ORIGINAL RESEARCH

Bagcilar Med Bull
**DOI:** 10.4274/BMB.galenos.2025.78941

# Evaluation of ChatGPT's Performance in Making-decision of Dialysis in Acute Kidney Injury

## Akut Böbrek Hasarında Diyaliz Kararı Almada ChatGPT Performansının Değerlendirilmesi

Başak Can[1], Esra Deniz Kahvecioğlu[1], Fatih Palit[2], Mehmet Küçük[3], Zeynep Karaali[1]

[1]University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital, Department of Internal Medicine, İstanbul, Turkey
[2]University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital, Department of Nephrology, İstanbul, Turkey
[3]University of Health Sciences Turkey, Prof. Dr. Cemil Taşcıoğlu City Hospital, Department of Nephrology, İstanbul, Turkey

## Abstract

**Objective:** Artificial intelligence chatbots have begun to be widely used in medicine. We aimed to evaluate the performance of ChatGPT in identifying patients in need of dialysis.

**Method:** A total of 100 patients who presented with acute kidney injury and were treated either with dialysis or without dialysis at the internal medicine clinic were retrospectively reviewed. Patient histories were created, consisting of demographic data, physical examination, and some laboratory tests. These patient histories were input into ChatGPT, and we requested a clinical evaluation along with recommendations categorizing them as low, medium, or high risk for dialysis treatment. The responses from ChatGPT were compared with the actual dialysis status of the patients. Additionally, ChatGPT responses were evaluated and scored by two nephrologists who were unaware of the dialysis status.

**Results:** The sensitivity of ChatGPT in recommending patients' need for dialysis was calculated as 94%, 97%, and 97% for ChatGPT 1, 2, and 3 answers, respectively. Specificity for ChatGPT responses 1, 2, and 3 was calculated as 81%, 76%, and 78%, respectively ($p<0.001$). The mean clinical evaluation scores were 4.71±0.4 and 4.67±0.4, and treatment recommendation scores were 4.45±0.7 and 4.39±0.7 for nephrologist 1 and nephrologist 2 ($p=0.002$) ($p<0.001$).

**Conclusion:** ChatGPT can be used as a decision support tool to identify patients who may need dialysis. Nevertheless, healthcare professionals should remain part of the decision-making process at present.

**Keywords:** Acute kidney injury, artificial intelligence, ChatGPT, dialysis

## Öz

**Amaç:** Yapay zeka sohbet botları tıpta yaygın olarak kullanılmaya başlanmıştır. Çalışmamızda, ChatGPT'nin diyaliz ihtiyacı olan hastaları belirlemedeki performansını değerlendirmeyi amaçladık.

**Yöntem:** Akut böbrek hasarı nedeniyle dahiliye kliniğine başvuran ve diyalizle veya diyalizsiz tedavi edilen toplam 100 hasta retrospektif olarak incelendi. Hastaların demografik verileri, fizik muayene bulguları ve bazı laboratuvar testlerini içeren hasta öyküleri oluşturuldu. Bu hasta öyküleri ChatGPT'ye girilerek hastalar için klinik bir değerlendirme yapılması ve diyaliz tedavisi gerekliliğine göre düşük, orta veya yüksek risk kategorilerine ayrılması istendi. ChatGPT'nin yanıtları, hastaların gerçek diyaliz durumlarıyla karşılaştırıldı. Ayrıca ChatGPT'nin yanıtları, diyaliz gereksiniminden habersiz olan iki nefrolog tarafından değerlendirilerek puanlandı.

**Bulgular:** ChatGPT'nin hastaların diyaliz ihtiyacını belirlemedeki duyarlılığı, ChatGPT 1, 2 ve 3 yanıtları için sırasıyla %94, %97 ve %97 olarak hesaplandı. Spesifiklik ise ChatGPT 1, 2 ve 3 yanıtları için sırasıyla %81, %76 ve %78 olarak belirlendi ($p<0,001$). Klinik değerlendirme puan ortalamaları nefrolog 1 ve nefrolog 2 için sırasıyla 4,71±0,4 ve 4,67±0,4 olarak hesaplandı. Tedavi önerisi puan ortalamaları ise sırasıyla 4,45±0,7 ve 4,39±0,7 olarak bulundu ($p=0,002$ ve $p<0,001$).

**Sonuç:** ChatGPT, diyaliz ihtiyacı olabilecek hastaları belirlemede bir karar destek aracı olarak kullanılabilir. Bununla birlikte, sağlık profesyonellerinin karar verme sürecinde belirleyici bir rol oynamaya devam etmesi gerekmektedir.

**Anahtar kelimeler:** Akut böbrek hasarı, ChatGPT, diyaliz, yapay zeka

**Address for Correspondence:** Başak Can, University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital, Department of Internal Medicine, İstanbul, Turkey
**E-mail:** basakozkurt@gmail.com **ORCID:** orcid.org/0000-0003-1910-9922

**Received:** 05.03.2025 **Accepted:** 11.06.2025 **Epub:** 17.06.2025

**Cite this article as:** Can B, Kahvecioğlu ED, Palit F, Küçük M, Zeynep Karaali Z. Evaluation of ChatGPT's performance in making-decision of dialysis in acute kidney injury. Bagcilar Med Bull. [Epub Ahead of Print]

## Introduction

Acute kidney injury (AKI) is defined as a syndrome characterized by impairment in kidney function resulting from a pathophysiological process caused by various etiologies. Estimates of AKI prevalence vary widely, ranging from less than 1% to as high as 66% (1). The primary causes of AKI include post-surgical or diagnostic interventions, sepsis, volume depletion, exposure to toxins, pregnancy-related complications, and iatrogenic factors (2,3). In AKI, impaired electrolyte balance and accumulation of waste products can induce a systemic inflammatory response and affect distant organs.

Uremic encephalopathy, pericarditis, life-threatening hyperkalemia, refractory acidosis, and hypervolemia causing end-organ complications are some indications for immediate dialysis (4). In some patients, dialysis treatment should be planned urgently to prevent complications such as permanent nephron loss. However, given the growing workload of clinicians and the challenges faced by healthcare systems in many countries, implementing systems that assist with patient treatment decisions may make a significant contribution.

Chat generative pretrained transformer (ChatGPT) is an artificial intelligence chatbot specializing in conversation. ChatGPT was first introduced into daily practice in late 2022. In a short period, it started being used in numerous areas of life, spanning from economics to education, and from engineering to medicine. Recently, numerous studies have demonstrated ChatGPT's ability in medical research (5-7). In this study, we aimed to evaluate the decision-making ability of ChatGPT in determining the need for dialysis in patients presenting to the hospital with AKI.

## Materials and Methods

We conducted a retrospective review of 100 consecutive patients who presented with AKI at the internal medicine clinic between January 2023 and May 2023. Demographic features, dialysis status, blood gas analysis, creatinine levels, hypervolemia, and uremic symptoms were retrospectively recorded for each patient. The study protocol was approved by the Local Ethics Committee of University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital and conducted in accordance with the Declaration of Helsinki (approval no: 2024.03.226, date: 22.04.2024). Patients diagnosed with chronic kidney disease (CKD), those with a history of dialysis, and those who may require intensive care were excluded from the study.

In the present study, the free version of ChatGPT 4, as of June 2024, was utilized to assess patient data. Before asking ChatGPT patient-related clinical questions, all personal browser data was cleared to prevent bias. New accounts were created. An entry was created for each patient that summarized the patient's demographic characteristics and clinic. We asked ChatGPT to evaluate the patients' need for dialysis. The exact question to ChatGPT was ”Hi, can I give you a patient story, where AKI is detected, and can you predict the risk stage as high, moderate, or low for immediate dialysis?”. Then, we entered patient information during the application (within a few hours), including age, gender, comorbidities, blood pressure, volume status (such as hypervolemia), uremic symptom status, urine output (in cc/h), blood gas analysis and levels of urea and creatinine. For example, “a 68-year-old woman has diabetes mellitus, hypertension, and asthma. Blood pressure: 122/155 mmHg, pericardial effusion (+), pleural effusion (+), nausea (+), vomiting (+), urine output: 83 cc/h. In venous blood gas, pH=7.37 $HCO_3$=17, lactate=1, potassium=4.7 creatinie=8.3, urea=214”.

All prompts were formulated in English. ChatGPT answered the questions for each patient with a clinical evaluation and a risk stratification as high, moderate, and low. To account for variability, ChatGPT was queried weekly for three weeks, yielding three responses per patient (1 day, 7 day, 14 day). Thus, ChatGPT's answers were labeled as 1, 2, and 3.

We evaluated the compatibility between these three ChatGPT's answers. Additionally, the study evaluated whether patients identified by ChatGPT as high-risk, for the need for emergency dialysis, actually received dialysis in real life.

ChatGPT answers were also evaluated by two experienced nephrologists. Two nephrologists independently evaluated ChatGPT responses in separate settings. The nephrologists scored the ChatGPT answers based on clinical evaluation and treatment recommendations. The nephrologists used a 5-point Likert scale for assessment. The meaning of points was (1) Strongly Disagree; (2) Disagree; (3) Neither Agree nor Disagree; (4) Agree; (5) Strongly Agree. The agreement between nephrologists' scores was also evaluated. Nephrologists were blinded to patients' actual dialysis status during the evaluation of ChatGPT responses.

Since the nephrologists in the study are the clinicians, we did not find it necessary to directly compare ChatGPT's predictions with the independent predictions made by nephrologists.

### Statistical Analysis

Statistical analysis was performed using the Statistical Package for the Social Sciences version 21.0 (SPSS Inc, Chicago, IL, USA). Continuous variables were presented as mean ± standard deviation. Cohen's kappa test was utilized to assess the agreement between the ChatGPT answers 1, 2, and 3. Similarly, the agreement of the nephrologists was evaluated with Cohen's kappa test. The relationship between patients recommended as high risk by ChatGPT and those who actually received dialysis was evaluated using the chi-squared test. A p-value of <0.05 was considered statistically significant.

## Results

A total of 100 patients with AKI were included in the study. General information of patients with AKI was presented in Table 1. The mean age of the patients was 68.07±16.3 years. Out of the patients, 44 were male and 56 were female.

Immediate dialysis treatment was administered to 36 of the patients presenting with AKI. On the other hand, 64 patients received medical treatment without dialysis. Out of 100 patients, 26 exhibited signs of hypervolemia, and 28 patients experienced uremic symptoms. The mean creatinine level was 5.4±3.5 mg/dL. The mean hospitalisation time was 13.7±9.9 days. All patients were discharged in stable condition following their treatment. In the blood gas analysis, the mean pH level was 7.2±0.09, and the mean $HCO_3$ level was 18.7±6.4 mEq/L. The mean lactate level was 38.1±8.4 mg/dL.

**Table 1. General information of acute kidney injury patients**

| | n=100<br>acute kidney injury patients |
|---|---|
| **Age (mean ± SD)** | 68.07±16.3 |
| **Gender (male/female)** | 44/56 |
| **Urea (mean ± SD) (mg/dL)** | 164.8±78.5 |
| **Creatinine (mean ± SD) (mg/dL)** | 5.4±3.5 |
| **Hypervolemia sign* (yes/no)** | 26/74 |
| **Uremic symptom# (yes/no)** | 28/72 |
| **Immediate dialysis (yes/no)** | 36/64 |
| **Hospitalization (mean ± SD) (day)** | 13.7±9.9 |
| **Blood gas analysis** | |
| **pH (mean ± SD)** | 7.2±0.09 |
| **$HCO_3$ (mean ± SD) (mEq/L)** | 18.7±6.4 |
| **Lactate (mean ± SD)** | 38.1±8.4 |

pH: Potential of hydrogen, $HCO_3$: Bicarbonate, SD: Standard deviation, *: Pleural effusion or pericardial effusion or pretibial edema, #: Nausea or vomiting or encephalopathy

According to Cohen's kappa test, the pairwise consistency between ChatGPT answer 1 and 2, ChatGPT answer 2 and 3, and ChatGPT answer 1 and 3 was all statistically significant (p<0.001). The relationship between ChatGPT treatment recommendation and dialysis status is presented in Table 2. Based on all ChatGPT results, there was a statistically significant association between patients predicted to be at high risk for dialysis and patients who actually received dialysis (for all ChatGPT answers p<0.001).

The results of nephrologists' evaluation of ChatGPT answers are presented in Figure 1. The mean clinical evaluation scores were 4.71±0.4 and 4.67±0.4 for nephrologist 1 and nephrologist 2 (p=0.002, respectively). The mean treatment recommendation scores were 4.45±0.7 and 4.39±0.7 for nephrologist 1 and nephrologist 2 (p<0.001) (respectively). The consistency of nephrologists was statistically significant for both clinical evaluation and treatment recommendation.

The sensitivity of the artificial intelligence chatbot in recommending patients' need for dialysis was calculated as 94%, 97%, and 97% for ChatGPT answers 1, 2, and 3,

**Table 2. The relationship between ChatGPT treatment recommendation and dialysis status**

| | | Dialysis (+)<br>n=36 | Dialysis (-)<br>n=64 | p-value |
|---|---|---|---|---|
| **ChatGPT answer 1** | High risk | 34 | 12 | <0.001 |
| | Non-high risk | 2 | 52 | |
| **ChatGPT answer 2** | High risk | 35 | 15 | <0.001 |
| | Non-high risk | 1 | 49 | |
| **ChatGPT answer 3** | High risk | 35 | 14 | <0.001 |
| | Non-high risk | 1 | 50 | |

ChatGPT: Chat generative pre-trained transformer

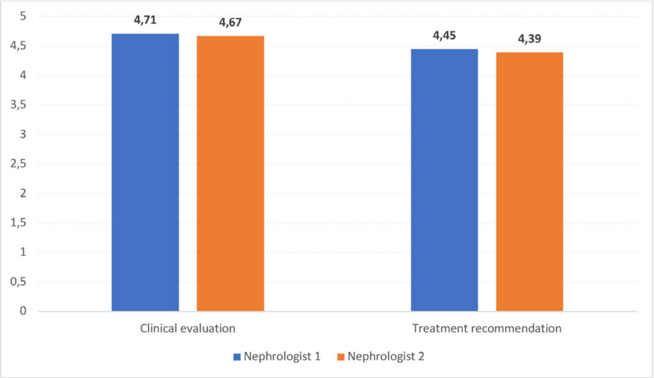

**Figure 1.** Rating of the performance of ChatGPT in two categories by the two reviewers: Clinical evaluation and treatment recommendation

*ChatGPT: Chat generative pre-trained transformer* respectively. The specificity was calculated as 81%, 76%, and 78% for ChatGPT answers 1, 2, and 3, respectively.

## Discussion

The present study revealed that ChatGPT has high sensitivity in suggesting patients' need for dialysis. Additionally, accuracy rates were verified by two independent nephrologists.

In recent years, artificial intelligence technology has been increasingly utilized in medicine (8). However, there are many questions regarding the reliability and efficacy of this new technology in medicine. In the field of medicine, ongoing research is focused on the use of AI in various areas including medical counselling, diagnosis, screening, surgery, patient management, and documentation (6,9). In our study, we evaluated the performance of ChatGPT in making decisions regarding the need for dialysis in patients. To the best of our knowledge, this is the first study assessing ChatGPT for this purpose.

AKI is a complex disease that can be apparent with various symptoms and signs, sometimes causing disruptions in multiple organ systems. Clinicians commonly use clinical semiology when evaluating the symptoms, signs, and clinical history of patients with conditions like AKI. However, our study demonstrated that if the relevant clinical findings and laboratory values of a patient with AKI are known, an artificial intelligence chatbot can determine the need for dialysis with up to a 97% accuracy rate.

On the other hand, using ChatGPT is not always optimal and straightforward. How healthcare professionals will use the chatbot is also an important consideration. When we ask ChatGPT about a patient's clinical details and treatment suggestions, it provides general recommendations and emphasizes the importance of seeking professional medical help. In this regard, as researchers gain a better understanding of how chatbot models like ChatGPT respond, they will learn how to formulate their future questions more effectively. We developed a question pattern through iterative trials to elicit correct answers and conducted the study using it. As stated in the methodology section, the question framework begins with the patient's age, sex, and comorbidities. The sentence is incomplete and lacks a subject and main verb to convey a complete thought. The effectiveness of this question framework model in decision-making processes for other nephrological diseases will be determined by future studies.

It is known that some questions asked of ChatGPT remain unanswered or may give incorrect answers (10). Similarly, in the study of Morath et al. (11), ChatGPT provided incorrect or incomplete answers to most of the 50 questions regarding drug information. These studies indicate that ChatGPT's performance may not be suitable for every medical topic at the present time. Therefore, due to the possibility of ChatGPT's responses being incorrect or inconsistent, we repeated the same clinical data and questions on the first day, on day 7, and on day 14, and requested treatment recommendations and patient evaluations from ChatGPT. The consistency and high accuracy rate of ChatGPT responses in this study showed that ChatGPT does not exhibit these limitations in making dialysis decisions.

The accuracy of information in healthcare services is critically important because errors or inaccuracies can lead to serious and irreversible consequences. A rigorous human review process, as well as human involvement at any stage of the workflow, can be crucial to the ChatGPT decision process. Although our study demonstrates that ChatGPT effectively manages the cases with high sensitivity and specificity, we believe that blindly relying on ChatGPT recommendations may entail clinical risks. However, in cases where access to a nephrologist is limited, initial guidance on treatment can be provided, and necessary cases can be shared with the nephrologist. Thus, the workload of nephrologists can be reduced. There is a lack of studies in the literature evaluating the decision-making capability of ChatGPT in nephrological diseases. In our previous study, we demonstrated that ChatGPT provided highly accurate answers to CKD-related questions aimed at informing patients (12). Nevertheless, further studies are needed in this area related to other nephrological diseases.

ChatGPT's training is based on large datasets of text sourced from the Internet. Therefore, as the size of chatbot models increases and they are trained on larger datasets, they have the potential to provide more accurate and detailed answers to the questions asked. Specialized chatbots tailored for specific healthcare purposes can be researched and developed, such as an AKI chatbot or a thyroid disease chatbot.

To reduce the possibility of bias in the datasets on which the model is trained, we cleared the browsers, created new users, input data into ChatGPT, and then queried the patient clinics. Nevertheless, reducing this potential bias to zero may be nearly impossible. After accumulating data, we believe that more comprehensive studies involving machine learning will improve the results.

### Study Limitations

Our study has identified certain limitations associated with ChatGPT, particularly related to patient privacy and the risk of information misuse. Therefore, meticulous data storage and access management are crucial aspects that require strict regulation and oversight (13). Another limitation of the study is the potential bias arising from its retrospective nature. Although the inclusion of only 100 patients limits the statistical power and generalizability of the findings, we believe it is sufficient for this preliminary study. Healthcare professionals must be aware of ChatGPT's limitations to use it effectively and responsibly.

## Conclusion

ChatGPT can serve as a decision-support tool to assist in identifying patients who may require dialysis. Additionally, software can be integrated into Hospital Health Information Management Systems to assist healthcare professionals in the initial management of patients with AKI. Although artificial intelligence cannot yet replace clinical judgment in dialysis decisions, its potential for future integration appears promising. Prospective studies with larger sample sizes are needed to strengthen the validity of these findings.

**Information:** A preprint version of our article is available on Research Square.

### Ethics

**Ethics Committee Approval:** The study protocol was approved by the Local Ethics Committee of University of Health Sciences Turkey, Başakşehir Çam and Sakura City Hospital and conducted in accordance with the Declaration of Helsinki (approval no: 2024.03.226, date: 22.04.2024).

**Informed Consent:** Retrospective study.

### Footnotes

### Authorship Contributions

Surgical and Medical Practices: B.C., F.P., M.K., Z.K., Concept: B.C., E.D.K., Design: B.C., M.K., Z.K., Data Collection or Processing: B.C., E.D.K., F.P., Z.K., Analysis or Interpretation: B.C., F.P., M.K., Literature Search: B.C., E.D.K., Writing: B.C.

**Conflict of Interest:** No conflict of interest was declared by the authors.

**Financial Disclosure:** The authors declared that this study received no financial support.

## References

1. Hoste EAJ, Kellum JA, Selby NM, Zarbock A, Palevsky PM, Bagshaw SM, et al. Global epidemiology and outcomes of acute kidney injury. Nat Rev Nephrol. 2018;14(10):607-625.
2. Jha V, Parameswaran S. Community-acquired acute kidney injury in tropical countries. Nat Rev Nephrol. 2013;9(5):278-290.
3. Olowu WA, Niang A, Osafo C, Ashuntantang G, Arogundade FA, Porter J, et al. Outcomes of acute kidney injury in children and adults in sub-Saharan Africa: A systematic review. Lancet Glob Heal. 2016;4(4):e242-e250.
4. Murdeshwar H, Anjum F. Hemodialysis. [Accessed on: 2023 Apr 27] [Internet]. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK563296/%0A
5. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus. 2023;15(4):e37589.
6. Sorin V, Klang E, Sklair-Levy M, Cohen I, Zippel DB, Balint Lahat N, et al. Large language model (ChatGPT) as a support tool for breast tumor board. npj Breast Cancer. 2023;9(1):44.
7. Sievert M, Conrad O, Mueller SK, Rupp R, Balk M, Richter D, et al. Risk stratification of thyroid nodules: assessing the suitability of ChatGPT for text-based analysis. Am J Otolaryngol - Head Neck Med Surg. 2024;45(2):104144.
8. Wen Z, Huang H. The potential for artificial intelligence in healthcare. J Commer Biotechnol. 2022;27(4):217-224.
9. Del Vecchio D, Stein MJ, Dayan E, Marte J, Theodorou S. Nanotechnology and artificial intelligence: an Emerging Paradigm for Postoperative Patient Care. Aesthetic Surg J. 2023;43(7):748-757.
10. Mediboina A, Badam RK, Chodavarapu S. Assessing the accuracy of information on medication abortion: a comparative analysis of ChatGPT and Google Bard AI. Cureus. 2024;16(1):e51544.
11. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. Eur J Hosp Pharm. 2024;31(6):491-497.
12. Can B, Kahvecioğlu ED, Palıt F, Cebeci E, Küçük M, Karaali Z. Assessing the performance of chat generative pretrained transformer (ChatGPT) in answering chronic kidney disease-related questions. Ther Apher Dial. 2024;1-6.
13. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33.