Odontology
https://doi.org/10.1007/s10266-026-01339-x

ORIGINAL ARTICLE

Clinical decision accuracy in endodontic treatment of patients with systemic diseases: a comparative analysis using different artificial intelligence models

Ayşegül Eroğlu[1] · İpek Eraslan Akyüz[1] · Emre Yılmaz[2] · Salih Düzgün[1]

Received: 6 January 2026 / Accepted: 28 January 2026
© The Author(s) 2026

Abstract
This study aims to compare the clinical decision-making accuracy of different artificial intelligence (AI) models in endodontic treatment planning for patients with systemic diseases. A scenario-based, cross-sectional educational study was conducted using 40 standardized clinical scenarios representing ten commonly encountered systemic conditions affecting endodontic care. Scenarios were developed based on international endodontic and medical guidelines and reviewed by medical specialists and experienced endodontists. Four AI models, ChatGPT-5.1, Gemini 2.5 Pro, Gemini 2.5 Flash, and ChatGPT-3.5, were queried using identical, standardized prompts within fully isolated interaction environments to prevent contextual memory effects. AI-generated responses were independently evaluated by two calibrated endodontists using a predefined 10-point scoring system across four clinical domains. Clinical accuracy was categorized as high, partial, or incorrect. Nonparametric statistical analyses were performed. No statistically significant differences were observed among AI models in overall clinical decision accuracy or domain-specific scores (Friedman test, $p > 0.05$). Although categorical analysis revealed an overall difference in the proportion of high-accuracy responses (Cochran's Q, $p = 0.007$), post hoc comparisons did not demonstrate significant pairwise differences. Deviation analysis revealed comparable proximity of all models to the expert-defined optimal decisions, with greater variability observed for the Gemini 2.5 Flash. Current AI models demonstrate comparable clinical decision-making performance in endodontic scenarios involving medically compromised patients. While descriptive trends were observed, no single model consistently outperformed others. AI systems may serve as supportive decision-making tools when used under professional supervision, but should not replace clinical judgment.

Keywords Artificial intelligence · Endodontics · Large language models · Systemic diseases

Introduction

Endodontic treatment is a cornerstone of contemporary dental practice, aiming to eliminate pulpal infection, preserve tooth structure, and maintain oral function [1, 2]. The presence of systemic diseases can significantly influence both the pathophysiology of endodontic infections and the clinical management strategies, potentially increasing the treatment complexity and patient risk [3, 4]. Conditions such as cardiovascular disease, diabetes mellitus, chronic kidney disease, hepatic disorders, and immunosuppressive therapy are prevalent in the general population. They can affect hemostasis, wound healing, and pharmacologic responses during endodontic treatment [5, 6]. For example, diabetes mellitus has been shown to increase the risk of apical periodontitis and reduce the retention of root-filled teeth. A meta-analysis reported that diabetic patients had a significantly higher odds of extracted root-filled teeth compared to non-diabetics [7]. Moreover, diabetic patients often exhibit delayed healing, reduced immune response, and greater susceptibility to persistent periapical lesions [8]. In patients on anticoagulant therapy, the risk of bleeding is a critical concern. Studies indicate that direct oral anticoagulants and vitamin K antagonists are associated with an increased risk of bleeding during dental and oral surgery; however, most

✉ Salih Düzgün
dtsalihduzgun@gmail.com

1 Department of Endodontics, Faculty of Dentistry, Erciyes University, 38039 Melikgazi, Kayseri, Türkiye

2 Department of Family Medicine, Faculty of Medicine, Erciyes University, Kayseri, Türkiye

Published online: 11 February 2026

bleeding events are mild and manageable with local hemostatic measures [9].

Despite extensive clinical guidelines, variability in treatment decisions exists, particularly in medically compromised patients, leading to potential over- or undertreatment. Recently, artificial intelligence (AI) has emerged as a promising adjunct in dental decision-making, offering rapid, data-driven recommendations based on large-scale clinical knowledge [10, 11]. However, the accuracy of AI-assisted clinical decision-making, especially in patients with systemic conditions, remains largely unexplored.

The present study aims to evaluate the performance of four artificial intelligence models (ChatGPT-5.1, Gemini 2.5 Pro, Gemini 2.5 Flash, and ChatGPT-3.5) in endodontic treatment decision-making for patients with systemic diseases. The study examines whether the AI models differ from expert endodontists in terms of accuracy and clinical decision-making. The null hypothesis of this study is that there is no significant difference among the artificial intelligence models in terms of the accuracy and clinical evaluation of endodontic treatment decisions.

Materials and methods

Study design

This scenario-based, cross-sectional educational study was conducted to evaluate the performance of artificial intelligence models in endodontic treatment decision-making for patients with systemic diseases. The study specifically aimed to assess the degree of alignment between AI-generated recommendations and expert endodontist decisions across different systemic scenarios. The study was reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines [12]. Ethical approval was not required, as only AI-generated data were used without involving human or animal participants. However, the study adhered to general ethical principles, maintaining research integrity and data confidentiality. A total of 40 different clinical scenarios representing the most commonly encountered systemic conditions in patients presenting to an endodontic clinic were developed. Scenario construction was based on current international guidelines and position statements addressing the dental and endodontic management of patients with systemic diseases, including recommendations from the American Association of Endodontists (AAE) [13], the European Society of Endodontology (ESE) [14], and relevant medical guidelines for systemic disease management [15]. These scenarios were created through a collaborative effort involving five medical specialists from various fields (cardiology, endocrinology, nephrology, neurology, and internal medicine), as well as three experienced endodontists. All scenarios underwent pretesting by endodontists and relevant medical specialists to ensure clinical validity and accuracy.

Each scenario was structured to reflect realistic clinical practice, including patient age, sex, medical history, current medications, dental complaints, radiographic and clinical endodontic findings, and the planned endodontic procedure. Particular attention was given to how systemic conditions and medications might influence treatment planning and outcomes.

Systemic conditions

The study focused on ten systemic conditions that are frequently encountered and can affect endodontic treatment:

- Diabetes mellitus
- Hypertension
- Thyroid disorders
- Cardiovascular system diseases
- Asthma
- Epilepsy
- Pulmonary embolism (PE)
- Chronic kidney disease (CKD)
- Osteoporosis
- Hepatitis (chronic hepatitis B and C)

For each condition, four clinical scenarios were created, resulting in a total of 40 scenarios.

AI models and prompting

To ensure methodological rigor and to minimize potential carryover effects from prior responses, each clinical scenario was evaluated within a fully isolated interaction environment. For every model, each scenario was treated as an independent diagnostic event, and a new conversation window was initiated before each evaluation. To further prevent any influence of previous outputs, chat histories and messages were manually cleared between sessions, and the "clear chat history" command was applied, ensuring that the models' contextual memory was fully reset before processing each case. All scenarios involved patients diagnosed with symptomatic apical periodontitis, for whom nonsurgical root canal treatment was planned as the definitive dental intervention. An example of each systemic disease included in the study is presented in Table 1. This approach guaranteed the independence of all responses and prevented information transfer between scenarios.

A total of 40 clinical scenarios were evaluated using four artificial intelligence models: ChatGPT-5.1 (OpenAI, USA), Gemini 2.5 Pro (Google DeepMind, USA), Gemini 2.5 Flash (Google DeepMind, USA), and ChatGPT-3.5

Table 1 Examples of systemic diseases included in the study

No	Systemic diseases	Clinical scenarios
1	Diabetes mellitus	A 20-year-old female patient presented to the endocrinology clinic with complaints of dry mouth, excessive thirst (polydipsia), frequent urination (polyuria), and sudden weight loss at the age of 6. Laboratory investigations revealed fasting plasma glucose ≥ 126 mg/dL, HbA1c ≥ 6.5%, and decreased C-peptide levels. Based on these findings, the patient was diagnosed with type 1 diabetes mellitus. Multiple-dose insulin therapy (basal-bolus regimen) was initiated, and education on diabetes management was provided. The patient then presented to the endodontics clinic with pain in her lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, was planned for the affected tooth
2	Hypertension	A 50-year-old female patient presented to the internal medicine clinic with complaints of morning headaches and easy fatigue. Blood pressure measurements showed an average of 155/95 mmHg, and secondary causes were ruled out. The patient was diagnosed with primary hypertension stage 2. In addition to lifestyle changes, the patient was started on monotherapy with antihypertensive medication (ACE inhibitor). The patient then presented to the endodontics clinic with a complaint of pain in her lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, was planned for the affected tooth
3	Thyroid disorders	A 40-year-old male patient presented to the internal medicine outpatient clinic with complaints of palpitations, sweating, hand tremors, and weight loss. Laboratory investigations revealed TSH levels of < 0.01 µIU/mL, high free T4, and elevated T3 levels. Anti-TPO and TSH receptor antibodies were positive. The patient was diagnosed with hyperthyroidism due to Graves' disease. He was started on thiamazole (antithyroid drug) treatment, and clinical and laboratory follow-up was recommended. The patient then presented to the endodontics clinic with a complaint of pain in his lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a bleeding and infectious procedure, was planned for the affected tooth
4	Cardiovascular system diseases	A 65-year-old male patient presented to the cardiology outpatient clinic with complaints of significant shortness of breath, easy fatigue, and peripheral edema during daily activities. Physical examination revealed bilateral ankle edema and hepatomegaly. Echocardiography showed an LVEF of 45%, and radiography revealed pulmonary congestion. Laboratory findings showed NT-proBNP levels of 850 pg/mL and serum creatinine levels of 1.2 mg/dL. The patient was diagnosed with NYHA Class II heart failure and started on a combination of diuretic (Furosemide 40 mg/day) and ACE inhibitor (Lisinopril 10 mg/day). The patient then presented to the endodontics clinic with pain in his lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, was planned for the affected tooth
5	Asthma	A 50-year-old male patient presented with complaints of daily shortness of breath, cough, and exercise restriction. He had experienced three acute exacerbations in the past year. Spirometry revealed FEV_1 of 65% and PEF variability of 35%. The patient was diagnosed with moderate persistent (stage 3) asthma, and combination therapy was initiated: medium-dose inhaled corticosteroid (ICS) + long-acting beta-2 agonist (LABA): Budesonide/Formoterol 320/9 µg, twice daily, and SABA (Salbutamol) as needed. The patient presented to the endodontics clinic with a complaint of pain in his lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a bleeding and infectious procedure, was planned for the affected tooth
6	Epilepsy	A 28-year-old female patient experienced generalized tonic–clonic seizures for the first time 2 years ago and was newly diagnosed with levetiracetam 500 mg/day. She has experienced only one seizure in the last 2 months and is under neurological follow-up. Physical examination revealed no abnormalities. The patient presented to the endodontics clinic with pain in her lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, was planned for the affected tooth
7	Pulmonary embolism	A 58-year-old female patient presented to the emergency department 18 months prior with complaints of sudden onset back pain and weakness. Evaluation revealed stable vital signs, but mild tachycardia (HR: 105/min) and oxygen saturation of 94%. CTPA was performed, revealing emboli in the segmental and sub-segmental pulmonary arteries. Laboratory investigations showed elevated D-dimer levels, while troponin levels were within normal limits. Due to hemodynamic stability, the patient was initially treated with IV heparin or LMWH, followed by DOAC. No complications were observed during follow-up. After 18 months of follow-up, the patient presented to the endodontics clinic with complaints of pain in her lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, has been planned for the affected tooth

Table 1 (continued)

No	Systemic diseases	Clinical scenarios
8	Chronic kidney disease	A 60-year-old female patient presented to the nephrology outpatient clinic with complaints of fatigue, loss of appetite, and frequent urination. Her history included 8 years of type 2 diabetes and hypertension; she was being treated with metformin and an ACE inhibitor. Laboratory findings: BUN: 48 mg/dL, creatinine: 2.0 mg/dL, GFR: 35 mL/min/1.73 m^2, HbA1c: 8.0%. Kidney ultrasound showed bilateral mild hypertrophy and parenchymal thinning. Given her GFR range of 30–59 mL/min/1.73 m^2 and the available ultrasound findings, the patient was diagnosed with stage 3 chronic kidney disease. The patient also presented to the endodontics clinic with pain in her lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, has been planned for the affected tooth
9	Osteoporosis	A 50-year-old male patient presented to the outpatient clinic with complaints of increasing diffuse bone pain in recent months, tenderness in the lumbar and hip regions, particularly noticeable after prolonged standing, as well as height loss and postural changes. Physical examination revealed pain on palpation in the thoracolumbar region, a slightly kyphotic posture, and decreased muscle tone. Investigations based on these clinical findings showed a previously recorded lumbar T-score of -2.4. This value, being below -1, suggests clinically significant bone mineral density loss, and being very close to -2.5 suggests an increased risk of fracture, which may lead to a diagnosis of osteoporosis. Therefore, the patient received intravenous Zoledronic Acid treatment, which was completed 5 years ago. Laboratory results supporting the diagnosis revealed low 25-OH vitamin D levels (18 ng/mL), serum calcium levels (8.6 mg/dL) at the lower limit of normal, and a slightly elevated PTH level (92 pg/mL), all consistent with the pathophysiology of osteoporosis. The patient presented to the endodontics clinic with pain in the lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, was planned for the affected tooth
10	Chronic hepatitis	A 45-year-old female patient has been experiencing mild jaundice, weakness, and loss of appetite for the past month. Liver ultrasonography revealed a slightly enlarged liver with heterogeneous parenchyma and mild hepatomegaly. Laboratory values were measured as AST 95 U/L, ALT 108 U/L, ALP 155 U/L, and total bilirubin 1.8 mg/dL. She was identified as HBsAg positive, but had not yet started antiviral treatment. The patient presented to the endodontics clinic with pain in her lower left first molar (tooth number 36). Following an examination by an endodontist, root canal treatment, a procedure involving bleeding and infection, was planned for the affected tooth

HbA1c; glycated hemoglobin; *ACE inhibitor* angiotensin-converting enzyme inhibitor; *TSH* thyroid-stimulating hormone; *Anti-TPO* anti-thyroid peroxidase antibody; *LVEF* left ventricular ejection fraction; *NT-proBNP* N-terminal pro-B-type natriuretic peptide; FEV_1 forced expiratory volume in one second; *PEF* peak expiratory flow; *ICS* inhaled corticosteroid; *LABA* long-acting beta-2 agonist; *SABA* short-acting beta-2 agonist; *CTPA* computed tomography pulmonary angiography; *LMWH* low-molecular-weight heparin; *DOAC* direct oral anticoagulant; *BUN* blood urea nitrogen; *GFR* glomerular filtration rate; *PTH* parathyroid hormone

(OpenAI, USA). All models were queried using standardized and identical prompts to maintain consistency across evaluations.

For every scenario, the following four prompts were applied sequentially (Table 2).

All responses generated by the artificial intelligence models were recorded verbatim and stored for subsequent scoring and statistical analysis.

Scoring and evaluation

The responses generated by each AI model were independently evaluated by two experienced endodontists using a structured scoring framework designed to capture the essential components of clinical decision-making in medically compromised patients. The evaluation was conducted in a double-blind design to minimize assessor bias, and discrepancies were resolved by consensus. The scoring system consisted of four domains: (1) evaluation of the

Table 2 Sequential prompts applied to large language models for each clinical scenario

Prompt No.	Prompt content
Prompt 1	Considering this patient's systemic condition, how should the endodontic situation be evaluated?
Prompt 2	What potential risks or complications related to the patient's systemic condition should be considered before endodontic treatment?
Prompt 3	Should endodontic treatment be deferred or continued, and what precautions or modifications, if any, are required?
Prompt 4	Which drug interactions, analgesic or antibiotic choices should be considered, and is medical consultation indicated?

patient's systemic condition and its relevance to endodontic treatment (1–3 points), (2) appropriateness of the treatment approach and the decision to proceed, modify, or defer care (1–3 points), (3) accuracy and safety of medication-related recommendations, including analgesics, antibiotics, and drug–disease or drug–drug interactions (0–2 points), and (4) the justification and necessity of medical consultation when indicated (0–2 points).

Given the multidimensional nature of clinical decision-making, different Likert-type scoring approaches were deliberately applied across domains, in accordance with previously published methodologies evaluating AI-generated clinical recommendations.

To assess the completeness and informational adequacy of the responses provided by large language models, a three-point Likert scale was applied to domains related to clinical reasoning and treatment planning, consistent with prior expert-based evaluations of AI performance in healthcare [16].

1 = Incomplete: the response addresses only some aspects of the clinical scenario and omits critical information.

2 = Adequate: the response covers all essential aspects of the clinical scenario with sufficient information.

3 = Comprehensive: the response thoroughly addresses all aspects of the scenario and exceeds basic requirements by providing additional relevant context or justification.

Domains related to medication-related recommendations and the indication for medical consultation were evaluated using a three-category Likert-type ordinal scale (0–2), reflecting the graded nature of clinical safety and appropriateness judgments. This approach has been adopted in prior studies assessing the safety and acceptability of AI-generated medical advice, allowing differentiation between unsafe, partially appropriate, and entirely appropriate recommendations [17].

0 = Inappropriate/unsafe: the recommendation is incorrect or clinically unsafe, fails to consider relevant drug–disease or drug–drug interactions, or omits a clearly indicated medical consultation.

1 = Partially appropriate: the recommendation demonstrates partial clinical appropriateness but contains omissions, insufficient justification, or incomplete consideration of medication safety or consultation requirements.

2 = Appropriate/safe: the recommendation is clinically appropriate, secure, and entirely consistent with current medical and dental practice guidelines, including proper consideration of drug interactions and the need for medical consultation when indicated.

- Scores between 8 and 10 were classified as high accuracy, indicating that the model's reasoning and recommendations closely aligned with expert consensus.
- Scores from 5 to 7 were considered partially correct, reflecting answers that were generally acceptable but incomplete, overly cautious, or missing key clinical considerations.
- Scores from 0 to 4 were categorized as incorrect or potentially unsafe, typically involving errors in systemic risk interpretation, inappropriate treatment decisions, or clinically significant medication-related inaccuracies.

This categorical accuracy variable was used to determine the proportion of scenarios in which each model provided clinically reliable recommendations. All assessments were conducted independently by two calibrated endodontists, and any discrepancies were resolved through consensus discussion to ensure consistency and methodological rigor. In this study, clinical decision accuracy was operationally defined as the degree of alignment between AI-generated recommendations and expert consensus, rather than diagnostic accuracy based on patient outcomes.

Gold standard reference

The gold standard responses for each clinical scenario were established through a structured consensus process involving an experienced endodontist and medical specialists with expertise across the five selected systemic condition groups. The panel members independently reviewed each scenario, and a consensus meeting was then conducted to resolve any inconsistencies until an agreement was reached. The finalized gold standard responses provided detailed guidance on systemic risk assessment, the appropriate endodontic treatment approach, medication management, and the indication for medical consultation. All AI-generated responses were subsequently evaluated against these gold standard references using the predefined 10-point scoring system.

Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics software (Version 25; IBM Corp., Armonk, NY, USA). Normality of distributions was assessed using the Shapiro–Wilk test and revealed non-normal distributions for all models; therefore, nonparametric statistical methods were applied throughout the analysis.

Overall clinical decision accuracy scores and domain-specific scores (systemic condition evaluation, treatment decision, medication management, and consultation requirement) were compared among AI models across identical clinical scenarios using the Friedman test for related samples.

Clinical accuracy was additionally categorized to facilitate interpretation. Total scores between 8 and 10 were classified as high clinical accuracy, whereas scores of ≤ 7 were

considered to have non-high accuracy. Differences in the proportion of high-accuracy classifications among AI models were evaluated using Cochran's Q test. When Cochran's Q test indicated statistical significance, post hoc pairwise comparisons were conducted using McNemar tests with Bonferroni correction.

To assess the proximity of AI-generated decisions to expert consensus, deviation scores were calculated by subtracting each model's total score from the predefined maximum score (10 points). Lower deviation values indicated closer alignment with expert-defined optimal clinical decisions. Deviation scores were compared among models using the Friedman test. Statistical significance was set at $p < 0.05$.

Results

A comparative analysis of overall clinical decision accuracy scores revealed no statistically significant differences among the AI models across all clinical scenarios (Friedman test, $p = 0.298$) (Table 3, Fig. 1A). Similarly, no significant inter-model differences were observed in domain-specific analyses, including systemic condition evaluation, treatment decision-making, medication management, and consultation requirement (all $p > 0.05$) (Fig. 2).

When clinical accuracy was categorized, Cochran's Q test revealed a statistically significant overall difference in the proportion of high-accuracy classifications among AI models ($p = 0.007$). However, subsequent Bonferroni-adjusted McNemar post hoc analyses did not identify statistically significant differences between individual model pairs. Although no statistically significant pairwise differences were detected, descriptive analyses suggested relative performance trends among the models. ChatGPT-3.5 and Gemini 2.5 Pro demonstrated the highest proportions of high clinical accuracy (92.5%), whereas Gemini 2.5 Flash exhibited a lower high-accuracy rate (70.0%) (Fig. 1B).

Deviation analysis from expert-defined optimal total scores showed no statistically significant differences among AI models (Friedman test, $p = 0.298$) (Table 3, Fig. 1C). All models demonstrated comparable proximity to expert consensus; however, descriptive inspection revealed greater

Table 3 Overall and domain-specific clinical decision accuracy scores of artificial intelligence models across all clinical scenarios, analyzed using the Friedman test

Outcome/model	ChatGPT-3.5	ChatGPT-5.1	Gemini Pro	Gemini Flash	Friedman p value
Overall score (mean ± SD)	9.30 ± 1.31	9.25 ± 1.28	9.18 ± 1.20	8.80 ± 1.51	0.298
Median (IQR)	10 (1)	10 (1)	10 (2)	9 (2)	—
Systemic condition evaluation (mean ± SD)	2.95 ± 0.22	2.95 ± 0.22	2.93 ± 0.26	2.90 ± 0.31	1.000
Treatment decision (mean ± SD)	2.80 ± 0.41	2.78 ± 0.44	2.75 ± 0.48	2.60 ± 0.63	0.103
Medication management (mean ± SD)	1.85 ± 0.36	1.83 ± 0.38	1.80 ± 0.40	1.72 ± 0.45	0.392
Consultation requirement (mean ± SD)	1.78 ± 0.51	1.75 ± 0.53	1.70 ± 0.56	1.65 ± 0.58	0.944

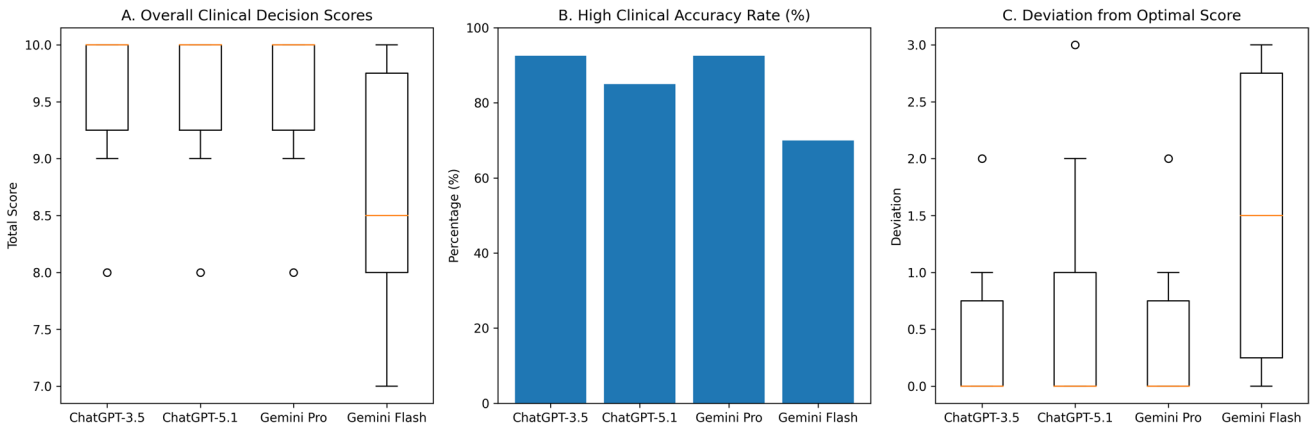


Fig. 1 Integrated comparison of overall performance, high clinical accuracy, and deviation from expert consensus among artificial intelligence models. **A** Distribution of overall clinical decision accuracy scores across all clinical scenarios, shown as boxplots with medians and interquartile ranges. **B** Proportion of scenarios in which each model achieved high clinical accuracy, defined as total scores between 8 and 10. **C** Deviation from the expert-defined optimal total score for each model, where lower values indicate closer agreement with expert consensus

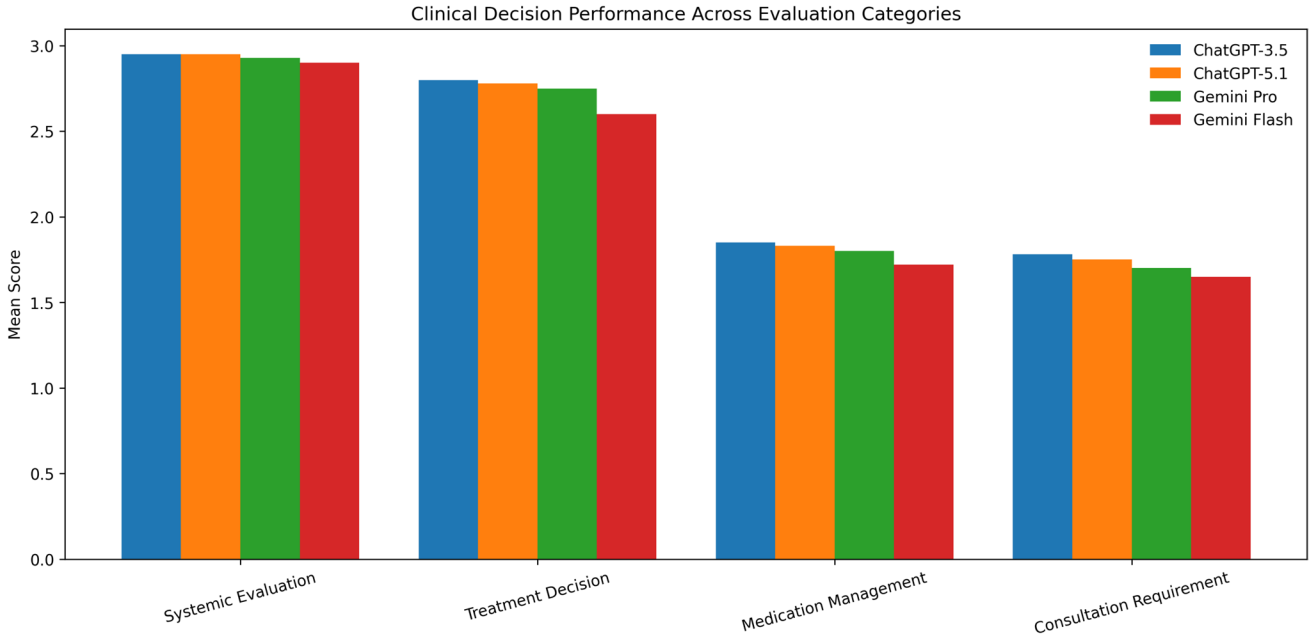


Fig. 2 Clinical decision performance of artificial intelligence models across evaluation categories

variability and higher dispersion of deviations for Gemini 2.5 Flash compared with the other models.

Overall, these findings indicate comparable performance among AI models in medically compromised endodontic scenarios, with observable descriptive trends despite the absence of statistically significant pairwise differences.

Discussion

The management of endodontic patients with systemic diseases represents one of the most complex aspects of contemporary dental practice, requiring integration of dental pathology with systemic risk assessment and medical considerations [18]. Systemic conditions, such as diabetes mellitus, cardiovascular diseases, chronic kidney disease, hepatic disorders, and thromboembolic conditions, are known to influence the immune response, tissue healing, bleeding tendency, and pharmacokinetics of commonly prescribed medications [15]. Consequently, endodontic treatment planning in these patients often necessitates modifications in treatment timing, procedural approach, medication selection, and interdisciplinary consultation.

Unlike most existing studies, which primarily evaluate the diagnostic performance of artificial intelligence systems in dentistry [5, 16], the present study focuses on the more complex and clinically consequential task of treatment-related decision-making in medically compromised endodontic patients. This includes not only procedural considerations, but also systemic risk assessment, medication safety, and the need for interdisciplinary consultation. By addressing these domains simultaneously, this study expands the scope of AI evaluation beyond pattern recognition and diagnostic classification, offering a more realistic representation of how AI tools may function as clinical decision support systems in everyday endodontic practice.

Previous studies have demonstrated that systemic diseases can affect both the prevalence and healing of apical periodontitis and may influence the long-term survival of endodontically treated teeth [7, 15]. In particular, diabetes mellitus has been associated with delayed periapical healing and increased tooth loss [7], while anticoagulant therapy poses challenges related to bleeding risk and perioperative management [9]. These factors collectively contribute to variability in clinical decision-making, even when established guidelines are available.

In this clinically complex context, decision-making is often challenging and subject to variability, even among experienced clinicians [19]. Artificial intelligence-based clinical decision support systems have, therefore, gained attention as potential adjunctive tools capable of integrating medical and dental information in a structured manner [20]. Through advanced natural language processing techniques and large language models, these systems can synthesize information from clinical guidelines, treatment protocols, and current scientific literature, providing researchers with structured, accessible, and clinically relevant content [21]. However, evidence regarding their reliability in medically compromised endodontic patients has been limited [11]. The present study directly addresses this gap by systematically

evaluating the clinical decision accuracy of different AI models across standardized endodontic scenarios involving common systemic diseases.

This study presents a profile that extends beyond simple diagnostic questions to evaluate the multidimensional nature of clinical decision-making in endodontics. While the existing literature primarily focuses on the ability of AI systems to identify pathology [4], the clinical success of endodontic treatment in patients with systemic diseases depends on the accurate integration of complex medical information [15]. By including 40 standardized clinical scenarios, this study offers a comprehensive challenge test for AI systems. This approach enables the evaluation of how well these models prioritize systemic contraindications in routine endodontic practice, contributing uniquely to the literature on artificial intelligence in endodontics [18].

The selection of ChatGPT-5.1, ChatGPT-3.5, Gemini 1.5 Flash, and Gemini 2.5 Pro models was a deliberate strategy to evaluate the influence of diverse algorithmic frameworks on clinical decision-making. Unlike static logic-based systems, these LLMs operate on stochastic and probabilistic frameworks, which inherently introduce variability in clinical advice [22, 23]. To ensure methodological consistency, we used independent chat sessions for each query to prevent contextual memory effects. Furthermore, the 10-point scoring system validated by expert consensus serves as a robust benchmark, moving beyond binary evaluations to provide a nuanced assessment of clinical decision accuracy.

Another fundamental aspect of this study is the harmonization of dental procedures with the patient's systemic status, treating the case as an integrated whole. Endodontic procedures are not isolated interventions; instead, they are profoundly influenced by the patient's systemic status [24]. By necessitating the integration of pharmacological interactions and anesthetic risks into treatment planning, this study aligns with the 'biological approach' in modern endodontics, which advocates for a balance between procedural requirements and patient safety through the preservation of vital tissues [14, 25]. Furthermore, the high internal consistency observed across models, despite their diverse training datasets and architectures, suggests that AI models have reached a level of maturity that enables repeatable, though not yet infallible, clinical decision support.

In this study, the AI models demonstrated a similar level of performance in both overall and domain-specific clinical decision-making when compared with expert-defined reference standards. The absence of statistically significant differences in the Friedman analyses indicates that none of the evaluated models showed a clear advantage over the others. Rather than reflecting limited performance, this finding suggests that current AI systems tend to reach comparable conclusions when addressing core clinical tasks such as systemic risk evaluation, treatment planning, medication-related considerations, and the need for medical consultation. From a clinical perspective, this consistency is a sign of stable decision-support capability across different AI platforms. This situation can be explained by the fact that AI models share similar training paradigms, are primarily based on standard clinical guidelines and literature, and present scenarios in a structured manner. Furthermore, the lack of statistically significant differences suggests that the quality of AI-assisted clinical decisions may be more closely related to the clinical framework and presentation of the questions than to the platform used.

The statistical analyses showed no significant performance differences between the AI models. Although an overall difference emerged when models were grouped according to high clinical accuracy, this finding was not supported by pairwise post hoc comparisons, indicating that none of the models consistently outperformed the others. These results suggest that paid versions may offer particular advantages in terms of response consistency and content; however, these differences do not result in a statistically significant and consistent superiority when viewed in terms of clinical decision accuracy. When the results were examined descriptively, ChatGPT-3.5 and Gemini 2.5 Pro were more frequently classified in the high-accuracy category, whereas Gemini 2.5 Flash showed lower high-accuracy rates and greater variability. Together, these findings suggest that statistical significance alone may not fully capture performance patterns and that descriptive evaluation provides important complementary insight in clinical decision-making research.

An additional practical consideration relates to the distinction between paid and free versions of contemporary AI models. While paid platforms may offer advantages in response stability, extended context handling, or update frequency, the present findings indicate that free-access models can still achieve clinically comparable decision-making accuracy when used within a structured, well-defined clinical framework. Furthermore, to minimize time-dependent response variability and the impact of potential model updates in comparisons between AI platforms, all AI assessments were performed simultaneously on different computers on the same day and at the same time. This approach strengthened the internal validity of the findings by controlling for time-dependent variation in responses in cross-platform comparisons. Given all these findings, free AI platforms, when used under appropriate professional supervision, can support clinical decision-making processes, especially in academic research settings and clinical scenarios with limited resources.

One of the key strengths of this study lies in the analysis of deviations from expert-defined optimal scores. Rather than relying solely on comparisons between AI models, this approach quantifies the degree to which AI-generated recommendations align with expert clinical judgment,

thereby offering a more clinically meaningful perspective. The similarity in deviation scores across models indicates that their recommendations generally remained close to expert-defined optimal decisions. Clinically, this finding suggests that current AI systems provide decision support that is broadly compatible with established standards of care for medically compromised patients, while maintaining consistency in their clinical reasoning.

The null hypothesis was accepted, as no statistically significant difference was observed in clinical decision accuracy among the evaluated artificial intelligence models.

Several limitations of this study should be acknowledged. First, the analyses were based on standardized clinical scenarios rather than real patient encounters. Although these scenarios were carefully constructed and reviewed by both endodontic and medical experts, clinical practice involves dynamic patient-related factors that cannot be fully replicated in predefined cases. Second, the scoring system was developed through expert consensus and followed a structured framework; however, it inevitably relied on clinical judgment and therefore cannot be considered entirely free of subjectivity. Finally, the evaluation was limited to a small number of AI models, and rapid model updates or architectural changes may affect performance in future applications.

Another limitation of this study is the equal weighting of all scoring domains within the composite clinical decision accuracy score, which sums multiple decision-making domains, including systemic risk assessment, treatment planning, medication-related recommendations, and the indication for medical consultation. Although systemic risk assessment, treatment planning, medication-related recommendations, and the indication for medical consultation do not necessarily carry identical clinical weight in real-world practice, equal weighting was deliberately applied to ensure methodological transparency and reproducibility across models. Future studies may benefit from exploring weighted scoring systems that reflect the relative clinical impact of specific decision domains, particularly medication safety and consultation requirements in medically compromised patients.

Despite these limitations, the findings of this study have important clinical implications. AI-based decision support systems can serve as valuable tools in managing endodontic patients with systemic conditions, assisting in risk evaluation and informing decisions related to medication use and the need for medical consultation. These systems should not be regarded as substitutes for clinical judgment, but rather as complementary resources that operate alongside clinician expertise. When applied with appropriate professional oversight, AI has the potential to promote safer, more consistent, and evidence-informed decision-making in the endodontic care of medically complex patients.

Conclusion

This study demonstrated that contemporary artificial intelligence models exhibit comparable performance in clinical decision-making for endodontic treatment in patients with systemic diseases. Across standardized scenarios, AI-generated recommendations showed a high degree of alignment with expert consensus, particularly in systemic risk assessment, treatment planning, medication considerations, and consultation requirements. Significantly, no AI model consistently outperformed the others in statistically robust analyses, supporting the null hypothesis.

These findings suggest that AI-assisted decision support systems, regardless of platform, can provide clinically meaningful guidance when applied within a structured and evidence-based framework. However, observed variability in descriptive performance highlights the importance of cautious interpretation and reinforces the necessity of clinician oversight. AI outputs should be regarded as adjunctive resources rather than definitive clinical directives.

From a clinical perspective, AI systems hold promise for enhancing consistency and supporting complex decision-making in medically compromised endodontic patients, particularly in educational and research settings or environments with limited access to multidisciplinary consultation. Future studies incorporating real-world clinical data and longitudinal outcomes are warranted further to clarify the role of AI in endodontic decision support.

Acknowledgements The authors deny any conflicts of interest related to this study.

Author contributions S.D. and İ.E.A.: conceptualization, study design, writing—review and editing, critical revision of the manuscript. A.E.: conceptualization, study design, investigation, data curation, writing—original draft, writing—review and editing, data interpretation. E.Y.: investigation, study design, data collection. İ.E.A.: data analysis, writing—original draft, writing—review and editing. All authors: approved the final version of the manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Data availability The data of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Consent for publication Not applicable.

Ethical approval Ethical approval was not required for this study, as it did not involve human participants, human data, or animal subjects. The study was based solely on artificially generated clinical scenarios and complied with general ethical principles of scientific research, including research integrity and data privacy.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Siqueira JJF. Endodontic infections: concepts, paradigms, and perspectives. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2002;94(3):281–93. https://doi.org/10.1067/moe.2002.126163.
2. Siqueira JJF, Rôças IN. Present status and future directions: microbiology of endodontic infections. Int Endod J. 2022;55:512–30. https://doi.org/10.1111/iej.13677.
3. Chen Y-w, Stanley K, Att W. Artificial intelligence in dentistry: current applications and future perspectives. Quintessence Int. 2020;51(3):248–57. https://doi.org/10.3290/j.qi.a43952.
4. Schwendicke Fa, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769–74. https://doi.org/10.1177/0022034520915714.
5. Gao S, Wang X, Xia Z, Zhang H, Yu J, Yang F. Artificial intelligence in dentistry: a narrative review of diagnostic and therapeutic applications. Med Sci Monit. 2025;31:e946676. https://doi.org/10.12659/MSM.946676.
6. Shan T, Tay F, Gu L. Application of artificial intelligence in dentistry. J Dent Res. 2021;100(3):232–44. https://doi.org/10.1177/0022034520969115.
7. Cabanillas-Balsera D, Martín-González J, Montero-Miralles P, Sánchez-Domínguez B, Jiménez-Sánchez M, Segura-Egea J. Association between diabetes and nonretention of root filled teeth: a systematic review and meta-analysis. Int Endod J. 2019;52(3):297–306. https://doi.org/10.1111/iej.13011.
8. Alshargawi W, Alghamdi F, Alziyad S, Ramadan F, Alamoudi A, Alotni S, et al. Impact of diabetes mellitus on endodontic treatment outcomes. Int J Community Med Public Health. 2024. https://doi.org/10.18203/2394-6040.ijcmph20242687.
9. Dou K, Shi Y, Yang B, Zhao Z. Risk of bleeding with dentoalveolar surgery in patients taking direct oral anticoagulants or vitamin K antagonists: a systematic review and meta-analysis. Jpn Dent Sci Rev. 2025;61:188–99. https://doi.org/10.1016/j.jdsr.2025.08.002.
10. Aminoshariae A, Kulild J, Nagendrababu V. Artificial intelligence in endodontics: current applications and future directions. J Endod. 2021;47(9):1352–7. https://doi.org/10.1016/j.joen.2021.06.003.
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.
12. Malta M, Cardoso LO, Bastos FI, Magnanini MMF, Silva CMFPd. STROBE initiative: guidelines on reporting observational studies. Rev Saude Publica. 2010;44:559–65. https://doi.org/10.1590/s0034-89102010000300021.
13. Dahlkemper P. American Association of Endodontists: Guide to clinical endodontics. 2016;Chicago, USA.
14. Endodontology ESo. Quality guidelines for endodontic treatment: consensus report of the European Society of Endodontology. Int Endod J. 2006;39(12):921–30. https://doi.org/10.1111/j.1365-2591.2006.01180.x
15. Segura-Egea JJ, Cabanillas-Balsera D, Martín-González J, Cintra LT. Impact of systemic health on treatment outcomes in endodontics. Int Endod J. 2023;56:219–35. https://doi.org/10.1111/iej.13789.
16. Taşyürek M, Adıgüzel Ö, Ortaç H. Comparative Evaluation of Responses from ChatGPT-5, Gemini 2.5 Flash, Grok 4, and Claude Sonnet-4 Chatbots to Questions About Endodontic Iatrogenic Events. Healthcare (Basel). 2025. https://doi.org/10.3390/healthcare13202615.
17. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. 2023:rs. 3. rs-2566942. https://doi.org/10.21203/rs.3.rs-2566942/v1
18. Xu X, Zheng X, Lin F, Yu Q, Hou B, Chen Z, et al. Expert consensus on endodontic therapy for patients with systemic conditions. Int J Oral Sci. 2024;16(1):45. https://doi.org/10.1038/s41368-024-00312-0.
19. de Araujo BMdM, de Miranda BM, Kowaltschuk TC, Gonçalves FM, Schroder AGD, Kuchler EC, et al. Impact of chronic diseases on the periapical health of endodontically treated teeth: A systematic review and meta-analysis. PLoS ONE. 2024;19(2):e0297020. https://doi.org/10.1371/journal.pone.0297020
20. Bonny T, Al Nassan W, Obaideen K, Al Mallahi MN, Mohammad Y, El-Damanhoury HM. Contemporary role and applications of artificial intelligence in dentistry. F1000Res. 2023;12:1179. https://doi.org/10.12688/f1000research.140204.1
21. Hamada M, Kikuchi S, Akitomo T, Kusaka S, Iwamoto Y, Nomura R. Applications and potential of ChatGPT in dentistry: Scoping review of research perspectives. J Dent Sci. 2025. https://doi.org/10.1016/j.jds.2025.08.016.
22. Bender EM, Gebru T, McMillan-Major A, Shmitchell S, editors. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; 2021. https://doi.org/10.1145/3442188.3445922
23. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2.
24. Aminoshariae A, Kulild JC, Mickel A, Fouad AF. Association between systemic diseases and endodontic outcome: a systematic review. J Endod. 2017;43(4):514–9. https://doi.org/10.1016/j.joen.2016.11.008.
25. Little JW, Miller C, Rhodus NL. Dental Management of the medically compromised Patient-E-Book. Elsevier Health Sciences; 2017.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.