





ORIGINAL ARTICLE

The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland

Jan Nicikowski ^{1,2}, Mikołaj Szczepański ^{1,2}, Miłosz Miedziaszczyk ³
and Bartosz Kudliński ²

¹University of Zielona Gora, Faculty of Medicine and Health Sciences, Student Scientific Section of Clinical Nutrition, Zielona Góra, Poland, ²University of Zielona Góra, Faculty of Medicine and Health Sciences, Department of Anaesthesiology, Intensive Care and Emergency Medicine, Zielona Góra, Poland and ³Poznan University of Medical Sciences, Department of General and Transplant Surgery, Poznan, Poland

Correspondence to: Jan Nicikowski; E-mail: jan.nicikowski@gmail.com

ABSTRACT

Background. In November 2022, OpenAI released a chatbot named ChatGPT, a product capable of processing natural language to create human-like conversational dialogue. It has generated a lot of interest, including from the scientific community and the medical science community. Recent publications have shown that ChatGPT can correctly answer questions from medical exams such as the United States Medical Licensing Examination and other specialty exams. To date, there have been no studies in which ChatGPT has been tested on specialty questions in the field of nephrology anywhere in the world.

Methods. Using the ChatGPT-3.5 and -4.0 algorithms in this comparative cross-sectional study, we analysed 1560 single-answer questions from the national specialty exam in nephrology from 2017 to 2023 that were available in the Polish Medical Examination Center's question database along with answer keys.

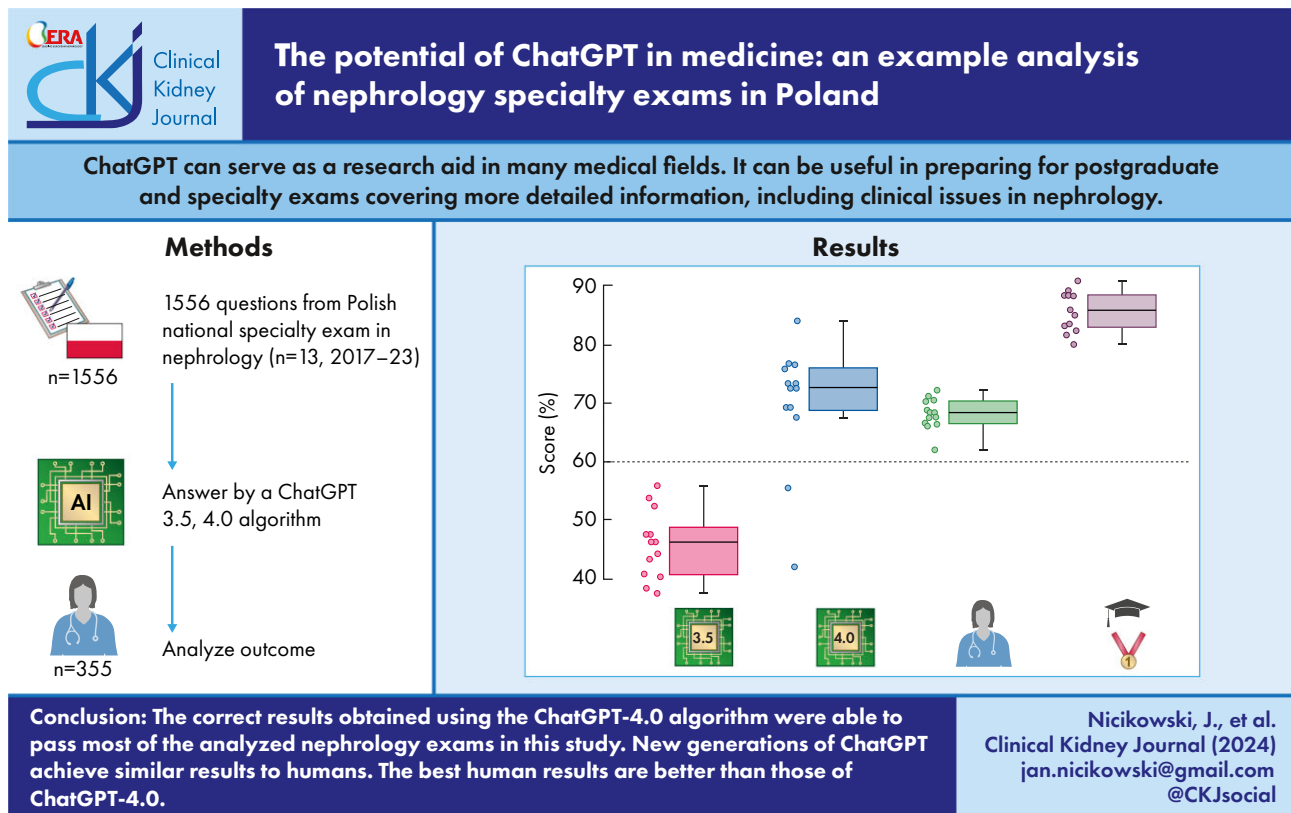
Results. Of the 1556 questions posed to ChatGPT-4.0, correct answers were obtained with an accuracy of 69.84%, compared with ChatGPT-3.5 (45.70%, $P = .0001$) and with the top results of medical doctors (85.73%, $P = .0001$). Of the 13 tests, ChatGPT-4.0 exceeded the required $\geq 60\%$ pass rate in 11 tests passed, and scored higher than the average of the human exam results.

Conclusion. ChatGPT-3.5 was not spectacularly successful in nephrology exams. The ChatGPT-4.0 algorithm was able to pass most of the analysed nephrology specialty exams. New generations of ChatGPT achieve similar results to humans. The best results of humans are better than those of ChatGPT-4.0.

Received: 24.12.2023; Editorial decision: 17.5.2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the ERA. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

GRAPHICAL ABSTRACT



Keywords: artificial intelligence, ChatGPT, clinical nephrology, machine learning, medical education

KEY LEARNING POINTS

What was known:

- ChatGPT can serve as a research aid in many medical fields, for example in preparation for postgraduate and specialty exams covering more detailed information, including clinical issues in nephrology.

This study adds:

- In this study, the first analysis of the nephrology specialty exam taking place in Poland between 2017 and 2023 was conducted.
- This study shows the capabilities of the different versions of the ChatGPT-3.5 and -4.0 algorithm in relation to the average and leading human performance.

Potential impact:

- To date, according to our knowledge there has been no study comparing the capabilities of the different versions of ChatGPT as a study aid for the nephrology specialty exam in Poland or anywhere in the world.
- Newer versions of ChatGPT, especially 4.0, perform better in some exams than the average 'human' group taking the test.

INTRODUCTION

The academic medical community has taken a significant interest in the topic of artificial intelligence (AI) and its use to improve education, medical knowledge and information retrieval in a more efficient and easier way in clinical practice since the public release of ChatGPT by OpenAI (OpenAI, San Francisco, CA, USA) [1–4]. Chat Generative Pre-trained Transformer (ChatGPT) is large language model (LLM) that generates human-like text.

A recent study has shown that AI can pass the United States Medical Licensing Examination (USMLE) [2]. Furthermore, it has been proven that ChatGPT can score 58.8% on the European Exam in Core Cardiology and American College of Gastroenterology self-assessment tests, averaging 63.75% [5, 6]. In a similar study by Giannos (2023), scores of 42% and 57% on the British Specialty Certificate Examination in neurology were achieved using the ChatGPT-4.0 and -3.5 algorithms [7].

The aim of our study was to determine whether the ChatGPT-3.5 and -4.0 algorithms would be able to correctly answer questions from the Polish National Specialty Examination (PES) in Nephrology and what percentage of the results it would be able to produce. We also wanted to determine whether ChatGPT-3.5 and/or -4.0 would be helpful to residents when studying for the specialty exam with the obtained results. Out of 56 possible specialties, we chose nephrology because it has some of the most complex, clinically challenging cases in this field of medicine [8]. In this short study, we explore the use of ChatGPT on a large pool of specialty questions in nephrology, filling a gap in the literature and providing insight into the potential educational applications and the possible opportunities and challenges of using AI-based language models in nephrology [9]. To the best of our knowledge, this is the first study in the world to evaluate the capabilities of AI in nephrology.

MATERIALS AND METHODS

Specifics of the Polish National Specialty Examination

The PES is organized in two sessions, the first taking place in the Spring and the second in the Fall. Any doctor in Poland wishing to become a specialist in a particular field of medicine must report his or her willingness to take the exam to the Center for Medical Examinations. This is the main body responsible for verifying the training of physicians in Poland. Each specialty exam in Poland consists of two parts: written and oral. The written part consists of 120 closed questions, with five answer options, with one correct answer. Responses are scored using a uniform key for each session exam. According to the organizational guidelines, two forms of exam questions are allowed: type A—with one correct answer; and type K—compound questions, where the examinee must choose a single set of correct statements [10]. A score of 60%, or 72 points, is required to pass this part of the exam and proceed to the second part. As of 31 December 2022, there has been a change regarding the conduct of the oral part of the specialty exam. This change pertains to the fact that any entrant who achieves at least 75% from the written part is exempted from the oral part of the National Specialty Examination, which takes place after the test form.

ChatGPT-3.5 and -4.0

In this comparative cross-sectional study, two independent researchers conducted the introduction of individual questions from the PES from 22 July to 5 December 2023. Questions that were withdrawn by the examination board were not included in the analysis. ChatGPT-3.5 and -4.0 were used to answer PES in Nephrology self-assessment exams (www.cem.edu.pl/) from 2017 to 2023 (total of 13 tests), including 1560 questions. There are two examination sessions each year—a Spring session and a Fall session. Two exams from each year were included in the study, except for the 2023 session, where only the Spring session was conducted. The exact questions were entered into the 3.5 version of ChatGPT (<https://chat.openai.com/>) and the 4.0 version (<https://www.bing.com/search>) [11, 12]. We used the Bing search engine for the study because it has the same version of OpenAI's software implemented as ChatGPT-4.0. In addition, it is open source, free for any user and does not require a subscription fee to use it, as is the case with the ChatGPT-4.0 version on OpenAI's website. Furthermore, Bing and ChatGPT-4.0 share a common infrastructure, which is provided by Microsoft, as can be read on the manufacturers' websites for both language

models. The main difference between the Bing search engine and ChatGPT is that the former is integrated with Microsoft's search engine which works in a similar way to Google's search engine, the most popular search engine in Poland [13], but also in the world [14]. ChatGPT, on the other hand, is an isolated interface, i.e. it requires opening a separate website or downloading an application to a computer or mobile device. However, using the ChatGPT-4.0 version based on the latest GPT-4.0 model is possible only after paying a fee. A sample question search using the Bing search engine is included in Supplementary data (Supplementary, 1). The main difference between the analysed versions of the language models is their ability to process information [15–17]. The more modern version of the GPT-4.0 language model, is a multimodal model. This allows it to process different types of data, both text and images. The older version (GPT-3.5), on the other hand, is a model that can only interpret textual data. In addition, the two versions differ in the validity of the data they use. The GPT-3.5 version uses data available up to June 2021, while GPT-4.0 covers data up to September 2021, but has selected information incorporated from a later period. In terms of the performance of a given language model, the topic is of interest to many researchers [15, 16]. A growing number of publications suggest that the GPT-4.0 version outperforms its predecessor GPT-3.5 in numerous respects, which aligns with OpenAI's disclosure that GPT-4.0's inference capabilities are more advanced than those of the older version [15, 18–20].

Statistical analysis

The chi-square test was used to assess the statistical significance of nominal variables when comparing the study groups. The Shapiro–Wilk test was used to verify whether the results of differences in pairs of study groups were normally distributed for interval scale. For normal distribution variables ($P > .05$), the paired Student's *t*-test was applied to estimate the significance of differences between the two analysed groups. Parameters that were significantly different from the normal distribution ($P < .05$) were analysed using the paired Wilcoxon signed-rank test. Statistical analysis was conducted using MedCalc® Statistical Software version 20.106 [21]. RStudio version 2023.12 [22] was used to edit and present the graphical results.

RESULTS

ChatGPT-3.5 achieved a total score of 45.70% based on 1556 included questions. The lowest score by ChatGPT-3.5 was 37.50% (Fall 2019) and the highest was 55.83% (Fall 2017) among the 13 tests. Four exam questions were withdrawn in the answer key. In the three highest scores achieved, the algorithm passed the test with a score of 55%, 53% and 52%, respectively (Table 1). ChatGPT-3.5 did not score enough points in the 13 exams to achieve satisfactory results to pass the nephrology specialty exam. ChatGPT-4.0 achieved a total score of 69.84% based on 1556 included questions. The lowest score among the 13 exam sessions obtained by ChatGPT-4.0 was 42.02% (Spring 2017) and the highest was 84.03% (Spring 2018). Four exam questions were withdrawn in the answer key. In the three highest scores achieved, the ChatGPT-4.0 algorithm passed the test with a score of 84%, 76% and 76% (Table 1). Considering the best human scores (85.73%), ChatGPT-4.0 was not able to surpass the best score achieved by a human in any session, and the closest it came was in the Spring 2018 session. ChatGPT-4.0 demonstrated sufficient performance in 11 examinations to

Table 1: Summary and statistical analysis of scores (%) obtained by ChatGPT-3.5 and -4.0 and the top human scores obtained by physicians between 2017 and 2023 in the PES in Nephrology.

Exam session	ChatGPT 3.5	ChatGPT 4.0	Top human results	P-value		
				ChatGPT-3.5 vs top human result	ChatGPT-4.0 vs top human result	ChatGPT-3.5 vs ChatGPT-4.0
Spring 2017 ^a	40.34	42.02	82.35	.455	.932	.116
Autumn 2017	55.83	73.33	85.83	.191	.001	.235
Spring 2018 ^a	53.78	84.03	88.24	.045	.004	.268
Autumn 2018 ^a	46.22	55.46	90.76	.958	.949	.017
Spring 2019	52.5	72.5	88.33	.183	.590	.001
Autumn 2019	37.5	75.83	80.00	.639	.524	.089
Spring 2020	47.5	76.67	89.17	.002	.174	.322
Autumn 2020	43.33	73.33	85.00	.681	.016	.109
Spring 2021	40.83	69.17	88.33	.323	.024	.004
Autumn 2021 ^a	46.22	76.47	83.19	.272	.866	.003
Spring 2022	38.33	72.5	83.33	.867	.413	.050
Autumn 2022	47.5	69.17	81.67	.249	.912	.001
Spring 2023	44.17	67.5	88.33	.500	.001	.005
Total	45.70	69.84	85.73	.0029	.0001	.0001

^aMaximum possible score of 119 due to a withdrawn question in the exam. Statistically significant values are presented in bold.

Table 2: Statistical parameters to assess the differences between average scores from the 2017–23 exams study groups.

Results for selected groups	ChatGPT-3.5 vs ChatGPT-4.0	ChatGPT-3.5 vs humans	ChatGPT-3.5 vs top human	ChatGPT-4.0 vs humans	ChatGPT-4.0 vs top human	Humans vs top human
Result (mean ± SD)	44.93 ± 7.32 vs 69.84 ± 10.62	44.93 ± 7.32 vs 68.13 ± 6.88	44.93 ± 7.32 vs 85.73 ± 11.43	Q1: 42.02 Q2: 72.50 Q3: 84.03	Q1: 42.02 Q2: 72.50 Q3: 84.03	68.13 ± 6.88 vs 85.73 ± 11.43
				vs Q1: 61.94 Q2: 68.38 Q3: 72.12	vs Q1: 80.00 Q2: 85.83 Q3: 90.76	
Shapiro–Wilk test	P = .7804	P = .6778	P = .7691	P = .0161	P = .0266	P = .7993
95% CI	18.07 to 31.77	18.86 to 27.54	36.84 to 44.78	-7.21 to 5.62	9.20 to 23.53	16.05 to 19.16
t-test for dependent groups	P = .0001	P = .0001	P = .0001	P = .1677 ^a	P = .0015^a	P = .0001

The results of 13 examination sessions were compared between groups. Statistically significant values are presented in bold.

^aThe paired Wilcoxon signed-rank test. SD, standard deviation.

achieve a satisfactory score and pass the nephrology specialty examination (Tables 1 and 2).

Between 2017 and 2023, physicians ($n = 355$) taking the specialty exam achieved an average score of 81.42 points (68.02%). The lowest score achieved by physicians was 31 points (26.05%), and the highest score was 108 points (90.75%). The presented percentage is the final result and takes into account the cancellation of questions by the examination board, therefore it does not represent the conversion of points in relation to the maximum possible value of 120 points. There is some variation in the results, which may indicate differences in the level of preparation of those taking the exam. The Supplementary data shows graphically (Supplementary, 1) how the questions were presented to the algorithm. The ranked results of AI are attached in the file (Supplementary, 2).

In Tables 1 and 2 and Figs 1 and 2, we included the top human results as the best result obtained by a doctor in a given examination session, the average result of the best results obtained among doctors, and the average result of all doctors taking individual examination sessions. Such comparisons were conducted to compare the obtained results versus AI using the ChatGPT-3.5 and -4.0 algorithms.

Table 1 shows the results of each examination session by the ChatGPT-3.5 and -4.0 algorithms, the average score obtained by doctors (humans) and the best score obtained by a doctor, together with a comparison of statistical significance. In Fig. 1, the graphs show the results graphically by exam session with groups: 3.5 and 4.0 algorithm, and best human score with statistical significance. Figure 2 shows the results of average scores from the 2017–23 exams depending on the study group: 3.5 and

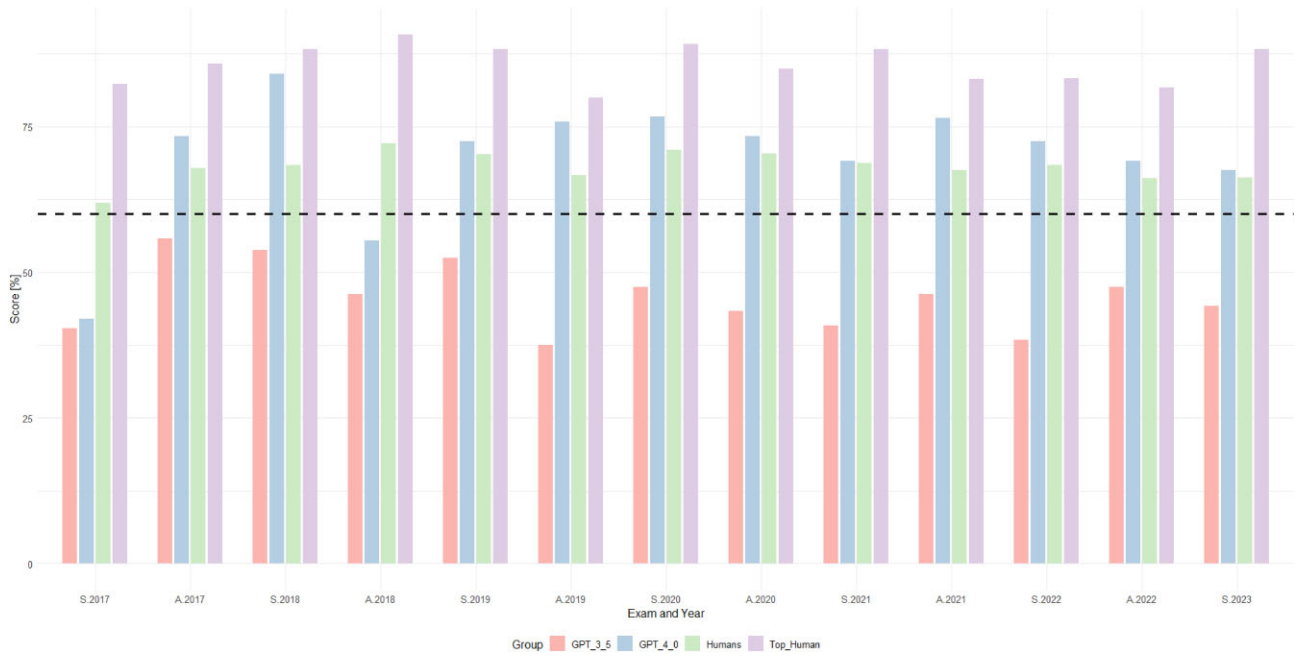


Figure 1: Scores obtained from individual 13 examinations using ChatGPT-3.5 and -4.0 between 2017 and 2023 (Spring and Fall parts) compared with the main and the highest score achieved by a medical doctor taking the nephrology specialty exam. Total number of questions: 1556. The dotted line in the chart includes the exam passing threshold of 60%.

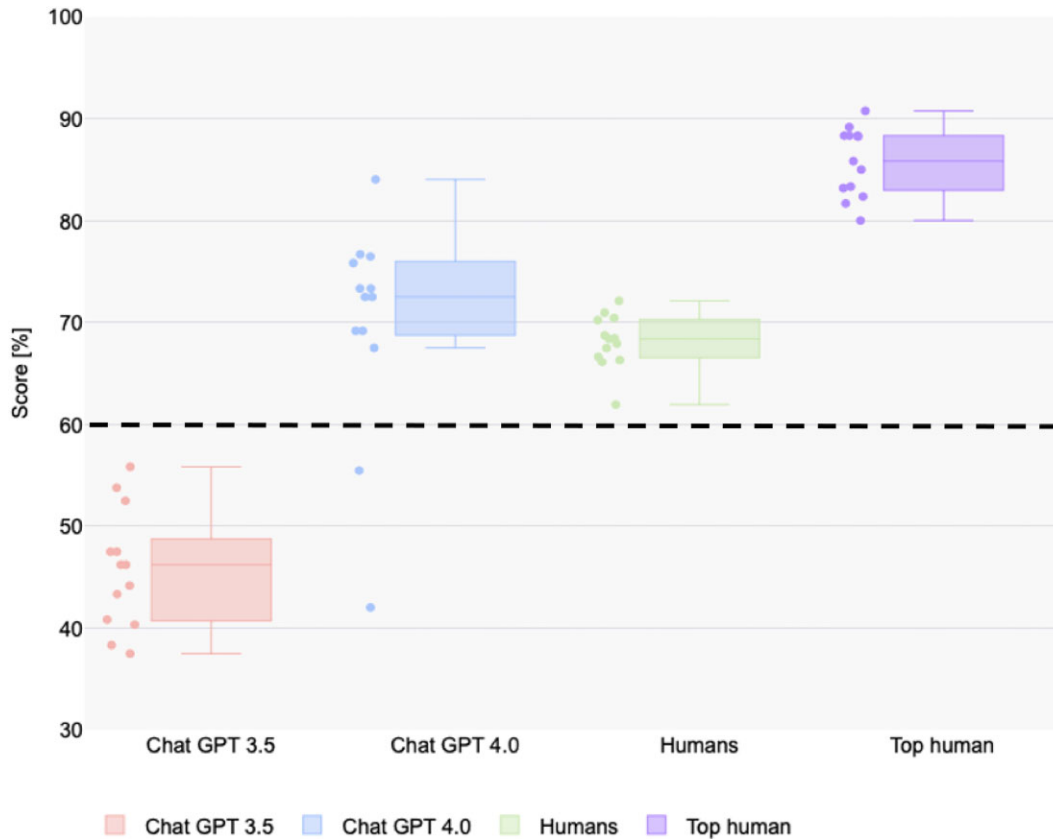


Figure 2: The figure shows the results of average scores from the 2017–23 exams for the studied groups. The lower part of the box is defined by the first quartile, the upper part by the third quartile. The horizontal line inside the box represents the median value. The upper end of the line is the highest value in the group, while the lower end of the line is the lowest value. The points show the results from each exam session (13 exam sessions, 2017–23). The dotted line in the chart includes the exam passing threshold of 60%.

4.0 algorithm, average human scores and the best human score with statistical significance.

DISCUSSION

Recent years have seen astonishing developments in the field of AI. We have reached a stage where AI is surpassing human abilities in tests from different countries and medical fields [3, 23]. More and more researchers are comparing the capabilities of the two versions of ChatGPT in various medical fields from different corners of the world including Japan, Poland and the USA [24–27]. In our study, the ChatGPT-3.5 version approached the passing threshold in only 3 of the 13 selected Medical Examination Center exams. Despite the developed answers with provided comments, the algorithm failed to cope with the presented questions. The results obtained by ChatGPT-3.5 were significantly lower than the results of medical doctors (44.93 ± 7.32 vs 68.13 ± 6.88 , $P = .0001$) (Table 2). Similar results were obtained in the first input data, as in the work of Liévin et al. (2023), with a 46% accuracy for the algorithm, with zero suggestions, as well as in neurology exam answers [7, 28]. This is also a similar result to a recent paper published by Suwała et al. (2023), whose mean algorithm scores from the internal medicine exam ranged from 47.5% to 53.33% (median 49.37%) [4].

In contrast, the ChatGPT-4.0 algorithm exceeded this threshold ($\geq 60\%$) in 11 out of 13 nephrology specialty tests. The latest version, ChatGPT-4.0, received an average of 29 (24%) more correct answers than the 3.5 version (44.93 ± 7.32 vs 69.84 ± 10.62 , $P = .0001$) (Table 2). It cannot be ruled out that the improvement in performance may be due to an improvement in the ability of the newer version of the language model to recognize Polish. This fact is emphasized by software manufacturers in their March 2023 report, which includes information that the language recognition abilities of the GPT-4.0 version are higher relative to previous versions [11]. Using Moshirfar et al. (2023) as an example, the researchers focused on analysing ophthalmology questions based on the StatPearls database [3]. However, it should be emphasized that the average score for humans does not differ significantly from the score obtained by the GPT-4.0 model. Moreover, the best result obtained by a human is significantly higher than the GPT-4.0 model's result.

In medicine, knowledge and analysis of information take time and experience to be able to work freely with patients. Equally important is proper communication with the patient and the relationship between the patient and the healthcare professional. This aspect also appears among the questions on the USMLE [29]. Brin et al. (2023) explored the possibilities of using ChatGPT to develop communication skills, ethics, empathy and professionalism by analysing questions from the USMLE and the AMBOSS question database [30]. In their study, ChatGPT-4.0 scored higher than its older versions (90.0% vs 62.5%). Brin et al. (2023) subjected ChatGPT to a fidelity test of their choice [30]. This test was based on casting doubt on the answer given, even if the initial answer was either correct or incorrect. It turns out that the older version of the model, when asked to revise its answer, changed its mind 82.5% of the time and indicated a different answer with the correct answer 53.8% of the time. In contrast, the ChatGPT-4.0 version did not change its answer in any case. Even when it provided a wrong answer it stood by its original answer [30].

A recent study by Eriksen et al. (2023) examined the efficacy of ChatGPT-4.0 in answering clinical questions. The study in-

involved 38 clinical cases and found that ChatGPT-4.0 correctly diagnosed 57% of cases, while a group of readers ($n = 10\,000$) correctly diagnosed 36% of cases [31]. The study is interesting because the researchers provided the algorithm with a complete patient history with test results and proposed diagnoses. However, limitations such as the relatively small number of cases to be verified by the AI may deviate from the actual clinical accuracy of the algorithm. The responding group of readers was of unknown medical skill level, with no exact information on how many doctors answered the questions correctly [31].

ChatGPT is a valuable tool in the education process, serving both students and teachers. Its ability to create, transform and translate content in real time makes it an essential tool in learning processes [32–34].

In their work, Dunlosky et al. (2013) demonstrate that one of the most effective methods of learning is solving quizzes and tests [35]. ChatGPT, with its potential to generate such tests in real time, supports not only the student in the practical retrieval of information, but also the teacher responsible for verifying knowledge.

In addition, ChatGPT is being used as a simulator for patient–doctor conversations, an important skill for any healthcare professional. Consequently, students who lack proficiency in conversational skills may use ChatGPT to overcome their limitations and enhance their communication abilities. This ability to simulate patient interactions can help to improve the quality of healthcare, if only by better understanding the patient's needs [34, 36]. However, given the effectiveness of ChatGPT, as a user one should be critical of the information retrieved by the tool. Sometimes the information it provides is not true, which some authors in the literature refer to as 'AI hallucinations' [37, 38]. Alkaissi et al. (2023) explained this as confident responses that seemed faithful and nonsensical when viewed in the context of the common knowledge in these areas [37]. On the other hand, methods already exist to reduce the aforementioned inadequacy of LLMs. One such approach is the implementation of a retrieval augmented generation system. This involves enriching the model's database with information from external sources, such as guidelines from scientific societies [39]. As a result, the user of such a model has the ability to generate answers that are subject to fewer errors and contain more up-to-date information [39]. The model presented by Miao et al. (2024) incorporating the KDIGO 2023 guidelines for chronic kidney disease can provide an additional tool in clinical decision-making and the education of healthcare professionals in the field of nephrology. However, in order to take full advantage of the capabilities of such an LLM, it is necessary to prepare appropriate user instructions and collaborate with AI experts [40].

It appears that replacing popular search engines with AI tools will be possible in the near future, but more research on language models and their potential for the general population is still necessary.

Our study has some limitations, as the analysis was based solely on ChatGPT's indication of the correct answer. We did not grade the questions to take into account, for example, the complexity of the questions, or their length. Another limitation of this study is that it did not take into account the number of physicians who achieved below-average numbers on the scale of the individual tests, and how many medical doctors in each session exceeded the threshold of 72 points (60%) needed to pass the exam.

CONCLUSIONS

We do not recommend the use of AI in direct medical education in nephrology in its current form, i.e. the ChatGPT-3.5 algorithm. However, we believe that ChatGPT-4.0 may be able to assist in analysing responses.

The nephrology specialty exam is one of the most demanding exams, and a candidate who takes it has completed 6 years of medical school, one internship and 5 years of specialty training in nephrology (a 3-year core module in internal medicine, followed by a 2-year nephrology specialty module). In addition, candidates usually spend months studying before the exam. Specialist vocabulary and knowledge is often misunderstood by the employed algorithm, which confuses many of the fundamental aspects of basic and preclinical medical science. We speculate that the algorithm's explanation of the answers may be helpful in the future for residents preparing for nephrology exams using ChatGPT-3.5 or other similar AI algorithms. The poorer performance compared with the USMLE may indicate that the specialty exam under study is more context-dependent and clinical case-dependent in nature as asked in the question. The questions do not focus on memory-based response models [29]. In future updates, we will be able to compare the existing algorithms with new AI tools. The results presented in this study can be used by other researchers, physicians and medical students who are interested in comparing the results of nephrology specialty examinations with AI.

SUPPLEMENTARY DATA

Supplementary data are available at [Clinical Kidney Journal](#) online.

ACKNOWLEDGEMENTS

We would like to thank Mr Bernard Baker for his contribution to the proofreading of the text.

FUNDING

This work was supported by funds from the University of Zielona Góra (No. 2023/2024 Ministry of Science and Higher Education, Poland).

AUTHORS' CONTRIBUTIONS

J.N.: conceptualization, software, formal analysis, original draft preparation, writing, data, investigation. M.S.: investigation, writing. M.M.: statistical validation. B.K.: supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

All data generated for this research can be found within the article and the Supplementary data.

CONFLICT OF INTEREST STATEMENT

The author has received no financial support for the research, authorship and/or publication of this article.

REFERENCES

- Cascella M, Montomoli J, Bellini V et al. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023;47:33. <https://doi.org/10.1007/s10916-023-01925-4>
- 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:e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
- Moshirfar M, Altaf AW, Stoakes IM et al. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. *Cureus* 2023;15:e40822. <https://doi.org/10.7759/cureus.40822>
- Suwała S, Szulc P, Dudek A et al. ChatGPT fails the internal medicine state specialization exam in Poland: artificial intelligence still has much to learn. *Polish Arch Intern Med* 2023;133:16608. <https://doi.org/10.20452/pamw.16608>
- Skalidis I, Cagnina A, Luangphiphat W et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? *Eur Heart J Digit Health* 2023;4:279–81. <https://doi.org/10.1093/ehjdh/ztad029>
- Suchman K, Garg S, Trindade AJ. Chat generative pre-trained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. *Am J Gastroenterol* 2023;118:2280–2. <https://doi.org/10.14309/ajg.0000000000002320>
- Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurol Open* 2023;5:e000451. <https://doi.org/10.1136/bmjno-2023-000451>
- Tonelli M, Wiebe N, Manns BJ et al. Comparison of the complexity of patients seen by different medical subspecialists in a universal health care system. *JAMA Netw Open* 2018;1:e184852. <https://doi.org/10.1001/jamanetworkopen.2018.4852>
- Glasscock RJ. Artificial intelligence in medicine and nephrology: hope, hype, and reality. *Clin Kidney J* 2024;17:sfae074. <https://doi.org/10.1093/ckj/sfae074>
- Polish Center for Medical Examinations (CEM). Warunki, jakie powinny spełniać pytania testowe przesyłane do Centrum Egzaminów Medycznych (CEM) [Conditions that test questions sent to the Center for Medical Examinations (CEM) should meet]. 2023. Available from: <https://www.cem.edu.pl/spz/Instrukcja2016.pdf> (22 June 2023, date last accessed).
- OpenAI. ChatGPT. 2023. Available from: <https://openai.com/chatgpt> (22 June 2023, date last accessed).
- Microsoft. Bing Chat. 2023. Available from: <https://www.microsoft.com/en-us/edge/features/bing-chat?form=MT00D8> (22 June 2023, date last accessed).
- Adriana S. Most popular PC web browsers in Poland from June 2019 to May 2023, based on share of views. 2023. Available from: <https://www.statista.com/statistics/957745/poland-most-popular-pc-web-browsers/> (31 June 2023, date last accessed).
- Fleck A. Google's Chrome Has Taken Over the World. 2023. Available from: <https://www.statista.com/chart/30734/browser-market-share-by-region/> (1 September 2023, date last accessed).
- OpenAI Research GTP-4. Available from: <https://openai.com/research/gpt-4> (18 April 2024, date last accessed).
- Koubaa A. GPT-4 vs. GPT-3. TechRxiv. 2023. Available from: <https://doi.org/10.36227/techrxiv.22312330.v2> (7 April 2023, date last accessed).

17. OpenAI, Achiam J, Adler S et al. GPT-4 Technical Report. 2023. Available from: <https://doi.org/10.48550/ARXIV.2303.08774> (4 March 2024, date last accessed).
18. Tao BK-L, Hua N, Milkovich J et al. ChatGPT-3.5 and Bing Chat in ophthalmology: an updated evaluation of performance, readability, and informative sources. *Eye* 2024;**38**:1897–1902. <https://doi.org/10.1038/s41433-024-03037-w>
19. Miao J, Thongprayoon C, Garcia Valencia OA et al. Performance of ChatGPT on nephrology test questions. *Clin J Am Soc Nephrol* 2024;**19**:35–43. <https://doi.org/10.2215/CJN.0000000000000330>
20. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German Medical Licensing Examination: observational study. *JMIR Med Educ* 2024;**10**:e50965. <https://doi.org/10.2196/50965>
21. MedCalc. *MedCalc® Statistical Software version 20.106*. Ostend, Belgium: MedCalc Software Ltd. 2022.
22. RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. 2020.
23. Takagi S, Watari T, Erabi A et al. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. *JMIR Med Educ* 2023;**9**:e48002. <https://doi.org/10.2196/48002>
24. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? *Am J Med Sci* 2023;**366**:291–5. <https://doi.org/10.1016/j.amjms.2023.08.001>
25. Rosoł M, Gašior JS, Łaba J et al. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Sci Rep* 2023;**13**:20512. <https://doi.org/10.1038/s41598-023-46995-z>
26. Kasai J, Kasai Y, Sakaguchi, K et al. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. 2023. Available from: <https://doi.org/10.48550/ARXIV.2303.18027> (17 December 2023, date last accessed).
27. Taloni A, Borselli M, Scarsi V et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Sci Rep* 2023;**13**:18562. <https://doi.org/10.1038/s41598-023-45837-2>
28. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns (N Y)* 2024;**5**:100943. <https://doi.org/10.1016/j.patter.2024.100943>
29. Mbakwe AB, Lourentzou I, Celi LA et al. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digit Health* 2023;**2**:e0000205. <https://doi.org/10.1371/journal.pdig.0000205>
30. Brin D, Sorin V, Vaid A et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep* 2023;**13**:16492. <https://doi.org/10.1038/s41598-023-43436-9>
31. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI* 2023;**1**:Aip2300031. <https://doi.org/10.1056/Aip2300031>
32. Augustin M. How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. *Yale J Biol Med* 2014;**87**:207–12.
33. Songsienchai S, Sereerat B, Watananimitgul W. Leveraging artificial intelligence (AI): Chat GPT for effective English language learning among Thai students. *ELT* 2023;**16**:68. <https://doi.org/10.5539/elt.v16n11p68>
34. Shorey S, Mattar C, Pereira TL-B et al. A scoping review of ChatGPT's role in healthcare education and research. *Nurse Educ Today* 2024;**135**:106121. <https://doi.org/10.1016/j.nedt.2024.106121>
35. Dunlosky J, Rawson KA, Marsh EJ et al. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol Sci Public Interest* 2013;**14**:4–58. <https://doi.org/10.1177/1529100612453266>
36. Holderried F, Stegemann-Philipps C, Herschbach L et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. *JMIR Med Educ* 2024;**10**:e53961. <https://doi.org/10.2196/53961>
37. Alkaiissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 2023;**15**:e35179. <https://doi.org/10.7759/cureus.35179>
38. Suppadungsuk S, Thongprayoon C, Krisanapan P et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. *J Clin Med* 2023;**12**:5550. <https://doi.org/10.3390/jcm12175550>
39. Wang C, Ong J, Wang C et al. Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. *Ann Biomed Eng* 2024;**52**:1115–8. <https://doi.org/10.1007/s10439-023-03327-6>
40. Miao J, Thongprayoon C, Suppadungsuk S et al. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. *Medicina (Kaunas)* 2024;**60**:445. <https://doi.org/10.3390/medicina60030445>