VIA MEDICA

RESEARCH PAPER
PRACA ORYGINALNA

Child Neurology — Neurologia Dziecięca
2024, vol. 34, no. 62, 39–44

ISSN 1230–3690, e-ISSN 2451–1897

DOI: 10.5603/cnne.100681

Evaluation of Chat GPT's performance in Polish paediatric neurology State Specialisation Examination

Ocena skuteczności Chatu GPT w rozwiązywaniu Państwowego Egzaminu Specjalizacyjnego z neurologii dziecięcej

Kamil Dzwilewski[1], Karolina Wangin[1], Natalia Jasińska[1], Marta Zawadzka[1], Przemysław Waszak[2], Maria Mazurkiewicz-Bełdzińska[1]

[1]Department of Developmental Neurology, Medical University of Gdańsk, Gdańsk, Poland
[2]Department of Hygiene and Epidemiology, Medical University of Gdańsk, Gdańsk, Poland

ABSTRACT

Introduction: Dynamic technological progress has contributed to significant advances in the field of Artificial Intelligence (AI). Its potential is already being used in many aspects of life, including medicine. The aim of this article was to focus on analyzing the effectiveness of AI-based language models in the context of tackling the Polish State Specialisation Examination (SSE) in paediatric neurology.

Material and methods: The study evaluated the effectiveness of two language models i.e., Chat GPT 3.5 and Chat GPT 4.0 in solving two past papers of SSE in paediatric neurology, i.e., those set in spring and autumn 2023. The point scores of both models were compared to the results of physicians taking the SSE at these two sessions. For the study, questions were divided into six thematic groups.

Results: Chat GPT 4.0 achieved a pass score (60%) in both examination sessions. Considering the total points obtained in both examination sessions, Chat GPT 4.0 achieved similar scores (72%) to physicians (74%). Significant differences were demonstrated between the results achieved by the older (48%) and newer (72%) versions of Chat GPT.

Conclusions: The results presented in our study may indicate the potential utilization of artificial intelligence in the practice of paediatric neurologists. Despite promising results, the use of AI in medicine poses serious ethical and practical challenges for physicians. Our article emphasizes the importance of further research on the use of AI in paediatric neurology and the need for continuous assessment and development of these technologies, raising issues regarding their potential applications and challenges associated with their implementation in clinical practice.

Keywords: child neurology, artificial intelligence, AI, Chat GPT

STRESZCZENIE

Wstęp: Dynamiczny postęp technologiczny przyczynił się do znamiennego rozwoju w zakresie Sztucznej Inteligencji (AI). Jej potencjał wykorzystywany jest w wielu sferach życia człowieka, w tym również w medycynie. W artykule skupiono się na analizie efektywności modeli językowych opartych na AI w kontekście rozwiązywania polskiego Państwowego Egzaminu Specjalizacyjnego (PES) z neurologii dziecięcej.

Materiał i metody: W badaniu oceniono skuteczność 2 modeli językowych: Chat GPT 3.5 oraz 4.0 w rozwiązywaniu dwóch arkuszy PES w dziedzinie neurologii dziecięcej — z wiosennej oraz jesiennej sesji z 2023 roku. Wyniki punktowe obu modeli porównano z wynikami lekarzy zdających PES w danych sesjach. Na potrzeby badania pytania podzielono na 6 grup tematycznych.

Wyniki: Chat GPT 4.0 w obu sesjach egzaminacyjnych uzyskał wynik wymagany do zdania egzaminu (60%). Biorąc pod uwagę sumę punktów uzyskanych w obu sesjach egzaminacyjnych, Chat GPT 4.0 osiągnął wyniki zbliżone (72%) do lekarzy (74%). Wykazano również znaczące różnice między wynikami osiągniętymi przez starszą (48%) oraz nowszą (72%) wersją Chatu GPT.

Address for correspondence:
Kamil Dzwilewski
Department of Developmental Neurology, Medical University of Gdańsk
ul. Dębinki 7, 80–211, Gdańsk, Poland
e-mail: kamil.dzwilewski@gmail.com

Received: 13.05.2024 Accepted: 27.01.2025 Published: 12.02.2025

Podsumowanie: Wyniki przedstawione w niniejszej pracy mogą wskazywać na potencjalne wykorzystanie sztucznej inteligencji w praktyce neurologów dziecięcych. Pomimo obiecujących wyników, wykorzystanie AI w medycynie stawia przed lekarzami wyzwania etyczne i praktyczne. Artykuł podkreśla znaczenie dalszych badań nad wykorzystaniem AI w neurologii dziecięcej oraz potrzebę ciągłej oceny i rozwoju tych technologii, podnosząc kwestie ich potencjalnych zastosowań oraz wyzwań związanych z ich implementacją w praktyce klinicznej.

Słowa kluczowe: neurologia dziecięca, sztuczna inteligencja, AI, Chat GPT

INTRODUCTION

Ceaseless dynamic technological development continues to contribute to significant progress in the area of artificial intelligence (AI). Its potential is already being used in various fields of life, including industry, education, and medicine [1].

Continuous work on AI has led to the development of highly advanced language models based on machine learning algorithms, such as Chat GPT. Chat GPT is trained on vast amounts of text containing a huge variety of information, styles of writing, and topics, allowing the model to learn, understand and generate human language. When a user provides a prompt, the model analyses it to understand the context and intentions behind the question. On this basis, the model generates a response aimed at being as relevant and informative as possible. It predicts the text that best fits as a continuation of the given phrase. Advanced technologies such as natural language processing (NLP) and neural networks optimize its ability to generate appropriate and coherent responses. Chat GPT has also been designed to ensure that the content it generates is not harmful or inappropriate [2].

In the literature can be found a variety of reports on the application of language models based on artificial intelligence in various medical specialties, including radiology [3], cardiology [4], and urology [5]. Artificial intelligence has demonstrated its usefulness in teaching [6], clinical reasoning, and even in the writing of medical articles [7]. However, knowledge regarding its use in paediatric neurology is very limited.

For language models to be routinely used in medical practice, they must undergo a thorough evaluation of their reliability and efficiency. In order to determine whether Chat GPT might be a useful tool in the field of child neurology, its medical 'knowledge' and deduction skills need to be compared to the performance of doctors.

Poland's State Specialisation Examination (SSE) in paediatric neurology is a mandatory exam taken at the end of the specialty training that verifies the fundamental medical knowledge of future paediatric neurology specialists. In this study, we have compared the results of paediatric neurology SSEs obtained by doctors to those obtained by Chat GPT 3.5, and its newer version 4.0.

MATERIAL AND METHODS

The study used two examination papers of the SSE in paediatric neurology: the spring 2023 paper and the autumn 2023 paper, which were published on Poland's Medical Examination Centre's official website [8]. Each exam consisted of 120 multiple-choice questions, each with five possible answers from which to choose, of which only one was correct. Each answer scored 1 point if it was correct, or 0 points if not. Thus, a maximum of 120 points could be scored on each exam. The SSE is considered passed if a minimum of 60% correct answers are provided. To each question there was assigned a percentage score of correctly answered questions by doctors taking the exam in a controlled examination setting. For each question, the difficulty index of the question was calculated according to the formula:

$$IDI = \frac{Ns + Ni}{2n}$$

where n represented the number of examinees in each extreme group (an extreme group comprises 27% of the highest scorers and 27% of the lowest scorers in the whole test), Ns was the correct answer to the analysis task in the highest scorers group, and Ni was the correct answer to the analysis task in the lowest scorers group. This index takes values from 0 (for extremely difficult questions) to 1 (for extremely easy questions). It is not identical to the percentage of correct answers, as it does not consider answers provided by examinees with average test scores. The test difficulty index is calculated as the average value of the difficulty indices of individual tasks [9]. The content of the questions was not used in the study. The questions covered the diverse topics covered by the specialty training programme in paediatric neurology in Poland. For the purposes of our study, the questions were divided into six thematic groups:

1) Epilepsy (60 questions),
2) Child development (22 questions),
3) Metabolic disorders and other rare diseases (103 questions),
4) Headaches and CNS tumours (21 questions),
5) Cerebrovascular diseases (11 questions),
6) Neuroinfections and CSF examination (22 questions).

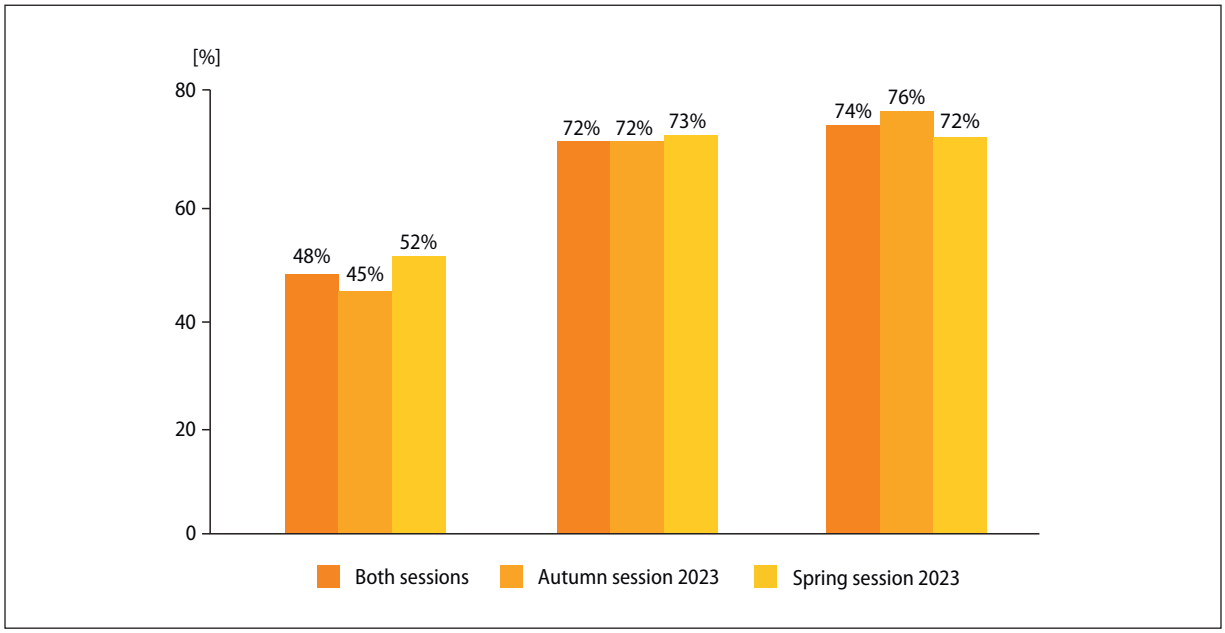

Figure 1. Results of Polish State Specialisation Examination in paediatric neurology in spring 2023 and autumn 2023

Rycina 1. *Wyniki Państwowego Egzaminu Specjalizacyjnego z neurologii rozwojowej w sesji wiosennej 2023 r. i jesiennej 2023 r.*

In the study, two language models were used: Chat GPT 3.5 and Chat GPT 4.0 [10]. The total scores from all questions combined, as well as for individual thematic categories, were calculated. The responses provided by both models were compared to those given by the doctors taking the SSE in each session. The results obtained by the language models for individual questions were also compared against the difficulty index of the questions and the average percentage scores obtained by the doctors.

Statistics of the results obtained by the Chat GPT language models and the doctors were prepared. Both the cumulative final results of the SSE and the percentages of correct answers, categorized by question types, were processed. A correlation analysis was also conducted using Spearman's rank correlation (the data did not meet the criteria for normal distribution). A significance level of $p < 0.05$ was adopted.

RESULTS

Figure 1 presents the percentage results of the SSE in paediatric neurology obtained by doctors and both GPT Chat language models in the spring 2023 exam, the autumn 2023 exam, and the average of both sessions. In the spring examination, Chat GPT 4.0 scored the highest number of points (73% correct answers), while in the autumn doctors achieved the highest score (76% correct answers). In the combined scoring, doctors performed the best, obtaining 74% correct answers, whereas Chat GPT 3.5 and Chat GPT 4.0 achieved 48% and 72% respectively.

Table 1 presents the distribution of percentage points obtained in individual thematic categories from both examination sessions.

Doctors achieved the highest scores in categories 1, 2, 5, and 6, while GPT 4.0 performed best in categories 3 and 4. Chat GPT 3.5 obtained the lowest scores in every category. The most significant statistical differences were observed in questions related to epilepsy, where doctors outperformed Chat GPT 4.0 ($p < 0.01$), which in turn performed better than Chat GPT 3.5 ($p = 0.02$), and in questions from category 3, where Chat GPT 4.0 performed better than Chat GPT 3.5 ($p < 0.01$) and doctors ($p < 0.01$). In the remaining thematic categories, statistically significant p-values were not achieved.

The calculated Spearman's rank correlation coefficient regarding the difficulty index of questions (discussed in the study methodology above) and the responses provided by Chat GPTs and doctors is presented in Table 2.

The strongest correlation with the difficulty coefficient was observed among the responses of doctors (0.83), and the weakest was among the responses of Chat GPT 3.5 (0.24) ($p < 0.01$). Spearman's rank correlation coefficient was also calculated regarding the percentage of correct answers from doctors to individual questions and answers provided by both language models (Tab. 3).

The Spearman's rank correlation coefficient was higher for the responses of Chat GPT 4.0 (0.33) than for the responses given by Chat GPT 3.5 (0.27).

Table 1. Percentage results categorised by thematic categories
Tabela 1. *Wyniki procentowe z podziałem według kategorii tematycznych*

Thematic category	Chat GPT 3.0	Chat GPT 4.0	Doctors
1. Epilepsy	50%	72%	75%
2. Child development	45%	55%	67%
3. Metabolic disorders and other rare diseases	49%	74%	73%
4. Headaches and CNS tumours	48%	81%	74%
5. Cerebrovascular diseases	36%	64%	74%
6. Neuroinfections and CSF examination	55%	82%	85%

Table 2. Spearman's rank correlation coefficient regarding difficulty coefficient
Tabela 2. *Współczynnik korelacji rang Spearmana dotyczący współczynnika trudności*

	Valid N	Spearman	t(N-2)	p-value
Chat 3.5's points & difficulty index	239	0.235065	3.72311	0.000246
Chat 4.0's points & difficulty index	239	0.291074	4.68384	0.000005
Doctors' points & difficulty index	239	0.827106	22.65498	0.000000

Table 3. Spearman's rank correlation coefficient regarding percentage of correct answers from doctors to individual questions and answers provided by both language models
Tabela 3. *Współczynnik korelacji rang Spearmana dotyczący odsetka poprawnych odpowiedzi lekarzy na poszczególne pytania i odpowiedzi udzielonych przez oba modele językowe*

	Valid N	Spearman	t(N-2)	p-value
Chat GPT 3.5's points & percentage score of doctors	239	0.268830	4.296763	0.000025
Chat GPT 4.0's points & percentage score of doctors	239	0.326169	5.311807	0.000000

DISCUSSION

In the literature, several studies have described the performance of Chat GPT in solving medical exams. Kung et al. [11] presented the effectiveness of Chat GPT in solving the United States Medical Licensing Exam (USMLE). Depending on the examination paper, Chat GPT achieved a sufficient or nearly sufficient score to pass the exam. Giannos [12] compared the results obtained by Chat GPT 3.5 and 4.0 while tackling the UK's Neurology Specialty Certificate Examination from 2022 in his work. He demonstrated that Chat GPT achieved 64% correct answers, which was a sufficient result to pass the aforementioned exam.

Based on the presented results, it can be observed that the older version of Chat GPT (3.5) did not achieve the minimum number of points required to pass either session of the State Specialist Examination in paediatric neurology (minimum 60% correct answers required). However, its successor, Chat GPT 4.0, achieved a score above 60% in both exams, and its results were very close to those obtained by doctors. Furthermore, in the spring exam, it actually scored higher than the average score of the doctors taking the exam.

This demonstrates how effectively the issue of language models is being developed, and that their potential use in medical practice seems promising. Chat GPT 4.0 outperformed its predecessor, probably due to significant enhancements, such as more advanced contextual understanding, greater language fluency, and a much larger base of learned information [13]. However, assessing the potential reasons behind the performance differences between GPT 3.5 and GPT 4.0 is challenging due to the confidential nature of GPT 4.0's architecture, making it difficult to conduct research on future applications [14].

Interpreting the Spearman rank correlation coefficient values, we noted that there was only a moderate correlation between the results obtained by both language models and the difficulty coefficient, as well as the percentage of correct answers obtained by doctors. This suggests that questions potentially challenging for doctors do not necessarily have to be difficult to solve for language models, and vice versa.

Making a thorough comparison between human candidates and Chat GPT is inherently complex. Chat GPT generates structured responses based on probabilistic patterns in its training data. However, it is prone to producing inaccurate statements presented as factual, and is lacking in self-awareness. This limitation may cause problems in the clinical decision-making required during answering specific questions [15]. Moreover, the questions in the SSE are composed in Polish, whereas Large Language Models typically favour languages that are more represented in their training database. Since Chat GPT seems to perform best with inputs in English, the Polish language becomes a valid limitation of its effectiveness [16]. Furthermore, human candidates are assessed on their ability to recall and synthesize information from memory in real time during board examinations, while Chat GPT leverages vast repositories of pre-processed online data to optimize its responses. If human candidates were permitted unrestricted access to electronic resources during their examinations, their performance might significantly surpass the results observed in this study [15]. On the other hand, the differences might suggest that human test-takers lack sufficient exposure to some topics related to the nature or structure of the questions [14].

The variations in Chat GPT's performance in different categories may be a result of inadequate modelling by the engineers and differences in availability of specialism-specific materials in the training database [15]. Additionally, difficulties could stem from the presence of specific innovations in the field of developmental neurology that may not be updated in Chat GPT's training database. Alternatively, qualified neurologists may be more familiar with particular areas of knowledge, especially those associated with clinical practice [13].

Despite the promising results achieved by language models, it is important to remember that open access to language models could pose a potential threat to patients, as instead of seeking professional medical advice, they might choose to look for solutions to their health problems using Chat GPT.

It is important to underline that Chat GPTs are constantly being improved and developed over the years. Although the model does not learn from every interaction in real time, data is collected that can be used for future model improvements, so continuous verification of its abilities will be required for language models to become reliable tools in medical practice [17].

It is also worth mentioning that large language models are just one of the possible implications of AI in developmental neurology. It is also being tested in the field of diagnosis and treatment of epilepsy and autism [18–21]. Research is also underway regarding the use of AI in neuroimaging of brain tumours and focal cortical dysplasia [22, 23].

The use of AI in medicine poses a distinct ethical dilemma for doctors [24]. Its proponents believe that it is an innovative tool that will facilitate further development of medicine. However, some members of the medical community believe that it poses a serious threat to the integrity and safety aspects of medicine [25].

LIMITATIONS

This study was conducted based only on two SSE examination sheets. To increase the objectivity of the results, it would be necessary to conduct them on a larger number of questions. Each question was answered by the language models only once. Perhaps in attempting to address the same questions again, Chat GPT would generate a different response. There is a lack of methodological studies evaluating the consistency and reliability of information provided by Chat GPT in multiple inquiries on the same topic. There is a lack of methodological frameworks for conducting studies analysing trends or patterns of Chat GPT usage. Some questions covered very broad topics and involved several issues, making it difficult to categorise them thematically. Additionally, there was a discrepancy in the number of questions in each category. We also limited our study to only one artificial intelligence language model; future research in this area should also consider using other models such as Google Bard.

CONCLUSIONS

Artificial Intelligence, including large language models, is evolving rapidly, finding applications in many fields of medicine. The results presented in our study may indicate the potential use of artificial intelligence in the practice of paediatric neurologists. Comparing the SSE results obtained by Chat GPT 3.5 and 4.0 to the results of doctors demonstrates that the newer version of Chat GPT (4.0) achieves comparable results to those of doctors, demonstrating significant progress in AI development. There are significant differences in the results achieved between the older (3.5) and newer (4.0) versions of Chat GPT, again indicating rapid progress in the development of these technologies.

Despite promising results, the use of AI in medicine poses ethical and practical challenges for doctors, including issues related to the integrity and safety of medicine. Further research on the use of AI in paediatric neurology and other medical fields, including diagnosis, neuroimaging, and therapy for patients with developmental disorders, is warranted.

In summary, our study emphasizes the importance of continuous development and evaluation of AI models in medicine, raising issues surrounding their potential applications, and challenges associated with their implementation in clinical practice.

ARTICLE INFORMATION AND DECLARATIONS

Ethics statement

Not applicable.

Authors' contributions

First author, corresponding author: KD. First coauthor: KW. Second coauthor: NJ. Final approval of manuscript: MZ, MM-B. Statistics, final approval of manuscript: PW.

Funding

The authors did not receive any funding.

Conflicts of interest

The authors have no conflict of interest to declare.

REFERENCES

1. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017; 2(4): 230–243, doi: 10.1136/svn-2017-000101, indexed in Pubmed: 29507784.
2. Manickam P, Mariappan SA, Murugesan SM, et al. Artificial Intelligence (AI) and Internet of Medical Things (IoMT) assisted biomedical systems for intelligent healthcare. Biosensors (Basel). 2022; 12(8), doi: 10.3390/bios12080562, indexed in Pubmed: 35892459.
3. Rao A, Kim J, Kamineni M, et al. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. 2023, doi: 10.1101/2023.02.02.23285399, indexed in Pubmed: 36798292.
4. Danilov A, Aronow WS. Artificial Intelligence in cardiology: applications and obstacles. Curr Probl Cardiol. 2023; 48(9): 101750, doi: 10.1016/j.cpcardiol.2023.101750, indexed in Pubmed: 37088174.
5. Talyshinskii A, Naik N, Hameed BM, et al. Potential of AI-driven chatbots in urology: revolutionizing patient care through artificial intelligence. Curr Urol Rep. 2024; 25(1): 9–18, doi: 10.1007/s11934-023-01184-3, indexed in Pubmed: 37723300.
6. Ghorashi N, Ismail A, Ghosh P, et al. AI-powered chatbots in medical education: potential applications and implications. Cureus. 2023; 15(8): e43271, doi: 10.7759/cureus.43271, indexed in Pubmed: 37692629.
7. Hutson M. Could AI help you to write your next paper? Nature. 2022; 611(7934): 192–193, doi: 10.1038/d41586-022-03479-w, indexed in Pubmed: 36316468.
8. Centrum Egzaminów Medycznych w Łodzi. https://www.cem.edu.pl/index.php.
9. Opis statystyk. Centrum Egzaminów Medycznych w Łodzi. https://www.cem.edu.pl/aktualnosci/opis_statystyk.pdf.
10. https://chat.openai.com/c/3674b438-d7ab-41c0-ad95-691fbf444f58 (13.04.2024).
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023; 2(2): e0000198, doi: 10.1371/journal.pdig.0000198, indexed in Pubmed: 36812645.
12. 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(1): e000451, doi: 10.1136/bmjno-2023-000451, indexed in Pubmed: 37337531.
13. 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(6): e40822, doi: 10.7759/cureus.40822, indexed in Pubmed: 37485215.
14. Thirunavukarasu AJ, Ting DS, Elangovan K, et al. Large language models in medicine. Nat Med. 2023; 29(8): 1930–1940, doi: 10.1038/s41591-023-02448-8, indexed in Pubmed: 37460753.
15. Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open. 2023; 5(2): e000530, doi: 10.1136/bmjno-2023-000530, indexed in Pubmed: 37936648.
16. Lost in Translation: Large Language Models in Non-English Content Analysis. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.
17. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? [preprint]. 2023, doi: 10.48550/arXiv.2307.09009.
18. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia. 2019; 60(10): 2037–2047, doi: 10.1111/epi.16333, indexed in Pubmed: 31478577.
19. Lampos V, Mintz J, Qu X. An artificial intelligence approach for selecting effective teacher communication strategies in autism education. NPJ Sci Learn. 2021; 6(1): 25, doi: 10.1038/s41539-021-00102-x, indexed in Pubmed: 34471124.
20. Chaddad A, Li J, Lu Q, et al. Can autism be diagnosed with Artificial Intelligence? A narrative review. Diagnostics. 2021; 11(11): 2032, doi: 10.3390/diagnostics11112032.
21. Al-Nafjan A, Alhakbani N, Alabdulkareem A. Measuring engagement in robot-assisted therapy for autistic children. Behav Sci (Basel). 2023; 13(8), doi: 10.3390/bs13080618, indexed in Pubmed: 37622758.
22. Peng J, Kim DD, Patel JB, et al. Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. Neuro Oncol. 2022; 24(2): 289–299, doi: 10.1093/neuonc/noab151, indexed in Pubmed: 34174070.
23. AI algorithm that detects brain abnormalities could help cure epilepsy. https://www.ucl.ac.uk/news/2022/aug/ai-algorithm-detects-brain-abnormalities-could-help-cure-epilepsy (13.04.2024).
24. Morley J, Machado CCV, Burr C, et al. The ethics of AI in health care: a mapping review. Soc Sci Med. 2020; 260: 113172, doi: 10.1016/j.socscimed.2020.113172, indexed in Pubmed: 32702587.
25. Layard Horsfall H, Palmisciano P, Khan DZ, et al. Attitudes of the surgical team toward Artificial Intelligence in neurosurgery: international 2-stage cross-sectional survey. World Neurosurg. 2021; 146: e724–e730, doi: 10.1016/j.wneu.2020.10.171, indexed in Pubmed: 33248306.