in the present study (20). Likewise, in another study presenting HV measurements from radiographic images via landmark detection with a deep convolutional neural network, 18 points were manually annotated. Although angular differences ranging from 3° to 5.08° were observed between predicted values and the reference standard, the IMA and IPA values were consistent with those of the patient group in this study (16).

Another paper aiming to compare deep learning and manual measurements of HV and IMA on radiographs developed an automated measurement system with low error rates. The IMA results were similar to those of the patient group in the present study (21). Yet a study that evaluated radiological findings using deep learning for HV diagnosis compared algorithmic performance with clinicians' manual performance. The findings which also formed the basis of patient control discrimination in the present study, demonstrated that the model accuracy for HV was sufficient (22). In further research designed to automatically classify the position of the hallux sesamoid on radiographs using a deep neural network, a kappa value above 0.6 was achieved. The developed model provided a high accuracy rate consistent with the present study and the related literature (23).

In a report aimed at developing predictive models for HV recurrence, patients who had completed treatment were classified as recurrent or non-recurrent based on one-year follow-up results. MLA models applied to pre- and postoperative radiographs achieved an accuracy of 75.6%. The angular measurement results, consistent with those of the present study, enable the identification of patients at risk of HV recurrence (5). An MLA study evaluating the relationship between structural deformity and foot pain in HV patients using radiographic findings found higher angular values in the painful group. According to the findings of the present study, the MTTA, MTAA, MTPA, and IPA variables, which are essential for HV diagnosis, were also reported to perform well in classifying painful versus non-painful feet (24). In a study designed to identify the presence of HV using MLAs applied to photographic images of participants' feet, two schemes were developed and evaluated with a convolutional neural network model. Consistent with the present study, the study yielded accuracy rates ranging from 62% to 96% (4). Finally, a subsequent study that evaluated the effects of age, sex, body mass index, and foot arch height on HV using a decision tree approach to construct a predictive model reported that these parameters influenced the presence of HV. According to the model's results, being over 50 years of age and female sex were associated with HV-related angles. In the present study, the same risk group can likewise be said to be more prevalent in the patient group (25).

## CONCLUSION

Radiographic images, accepted as the standard method for diagnosing and monitoring foot deformities, are widely used. The present study is consistent with research employing artificial intelligence tools for HV diagnosis and classification. We believe that the study's findings will contribute to databases in this field and will guide developments in forefoot surgery and improve treatment options.

## Ethical approval

This study has been approved by the Bolu Abant İzzet Baysal University Non-Interventional Clinical Research Ethics Committee (approval date 05/05/2026, number 2026/217).

## Author contribution

Concept: MKu, AR; Design: MKu, GR; Data Collection or Processing: AR, MKa; Analysis or Interpretation: AR, İK; Literature Search: MKu; Writing: MKu, GR. All authors reviewed the results and approved the final version of the article.

## Source of funding

The authors declare the study received no funding.

## Conflict of interest

The authors declare that there is no conflict of interest.

## Use of artificial intelligence

The authors declare that artificial intelligence was used in this study solely during the data analysis phase. Specifically, machine learning algorithms were employed to classify and evaluate the angular measurements derived from foot X-ray images for the diagnosis of hallux valgus. No generative artificial intelligence (generative AI) tools were used in the design of the study, the writing or editing of the manuscript, the generation or selection of references, or the interpretation of the results. The authors have reviewed all outputs and assume full responsibility for the entire content and findings of the manuscript.

## REFERENCES

1. Kwolek K, Gądek A, Kwolek K, Lechowska-Liszka A, Malczak M, Liszka H. Artificial intelligence-based diagnosis of hallux valgus interphalangeus using anteroposterior foot radiographs. *World J Orthop.* 2025; 16(6): 103832. [\[Crossref\]](#)
2. Açıkgoz AK, Mutluay ŞD, Binokay F, Bozkir MG. Evaluation of the relationship between hallux valgus and foot measurements in radiographic images of adult female. *Surg Radiol Anat.* 2022; 44(9): 1281-8. [\[Crossref\]](#)
3. Nebhani N, Yatin Trivedy M, Heyes G, Molloy A, Mason L. Hallux valgus interphalangeus measurement: comparison of the 2 radiographic methods. *Foot Ankle Spec.* 2024; 17(4): 323-8. [\[Crossref\]](#)
4. Hida M, Eto S, Wada C, et al. Development of hallux valgus classification using digital foot images with machine learning. *Life (Basel).* 2023; 13(5): 1146. [\[Crossref\]](#)
5. Zhao R, Wang G, Li F, et al. Developing machine learning-based predictive models for hallux valgus recurrence based on measurements from radiographs. *Foot Ankle Int.* 2024; 45(9): 1000-8. [\[Crossref\]](#)
6. Kwolek K, Gądek A, Kwolek K, Kolecki R, Liszka H. Automated decision support for hallux valgus treatment options using anteroposterior foot radiographs. *World J Orthop.* 2023; 14(11): 800-12. [\[Crossref\]](#)
7. Forin Valvecchi T, Marcolli D, De Cesar Netto C. Advanced three-dimensional assessment and planning for hallux valgus. *Foot Ankle Clin.* 2025; 30(2): 349-62. [\[Crossref\]](#)
8. Spindler FT, Ettinger S, D. A. F. Scientific committee, Baumbach SF. Classification of hallux valgus deformity-is there a standard? *Arch Orthop Trauma Surg.* 2024; 144(11): 4737-43. [\[Crossref\]](#)
9. Chen YS, Liang CH, Shih HT, Tu KC, Tang SC, Wang SP. Correlation between hallux valgus severity and the prevalence of metatarsus adductus in hallux valgus. *J Foot Ankle Res.* 2025; 18(2): e70049. [\[Crossref\]](#)
10. Li R, Tan K, Xie Y, Wang F. Morphology variations with medial cuneiform in hallux valgus. *Anat Sci Int.* 2024; 99(1): 59-67. [\[Crossref\]](#)
11. Bu P, Li C, Pu L, Ma X, Meng X, Xu Y. Radiographic assessment of relationship between medial cuneiform obliquity and hallux valgus. *J Foot Ankle Surg.* 2023; 62(3): 583-9. [\[Crossref\]](#)
12. Takeda R, Mizuhara H, Uchio A, et al. Automatic estimation of hallux valgus angle using deep neural network with axis-based annotation. *Skeletal Radiol.* 2024; 53(11): 2357-66. [\[Crossref\]](#)
13. Matsumoto T, Higuchi J, Maenohara Y, et al. The discrepancy between radiographically-assessed and self-recognized hallux valgus in a large population-based cohort. *BMC Musculoskelet Disord.* 2022; 23(1): 31. [\[Crossref\]](#)

14. Kwolek K, Liszka H, Kwolek B, Gądek A. Measuring the angle of hallux valgus using segmentation of bones on X-ray images. In: Tetko I, Kůrková V, Karpov P, Theis F, editors. *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. ICANN 2019. Lecture Notes in Computer Science, Vol 11731*. Cham: Springer; 2019: 327-38. [\[Crossref\]](#)
15. Izquierdo P, Calderon A, Jenkins GL, Thorne S, Mathieson I. Automatic angle recognition in hallux valgus. In: *Proceedings of the UKSim-AMSS 22nd International Conference on Computer Modelling and Simulation (UKSim2020)*; 2020; Cambridge, United Kingdom. [\[Crossref\]](#)
16. Li T, Wang Y, Qu Y, Dong R, Kang M, Zhao J. Feasibility study of hallux valgus measurement with a deep convolutional neural network based on landmark detection. *Skeletal Radiol*. 2022; 51(6): 1235-47. [\[Crossref\]](#)
17. Ray A, Ray G, Kürtül İ, Şenol GT. Sex estimation from the variables of talocrural joint by using machine learning algorithms. *J Forensic Leg Med*. 2025; 114: 102912. [\[Crossref\]](#)
18. Taşkın Şenol G, Kürtül İ, Ray A, Ahmetoğlu G, Seçgin Y, Öner Z. Determination of gender by machine learning algorithms, through using craniocervical junction parameters and dimensions of the cervical spinal canal. *Osmangazi Journal of Medicine*. 2023; 45(5): 672-7. [\[Crossref\]](#)
19. Ma S, Wang H, Zhao W, et al. An interpretable deep learning model for hallux valgus prediction. *Comput Biol Med*. 2025; 185: 109468. [\[Crossref\]](#)
20. Wang Q, Wang JH, Ji DD, Zhang Y, Zhao HM. Automated measurement of hallux valgus angles using a deep learning model: validation and comparison with surgeons of varying expertise. *Sci Rep*. 2026; 16(1): 12280. [\[Crossref\]](#)
21. Wang Q, Wang JH, Ji DD, Liang JQ, Yang XQ, Zhao HM. Deep learning versus manual measurement of hallux valgus angle and intermetatarsal angle on weight-bearing X-rays in hallux valgus. *BMC Musculoskelet Disord*. 2025; 27(1): 58. [\[Crossref\]](#)
22. Hussain A, Lee C, Hu E, Amirouche F. Deep learning automation of radiographic patterns for hallux valgus diagnosis. *World J Orthop*. 2024; 15(2): 105-9. [\[Crossref\]](#)
23. Takeda R, Uchio A, Iidaka T, et al. Automatic Hardy and Clapham's classification of hallux sesamoid position on foot radiographs using deep neural network. *Foot Ankle Surg*. 2025; 31(3): 220-6. [\[Crossref\]](#)
24. Wang C, Li C, Zhang R, et al. Identification of radiographic characteristics associated with pain in hallux valgus patients: a preliminary machine learning study. *Front Public Health*. 2022; 10: 943026. [\[Crossref\]](#)
25. Jung DY, Jung SH, Gwak GT. Contributions of age, gender, body mass index, and normalized arch height to hallux valgus: a decision tree approach. *BMC Musculoskelet Disord*. 2023; 24(1): 278. [\[Crossref\]](#)

# Comparison of thoracolumbar fascia flexibility and its relationship with trunk proprioception and spinal posture in individuals with nonspecific low back pain and healthy peers

Hüsna Nur Çördükçü<sup>1</sup>, Ramazan Kurul<sup>2</sup>

<sup>1</sup>İzzet Baysal Physical Therapy and Rehabilitation Training and Research Hospital, Bolu, Türkiye

<sup>2</sup>Department of Physiotherapy and Rehabilitation, Bolu Abant İzzet Baysal University, Bolu, Türkiye

**Cite as:** Çördükçü HN, Kurul R. Comparison of thoracolumbar fascia flexibility and its relationship with trunk proprioception and spinal posture in individuals with nonspecific low back pain and healthy peers. Northwestern Med J. 2026;6(4):343-353.

## Abstract

**Aim:** To investigate the relationship between thoracolumbar fascia flexibility, trunk proprioception, and spinal posture in individuals with non-specific low back pain.

**Materials and Methods:** A total of 80 participants were included in the study, comprising 40 individuals with nonspecific low back pain and 40 healthy controls. The Visual Analog Scale (VAS) was used to determine pain severity; the Oswestry Disability Index (ODI) to assess the impact of low back pain on daily life; and the International Physical Activity Questionnaire-Short Form (IPAQ) to determine physical activity levels. Thoracolumbar fascia flexibility was assessed with a goniometric platform; trunk proprioception with; and spinal posture with the Spinal Mouse device.

**Results:** There were significant differences in VAS and Oswestry Disability Index scores between the groups with and without nonspecific low back pain ( $p<0.05$ ). IPAQ scores were similar between groups ( $p=0.288$ ). Thoracolumbar fascia flexibility in the nonspecific low back pain group was lower than that in the healthy group ( $p<0.05$ ). When results from the posture evaluation were compared, the thoracic angle during maximum flexion was significantly smaller in the group with non-specific low back pain ( $p=0.014$ ). No significant results were found for the upright and maximum extension postures ( $p>0.05$ ). No correlation was found between thoracolumbar fascia flexibility and trunk proprioception.

**Conclusion:** Thoracolumbar fascia flexibility and trunk proprioception were adversely affected in individuals with nonspecific low back pain. Although TLF flexibility and proprioception differed between groups, TLF flexibility was not associated with proprioceptive error. These findings highlight the potential clinical value of incorporating assessments of thoracolumbar fascia flexibility and trunk proprioception into the evaluation of individuals with nonspecific low back pain.

**Keywords:** flexibility, low back pain, posture, proprioception, thoracolumbar fascia

**Corresponding author:** Hüsna Nur Çördükçü **E-mail:** husnanur1307@gmail.com

**Received:** 04.08.2026 **Accepted:** 13.10.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

## INTRODUCTION

Low back pain is the second leading cause of disability worldwide; however, accurately identifying its underlying cause remains challenging (1). Consequently, cases in which no specific underlying pathology can be identified are commonly classified as nonspecific low back pain. Therefore cannot be attributed to conditions such as osteoporosis, infection, fracture, structural deformity, cauda equina syndrome, tumor, radicular syndrome, or inflammatory disorders, is defined as non-specific low back pain (2). Low back pain may lead to postural impairments, deficits in proprioceptive acuity, and reductions in flexibility. Dysfunction within the myofascial system and the thoracolumbar fascia is considered one of the primary contributing factors to non-specific low back pain (3). The thoracolumbar fascia is an active connective tissue structure that plays a crucial role in the transmission of forces between the trunk muscles and the vertebral column (4). This connective tissue, within the deep epimysium and aponeurotic tissues, is rich in nerve fibers responsive to pain and low-threshold mechanical stimuli. Under conditions of mechanical loading, microtrauma, and fatigue-related stress, inflammatory responses may develop within the connective tissue (5). Abnormal connective tissue structure may predispose individuals to the development of low back pain or may occur as a consequence of injury and/or alterations in movement patterns resulting from chronic pain (6).

In individuals with low back pain, alterations in muscle activation patterns occur, leading to impairments in proprioceptive feedback (7). These individuals have been shown to exhibit greater lumbopelvic repositioning errors during standing or sitting postures (8). Fascia is richly innervated and plays a role in proprioception through receptors embedded within its tissue. Impairments in fascial mobility have been associated with deficits in proprioceptive acuity (9). Furthermore, it has been suggested that acute interventions targeting the fascia may lead to improvements in proprioceptive function (10). The thoracolumbar fascia (TLF) contributes to trunk stabilization and plays a key role in load distribution (5). Prolonged maintenance of inappropriate postures alters the tensioning patterns of the fascia, leading to tissue deformation. Studies conducted in individuals with chronic non-specific

low back pain have demonstrated structural changes in the TLF, including reduced flexibility and increased thickness. These alterations may contribute to changes in postural habits, affecting posture through mechanisms such as localized inflammation or fibrosis. Poor posture can impair the mechanical properties of the TLF, suggesting a bidirectional interaction between posture and fascia.

The flexibility of the TLF has been shown to decrease when sitting and in forward-flexed postures. Prolonged maintenance of these postures may reduce activation of the trunk and hip musculature, leading to deficits in postural control and increased lumbar pain (11). Consequently, examination of the thoracolumbar fascia represents an important step in elucidating the pathophysiology of low back pain. A review of the literature suggests that alterations in the thoracolumbar fascia (TLF) may play a role in the development of chronic low back pain. Adhesions within the fibrous structure of the TLF reduce its gliding capacity, resulting in increased stiffness and decreased flexibility (6). Anatomical studies in this field have demonstrated a high density of mechanoreceptors within the TLF, leading to the interpretation that fascia is not merely a passive connective tissue but also a structure capable of providing nociceptive and proprioceptive input (5).

A study assessing TLF stiffness using shear-wave elastography reported that increased stiffness was correlated with pain intensity and functional limitation; these findings suggest that mechanical alterations in the fascia may influence both motor control and proprioceptive function (12). Existing studies examining fascial flexibility in relation to proprioception and posture have generally involved small sample sizes and predominantly employed advanced measurement techniques such as ultrasound elastography. Research on non-imaging assessment methods that can be easily implemented in clinical practice remains limited. Furthermore, studies simultaneously investigating fascial flexibility together with both proprioception and posture are scarce, with most existing research focusing on a single parameter (13). This limitation impedes the ability to adequately explain the functional interactions among these variables.

In light of these considerations, we hypothesized that thoracolumbar fascial flexibility is reduced in individuals with non-specific low back pain and that this loss of flexibility adversely affects proprioceptive function and spinal posture. Therefore, this study aimed to compare thoracolumbar fascial flexibility, trunk proprioception, and spinal posture between individuals with non-specific low back pain and healthy controls and to determine the relationships between fascial flexibility and spinal posture, and between fascial flexibility and trunk proprioception. The findings of this study are expected to have clinical applicability and to contribute to the scientific literature.

## MATERIALS AND METHODS

### Study design and participants

This study was a randomized, controlled, comparative cross-sectional study involving 80 participants (Figure 1). Participants were assessed between October 2022 and April 2023. Ethical approval was obtained from the Bolu Abant İzzet Baysal University Clinical Researches Ethics Committee (approval date: 06 September 2022; decision number: 2022/226) and registered to clinicaltrials.gov (ID: NCT07039214 Date: 18/06/2025). Written informed consent was obtained from all participants who met the inclusion criteria prior to their enrollment in the study.

The inclusion criteria were: being between 35 and 65 years of age; having low back pain with an intensity greater than 3 on the Visual Analog Scale (VAS) that had persisted for at least 3 months; and having pain intensity below 3 on the VAS in all other body region (14). The exclusion criteria included a history of lumbar spine surgery; diagnosed osteoporosis, infection, fracture, structural deformity, cauda equina syndrome, tumor, radicular syndrome, or inflammatory disorder; being classified as highly physically active according to the International Physical Activity Questionnaire (IPAQ); generalized joint hypermobility (Beighton score  $\geq 4$ ) (15); the presence of a psychiatric disorder requiring treatment; pregnancy; or a diagnosed rheumatic disease.

Based on a similar study assessing fascial flexibility (16), a sample size calculation was performed

according to trunk proprioception, which was one of the primary hypotheses for the comparison between the two groups. Assuming a large effect size ( $d = 0.68$ ), a power of 90%, and a significance level of 5%, the required total sample size was determined to be 70 participants (35 per group). sample size calculations were performed using G\*Power 3.1.

### Outcome measurements

Demographic data were collected from all participants.

Pain intensity was assessed using a 10-cm horizontal Visual Analog Scale (VAS). The VAS is a scale with established validity and reliability for assessing pain in individuals with musculoskeletal disorders (14). The scale consists of a 10-cm line anchored by 0, indicating “no pain,” and 10, indicating “unbearable pain.” Participants were instructed to indicate the intensity of their perceived pain by placing a dot, drawing a line, or making a mark on the scale. The distance from the starting point (0) to the point marked by the participant was recorded as the participant’s pain intensity. VAS scores of 0–3.4 cm, 3.5–7.4 cm, and  $\geq 7.5$  cm were classified as mild, moderate, and severe pain, respectively.

The Oswestry Disability Index (ODI) was used to assess the impact of low back pain on daily activities. The Turkish validity and reliability of the questionnaire were established by Yakut et al. in 2004 (17). The questionnaire consists of 10 items assessing different aspects of functional status. Each item is scored from 0 to 5, with higher scores indicating greater disability.

The International Physical Activity Questionnaire (IPAQ) is a standardized questionnaire developed by researchers from several countries with support from the World Health Organization (WHO) to assess physical activity levels. Its Turkish validity and reliability were established by Sağlam et al (18). The original questionnaire consists of 27 items, whereas the short form comprises 7 items. Participants are asked to estimate the frequency and duration of their physical activities during the previous week. The duration of each activity is multiplied by its corresponding metabolic equivalent of task (MET) value, and the overall score is calculated by summing the MET values across all activity categories.

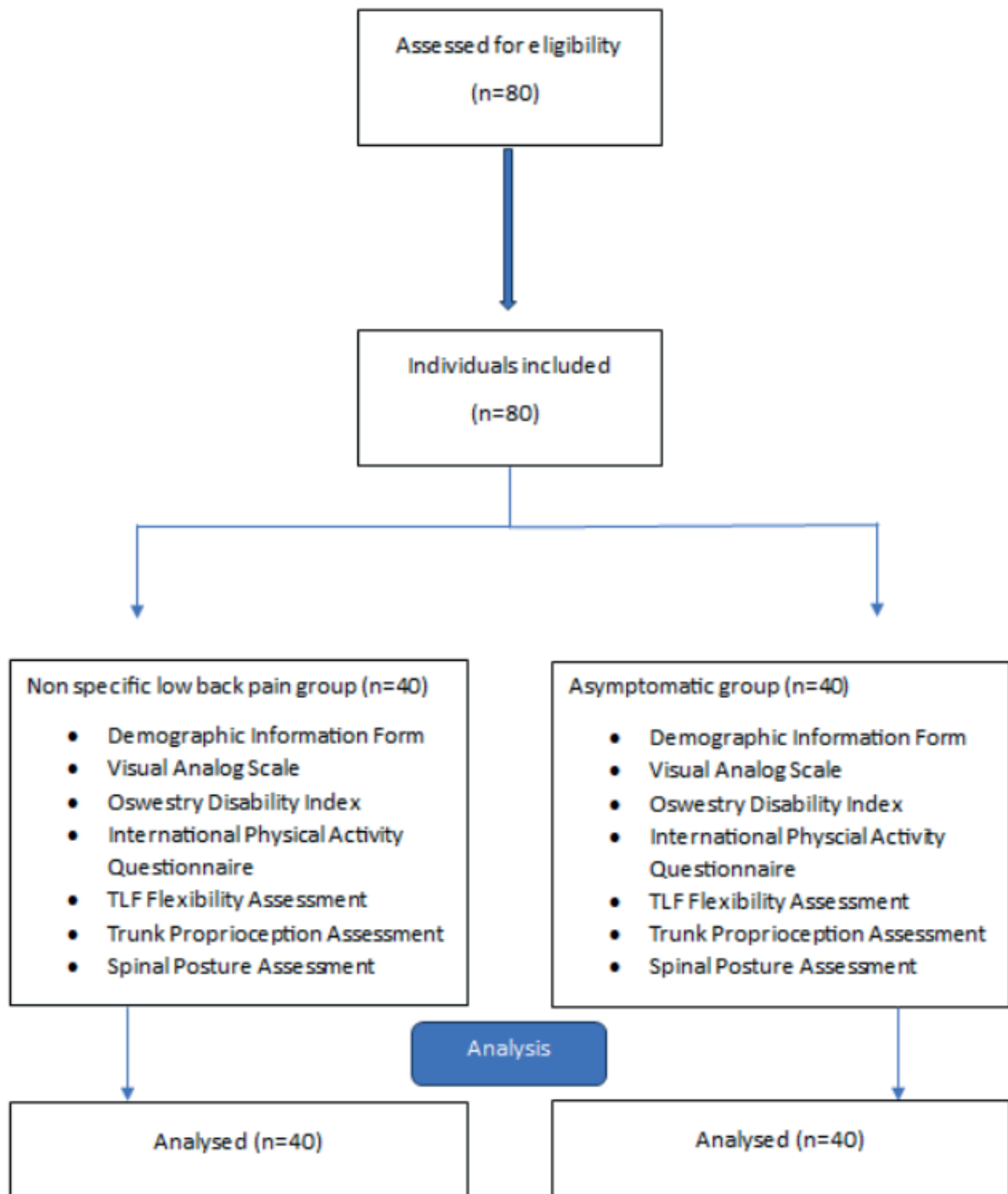


**Figure 1.** Flowchart of study.

The primary outcome was thoracolumbar fascia (TLF) flexibility. Thoracolumbar fascia (TLF) flexibility was assessed using a goniometric platform developed by Sarioğlu and Özünlü Pekyavaş (19), which has demonstrated good internal consistency, with reported Cronbach's  $\alpha$  coefficients of 0.80 for right rotation and 0.96 for left rotation. The goniometric platform was placed on a table positioned in front of the participants. Participants were seated on a chair without back support, with the lumbar spine in a neutral position and the hips and knees flexed to 90°. To ensure that rotation occurred primarily through the thoracic region, the researcher stabilized the pelvis by holding the posterior superior iliac spines. Participants were instructed to perform trunk rotation first to the right and then to the left. Verbal cues were provided to monitor and guide the rotation. At the end range of rotation, participants were asked to mark the platform positioned in front of them while maintaining 90° of shoulder flexion, with the hands clasped in front of the trunk, holding a wooden pencil between the hands. Each measurement was repeated three times on both sides, and the mean of the three trials was used for analysis.

Trunk proprioception was assessed using a mobile application, Clinometer (20). Participants stood in an upright position with their arms fixed alongside their bodies, heels positioned 3 cm apart, and their gaze directed toward a fixed point at eye level. The smartphone was placed vertically just above the iliac crest, along the mid-axillary line at the midpoint of the iliac spine, and secured with a belt. The clinometer was calibrated to 0°, with 0° defined as the starting position. The target movement range was set between 0° and 30°. Participants were first passively guided by the physiotherapist to the target angle. They were instructed to maintain this position for 10 seconds to memorize it, after which they actively returned to the starting position. Subsequently, with their eyes closed, participants were asked to actively reproduce the target angle three times. The mean absolute deviation of the three trials was recorded as the proprioception value (20).

Spinal posture was assessed using the Spinal Mouse device (Idiag, Voletswil, Switzerland; SM) (21). Prior to the measurements, the researcher demonstrated all test positions to the participants. Participants were

instructed to maintain each position for 1–2 seconds during the measurements. They were asked to remove clothing to allow full exposure of the spinous processes and to stand upright. The spinous processes from C7 to S3 were palpated and marked. The Spinal Mouse was then guided along the marked points, from the superior to the inferior. Participants were instructed to assume three positions sequentially: upright standing, maximal trunk flexion, and maximal trunk extension. For the upright position, participants stood comfortably with their feet shoulder-width apart, heads in a neutral position, and arms relaxed at their sides. For maximal flexion, participants were instructed to bend forward as far as possible without flexing the knees. To achieve maximal extension, participants were asked to cross their arms in front of their trunk and extend the trunk as far as possible without knee flexion.

### Statistical analysis

Statistical analyses were performed using SPSS version 25.0 (IBM SPSS Statistics 25; IBM Corp., Armonk, NY, USA). Continuous variables that were found to be normally distributed according to the Kolmogorov–Smirnov test were compared between groups using the independent samples t-test. The chi-square test was applied to examine the associations between categorical variables and group membership. Correlation analyses were conducted using Pearson's correlation coefficient. Descriptive statistics for continuous variables were presented as mean ( $\bar{X}$ ), standard deviation (SD), minimum and maximum values, while categorical variables were expressed as frequencies (n) and percentages (%). A p-value less than 0.05 was considered statistically significant.

## RESULTS

A total of 80 individuals aged between 35 and 65 years were included in this study. The non-specific low back pain group consisted of 40 individuals (24 females, 16 males), while the healthy control group included 40 individuals (23 females, 17 males).

The normality of data distribution for all outcome measures in both groups was assessed using the Shapiro-Wilk test. The results indicated that all parameters were normally distributed ( $p > 0.05$ ).

In this study, the primary outcome measure was the difference in thoracolumbar fascia (TLF) flexibility between groups. Effect sizes, calculated using Cohen's *d*, were 1.63 for right rotation and 1.57 for left rotation, indicating large effects.

The mean age of the participants was  $52.45 \pm 7.00$  years in the non-specific low back pain group and  $52.75 \pm 7.62$  years in the healthy group. Body mass index was  $24.04 \pm 0.96$  kg/m<sup>2</sup> in the non-specific low back pain group and  $27.65 \pm 2.18$  kg/m<sup>2</sup> in the asymptomatic group (Table 1).

In between-group comparisons, significant differences were observed in Visual Analog Scale (VAS) scores ( $p < 0.001$ ), Oswestry Disability Index scores ( $p < 0.001$ ), thoracolumbar fascia (TLF) flexibility measurements on the right and left (both  $p < 0.001$ ), and proprioception values ( $p = 0.010$ ). In contrast, no significant differences were found between groups in International Physical Activity Questionnaire (IPAQ) scores ( $p = 0.288$ ) (Table 2).

Thoracolumbar fascia (TLF) flexibility on the right side was  $47.60 \pm 10.41$  in the non-specific low back pain group and  $64.30 \pm 10.07$  in the healthy group. For the left side, TLF flexibility was  $48.28 \pm 11.38$  in individuals with non-specific low back pain and  $65.38 \pm 10.31$  in healthy individuals. On both sides, TLF flexibility was significantly higher in the healthy group ( $p < 0.001$ ; Table 2).

The mean deviation value obtained from the proprioception assessment was  $13.01 \pm 6.97$  in the non-specific low back pain group and  $9.49 \pm 4.82$  in the healthy group. A significant difference in proprioception values was observed between the groups ( $p = 0.01$ ) (Table 2).

Postural assessment data indicated that thoracic curvature was significantly lower in the non-specific low back pain group than in the healthy group ( $p = 0.014$ ). No significant differences were observed between the groups for the remaining spinal angle parameters ( $p > 0.05$ ) (Table 3).

**Table 1.** Individuals' age, body mass index information.

|                                      | Non-Specific Low Back Pain (n=40) | Healthy (n=40) |        |       |
|--------------------------------------|-----------------------------------|----------------|--------|-------|
|                                      | X±SD                              | X±SD           | t      | p     |
| Age (year)                           | 52.45±7.00                        | 52.75±7.62     | -0.183 | 0.855 |
| Body Mass Index (kg/m <sup>2</sup> ) | 24.04±0.96                        | 27.65±2.18     | 1.330  | 0.188 |

t: Independent t-Test, \* $p < 0.05$ , X±SD: mean±standart deviation

**Table 2.** Independent t-test results of mean differences between groups.

|                         | Non-Specific Low Back Pain (n=40) | Healthy (n=40) |        |        |                  |
|-------------------------|-----------------------------------|----------------|--------|--------|------------------|
|                         | X±SD                              | X±SD           | t      | p      | CI %95 (min-max) |
| VAS                     | 7.08±1.78                         | 0.06±0.26      | 24.56  | <0.001 | 6.45-7.59        |
| IPAQ                    | 340.06±291.80                     | 266.66±321.42  | 1.069  | 0.288  | -63.25 – 210.05  |
| Oswestry                | 46.40±13.17                       | 0.55±1.69      | 21.830 | <0.001 | 41.66 – 50.03    |
| TLF Flexibility (Right) | 47.60±10.41                       | 64.30±10.07    | -7.292 | <0.001 | -21.26 – -12.16  |
| TLF Flexibility (Left)  | 48.28±11.38                       | 65.38±10.31    | -7.043 | <0.001 | -21.94 – -12.27  |
| Proprioception          | 13.01±6.97                        | 9.49±4.82      | 2.625  | 0.010  | 0.85 – 6.19      |

VAS: Visual Analog Scale, IPAQ: International Physical Activity Questionnaire, TLF: thoracolumbar fascia, t: Independent t-Test, X±SD: mean±standart deviation, \* $p < 0.05$ , CI %95: Confidence Interval of the Difference.

**Table 3.** Comparison of spinal postures of individuals.

|     |                | Non-Specific Low Back Pain (n=40) | Healthy (n=40) |        |        |
|-----|----------------|-----------------------------------|----------------|--------|--------|
|     |                | X±SD                              | X±SD           | p      | t      |
| UP  | Sacrum-Hip     | 32.00 ± 39.70                     | 23.65 ± 14.31  | 0.226  | 1.220  |
|     | Thoracic Angle | 53.89 ± 10.25                     | 58.63 ± 49.23  | 0.559  | -0.588 |
|     | Lumbar Angle   | -38.10 ± 9.29                     | 36.05 ± 22.30  | 0.598  | -0.529 |
| MFP | Sacrum-Hip     | 75.15 ± 50.17                     | 70.57 ± 14.66  | 0.591  | 0.540  |
|     | Thoracic Angle | 48.64 ± 14.76                     | 56.52 ± 12.57  | 0.014* | -2.520 |
|     | Lumbar Angle   | 12.61 ± 18.14                     | 17.07 ± 10.97  | 0.197  | -1.302 |
| MEP | Sacrum-Hip     | 19.17 ± 12.10                     | 37.52 ± 69.79  | 0.110  | -1.617 |
|     | Thoracic Angle | 28.87 ± 15.18                     | 25.07 ± 15.89  | 0.288  | 1.071  |
|     | Lumbar Angle   | -48.43 ± 19.43                    | -56.97 ± 30.43 | 0.145  | 1.471  |

Independent t-Test \*p<0.05, X±SD: mean±standart deviation UP: Upright Posture; MFP: Maximum Flexion Posture; MEP: Maximum Extansion Posture.

**Table 4.** Correlation evaluations of thoracolumbar fascia flexibility according to proprioception in non-specific low back pain and healthy groups.

|                                   |                                         | Proprioception |       |
|-----------------------------------|-----------------------------------------|----------------|-------|
|                                   |                                         | r              | t     |
| Non-Specific Low Back Pain (n=40) | Thoracolumbar Fascia Flexibilty (Right) | 0.014          | 0.931 |
|                                   | Thoracolumbar Fascia Flexibilty (Left)  | 0.008          | 0.961 |
| Healthy (n=40)                    | Thoracolumbar Fascia Flexibilty (Right) | -0.070         | 0.666 |
|                                   | Thoracolumbar Fascia Flexibilty (Left)  | -0.044         | 0.786 |

Pearson Correlation Test \*p<0.05

A Pearson product-moment correlation was used to examine the relationship between thoracolumbar fascia flexibility and proprioception. No significant correlations were found between thoracolumbar fascia flexibility on the right or left and proprioception in either the non-specific low back pain group or the healthy group (Table 4).

## DISCUSSION

This study investigated the relationship between thoracolumbar fascia flexibility, trunk proprioception, and spinal posture in individuals with non-specific low back pain. A total of 80 participants were included: 40 individuals with non-specific low back pain and 40 healthy controls with comparable demographic characteristics. The findings demonstrated that, compared with healthy individuals, those with non-specific low back pain had significantly lower

thoracolumbar fascia flexibility and a significantly greater mean deviation in trunk proprioception. Spinal posture parameters were largely comparable between the groups. Furthermore, when the relationships among thoracolumbar fascia flexibility, proprioception, and spinal posture were examined, no significant correlations were identified among these variables in either group.

Physical activity is considered an important component in the management of non-specific low back pain, as prolonged physical inactivity has been shown to increase lumbar pain (22). However, no clear consensus exists that individuals with chronic low back pain consistently exhibit lower levels of physical activity than healthy individuals. A published review reported no clear association between physical activity levels and disability, pain, or being asymptomatic in individuals with non-specific low back pain (22). When

observational studies investigating the relationship between physical activity and low back pain were synthesized, moderate-level evidence supporting an association between these two parameters was identified (23). Heneweer et al. reported moderate evidence indicating that physical inactivity increases the risk of developing chronic low back pain in both individuals with and without existing low back pain (23). In the present study, no significant difference was found between the groups in terms of physical activity levels, and both groups were classified as sedentary based on questionnaire results. Although low back pain may reduce physical activity levels, it is difficult to conclude that sedentary behavior alone directly causes low back pain in individuals who are already inactive. In line with the findings of Heneweer et al. (23), being physically active or inactive does not appear to represent a definitive risk factor for low back pain, and the results of the present study are consistent with this perspective. The participants in this study were adults. A published review has demonstrated that, within the same adult age group, differences in physical activity are more likely to be observed in the distribution and intensity of activity throughout the day rather than in total activity levels (24). Accordingly, the similar physical activity levels observed between the groups in the present study may be interpreted in this context.

Previous studies have reported that the elastic modulus of the thoracolumbar fascia (TLF) in individuals with low back pain is approximately 25–30% lower than that observed in healthy individuals, and that this reduction is proportional to pain intensity, with decreasing flexibility being associated with increased pain levels (13). In a study evaluating TLF thickness using ultrasound in individuals with chronic non-specific low back pain, the fascia was found to be thicker in both longitudinal and transverse directions, indicating a loss of normal adaptive capacity (25). In the present study, TLF flexibility was reduced in individuals with non-specific low back pain. This finding may be attributed to an inability to maintain optimal positioning of the TLF because of pain, which may result in increased fiber thickness. Additionally, pain-related avoidance of movement aimed at protecting the lumbar region may reduce the fluid content of the TLF, leading to increased tissue adhesions and decreased flexibility. A reduction in fluid content may also decrease the

amount of hyaluronic acid within the fascia, alter fascial traction, and further reduce flexibility. These findings are consistent with our results and support the existing literature.

The literature has consistently emphasized that proprioceptive function is reduced in individuals with non-specific low back pain compared with healthy individuals. The majority of studies have assessed proprioception using repositioning error measures and have reported greater mean deviations in individuals with low back pain (7,26). A published meta-analysis demonstrated a significant and large deficit in repositioning accuracy in the low back pain group (26). In their systematic review, Tong et al. (27) suggested that chronic low back pain directly affects afferent mechanoreceptors responsible for proprioceptive control, leading to structural alterations in proprioceptors, while thixotropic adaptations of muscle spindles impair their signaling capacity. In individuals with low back pain, prolonged maintenance of static postures may cause muscle spindle stiffening through cross-bridge formation, thereby disrupting signal generation and transmission. In the present study, the mean deviation of repositioning error was significantly higher in individuals with non-specific low back pain than in healthy individuals. This finding may be explained by pain-related presynaptic inhibition, which disrupts the modulation of proprioceptors and leads to prolonged delays in muscle spindle feedback (28). Such impairments may reduce the accuracy of proprioceptive information and promote erroneous signal transmission, ultimately weakening the sense of repositioning in these individuals. Taimela et al. reported, in a biomechanical modeling study, that fatigue of the lumbar musculature reduces lumbar proprioception (29). Given that study participants had experienced low back pain for at least three months, prolonged pain-related changes in the lumbar muscles may have contributed to impaired proprioceptive function and, consequently, to increased mean deviation values in these participants.

Postural alterations are commonly observed in sedentary individuals, with physiological spinal curvatures often deviating from normal angular values. However, pain arising from musculoskeletal problems in this region is not always interpreted as

a direct consequence of postural deformities (30). A previous study reported that sedentary individuals tend to use similar spinal movement patterns during activities of daily living (31). Consequently, there is no clear consensus in the literature regarding the relationship between posture and the development of low back pain. In studies suggesting an association, postural assessments have typically been performed in specific or constrained positions (32). The postural findings of the present study are consistent with this perspective. In the present study, angular values measured in both groups were within normal ranges in the upright standing and maximal flexion positions, whereas values obtained in the maximal extension position exceeded normal limits. This finding may be attributed to the sedentary nature of both groups. Sedentary individuals have been shown to exhibit increased posterior pelvic tilt, which is associated with greater lumbar extension angles (33). Accordingly, the similar lumbar angles observed in both groups may be explained by comparable sedentary behavior. Although prolonged poor posture may result in tissue damage, not all tissue damage necessarily leads to pain; therefore, the absence of significant differences in postural angles between groups may be expected. Postural variations may influence muscle contraction patterns, optimal muscle length, and muscle activation, thereby contributing to the development of low back pain. However, given the lack of significant differences in postural parameters between the groups, the pain observed in the non-specific low back pain group in the present study is more likely attributable to reduced thoracolumbar fascia flexibility than to postural alterations.

Capsular or ligamentous tension in individuals with low back pain directly affects spinal mobility (34). The results of the present study demonstrated that the thoracic angle measured at maximal flexion was significantly greater in the healthy group. This finding indicates reduced thoracic mobility in individuals with non-specific low back pain.

Histological examination of the thoracolumbar fascia (TLF) has demonstrated the presence of mechanoreceptors and proprioceptors within its structure. However, a published review reported that these receptors are not uniformly distributed

throughout the deep layers of the fascia, suggesting that the TLF may play a proprioceptive role primarily through polymodal receptors rather than discrete proprioceptive endings (35). Although the TLF contains proprioceptive receptors, Stecco et al., in their investigation of deep fascia, indicated that not all regions of the TLF are innervated; therefore, sensory receptors are not present at every point within the fascia (36). The method used to assess fascial flexibility evaluated the fascia's mechanical properties at the macroscopic level. Proprioception, in contrast, was assessed using responses from peripheral tissues. However, proprioception is not solely dependent on peripheral input but also involves central processing mechanisms, including muscle spindles and Golgi tendon organs (37). This distinction may have contributed to the lack of correlation observed between TLF flexibility and proprioception. Furthermore, the goniometric platform used to assess fascial flexibility has limitations in accurately capturing the fascia's viscoelastic properties. Previous studies have reported that ultrasonographic techniques provide more objective and detailed assessments of fascial characteristics (14,25). Therefore, the measurement method used in this study may have limited the ability to detect a potential association. In addition, proprioception was evaluated in this study using repositioning error, which requires active participation. Repositioning accuracy represents only one aspect of proprioception. Factors such as fear-avoidance behaviors, postural control deficits, and kinesiophobia are also known to influence proprioceptive function (38). Consequently, the absence of a correlation between TLF flexibility and proprioception in our findings may be attributed to the influence of these additional factors, which contribute to reduced proprioceptive performance.

### Limitations

This study has several limitations that should be considered when interpreting the results. The method used to evaluate TLF flexibility assessed all fascial layers collectively, without distinguishing superficial from deep fascial layers. Furthermore, the study included only individuals with non-specific low back pain and healthy peers; other lumbar pathologies were not evaluated. Physical activity levels were assessed

using a self-reported questionnaire rather than an objective measurement method, which may have introduced reporting bias.

## CONCLUSION

This study evaluated whether thoracolumbar fascia flexibility is affected in individuals with non-specific low back pain, and whether it is associated with proprioception and spinal posture. The results demonstrated that fascia flexibility was significantly lower in individuals with non-specific low back pain than in healthy individuals; however, no association was observed between fascia flexibility and proprioception. Postural parameters were similar between the two groups. Additionally, this study found that individuals with low back pain had greater proprioceptive impairments than healthy individuals. The findings suggest that fascia should be considered in the assessment of individuals with low back pain and that interventions aimed at improving fascial flexibility may be an important component of rehabilitation programs to reduce symptoms.

## Ethical approval

This study has been approved by the Bolu Abant İzzet Baysal University Clinical Researches Ethics Committee (approval date 06/09/2022, number 2022/226). Written informed consent was obtained from the participants.

## Author contribution

Concept: HNC, RK; Design: HNC, RK; Data Collection or Processing: HNC, RK; Analysis or Interpretation: HNC, RK; Literature Search: HNC, RK; Writing: HNC, RK. All authors reviewed the results and approved the final version of the article.

## Source of funding

The authors declare the study received no funding.

## Conflict of interest

The authors declare that there is no conflict of interest.

## REFERENCES

1. Balagué F, Mannion AF, Pellisé F, Cedraschi C. Non-specific low back pain. *Lancet*. 2012; 379(9814): 482–91. [\[Crossref\]](#)
2. Casato G, Stecco C, Busin R. Role of fasciae in nonspecific low back pain. *Eur J Transl Myol*. 2019; 29(3): 8330. [\[Crossref\]](#)
3. Schleip R, Vleeming A, Lehmann-Horn F, Klingler W. Letter to the editor concerning "A hypothesis of chronic back pain: ligament subfailure injuries lead to muscle control dysfunction" (M. Panjabi). *Eur Spine J*. 2007; 16(10): 1733–5; author reply 1736. [\[Crossref\]](#)
4. Brandl A, Egner C, Reer R, Schmidt T, Schleip R. Immediate effects of myofascial release treatment on lumbar microcirculation: a randomized, placebo-controlled trial. *J Clin Med*. 2023; 12(4): 1248. [\[Crossref\]](#)
5. Langevin HM, Fox JR, Koptiuch C, et al. Reduced thoracolumbar fascia shear strain in human chronic low back pain. *BMC Musculoskelet Disord*. 2011; 12: 203. [\[Crossref\]](#)
6. Yang QH, Wang XQ. Lumbar joint position sense measurement of patients with low back pain. *EFORT Open Rev*. 2023; 8(8): 639–50. [\[Crossref\]](#)
7. Korakakis V, O'Sullivan K, Kotsifaki A, Sotiralis Y, Giakas G. Lumbo-pelvic proprioception in sitting is impaired in subgroups of low back pain-but the clinical utility of the differences is unclear. A systematic review and meta-analysis. *PLoS One*. 2021; 16(4): e0250673. [\[Crossref\]](#)
8. Langevin HM. Fascia mobility, proprioception, and myofascial pain. *Life (Basel)*. 2021; 11(7): 668. [\[Crossref\]](#)
9. Güneş M, Yana M. Acute effects of thoracolumbar fascia release techniques on range of motion, proprioception, and muscular endurance in healthy young adults. *J Bodyw Mov Ther*. 2023; 35: 145–50. [\[Crossref\]](#)
10. Chen B, Liu C, Lin M, Deng W, Zhang Z. Effects of body postures on the shear modulus of thoracolumbar fascia: a shear wave elastography study. *Med Biol Eng Comput*. 2021; 59(2): 383–90. [\[Crossref\]](#)
11. Willard FH, Vleeming A, Schuenke MD, Danneels L, Schleip R. The thoracolumbar fascia: anatomy, function and clinical considerations. *J Anat*. 2012; 221(6): 507–36. [\[Crossref\]](#)
12. Liu K, Zhao T, Zhang Y, et al. Shear wave elastography based analysis of changes in fascial and muscle stiffness in patients with chronic non-specific low back pain. *Front Bioeng Biotechnol*. 2024; 12: 1476396. [\[Crossref\]](#)
13. Tamartash H, Bahrpeyma F, Mokhtari Dizaji M. Ultrasound evidence of altered lumbar fascia in patients with low back pain. *Clin Anat*. 2023; 36(1): 36–41. [\[Crossref\]](#)
14. Boonstra AM, Schiphorst Preuper HR, Reneman MF, Posthumus JB, Stewart RE. Reliability and validity of the visual analogue scale for disability in patients with chronic musculoskeletal pain. *Int J Rehabil Res*. 2008; 31(2): 165–9. [\[Crossref\]](#)
15. Georgy EE. Lumbar repositioning accuracy as a measure of proprioception in patients with back dysfunction and healthy controls. *Asian Spine J*. 2011; 5(4): 201–7. [\[Crossref\]](#)

16. Carter C, Surgery JWTJ of B& J, 1964 U. Persistent joint laxity and congenital dislocation of the hip. *Bone Joint J*. 1964;46:40–5. [\[Crossref\]](#)
17. Yakut E, Düger T, Oksüz C, et al. Validation of the Turkish version of the Oswestry Disability Index for patients with low back pain. *Spine (Phila Pa 1976)*. 2004; 29(5): 581–5; discussion 585. [\[Crossref\]](#)
18. Saglam M, Arıkan H, Savcı S, et al. International Physical Activity Questionnaire: reliability and validity of the Turkish version. *Percept Mot Skills*. 2010; 111(1): 278–84. [\[Crossref\]](#)
19. Sarioglu K, Pekyavas NO. A comparison between people with and without subacromial impingement syndrome and a new method for measuring thoracolumbar fascia flexibility. *J Chiropr Med*. 2021; 20(1): 9–15. [\[Crossref\]](#)
20. Caña-Pino A, Espejo-Antúnez L, Adsuar JC, Apolo-Arenas MD. Test-retest reliability of an iPhone® inclinometer application to assess the lumbar joint repositioning error in non-specific chronic low back pain. *Int J Environ Res Public Health*. 2021; 18(5): 2489. [\[Crossref\]](#)
21. Mannion AF, Knecht K, Balaban G, Dvorak J, Grob D. A new skin-surface device for measuring the curvature and global and segmental ranges of motion of the spine: reliability of measurements and comparison with data reviewed from the literature. *Eur Spine J*. 2004; 13(2): 122–36. [\[Crossref\]](#)
22. Lin CWC, McAuley JH, Macedo L, Barnett DC, Smeets RJ, Verbunt JA. Relationship between physical activity and disability in low back pain: a systematic review and meta-analysis. *Pain*. 2011; 152(3): 607–13. [\[Crossref\]](#)
23. Heneweer H, Vanhees L, Picavet HSJ. Physical activity and low back pain: a U-shaped relation? *Pain*. 2009; 143(1–2): 21–5. [\[Crossref\]](#)
24. Lau KKL, Kwan KYH, Cheung JPY, et al. Normative values of spinal and peripheral proprioception in position sense among healthy adolescents and young adults. *Sci Rep*. 2024; 14(1): 31722. [\[Crossref\]](#)
25. Pratt RL. Hyaluronan and the fascial frontier. *Int J Mol Sci*. 2021; 22(13): 6845. [\[Crossref\]](#)
26. Laird RA, Gilbert J, Kent P, Keating JL. Comparing lumbo-pelvic kinematics in people with and without back pain: a systematic review and meta-analysis. *BMC Musculoskelet Disord*. 2014; 15: 229. [\[Crossref\]](#)
27. Tong MH, Mousavi SJ, Kiers H, Ferreira P, Refshauge K, van Dieën J. Is there a relationship between lumbar proprioception and low back pain? A systematic review with meta-analysis. *Arch Phys Med Rehabil*. 2017; 98(1): 120–136.e2. [\[Crossref\]](#)
28. Lalonde NR, Bui TV. Do spinal circuits still require gating of sensory information by presynaptic inhibition after spinal cord injury? *Current Opinion in Physiology*. 2021; 19: 113–8. [\[Crossref\]](#)
29. Taimela S, Kankaanpää M, Luoto S. The effect of lumbar fatigue on the ability to sense a change in lumbar position. A controlled study. *Spine (Phila Pa 1976)*. 1999; 24(13): 1322–7. [\[Crossref\]](#)
30. Kripa S, Kaur H. Identifying relations between posture and pain in lower back pain patients: a narrative review. *Bulletin of Faculty of Physical Therapy*. 2021; 26(1): 34. [\[Crossref\]](#)
31. Hemming R, Sheeran L, van Deursen R, Sparkes V. Non-specific chronic low back pain: differences in spinal kinematics in subgroups during functional tasks. *Eur Spine J*. 2018; 27(1): 163–70. [\[Crossref\]](#)
32. Swain CTV, Pan F, Owen PJ, Schmidt H, Belavy DL. No consensus on causality of spine postures or physical exposure and low back pain: a systematic review of systematic reviews. *J Biomech*. 2020; 102: 109312. [\[Crossref\]](#)
33. Bozorgmehr A, Ebrahimi Takamjani I, Akbari M, Salehi R, Mohsenifar H, Rasouli O. Effect of posterior pelvic tilt taping on abdominal muscle thickness and lumbar lordosis in individuals with chronic low back pain and hyperlordosis: a single-group, repeated-measures trial. *J Chiropr Med*. 2020; 19(4): 213–21. [\[Crossref\]](#)
34. Mellin G. Decreased joint and spinal mobility associated with low back pain in young adults. *J Spinal Disord*. 1990; 3(3): 238–43.
35. Suarez-Rodriguez V, Fede C, Pirri C, et al. Fascial innervation: a systematic review of the literature. *Int J Mol Sci*. 2022; 23(10): 5674. [\[Crossref\]](#)
36. Stecco C, Gagey O, Belloni A, et al. Anatomy of the deep fascia of the upper limb. Second part: study of innervation. *Morphologie*. 2007; 91(292): 38–43. [\[Crossref\]](#)
37. Tesarz J, Hoheisel U, Wiedenhöfer B, Mense S. Sensory innervation of the thoracolumbar fascia in rats and humans. *Neuroscience*. 2011; 194: 302–8. [\[Crossref\]](#)
38. Simsek S, Altindal F, Kilicarslan B. Comparison of affected lumbar proprioception due to different injuries with healthy controls: an observational study. *International Journal of Osteopathic Medicine*. 2023; 49: 100662. [\[Crossref\]](#)

# Peripheral blood inflammation indices versus age in predicting gastritis histopathology: A cross-sectional study

Muhammet Fatih Keyif<sup>1</sup>, Ferdi Bolat<sup>1</sup>

<sup>1</sup>Department of General Surgery, Faculty of Medicine, Bolu Abant İzzet Baysal University, Bolu, Türkiye

**Cite as:** Keyif MF, Bolat F. Peripheral blood inflammation indices versus age in predicting gastritis histopathology: A cross-sectional study. Northwestern Med J. 2026;6(4):354-363.

## Abstract

**Objective:** We evaluated a comprehensive panel of hematological inflammation indices in relation to histopathological features of gastritis and benchmarked them against age.

**Materials and Methods:** In this single-center, retrospective, cross-sectional study, 254 adults who underwent upper gastrointestinal endoscopy with gastric biopsy between June and December 2025 were analyzed. Conventional cellular ratios, composite inflammation indices, C-reactive protein- and nutrition-based measures, and exploratory immature granulocyte-based indices were computed from the pre-procedural hemogram. The primary endpoint was active inflammation; secondary endpoints were chronic inflammation, *Helicobacter pylori*, atrophy, and intestinal metaplasia. Discrimination was assessed by receiver operating characteristic analysis with DeLong comparison, multivariable logistic regression, and five-fold cross-validation; Benjamini-Hochberg correction was applied for multiple testing.

**Results:** Active inflammation was present in 100 patients (39.4%) and corresponded almost entirely to *Helicobacter pylori*-associated chronic active gastritis (kappa 0.98). No index differed with respect to any histopathological endpoint, nor did any distinguish gastritis subtypes. After correction, age was the only variable independently associated with active inflammation (an odds ratio of 0.65 per standard deviation, 95% confidence interval 0.50-0.86;  $p = 0.002$ ). Age discriminated active inflammation better than each other index (area under the curve 0.631 versus 0.51-0.55), and this difference was statistically significant for all comparisons except for the systemic immune-inflammation index (DeLong  $p = 0.096$ ). Adding the blood panel to age reduced the cross-validated discrimination from 0.632 to 0.594.

**Conclusions:** Peripheral hematological inflammatory indices did not reflect the histopathological status of the gastric mucosa. The only discriminating variable was age, which reflected the natural history of gastric disease rather than a systemic inflammatory signal.

**Keywords:** age factors, biomarkers, blood cell count, gastritis, *Helicobacter pylori*, neutrophils

**Corresponding author:** Muhammet Fatih Keyif **E-mail:** dr\_fatihkeyif@hotmail.com

**Received:** 25.08.2026 **Accepted:** 16.09.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

## INTRODUCTION

Gastritis is less a single disease than a convergence of distinct histopathological processes that may coexist in varying combinations within the same mucosa: chronic mononuclear inflammation, active neutrophilic inflammation, glandular atrophy and intestinal metaplasia (IM) (1). Their recognition and grading are based almost entirely on the histopathological examination of biopsies obtained during upper gastrointestinal (GI) endoscopy. The Updated Sydney System provides a semiquantitative framework for the severity and activity of inflammation, while the atrophy- and metaplasia-based OLGA (Operative Link on Gastritis Assessment) and OLGIM (Operative Link on Gastric Intestinal Metaplasia Assessment) staging systems translate these findings into a graded estimate of gastric cancer risk (2).

The principal driver of chronic gastritis is *Helicobacter pylori* (*H. pylori*), the recognized initiator of the Correa cascade, in which chronic active gastritis progresses over years through atrophy and IM toward dysplasia and adenocarcinoma. Within this sequence, active inflammation is the component most directly tied to an acute inflammatory response: defined by neutrophilic infiltration of the gastric mucosa, it is, of all the histopathological components, the one most plausibly mirrored by neutrophil-based systemic markers (3). Reactive gastritis (more accurately termed chemical gastropathy) stands in deliberate contrast. Arising from nonsteroidal anti-inflammatory drugs (NSAIDs), bile reflux or alcohol, it is characterized by foveolar hyperplasia with conspicuously sparse inflammation and is separated from chronic active gastritis by both pathogenesis and management (1). That two of the most common patterns encountered on a gastric biopsy should differ so fundamentally in their inflammatory content raises the clinically meaningful question of whether peripheral blood can distinguish between them.

Histopathology remains the diagnostic reference standard, yet it is invasive, costly, prone to sampling error, subject to interobserver variability and ill-suited to repeated assessment or population-level screening (2). These limitations have sustained a long search for

non-invasive surrogates of the mucosal state, of which serum pepsinogens, gastrin-17 and their combination with *H. pylori* serology are the most extensively studied (4). Against this backdrop, indices derived from the complete blood count have attracted particular interest, for they are inexpensive, universally available, and do not require tests beyond those already performed in nearly every patient presenting for endoscopy.

The biological rationale for these indices is straightforward. Systemic inflammation tends to raise neutrophil and platelet counts while depressing lymphocytes, and ratios that capture this shift are held to quantify the intensity of the inflammatory response in peripheral blood (5). The neutrophil-to-lymphocyte ratio (NLR), the platelet-to-lymphocyte ratio (PLR) and the systemic immune-inflammation index (SII), which integrates all three lineages, are the most widely applied; more recent composites such as the pan-immune-inflammation value (PIV) and the systemic inflammation response index (SIRI) extend the concept by incorporating the monocyte compartment (6). Across oncology, cardiovascular disease, and inflammatory bowel disease, these markers have shown prognostic and, less consistently, diagnostic value.

In gastroenterology, however, their evaluation has been unduly narrow. Most studies have examined the indices against the binary presence of *H. pylori*, and even here the literature is discordant: some reporting that NLR, PLR or SII track infection or its severity, others finding no such association (7). Related work has explored these ratios in gastric adenocarcinoma and in peptic ulcer and dyspepsia, but a common limitation persists: the histopathological components of gastritis are rarely dissected individually, and clinically divergent subtypes (*H. pylori*-associated chronic active gastritis versus chemically mediated reactive gastritis) are seldom set against one another (8,9). The implicit and largely untested assumption underlying this entire body of work is that a localized mucosal process leaves a faithful signature in the systemic circulation.

We sought to interrogate that assumption directly. In adults undergoing upper GI endoscopy with gastric biopsy, we examined the relationship between pre-

procedural hematological inflammation indices and the individual histopathological features of gastritis, with active (neutrophilic) inflammation as the primary endpoint and chronic inflammation, *H. pylori*, atrophy, and IM as secondary endpoints. We further investigated whether these indices could discriminate chronic active from reactive gastritis and, crucially, benchmarked their performance against age, a readily obtainable clinical variable, to determine not only whether the indices carry information about the gastric mucosa but also whether they provide information beyond that available from simpler measures.

## MATERIALS AND METHODS

### Study design and setting

This single-center, retrospective, cross-sectional observational study was conducted in the endoscopy unit of the Department of General Surgery at Bolu Abant İzzet Baysal University Faculty of Medicine Hospital. Consecutive adult patients who underwent upper GI endoscopy with gastric biopsy between 22 June 2025 and 22 December 2025 were identified through the hospital information management system and the electronic medical records. The study was approved by the Bolu Abant İzzet Baysal University Clinical Research Ethics Committee (approval date [07/04/2026], number [2026/159]) and conducted in accordance with the principles of the Declaration of Helsinki; given its the retrospective design, the requirement for written informed consent was waived.

### Participants

Patients were eligible if they were 18 years of age or older, had undergone gastric biopsy at the time of upper GI endoscopy with an available histopathological report, and had a complete blood count obtained within seven days of the endoscopic procedure. When more than one hemogram was available, the result closest to the procedure was used; when a patient had undergone more than one endoscopy, only the first was analyzed. Patients were excluded because of missing or insufficient clinical, laboratory, or histopathological data; clinical or laboratory evidence of active infection; a history of malignancy; a known hematological disorder; ongoing immunosuppressive therapy or

systemic corticosteroid use; presentation with acute GI bleeding; or any concurrent condition expected to perturb the systemic inflammatory response. After these criteria were applied, 254 patients constituted the final study population.

### Data collection and laboratory variables

Demographic data (age, sex) and pre-procedural laboratory parameters were recorded systematically. The components of the complete blood count included white blood cell count (WBC), neutrophil count, lymphocyte count, monocyte count, platelet count, hemoglobin, red cell distribution width (RDW), and immature granulocyte percentage (IG%) with its absolute count (IG#). Biochemical parameters included C-reactive protein (CRP), serum albumin, urea, and creatinine. CRP values reported by the analyzer as being below the assay threshold were entered using the actual numerical results retrieved from the laboratory information system, rather than imputed substitutes, thereby preserving the continuity of the distribution for the CRP-derived indices.

### Histopathological evaluation

Histopathological assessment was based on the reported biopsy findings, graded according to the Updated Sydney System (2). Chronic inflammation, active inflammation (neutrophilic activity), atrophy, IM, and *Helicobacter pylori* (*H. pylori*) were recorded dichotomously as present or absent. Active inflammation was defined by the presence of neutrophilic infiltration of the gastric mucosa. Reported diagnoses were used to classify patients into gastritis subtypes (chronic active gastritis, chronic inactive gastritis, and reactive gastritis) for comparison among subtypes.

### Hematological inflammation indices

Indices were computed from the pre-procedural hemogram according to established definitions (5,6): NLR, PLR, the monocyte-to-lymphocyte ratio (MLR; monocyte/lymphocyte), SII, SIRI and PIV. To capture the inflammation–nutrition axis, we derived the CRP-to-lymphocyte ratio (CLR), the CRP-to-albumin ratio (CAR), the prognostic nutritional index (PNI; albumin

+ 5 × lymphocyte), and the hemoglobin–albumin–lymphocyte–platelet (HALP) score. Finally, recognizing that immature granulocytes index emergency granulopoiesis (a plausible systemic correlate of mucosal neutrophilic activity), we constructed and evaluated a set of exploratory composite indices incorporating IG% and IG#. Cell counts were expressed as absolute values ( $\times 10^3/\mu\text{L}$ ) in the index formulas; the resulting composite indices (e.g., SII, SIRI, PIV) are dimensionless. Because this exploratory panel multiplied the number of statistical comparisons, all analyses were subjected to formal correction for multiple testing, as detailed below.

## Outcomes

The primary endpoint was histopathological evidence of active inflammation. Secondary endpoints were chronic inflammation, *H. pylori*, atrophy, and IM. A prespecified secondary objective was to determine the discriminatory value of the indices in distinguishing between the chronic active and reactive gastritis subtypes.

## Statistical analysis

The distribution of continuous variables was examined with the Shapiro–Wilk test and by inspection of histograms. Because most variables departed from normality, they were summarized as medians (interquartile ranges) and compared with the Mann–Whitney U test for binary outcomes and the Kruskal–Wallis test for comparisons across the three subtypes. Categorical variables are presented as counts and percentages and compared using the chi-square test or Fisher’s exact test; agreement between active inflammation and *H. pylori* status was quantified using Cohen’s  $\kappa$  coefficient. The discriminatory performance of each variable for the primary endpoint was assessed by the area under the receiver operating characteristic curve (AUC), and the AUCs for age and the inflammation indices were compared using the DeLong method. Variables associated with an endpoint in univariable testing, together with age and sex, were entered into multivariable binary logistic regression

models, and continuous predictors were standardized to permit comparison of effect estimates. To address the inflation of type I error arising from the breadth of the index panel and the number of endpoints, p-values were adjusted using the Benjamini–Hochberg false discovery rate procedure, and adjusted significance was reported alongside nominal values. The internal validity of the predictive comparison was further examined using five-fold cross-validated AUC. A two-sided p value below 0.05 was considered statistically significant. All statistical analyses were performed in Python (version 3.12; Python Software Foundation), using the SciPy and stats models libraries for hypothesis testing and regression modeling, and the scikit-learn library for receiver operating characteristic analysis and cross-validation. The DeLong test for comparing correlated areas under the curve and the Benjamini–Hochberg false discovery rate correction were implemented in this environment.

## RESULTS

### Cohort characteristics and the primary endpoint

Of the 254 patients analyzed, 100(39.4%) had histopathologically confirmed active inflammation and 154(60.6%) did not. The median age of the cohort was 58 years (interquartile range [IQR] 49–68) and 115 patients (45.3%) were men. Chronic inflammation was present in 212 patients (83.5%), IM in 49(19.3%), and atrophy in 13 (5.1%). *H. pylori* was identified in 102(40.2%). The distribution of pre-procedural laboratory parameters and derived indices according to active inflammation status is presented in Table 1.

A defining feature of the cohort emerged at the outset. Active inflammation and *H. pylori* status effectively constituted the same variable: every patient with active inflammation was *H. pylori*-positive (100%), whereas only 2 of 154 patients without active inflammation (1.3%) harbored the organism, yielding almost perfect agreement (Cohen  $\kappa$  = 0.98). All patients with active inflammation also showed chronic inflammation. Active inflammation, corresponded in practice, to *H. pylori*-associated chronic active gastritis.

**Table 1.** Demographic, laboratory, and hematological index values by active inflammation status.

| Variable                              | All (n = 254)       | Active (+) (n = 100) | Active (-) (n = 154) | p      |
|---------------------------------------|---------------------|----------------------|----------------------|--------|
| Age, years                            | 58.0 (49.0–68.0)    | 56.0 (45.0–62.2)     | 61.5 (51.2–70.0)     | <0.001 |
| Male, n (%)                           | 115 (45.3)          | 47 (47.0)            | 68 (44.2)            | 0.752  |
| WBC, $\times 10^3/\mu\text{L}$        | 7.2 (5.9–8.7)       | 7.4 (6.2–8.7)        | 7.1 (5.8–8.6)        | 0.187  |
| Neutrophil, $\times 10^3/\mu\text{L}$ | 4.2 (3.1–5.3)       | 4.3 (3.4–5.4)        | 4.0 (3.0–5.3)        | 0.133  |
| Lymphocyte, $\times 10^3/\mu\text{L}$ | 2.2 (1.8–2.8)       | 2.2 (1.8–2.8)        | 2.1 (1.8–2.8)        | 0.484  |
| Monocyte, $\times 10^3/\mu\text{L}$   | 0.6 (0.5–0.7)       | 0.6 (0.5–0.7)        | 0.6 (0.5–0.7)        | 0.735  |
| Platelet, $\times 10^3/\mu\text{L}$   | 250.5 (208.0–294.8) | 258.5 (227.3–304.3)  | 244.0 (201.0–286.5)  | 0.038  |
| Hemoglobin, g/dL                      | 14.0 (13.0–15.2)    | 14.1 (13.1–15.4)     | 14.0 (12.9–15.2)     | 0.640  |
| RDW, %                                | 13.2 (12.7–14.0)    | 13.1 (12.7–13.7)     | 13.3 (12.7–14.2)     | 0.026  |
| IG, %                                 | 0.3 (0.2–0.4)       | 0.3 (0.2–0.4)        | 0.3 (0.2–0.4)        | 0.497  |
| IG#, $\times 10^3/\mu\text{L}$        | 0.02 (0.01–0.03)    | 0.02 (0.01–0.03)     | 0.02 (0.01–0.03)     | 0.200  |
| CRP, mg/L                             | 3.2 (1.5–6.2)       | 3.5 (1.5–6.0)        | 3.1 (1.5–6.5)        | 0.899  |
| Albumin, g/L                          | 46.0 (44.0–48.0)    | 47.0 (44.0–48.0)     | 46.0 (44.0–48.0)     | 0.367  |
| Urea, mg/dL                           | 31.0 (25.0–37.0)    | 28.5 (23.0–36.2)     | 32.0 (25.2–38.0)     | 0.065  |
| Creatinine, mg/dL                     | 0.81 (0.67–0.95)    | 0.79 (0.66–0.91)     | 0.83 (0.68–0.98)     | 0.108  |
| NLR                                   | 1.78 (1.37–2.56)    | 1.93 (1.38–2.55)     | 1.74 (1.36–2.56)     | 0.651  |
| PLR                                   | 111.7 (86.6–147.9)  | 116.3 (90.3–148.1)   | 107.5 (86.0–147.3)   | 0.413  |
| MLR                                   | 0.24 (0.20–0.33)    | 0.23 (0.20–0.29)     | 0.25 (0.20–0.34)     | 0.220  |
| SII                                   | 452.8 (320.7–660.5) | 493.6 (340.8–686.1)  | 425.2 (305.1–632.7)  | 0.187  |
| SIRI                                  | 1.03 (0.72–1.53)    | 0.99 (0.71–1.44)     | 1.03 (0.75–1.63)     | 0.873  |
| PIV                                   | 248.3 (156.9–426.6) | 255.1 (170.2–424.6)  | 246.4 (150.4–428.7)  | 0.429  |
| CLR                                   | 1.33 (0.77–2.72)    | 1.35 (0.76–2.57)     | 1.33 (0.77–2.90)     | 0.739  |
| CAR                                   | 0.07 (0.03–0.14)    | 0.08 (0.03–0.14)     | 0.07 (0.03–0.14)     | 0.872  |
| PNI                                   | 57.5 (54.2–60.8)    | 58.0 (55.0–60.8)     | 57.3 (53.6–60.7)     | 0.261  |
| HALP                                  | 5.63 (4.15–7.89)    | 5.45 (4.12–7.80)     | 6.08 (4.17–8.07)     | 0.533  |

Data are median (IQR) unless otherwise stated. Mann-Whitney U test for continuous variables; chi square test for sex. Abbreviations: IQR, interquartile range; WBC, white blood cell; RDW, red cell distribution width; IG, immature granulocyte; IG#, absolute immature granulocyte count; CRP, C-reactive protein; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; MLR, monocyte-to-lymphocyte ratio; SII, systemic immune-inflammation index; SIRI, systemic inflammation response index; PIV, pan-immune-inflammation value; CLR, CRP-to-lymphocyte ratio; CAR, CRP-to-albumin ratio; PNI, prognostic nutritional index; HALP, hemoglobin-albumin-lymphocyte-platelet score.

### Hematological indices and active inflammation

None of the composite hematological inflammation indices differed between patients with and without active inflammation (Table 1). NLR ( $p = 0.651$ ), PLR ( $p = 0.413$ ), SII ( $p = 0.187$ ), SIRI ( $p = 0.873$ ), PIV ( $p = 0.429$ ), MLR ( $p = 0.220$ ), CLR ( $p = 0.739$ ), CAR ( $p = 0.872$ ), PNI ( $p = 0.261$ ) and HALP ( $p = 0.533$ ) were all statistically indistinguishable across the two groups. Among the individual parameters, only age

( $p < 0.001$ ), platelet count ( $p = 0.038$ ) and RDW ( $p = 0.026$ ) reached nominal significance. After correction for multiple testing across the full panel using the Benjamini-Hochberg false discovery rate, age was the only variable to retain significance (adjusted  $p = 0.011$ ); the associations for platelet count and RDW did not survive correction (adjusted  $p = 0.32$  for both), suggesting that they are probable artifacts of multiple comparisons.

**Table 2.** Discriminatory performance for active inflammation and DeLong comparison against age.

| Variable | AUC   | DeLong p (vs. age) |
|----------|-------|--------------------|
| Age      | 0.631 | —                  |
| SII      | 0.549 | 0.096              |
| PLR      | 0.530 | 0.044              |
| PIV      | 0.529 | 0.044              |
| NLR      | 0.517 | 0.028              |
| CLR      | 0.512 | 0.014              |
| CAR      | 0.506 | 0.010              |

DeLong test comparing the AUC of each index against that of age. Abbreviations: AUC, area under the receiver operating characteristic curve; SII, systemic immune-inflammation index; PLR, platelet-to-lymphocyte ratio; PIV, pan-immune-inflammation value; NLR, neutrophil-to-lymphocyte ratio; CLR, C-reactive protein-to-lymphocyte ratio; CAR, C-reactive protein-to-albumin ratio.

The exploratory composite indices incorporating immature granulocytes, constructed on the rationale that emergency granulopoiesis might mirror mucosal neutrophilic activity, also showed no association with active inflammation and did not survive correction for any endpoint (all adjusted  $p > 0.05$ ).

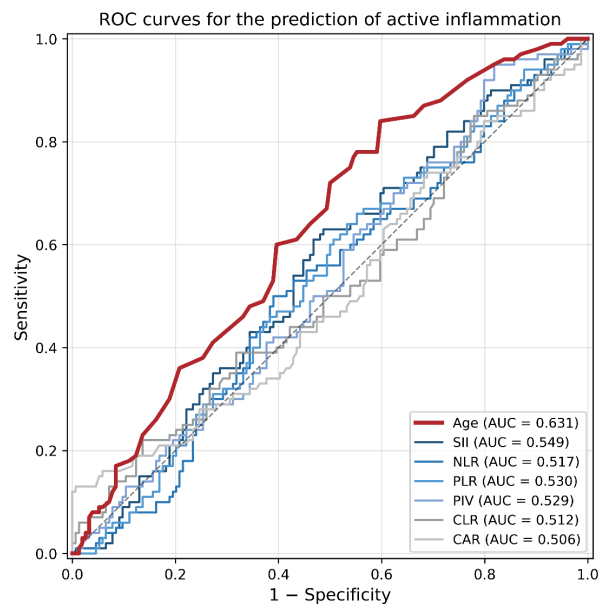
### Discriminatory performance: age versus the indices

Receiver operating characteristic (ROC) analysis reinforced this pattern (Table 2). Age discriminated active inflammation with an area under the curve (AUC) of 0.631; this value is modest in absolute terms, but the highest of any variable examined. Each composite index performed at or near chance level: SII = 0.549, PLR = 0.530, PIV = 0.529, NLR = 0.517, CLR = 0.512, CAR = 0.506. Receiver operating characteristic analysis was limited to these six indices; MLR, SIRI, PNI, and HALP showed no differences between groups in comparison testing (Table 1) and were not modeled further. Formal comparison using the DeLong test showed that age significantly outperformed NLR ( $p = 0.028$ ), PLR ( $p = 0.044$ ), PIV ( $p = 0.044$ ), CLR ( $p = 0.014$ ), and CAR ( $p = 0.010$ ); it was numerically superior to SII ( $p = 0.096$ ). Receiver operating characteristic analysis was performed on the six indices most frequently reported in the gastritis literature (SII, PLR, PIV, NLR, CLR, and CAR). MLR, SIRI, PNI, and HALP were evaluated by group comparisons only. The ROC curves illustrate

this divergence: the age curve rose well above the reference diagonal, whereas every composite index tracked it closely (Figure 1).

### Multivariable modeling

In a multivariable logistic regression model for active inflammation that included age, sex, platelet count, NLR, SII, and CRP (with continuous predictors standardized), age was the only independent predictor (odds ratio [OR] 0.65 per standard deviation, 95% confidence interval [CI] 0.50–0.86;  $p = 0.002$ ). None of the hematological variables retained any association after adjustment (platelet OR 1.21,  $p = 0.515$ ; NLR OR 0.93,  $p = 0.881$ ; SII OR 1.02,  $p = 0.964$ ; CRP OR 1.13,  $p = 0.363$ ). The inverse direction of the age estimate indicates that active inflammation was characteristic of younger rather than older patients.

**Figure 1.** Receiver operating characteristic (ROC) curves for the prediction of histopathological active inflammation.

Age (red) yielded the highest discrimination (AUC = 0.631), significantly exceeding NLR, PLR, PIV, CLR and CAR (DeLong  $p < 0.05$ ) and numerically exceeding SII. All composite indices clustered near the diagonal (AUC 0.51–0.55).

NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; SII, systemic immune-inflammation index; PIV, pan immune-inflammation value; CLR, C-reactive protein-to-lymphocyte ratio; CAR, C-reactive protein-to-albumin ratio.

The incremental value of the blood panel was evaluated using five-fold cross-validation. Age alone achieved a cross-validated AUC of 0.632; adding the entire hematological panel did not improve discrimination; rather, it degraded out-of-sample performance to an AUC of 0.594, consistent with the addition of noise rather than signal.

### Secondary endpoints and gastritis subtypes

For the remaining histopathological endpoints, the indices were similarly uninformative. Chronic inflammation was associated only with modestly higher neutrophil count (median 4.3 vs.  $3.7 \times 10^3/\mu\text{L}$ ;  $p = 0.021$ ) and WBC count ( $7.4$  vs.  $6.3 \times 10^3/\mu\text{L}$ ;  $p = 0.025$ ) in univariable analysis, neither of which remained significant after adjustment for age. No index discriminated *H. pylori* beyond what was already captured by age, and none was associated with atrophy or IM (all  $p > 0.05$ ); the small number of atrophy cases ( $n = 13$ ) precluded robust modeling of that endpoint.

The three diagnostic subtypes (chronic active gastritis [ $n = 100$ ], chronic inactive gastritis [ $n = 104$ ], and reactive gastritis [ $n = 50$ ]) differed in age but not in any hematological index (Table 3). Patients with chronic active gastritis were younger (median 56 years, IQR 45–62) than those with reactive gastritis (median

61 years, IQR 51–70) or chronic inactive gastritis (median 61.5 years, IQR 52.8–70.0; Kruskal–Wallis  $p = 0.002$ ; chronic active vs. reactive  $p = 0.019$ ). Across all composite indices, the Kruskal–Wallis test was not significant, indicating that systemic inflammatory markers did not distinguish between the clinically and pathogenetically divergent chronic active and reactive subtypes.

### DISCUSSION

In a cohort of 254 patients examined using a panel ranging from conventional ratios to newer composite and purpose-built indices, no hematological inflammation markers reflected the histopathological state of the gastric mucosa. None of the following could be distinguished by NLR, PLR, SII, SIRI, PIV, or any of the CRP- and nutrition-based indices: active or chronic inflammation; *H. pylori* infection; atrophy; IM; or any diagnostic subtype. The single variable that carried reproducible information about the mucosa was age, and it did so modestly but consistently better than the indices designed for the that purpose. The central message is, therefore, not that we failed to find the right index, but that peripheral blood, at least in the setting of chronic gastritis, does not appear to transcribe what is happening at the mucosal surface.

**Table 3.** Age and representative indices across gastritis subtypes.

| Variable   | Chronic active (n = 100) | Reactive (n = 50)   | Chronic inactive (n = 104) | p     |
|------------|--------------------------|---------------------|----------------------------|-------|
| Age, years | 56.0 (45.0–62.2)         | 61.0 (51.0–70.0)    | 61.5 (52.8–70.0)           | 0.002 |
| NLR        | 1.93 (1.38–2.55)         | 1.70 (1.37–2.66)    | 1.74 (1.36–2.53)           | 0.877 |
| PLR        | 116.2 (90.3–148.1)       | 102.1 (86.8–145.4)  | 109.2 (85.1–148.7)         | 0.712 |
| MLR        | 0.23 (0.20–0.29)         | 0.26 (0.21–0.33)    | 0.24 (0.20–0.36)           | 0.468 |
| SII        | 493.6 (340.8–686.1)      | 403.9 (314.0–577.7) | 436.4 (304.5–662.7)        | 0.316 |
| SIRI       | 0.99 (0.71–1.44)         | 0.99 (0.73–1.60)    | 1.05 (0.75–1.64)           | 0.820 |
| PIV        | 255.1 (170.2–424.6)      | 223.5 (150.4–389.1) | 259.6 (155.1–437.3)        | 0.542 |
| CLR        | 1.35 (0.76–2.57)         | 1.01 (0.77–2.49)    | 1.38 (0.82–3.56)           | 0.590 |
| CAR        | 0.08 (0.03–0.14)         | 0.05 (0.03–0.13)    | 0.07 (0.03–0.14)           | 0.559 |
| PNI        | 58.0 (55.0–60.8)         | 57.1 (54.1–60.6)    | 57.5 (53.2–60.7)           | 0.527 |
| HALP       | 5.45 (4.12–7.80)         | 6.26 (4.51–7.97)    | 6.08 (4.02–8.11)           | 0.712 |

Data are median (IQR). Kruskal–Wallis test. Abbreviations: IQR, interquartile range; NLR, neutrophil-to-lymphocyte ratio; PLR, platelet-to-lymphocyte ratio; MLR, monocyte-to-lymphocyte ratio; SII, systemic immune-inflammation index; SIRI, systemic inflammation response index; PIV, pan-immune-inflammation value; CLR, C-reactive protein-to-lymphocyte ratio; CAR, C-reactive protein-to-albumin ratio; PNI, prognostic nutritional index; HALP, hemoglobin–albumin–lymphocyte–platelet score.

This conclusion runs counter to the literature that is itself far less unanimous than its volume suggests. Studies relating hematological indices to *H. pylori* infection have produced contradictory results; the discordance extends even to the direction of the association. Guclu and Agan, for instance, found that NLR was significantly lower, not higher, in *H. pylori*-positive patients, with no accompanying difference in mean platelet volume (10). Erman, in a large retrospective cohort of dyspeptic patients with biopsy-confirmed *H. pylori* status, found no significant difference in either NLR or PLR between infected and uninfected groups (11). When the broader discordance of the field is taken into account, our findings are less an outlier than a clarification: where most studies have reported one significant marker among many and built a claim upon it, a comprehensively tested panel subjected to correction for multiple comparisons reveals that the underlying signal is, for practical purposes, absent (7).

That these indices succeed elsewhere but fail here is instructive rather than contradictory. In gastric adenocarcinoma, inflammatory ratios retain prognostic value and the same is true across numerous malignancies and inflammatory states (12). The distinction lies in the biology being interrogated. A tumor represents a sustained, high-volume systemic stimulus: necrosis, cytokine release, and a measurable shift in the circulating leukocyte pool. Mucosal gastritis, even in its active neutrophilic form, is a spatially confined process in which neutrophils are recruited locally to the epithelium; the amount of inflammatory traffic it generates is simply too small and too compartmentalized to register against the noise of peripheral blood. The indices are not invalid; they are being asked to detect a signal below their resolution.

The one finding that did emerge (the younger age of patients with active, *H. pylori*-associated gastritis) is best understood not as a biomarker but as a cross-sectional glimpse of the Correa cascade (3). The natural history of gastric disease is age-dependent: *H. pylori*-associated active inflammation predominates earlier and as the mucosa progresses toward atrophy and metaplasia the organism is frequently lost from a hostile, hypochlorhidric environment, so that

detectable infection paradoxically declines with age while atrophic and metaplastic changes accumulate (13). This age-stratified drift is well documented, with atrophy and OLGA stage rising in older cohorts and with the atrophy–metaplasia axis defining the trajectory toward neoplasia (14,15). Our data recapitulate this sequence in a contemporary surgical-endoscopic population. The clinical corollary is sobering for the index hypothesis: the limited discriminatory power for detecting active inflammation is captured by date of birth, and the blood adds nothing to it.

Perhaps the most pointed result concerns the immature granulocytes. We had reasoned that emergency granulopoiesis (the bone marrow's release of immature neutrophil precursors) might be the systemic event most likely to reflect mucosal neutrophilic activity, and we constructed several indices around it. Yet, immature granulocytes, whether raw or combined, were inert. The explanation is consistent with their established biology: immature granulocytes rise in response to substantial systemic infection, sepsis and major tissue injury, conditions categorically different in scale from the focal recruitment of neutrophils to gastric pits (16,17). In our view, the fact that even this most mechanistically plausible candidate failed constitutes the strongest single piece of evidence for compartmentalization: local mucosal and systemic compartments behaved independently.

Several features lend our null result more weight than is customary. The panel was deliberately broad, so that a true association would have had many opportunities to appear; the Benjamini–Hochberg correction guarded against the multiple-comparison artifacts that pervade this literature; the DeLong comparison established that age was statistically superior to all indices except SII, for which the advantage was only numerical; and fivefold cross-validation confirmed that adding the blood panel to age degraded, rather than improved, out-of-sample discrimination. Importantly, the dataset was not inert (it detected the age signal clearly), so the absence of an index effect cannot be dismissed as being due to insensitivity of the method.

These observations have practical implications. Hematological inflammation indices are inexpensive and reflexively ordered, and there is a temptation to

read mucosal significance into them. Our data counsel against this: a normal or elevated NLR or SII does not reliably indicate the presence of gastritis, its activity, *H. pylori* status or subtype, and should not influence the decision to biopsy or the interpretation of mucosal disease.

### Study limitations

This study has several limitations. Its retrospective, single-center, cross-sectional design cannot establish temporal or causal relationships. Clinical confounders relevant to reactive gastropathy (nonsteroidal anti-inflammatory drug use, proton pump inhibitor use, and bile reflux exposure) were not systematically recorded, which may have blurred the characterization of that subtype. The near-perfect collinearity between active inflammation and *H. pylori* ( $\kappa = 0.98$ ) means that, although listed as separate endpoints, they essentially interrogate a single biological entity. Atrophy was infrequent ( $n = 13$ ), precluding robust modeling of that endpoint. All blood counts and index values derived from a single hematology analyzer may not be directly transferable across platforms. Finally, these findings are hypothesis-clarifying rather than definitive and await confirmation in prospective, multicenter cohorts.

These findings carry three practical implications. Clinically, hematological inflammation indices should not be read as surrogates for mucosal disease: in a patient with dyspeptic symptoms, a raised NLR or SII conveys nothing about the presence or activity of gastritis, its *H. pylori* status, or its subtype, and endoscopy with biopsy remains irreplaceable for that purpose. Methodologically, the discordance between our null result and the positive associations reported elsewhere illustrates how readily an uncorrected panel of correlated indices generates significant findings. Studies in this field should report the full panel tested, apply corrections for multiple comparisons, and benchmark any candidate marker against simple demographic variables before claiming discriminatory value. For future research, prospective designs incorporating OLGA and OLGIM staging would allow the question to be reframed in terms of cumulative mucosal damage rather than a binary histopathological label; serial measurement before and after eradication

therapy could test whether a dynamic change carries information that a single cross-sectional value does not; and direct comparison with mucosa-specific serological markers such as pepsinogen I and II would clarify whether the failure of peripheral indices reflects the limitations of the blood count itself or the compartmentalization of mucosal inflammation more broadly.

### CONCLUSION

Across a broad, rigorously analyzed panel of hematological indices of inflammation, none reflected the histopathological state of the gastric mucosa: neither active nor chronic inflammation, nor *H. pylori* infection, nor atrophy, nor metaplasia, nor gastritis subtype. The only variable carrying reproducible discriminatory information was age, an expression of the age-dependent natural history of gastric disease rather than a systemic inflammatory signal, and one that the entire blood panel failed to improve upon. For the assessment of gastritis, the periphery is a poor mirror of the mucosa: a complete blood count—regardless of how its components are combined—is no substitute for histopathological examination. Whether any peripheral signature can capture localized mucosal inflammation remains an open question for prospective multicenter studies.

### Ethical approval

This study has been approved by the Clinical Research Ethics Committee of Bolu Abant İzzet Baysal University (approval date 07.04.2026, number 2026/159). Owing to the retrospective design of the study, the requirement for written informed consent was waived by the ethics committee. The study was conducted in accordance with the principles of the Declaration of Helsinki.

### Author contribution

Concept: MFK; Design: MFK; Data Collection or Processing: MFK, FB; Analysis or Interpretation: MFK; Literature Search: FB; Writing: MFK, FB. All authors reviewed the results and approved the final version of the article.

## Source of funding

The authors declare the study received no funding.

## Conflict of interest

The authors declare that there is no conflict of interest.

## REFERENCES

1. Zhang SL, Lollie TK, Chen Z, Narasimhalu T, Wang HL. Histopathologic diagnosis of gastritis and gastropathy: a narrative review. *Dig Med Res.* 2023; 6: 9. [\[Crossref\]](#)
2. Mansour-Ghanaei F, Joukar F, Yeganeh S, Sadeghi M, Daryakar A, Sepehrimanesh M. OLGA- and OLGIM-based staging in the patients with gastritis and endoscopy indications. *Turk J Gastroenterol.* 2022; 33(2): 95–102. [\[Crossref\]](#)
3. Zhou J, Li J, Chen J, et al. Decoding inflammatory mediators in the Correa's cascade: from chronic gastritis to carcinogenesis and targeted therapies. *Int Immunopharmacol.* 2025; 162: 115191. [\[Crossref\]](#)
4. Romańczyk M, Osmola M, Link A, et al. Non-invasive markers for the detection of gastric precancerous conditions. *Cancers (Basel).* 2024; 16(12): 2254. [\[Crossref\]](#)
5. Elgenidy A, Alomari O, Emad T, et al. Systemic immune-inflammation index: unveiling the diagnostic potential in ulcerative colitis through a comprehensive systematic review and meta-analysis. *Gastroenterol Endosc.* 2025; 3(1): 22–31. [\[Crossref\]](#)
6. Zhang Y, Yue Y, Sun Z, et al. Pan-immune-inflammation value and its association with all-cause and cause-specific mortality in the general population: a nationwide cohort study. *Front Endocrinol (Lausanne).* 2025; 16: 1534018. [\[Crossref\]](#)
7. Ordu M, Ozturk G, Ergin M. The correlation between systemic immune-inflammatory index and *Helicobacter pylori* infection and its severity. *Medical Bulletin of Haseki.* 2025; 63(3): 150–8. [\[Crossref\]](#)
8. Zhang J, Zhang L, Duan S, Li Z, Li G, Yu H. Single and combined use of the platelet-lymphocyte ratio, neutrophil-lymphocyte ratio, and systemic immune-inflammation index in gastric cancer diagnosis. *Front Oncol.* 2023; 13: 1143154. [\[Crossref\]](#)
9. Coşgun S, Aras Z. Assessment of the relationship between neutrophil-lymphocyte ratio and dyspeptic symptoms in patients with peptic ulcer diagnosed by endoscopy and patients without peptic ulcer. *Cureus.* 2023; 15(10): e46820. [\[Crossref\]](#)
10. Guclu M, Faruq Agan A. Association of severity of *Helicobacter pylori* infection with peripheral blood neutrophil to lymphocyte ratio and mean platelet volume. *Euroasian J Hepatogastroenterol.* 2017; 7(1): 11–6. [\[Crossref\]](#)
11. Boyuk B, Saydan D, Mavis O, Erman H. Evaluation of *Helicobacter pylori* infection, neutrophil-lymphocyte ratio and platelet-lymphocyte ratio in dyspeptic patients. *Gastroenterol Insights.* 2020; 11(1): 2–9. [\[Crossref\]](#)
12. Nechita VI, Al-Hajjar N, Leucuța DC, et al. Inflammatory ratios as survival prognostic factors in resectable gastric adenocarcinoma. *Diagnostics (Basel).* 2023; 13(11): 1910. [\[Crossref\]](#)
13. Chen S, Ying L, Kong M, Zhang Y, Li Y. The prevalence of *Helicobacter pylori* infection decreases with older age in atrophic gastritis. *Gastroenterol Res Pract.* 2013; 2013: 494783. [\[Crossref\]](#)
14. Xia J, Liu Z, Zhang K. Pepsinogen serology and gastritis OLGA staging in mucosal atrophy assessment: a cross-sectional study involving East China endoscopy population. *Gastroenterol Res Pract.* 2020; 2020: 2324505. [\[Crossref\]](#)
15. Burke E, Harkins P, Arumugasamy M. Incidence of gastric adenocarcinoma in those with gastric atrophy: a systematic review. *Cureus.* 2024; 16(10): e71768. [\[Crossref\]](#)
16. Alay GH, Tatlisuluoglu D, Bulut K, Fikri BI, Oztas A, Turan G. The relationship between immature granulocyte count and mortality in ARDS due to COVID-19. *Niger J Clin Pract.* 2022; 25(8): 1301–7. [\[Crossref\]](#)
17. Cetin N, Kocaturk E, Tufan AK, Kiraz ZK, Alatas O. Diagnostic values of immature granulocytes detected by the Sysmex XN 9000 hematology analyzer in children with urinary tract infections. *Saudi J Kidney Dis Transpl.* 2023; 34(Suppl 1): S133–41. [\[Crossref\]](#)

# Catamenial bilateral conjunctival hemorrhage associated with endometriosis: A rare case report

Elif Betül Esmer<sup>1</sup>

<sup>1</sup>Department of Obstetrics and Gynecology, İzzet Baysal State Hospital, Bolu, Türkiye

**Cite as:** Esmer EB. Catamenial bilateral conjunctival hemorrhage associated with endometriosis: A rare case report. Northwestern Med J. 2026;6(4):364-367.

## ABSTRACT

**Objective:** To present a rare case of catamenial bilateral conjunctival hemorrhage associated with endometriosis in a patient with ovarian endometrioma and to discuss its possible hormonal pathogenesis.

**Case:** A 24-year-old nulliparous woman presented with a 10-year history of severe dysmenorrhea and recurrent bilateral subconjunctival hemorrhage occurring exclusively during menstruation. Ophthalmologic examinations revealed no additional pathology. Pelvic ultrasonography demonstrated a 70×55-mm homogeneous cystic lesion with low-level internal echoes in the right ovary, consistent with an endometrioma. Serum CA-125 was mildly elevated (48.5 U/mL). Orbital magnetic resonance imaging was normal. Combined oral contraceptive therapy containing drospirenone and ethinylestradiol was initiated, resulting in marked improvement in both dysmenorrhea and ocular hemorrhage.

**Conclusion:** Catamenial conjunctival hemorrhage is a rare but clinically relevant manifestation that may be hormonally mediated. In patients presenting with cyclic ocular bleeding, underlying endometriosis should be considered.

**Keywords:** endometriosis, endometrioma, conjunctival hemorrhage, menstruation, dysmenorrhea

## INTRODUCTION

Endometriosis is a chronic, estrogen-dependent condition in which endometrial glands and stroma are found outside the uterine cavity. It affects approximately 10% of women of reproductive age (1). The disease mainly involves the pelvic organs, although endometriotic lesions can occasionally occur at extragenital sites (2). Ovarian endometrioma is a common manifestation of endometriosis and is

frequently associated with dysmenorrhea, chronic pelvic pain, and infertility (2). Ocular manifestations related to menstruation are uncommon and have mainly been reported as individual case reports (3). Recurrent subconjunctival hemorrhage has been described in association with the menstrual cycle, with the ocular findings appearing during menstruation and resolving afterward. In reported cases, ophthalmologic examination generally does not reveal an underlying ocular abnormality (3). The mechanism responsible for

**Corresponding author:** Elif Betül Esmer **E-mail:** betuloneren@gmail.com

**Received:** 11.05.2026 **Accepted:** 27.08.2026 **Published:** 25.10.2026

This manuscript has been previously published as a preprint on the Authorea preprint platform. <https://doi.org/10.22541/au.177080606.66731070/v1>

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

this cyclic bleeding is not well established. Hormonal changes during the menstrual cycle, including possible effects on vascular permeability and capillary fragility, have been suggested as one possible explanation (4).

In this report, we present a young patient with long-standing dysmenorrhea and recurrent bilateral conjunctival hemorrhage during menstruation, in whom a concomitant ovarian endometrioma, representing ovarian endometriosis, was identified, and we discuss the case in light of the current literature.

## CASE REPORT

A 24-year-old nulliparous woman presented with a 10-year history of severe dysmenorrhea and recurrent bilateral ocular redness that occurred exclusively during menstruation. The patient reported that ocular symptoms appeared only during menstrual periods and resolved spontaneously after completion of the cycle (Figure 1).

Repeated ophthalmologic examinations revealed bilateral subconjunctival hemorrhage. Visual acuity, anterior segment examination, and fundus evaluation were all normal. No additional ophthalmologic pathology, other than subconjunctival hemorrhage, was detected. Evaluation for systemic bleeding disorders revealed no abnormalities.

Pelvic ultrasonography performed during the menstrual period demonstrated an anteverted uterus with an endometrial thickness of 3 mm. The left ovary appeared normal, while the right ovary contained a well-circumscribed cystic lesion measuring approximately 70×55 mm, with homogeneous low-level internal echoes and regular walls, consistent with an ovarian endometrioma (2) (Figure 2). Laboratory evaluation revealed a mildly elevated serum CA-125 level of 48.5 U/mL, while all other biochemical and hematological parameters were within normal limits (5).

Orbital magnetic resonance imaging, performed to investigate the etiology of the ocular findings, revealed no pathological signal changes, masses, or vascular malformations. Based on these findings, secondary causes of conjunctival hemorrhage were excluded.

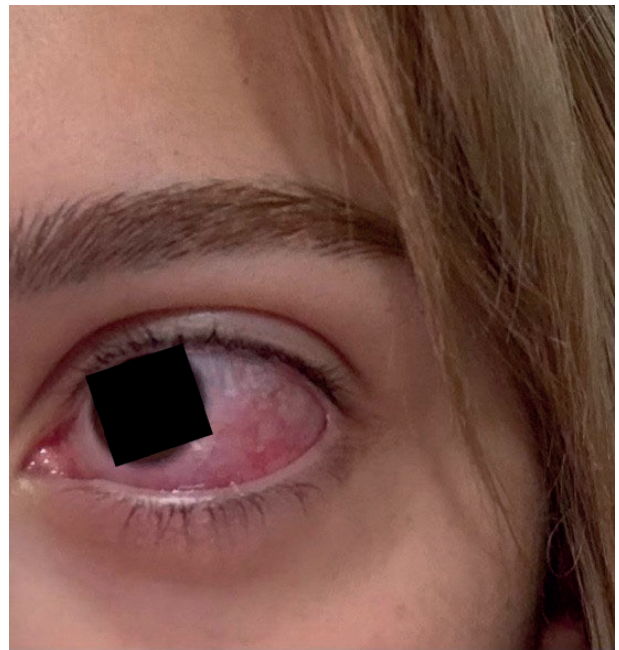


**Figure 1.** Conjunctival hemorrhage during menstruation.

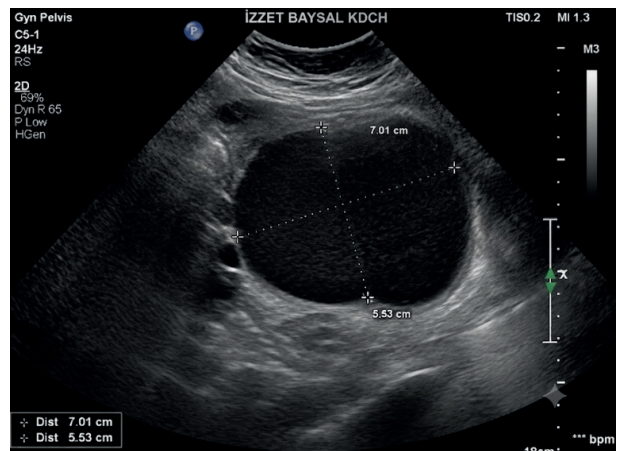


**Figure 2.** Transabdominal ultrasonographic image of the right ovarian endometrioma demonstrating homogeneous low-level internal echoes.

The patient was started on combined oral contraceptive therapy containing drospirenone and ethinylestradiol. From the third month of treatment onward, a marked improvement in dysmenorrhea and a significant reduction in menstruation-related conjunctival hemorrhage were observed (Figure 3).

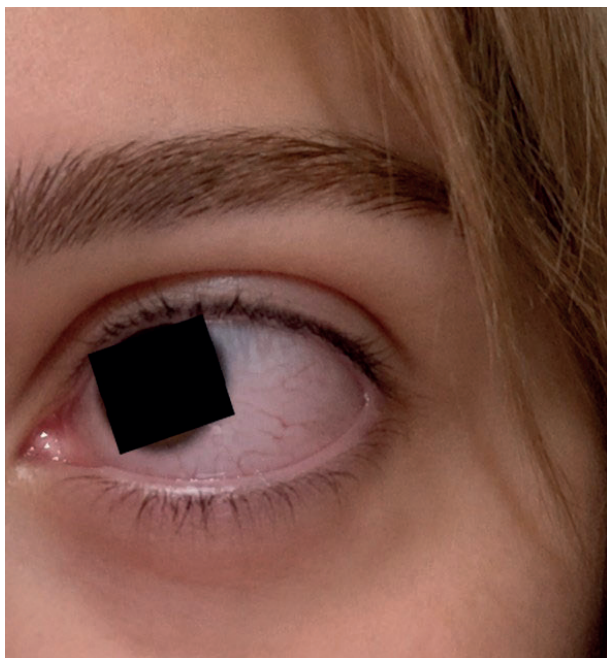


**Figure 3.** Marked regression of subconjunctival hemorrhage during menstruation after initiation of combined oral contraceptive therapy.

The patient remains under regular clinical follow-up. Written informed consent was obtained from the patient for publication of this case report and accompanying images.

The study was conducted in accordance with the ethical standards of the institutional research committee and the 1975 Declaration of Helsinki, as revised in 2000.

## DISCUSSION

Very few cases of catamenial conjunctival hemorrhage have been reported to date, and the existing literature is limited to isolated case reports (3,6–9). What links these cases is the close temporal relationship with the menstrual cycle; in most patients described, no ophthalmologic abnormality was identified beyond the subconjunctival hemorrhage. The first description came from Abboud and Hanna, who documented five adolescent girls with recurrent conjunctival bleeding that tracked their menstrual cycles and resolved on its own within days (6). Barat and Kwedar later described a 13-year-old girl whose recurrent cyclic subconjunctival

hemorrhage responded well to hormonal treatment (7). More recently, Ghosh et al. described a young woman with bilateral recurrent conjunctival hemorrhage but no identifiable ocular pathology (9), and Celebi and Aygun reported a case of recurrent unilateral subconjunctival hemorrhage that improved markedly with combined oral contraceptive therapy (3). Our patient's presentation fits this pattern: recurrent, menstruation-linked conjunctival hemorrhage without any accompanying ophthalmologic abnormality. What sets our case apart from most previously reported, however, is the presence of a concomitant ovarian endometrioma, pointing to underlying endometriosis and lending support to a possible link between pelvic endometriosis and this unusual ocular finding. Pelvic ultrasonography in our patient showed a homogeneous cystic lesion with low-level internal echoes, a well-recognized and highly specific sonographic feature of ovarian endometrioma (2). This was corroborated by a mildly elevated serum CA-125 level, lending further support to the diagnosis of endometriosis (5). The exact pathophysiological mechanism underlying catamenial conjunctival hemorrhage remains incompletely understood. It has been proposed that cyclical fluctuations in estrogen and progesterone increase conjunctival vascular permeability and capillary fragility, predisposing susceptible individuals to recurrent subconjunctival hemorrhage during menstruation (3,4). Furthermore, endometriosis is characterized by chronic inflammation, enhanced angiogenesis, and increased expression of vascular endothelial growth factor (VEGF), all of which may contribute to vascular instability. Another hypothesis suggests the presence of hormonally responsive ectopic endometrial tissue at extra-pelvic sites through hematogenous or lymphatic dissemination. However, histopathological confirmation is rarely available in ocular cases because conjunctival biopsy is generally not indicated in patients with self-limited disease (2–4).

Before attributing ocular bleeding to a catamenial pattern, other causes—trauma, coagulopathies, vascular malformations, infections, and orbital masses—must be ruled out. We addressed this in our patient with repeated ophthalmologic evaluations and orbital MRI, both of which were normal and effectively excluded these secondary causes. There is no single

treatment protocol; management must be tailored to the patient. That said, earlier reports point consistently to hormonal suppression, whether through combined oral contraceptives or GnRH analogues, as an effective way to control symptoms (3,4). Our patient's response supports the hypothesis that suppression of hormonal cycling attenuates the vascular changes underlying recurrent bleeding. The fact that both her dysmenorrhea and ocular symptoms improved after she started combined oral contraceptives further strengthens the case for a shared hormonal mechanism linking endometriosis to conjunctival hemorrhage. Catamenial conjunctival hemorrhage may be rare, but its occurrence alongside an ovarian endometrioma, as in our patient, suggests it is more than an isolated ocular curiosity; rather, it may reflect a broader, hormonally driven process. Therefore, endometriosis should be considered in the differential diagnosis of patients presenting with cyclic ocular bleeding. Catching this connection early carries practical value: it can prompt timely gynecological referral, spare patients from unnecessary ophthalmologic workups, and start appropriate hormonal treatment sooner.

### Ethical approval

Ethics committee approval was not required for this case report. Written informed consent was obtained from the patient for the publication of this case report and the accompanying images.

### Author contribution

Surgical and Medical Practices: EBE; Concept: EBE; Design: EBE; Data Collection or Processing: EBE; Analysis or Interpretation: EBE; Literature Search: EBE; Writing: EBE. The author reviewed the results and approved the final version of the article.

### Source of funding

The authors declare the study received no funding.

### Conflict of interest

The authors declare that there is no conflict of interest.

### Declaration on the Use of AI

During the preparation of this manuscript, the author(s) used an artificial intelligence-based language tool solely to improve the English language, grammar, and readability of the text. The tool was not used to generate scientific content, data, analysis, or conclusions. After using this tool, the author(s) reviewed and edited the content as needed and took full responsibility for the content of the publication.

### REFERENCES

1. Giudice LC, Kao LC. Endometriosis. *Lancet*. 2004; 364(9447): 1789-99. [\[Crossref\]](#)
2. Vercellini P, Viganò P, Somigliana E, Fedele L. Endometriosis: pathogenesis and treatment. *Nat Rev Endocrinol*. 2014; 10(5): 261-75. [\[Crossref\]](#)
3. Celebi ARC, Aygun EG. A rare cause of recurrent subconjunctival hemorrhage: ocular vicarious menstruation. *GMS Ophthalmol Cases*. 2023; 13: Doc05. [\[Crossref\]](#)
4. Bulun SE. Endometriosis. *N Engl J Med*. 2009; 360(3): 268-79. [\[Crossref\]](#)
5. Hirsch M, Duffy J, Davis CJ, Nieves Plana M, Khan KS, International Collaboration to Harmonise Outcomes and Measures for Endometriosis. Diagnostic accuracy of cancer antigen 125 for endometriosis: a systematic review and meta-analysis. *BJOG*. 2016; 123(11): 1761-8. [\[Crossref\]](#)
6. Abboud IA, Hanna LS. Bleeding from the conjunctiva. *Br J Ophthalmol*. 1971; 55(7): 487-91. [\[Crossref\]](#)
7. Barat M, Kwedar SA. Ocular vicarious menstruation. *J Pediatr Ophthalmol Strabismus*. 1988; 25(5): 254-5. [\[Crossref\]](#)
8. Dua P, Pointdujour R, Reich I, Lazzaro DR, Shinder R. Orbital vicarious menstruation. *Ophthalmic Plast Reconstr Surg*. 2014; 30(2): e35-7. [\[Crossref\]](#)
9. Ghosh S, Tale S, Handa N, Bhalla A. Rare case of red tears: ocular vicarious menstruation. *BMJ Case Rep*. 2021; 14(3): e237294. [\[Crossref\]](#)

# **NORTHWESTERN MEDICAL JOURNAL**

[www.nwmedj.org](http://www.nwmedj.org)

ISSN: 2979-9538