

Agreement Between Artificial Intelligence and Nephrologists in Addressing Common Patient Questions About Diabetic Nephropathy: Cross-Sectional Study

Niloufar Ebrahimi, Mehrbod Vakhshoori, Seigmund Teichman, Amir Abdipour

Submitted to: JMIR Diabetes
on: August 27, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	4
Supplementary Files.....	13
.....	13
0.....	13

Preprint
JMIR Publications

Agreement Between Artificial Intelligence and Nephrologists in Addressing Common Patient Questions About Diabetic Nephropathy: Cross-Sectional Study

Niloufar Ebrahimi¹; Mehrbod Vakhshoori¹ MD; Seigmund Teichman¹ MD; Amir Abdipour¹ MD

¹ Loma Linda University Medical Center Loma Linda US

Corresponding Author:

Amir Abdipour MD

Loma Linda University Medical Center
11234 Anderson St.
Loma Linda
US

Abstract

Diabetic nephropathy (DN) is one of the most frequent and severe complications of diabetes mellitus that requires early reduction for the individuals most at risk. Effective intervention and management of DN requires a patient's deep insight and helpful information. Patients require accurate information to obtain prevention advice about their kidney disease to make appropriate decisions; individuals may seek artificial intelligence (AI) as online sources instead of communicating with clinicians. Therefore, to evaluate ChatGPT and Google Bard's capabilities in providing accurate information regarding DN compared to nephrologists, we assessed their performance in answering questions related to DN. Our study revealed AI would not eliminate the need for open and detailed discussion with healthcare professionals regarding medical concerns and treatment plans, as they indicated inconsistencies in responses. Additionally, it is imperative to provide real-time health information updates to meet the needs of individuals with DN.

(JMIR Preprints 27/08/2024:65846)

DOI: <https://doi.org/10.2196/preprints.65846>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [http://](#)

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [https://](#)

Original Manuscript

Agreement Between Artificial Intelligence and Nephrologists in Addressing Common Patient Questions About Diabetic Nephropathy: Cross-Sectional Study

Niloufar Ebrahimi¹, MD; Mehrbod Vakhshoori¹, MD; Seigmund Teichman¹, MD; Amir Abdipour^{1*}; MD

1.Department of Medicine, Division of Nephrology, Loma Linda University Medical Center, Loma Linda, CA, USA

*Corresponding Author: Dr. Amir Abdipour, aabdipou@llu.edu

Keywords: Diabetic Nephropathy; ChatGPT; Google; Artificial Intelligence; Nephrologist

This research letter presents a cross-sectional analysis comparing the agreement of artificial intelligence models and nephrologists in responding to common patient questions about diabetic nephropathy.

Introduction

Diabetic nephropathy (DN) is one of the most frequent and severe complications of diabetes that needs early detection and management[1]. Patients with diabetes should receive accurate information from their providers on preventing kidney disease. However, many turn to artificial intelligence (AI) models like ChatGPT and Google Gemini as an online source for medical information[2-4]. To evaluate the capabilities of ChatGPT-4 and Google Gemini in providing accurate information about DN compared to nephrologists, their performance in answering DN-related questions most commonly raised by patients was assessed.

Methods

Collection of Questions

To generate patient-focused questions, the following query was prompted to AI models:

"What are the most frequently asked questions by individuals regarding diabetic nephropathy?"

The AI-generated responses were systematically reviewed. The final set of questions was refined and adjusted based on the principal investigator's experience in clinical practice, ensuring alignment with common patient concerns encountered in real-world practice.

Ultimately, ten questions covering various aspects of DN were developed. Questions 1, 3, and 7 were employed to evaluate DN's diagnosis, risk factors, and prevention, respectively.

Questions 2, 6, and 9 were used for DN management. Questions 8 and 10 were included to assess DN complications. Finally, to evaluate the progression and severity of DN, questions 4 and 5 were selected (**Table 1**).

Collecting Chatbot and Nephrologists Responses

To ensure consistency, a single investigator entered all questions into ChatGPT-4 and Google Gemini between September 23 and November 7, 2023. Each question was entered into ChatGPT-4 twice, initially and after 45 days, to assess changes in accuracy over time. Google Gemini was used once, concurrently with the second round of ChatGPT-4, and was limited to short-response tasks. Two experienced faculty nephrologists from Loma Linda University with clinical and academic experience also completed the questionnaire via an online survey. *Evaluation of Chatbot and Nephrologists Responses*

An independent reviewer, a Professor of Medicine, from the same academic center evaluated AI and nephrologists' responses. Each answer was graded as "Completely Inaccurate, Relatively Inaccurate, Irrelevant, Relatively Accurate, and Completely Accurate". To prevent any grading bias, the reviewer was not informed about the identity of the nephrologists.

Statistical Analysis

Analyses were conducted using RStudio (version 4.4.2), with P-values <0.05 considered significant.

Ethical considerations

As no patient data were involved, ethical approval was not required. The study adhered to ethical principles for research integrity and transparency.

Results

Table 1 presents the accuracy distribution of responses for each question assessed by reviewers. Notably, no responses were categorized as irrelevant or inaccurate; all answers were rated as either relatively or completely accurate.

Table 1. Distribution of answers according to each respondent.

Questions	ChatGPT-4	ChatGPT-4, 2 nd round	Google Gemini	Nephrologist 1	Nephrologist 2
1.What is the gold standard for diagnosis of diabetic nephropathy?	Completely Accurate	Completely Accurate	Completely Accurate	Completely Accurate	Completely Accurate
2.What is the current standard medication therapy for diabetic nephropathy?	Completely Accurate	Completely Accurate	Completely Accurate	Completely Accurate	Completely Accurate
3.Can diabetic nephropathy be prevented?	Completely Accurate	Relatively Accurate	Completely Accurate	Relatively Accurate	Relatively Accurate
4.Can tobacco use accelerate the progression of diabetic nephropathy?	Completely Accurate	Relatively Accurate	Completely Accurate	Completely Accurate	Completely Accurate
5.How is the severity of diabetic nephropathy determined?	Completely Accurate	Completely Accurate	Relatively Accurate	Relatively Accurate	Completely Accurate
6.How frequently should a patient be	Relatively Accurate	Completely Accurate	Completely Accurate	Relatively Accurate	Relatively Accurate

screened for diabetic nephropathy?					
7.What are the risk factors for the development of diabetic nephropathy?	Completely Accurate	Completely Accurate	Completely Accurate	Relatively Accurate	Relatively Accurate
8.What is the incidence of kidney failure in diabetic nephropathy?	Completely Accurate	Relatively Accurate	Completely Accurate	Relatively Accurate	Relatively Accurate
9.When should dialysis begin in diabetic nephropathy?	Relatively Accurate	Relatively Accurate	Relatively Accurate	Relatively Accurate	Completely Accurate
10.What is the most common cause of death in diabetic nephropathy?	Relatively Accurate	Completely Accurate	Relatively Accurate	Completely Accurate	Completely Accurate

ChatGPT: Chat Generative Pre-trained Transformer

Table 2 summarizes the inter-rater reliability indices among different respondents. Only the two nephrologists showed statistically significant agreement ($\kappa=0.61$, $P=.035$). ChatGPT-4 and Google Gemini had moderate but non-significant agreement ($\kappa=0.52$, $P=.098$). No significant agreement was found between either AI and the nephrologists ($P>.05$). Notably, ChatGPT-4 responses lacked consistency over time ($\kappa=-0.08$, $P=.778$). Further analysis showed negligible non-significant agreement among all respondents ($\kappa=0.083$, $P=.405$). Excluding ChatGPT-4's second-round responses did not alter the results ($\kappa=0.09$, $P=.451$), confirming the lack of significant agreement.

Table 2. Inter-rater reliability indices across different respondents

Respondents	ChatGPT-4	ChatGPT-4, 2 nd round	Google Gemini	Nephrologist 1	Nephrologist 2
ChatGPT-4	-	Kappa = -0.08 P-value = .778	Kappa = 0.52 P-value = .098	Kappa = 0.07 P-value = .778	Kappa = -0.08 P-value = .778
ChatGPT-4, 2nd round	Kappa = -0.08 P-value = .778	-	Kappa = -0.08 P-value = .778	Kappa = 0.23 P-value = .429	Kappa = 0.16 P-value = .598
Google Gemini	Kappa = 0.52 P-value = .098	Kappa = -0.08 P-value = .778	-	Kappa = 0.07 P-value = .778	Kappa = -0.52 P-value = .091
Nephrologist 1	Kappa = 0.07 P-value = .778	Kappa = 0.23 P-value = .429	Kappa = 0.07 P-value = .778	-	Kappa = 0.61 P-value = .035
Nephrologist 2	Kappa = -0.08 P-value = .778	Kappa = 0.16 P-value = .598	Kappa = -0.52 P-value = .091	Kappa = 0.61 P-value = .035	-

ChatGPT: Chat Generative Pre-trained Transformer

Inter-rater reliability was measured using Cohen's and Fleiss' kappa, with agreement classified as: 0.0–0.20 (none), 0.21–0.39 (minimal), 0.40–0.59 (weak), 0.60–0.79 (moderate), 0.80–0.90 (strong), and > 0.90 (almost perfect) [5].

Discussion

This study found that AI models generally provided accurate responses to DN-related questions, with moderate agreement between nephrologists on their accuracy. However, agreement between AI outputs and nephrologists' assessments was minimal, indicating a lack of standardized evaluation or clinical alignment. Moderate concordance between ChatGPT-4 and Google Gemini suggests similar underlying approaches. Improved agreement in ChatGPT-4's second round indicates potential learning and adaptability; however, limited alignment with nephrologists raises concerns regarding their clinical applicability. Despite that, interactive AI potentially enhances clinical processes by supporting patient education and facilitating communication between patients and clinicians regarding typical queries related to cardiovascular disease prevention[6]; the more questions lean

towards subspecialties, the less accurate AI responses tend to be[7].

In conclusion, while AI models can offer helpful responses about DN, they are not a substitute for thorough clinical discussions due to observed inconsistencies. Given the study's preliminary nature, findings should be interpreted cautiously. Further research with larger datasets is warranted to evaluate AI's reliability in clinical use.

Study Limitations

This study has several limitations. The AI models used were not specifically designed for medical applications, and the free version, intentionally selected to reflect typical patient use, may underperform compared to premium versions. Moreover, including only two nephrologists limits the diversity of clinical perspectives. Evaluations by a single senior nephrologist may introduce bias; future studies should include multiple reviewers to strengthen the reliability and validity of evaluations. Lastly, the study did not assess the clarity or helpfulness of AI responses from a patient perspective, highlighting the need for user-centered evaluations in future research.

Declarations

Acknowledgment: Niloufar Ebrahimi and Amir Abdipour are certified with the American Board of Artificial Intelligence in Medicine (ABAIM).[8]

Ethics Approval and Consent to Participate: This study did not involve patient data and was therefore exempt from ethical approval requirements.

Availability of Data and Materials: All data supporting the findings of this study are included within the manuscript, and no supplementary materials are provided.

Competing Interests: The authors declare no conflicts of interest.

Funding: This research received no external funding.

Authors' contributions: NE designed the study and drafted the manuscript, MV analyzed and interpreted the study data, ST reviewed and edited the answers, and AA reviewed and supervised the study. All authors read and approved the final manuscript.

References

- [1] N. Samsu, "Diabetic Nephropathy: Challenges in Pathogenesis, Diagnosis, and Treatment," *Biomed Res Int*, vol. 2021, p. 1497449, 2021, doi: 10.1155/2021/1497449.
- [2] J. Miao, C. Thongprayoon, and W. Cheungpasitporn, "Assessing the Accuracy of ChatGPT on Core Questions in Glomerular Disease," *Kidney Int Rep*, vol. 8, no. 8, pp. 1657-1659, Aug 2023, doi: 10.1016/j.ekir.2023.05.014.
- [3] "ChatGPT OpenAI." <https://help.openai.com/en/articles/6825453-chatgpt-release-notes> (accessed).
- [4] "Google Bard." <https://bard.google.com/updates> (accessed).
- [5] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem Med (Zagreb)*, vol. 22, no. 3, pp. 276-82, 2012.
- [6] A. Sarraju, D. Bruemmer, E. Van Iterson, L. Cho, F. Rodriguez, and L. Laffin, "Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model," *JAMA*, vol. 329, no. 10, pp. 842-844, Mar 14 2023, doi: 10.1001/jama.2023.1044.
- [7] J. T. Caranfa, N. K. Bommakanti, B. K. Young, and P. Y. Zhao, "Accuracy of Vitreoretinal Disease Information From an Artificial Intelligence Chatbot," *JAMA Ophthalmol*, vol. 141, no. 9, pp. 906-907, Sep 1 2023, doi: 10.1001/jamaophthalmol.2023.3314.
- [8] "ABAIM Educational Certification Alumni." <https://abaim.org/certification> (accessed).



Supplementary Files

Untitled.

URL: <http://asset.jmir.pub/assets/dfa31b813733e2964664929f4208b4c5.docx>

Untitled.

URL: <http://asset.jmir.pub/assets/d1cf064b4a399d24ee3647772c64b7d8.docx>