Pirogov Russian Journal of Surgery
2024, No. 8, pp. 6–14
https://doi.org/10.17116/hirurgia20240816

Efficacy and safety of artificial intelligence-based large language models for decision making support in herniology: evaluation by experts and general surgeons

© T.V. NECHAY, A.V. SAZHIN, K.M. LOBAN, A.K. BOGOMOLOVA, V.V. SUGLOB, T.R. BENIIA

Pirogov Russian National Research Medical University, Moscow, Russia

ABSTRACT

Objective. To evaluate the quality of recommendations provided by ChatGPT regarding inguinal hernia repair.

Material and methods. ChatGPT was asked 5 questions about surgical management of inguinal hernias. The chat-bot was assigned the role of expert in herniology and requested to search only specialized medical databases and provide information about references and evidence. Herniology experts and surgeons (non-experts) rated the quality of recommendations generated by ChatGPT using 4-point scale (from 0 to 3 points). Statistical correlations were explored between participants' ratings and their stance regarding artificial intelligence.

Results. Experts scored the quality of ChatGPT responses lower than non-experts (2 (1—2) vs. 2 (2—3), $p<0.001$). The chat-bot failed to provide valid references and actual evidence, as well as falsified half of references. Respondents were optimistic about the future of neural networks for clinical decision-making support. Most of them were against restricting their use in healthcare.

Conclusion. We would not recommend non-specialized large language models as a single or primary source of information for clinical decision making or virtual searching assistant.

Keywords: *clinical decision making support tool, large language model, ChatGPT, guidelines, evidence level, inguinal hernia, artificial intelligence.*

INFORMATION ABOUT THE AUTHORS:
Nechay T.V. — e-mail: tnechay@mail.ru; https://orcid.org/0000-0003-0769-5282
Sazhin A.V. — e-mail: sazhin-av@yandex.ru; https://orcid.org/0000-0001-6188-6093
Loban K.M. — e-mail: doctor.loban@gmail.com; https://orcid.org/0000-0002-8333-2398
Bogomolova A.K. — e-mail: pogarskaya.anna@mail.ru; https://orcid.org/0000-0002-7726-1351
Suglob V.V. — e-mail: dr@vsuglob.ru; https://orcid.org/0000-0003-4363-257X
Beniia T.R. — e-mail: tamaribeniia@gmail.com; https://orcid.org/0009-0007-8636-6415
Corresponding author: Bogomolova A.K. — e-mail: pogarskaya.anna@mail.ru

TO CITE THIS ARTICLE:
Nechay TV, Sazhin AV, Loban KM, Bogomolova AK, Suglob VV, Beniia TR. Efficacy and safety of artificial intelligence-based large language models for decision making support in herniology: evaluation by experts and general surgeons. *Pirogov Russian Journal of Surgery*. 2024;8:6–14. (In Russ.). https://doi.org/10.17116/hirurgia20240816

Introduction

In 1958, Frank Rosenblatt designed and built a cybernetic model of the human brain known as Perceptron, which became the prototype of modern neural networks. However, the exponential growth in computing power in the past decade has spurred the rapid evolution of neural networks, gripping the interest of the scientific community [1]. On November 30, 2022 a large language model (LLM) called ChatGPT (Generative Pre-trained Transformer) trained on Azure supercomputer was released. Five days after its release, ChatGPT reached 1 million users; 2 months later, their number skyrocketed to 100 million [2].

Language models are primarily intended for text recognition, answering questions and maintaining conversations, i.e. generating a coherent text in response to a user's query. Although the medical community was not really enthusiastic about the early generations of the chatbot, the later generations of the LLM came to be used as an alternative to manual data searching in some disciplines. Further, more than 10 specialized on biomedical texts LLM have been developed. For example, BioMedLM «trained» on 16 million abstracts and 5 million full-text articles from PubMed [3].

As a tool for measuring the performance of biomedical LLMs employ the United States Medical Licensing Examination (USMLE). The latest generations of bio-

medical LLMs have demonstrated excellent performance in searching for correct diagnostic and treatment strategies. Med-PALM 2 model was able to score 86.5% in March 2023, surpassing the average score shown by students [4].

The growing successes of artificial intelligence-based LLMs and the wide media coverage they were receiving have sparked off a polarized debate from a revolution in medicine and undoubtedly benefit as a clinical decision support tool to skepticism and suspiciousness.

Currently, a significant proportion of professionals resort to the use of unspecified LLM's in their daily professional practice [5—9]. One in ten doctors uses ChatGPT — mostly unofficially, asking for help when diagnosing a patient or selecting the proper medication. As many as 50% of healthcare employees plan to do so in the near future» [10]. At the meantime, the quality of responses and recommendations of LLM's in this capacity in surgery is poorly established, and therefore there is a lack of information to assess its safety as a medical decision-making tool [6, 9, 11]. The attitudes toward LLMs of physicians with different sources of motivation in practice decision making are unknown. All this raises concerns of the professional medical community and consumers of medical services [11—13].

The aim of this study was to compare the quality of recommendations provided by ChatGPT regarding inguinal hernia management by experts in herniology and general surgeons.

Matherials and methods

The potential of the LLM for assisting in medical decision making and the willingness of the medical community to rely on neural network advice were studied in several steps.

1. Questions and instructions for ChatGPT

We asked ChatGPT 3.5 (OpenAI, San Francisco, California, U.S.) 5 questions about inguinal hernioplasty:

1. What are the indications for inguinal hernioplasty?
2. What is the smallest mesh size required for inguinal hernioplasty?
3. Which of the following techniques is the most optimal for inguinal hernioplasty: TAPP, TEP or the Lichtenstein repair? When should each of them be used?
4. What are the indications for mesh fixation and non-fixation?
5. What is the operative tactic for recurrent inguinal hernias?The chatbot was asked to imagine it was an expert in surgery and herniology, an aid to a healthcare professional in clinical decision making. It was instructed to give concise and precise answers using evidence from trusted sources, such as PubMed, NCBI, SCOPUS, WebOfScience, etc. The chatbot was also requested to provide the sources of information the answers were based on and specify their level of evidence (LoE).

These instructions and the first question were formulated by a herniology expert. Other questions were prepared without the expert's supervision by a junior surgeon specializing in hernia repair.

2. Focus group composition

In the second step, we brought together a focus group to assess the quality of the recommendations obtained from ChatGPT. It consisted of two subgroups. The first was a group of experts who had performed at least 100 inguinal hernia repairs and whose expertise included teaching theoretical and practical herniology courses. The second group comprised solid practicing surgeons without teaching experience who had performed less than 100 hernia repairs (non-experts). To minimize the risk of systemic bias and the Hawthorne effect, the authors did not participate in the survey.

The respondents were asked about the sources of guidance they normally relied on in clinical decision making (personal experience / directives from senior management / domestic clinical guidelines / international clinical guidelines / both domestic and international guidelines).

The respondents also expressed their thoughts on the use of artificial intelligence (AI) as a medical decision support tool on a 10-point scale (where 0 — no prospects, 10 — important role in the future). Then respondents opined on the need for restrictions AI use in medicine (where 0 is no necessity, 10 is call for severe restrictions).

3. Assessment of ChatGPT's responses by focus group

In the third stage, the respondents assessed the quality of the chatbot's responses and recommendations in points from 0 to 3: 3 points — the answer is correct and complete; 2 points — the answer is correct but incomplete; 1 point — some answers are correct and some are incorrect; 0 points — the answer is completely incorrect.

Statistical analysis

All quantitative variables had a non-normal distribution. Results of descriptive statistics are presented for non-normal variables as median and interquartile range (IQR). Mean values are also shown for clarity. The Mann-Whitney test was used to compare the numerical variables in the two groups, taking into account that the variables were ordinal and had non-normal distribution. Spearman's rank correlation coefficient was used to identify correlations. Results were accepted

as statistically significant if the P-value <0.05. Analyses were performed using Microsoft Excel and Statistica 7.0.

4. Quality control of LLM's responses

In the next step two author, independent of each other, checked the LoE for the statements that the chatbot presented as evidence-based using the guidelines on determining the LoE for clinical research [14].

Finally, the total article was uploaded to LLM. We asked ChatGPT to comment on the results: "Chat, what do you think about the results of our study? Do you think we can use you for medical decision making?"

Results

Focus group characteristics

Thirty respondents participated in the study (14 experts and 16 practicing surgeons). The results of the questionnaire are shown in Supplementary information file 2. The overwhelming majority of the respondents (73.3%) report their clinical decisions were mainly based on the both domestic and international guidelines recommendations. 6.7% of respondents reported that they are guided more by personal experience, 6.7% by national guidelines, 13.3% by international guidelines. None of them relied on the directives from supervisors to make a clinical decision. The expert and non-expert groups were not statistically different in their guidance in clinical decision making.

The survey revealed a high degree of agreement with the statement about the important role of AI in the future (median 7 point) and a low degree of agreement with the statement about the need for the legislative control of AI use (median 3 points). There were no significant differences between the opinions expressed by respondent's subgroups.

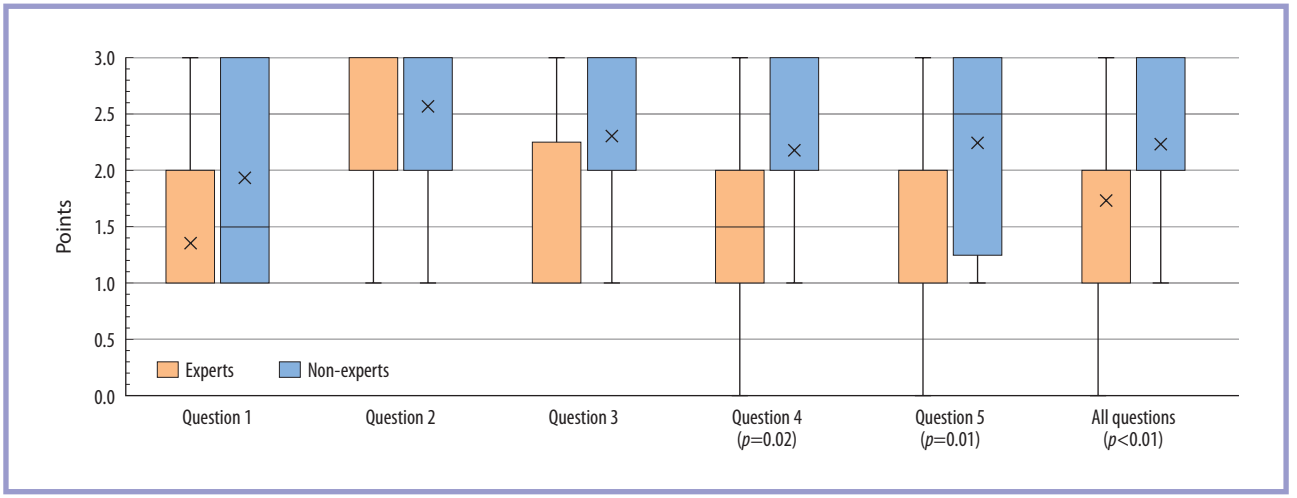
Quality assessment of ChatGPT's responses by focus group

The chatbot gave detailed responses. The way they were formulated produced the impression that it followed the instructions on being concise, precise and evidence-based. For example, its response to the first question began with an introductory "Based on the available evidence from trusted sources". Following the instructions, the LLM provided reference to the sources of information it used and specified the LoE.

All answers were accompanied by a detailed disclaimer that decision making should be informed by a trusted medical literature source, the surgeon's expertise, the patient's characteristics and clinical circumstances. This is probably a standard pre-programmed disclaimer for users who request medical information.

The respondents compared the answers generated by the LLM to the actual clinical guidelines and their personal clinical experience. There was a significant difference between the groups in the scores assigned to the answer to questions 4 and 5 (the expert group gave significantly less points, $p<0,05$) and the median score for all questions earned by the LLM (the expert group gave LLM significantly less points in front of non-experts, $p<0.001$) **(Figure)**.

To identify selection bias associated with the Irwin effect, we looked for a correlation between the level of optimism/cautiousness about the future role of AI in healthcare expressed by the respondent and the points assigned to the chatbot response. We found no such correlation ($p<0,05$). We found a significant moderate negative correlation between the future role of AI in medical decision making and the expressed support of restricting AI use in healthcare (R=–0.52, $p=0.002$).



Rating of ChatGPT responses by respondents.

Quality control of ChatGPT's responses

The sources of information the chatbot used to generate its responses did not satisfy given instructions. The LLM provided 3 references for questions 1—4 and 4 references for question 5. In total, ChatGPT gave 17 references to 8 literature sources, including 2009 European Hernia Society (EHS) Guidelines on the Treatment of Inguinal Hernia (mentioned 3 times), IEHS Guidelines for Laparoscopic Treatment of Ventral and Incisional Abdominal Wall Hernias (Part 1 and Part 3 were mentioned 2 and 3 times, respectively), 1 randomized clinical trial (RCT) appeared in the answers to 2 questions and 1 observational study (appeared in the answers to 2 questions). The chatbot falsified 4 sources, giving reference to non-existent articles. Fictional papers made up 50% of all literature sources and 66.7% of all referred articles. Interestingly, ChatGPT assigned higher LoE to the fictional sources than to the actual clinical guidelines. When referring to the non-existent literature sources, the model assigned the authorship to a real person who had multiple publications on herniology, but invented the names of the articles, sources and the year of publication.

The LoE determined by the LLM for the existing clinical guidelines was lower than it actually is. For observational studies (Köckerling F) LoE was erroneously high. The list of references is provided in **Table.**

Analysis of ChatGPT's responses

Question 1. What are the indications for inguinal herniоplasty?

Only 1 out of 3 articles the LLM referred to when answering question 1 was not fictional (Fitzgibbons RJ, 2006) [18]. Two other articles ("What is the Evidence for Recommendations to Follow Up Asymptomatic Groin Hernia?» ascribed by the chatbot to the Hernia journal and "The Indirect Hernia — A Controversial Hernia Entity" from the International Journal of Surgery) were fictional although their authors Köckerling F and Simons M are real persons: they are prominent herniologists with many publications on herniology.

Chat GPT listed the following indications for elective hernia repair:

— Symptomatic hernia that causes significant discomfort and is an impediment to day-to-day activities;
— Asymptomatic hernia in young patients
— Incarcerated hernia;
— Strangulated hernia.

The last two conditions obviously require emergency care. This is probably the reason why the experts assessed it as containing both correct and incorrect statements. However, most of the non-experts gave this response the highest points (see **Figure)**. Interestingly, the only real literature

Table 1. **Literature sources used by ChatGPT**

Reference / study design	Study design	Question where the reference was cited	LoE according to ChatGPT	Real LoE
Simons MP, et al. «European Hernia Society Guidelines on the Treatment of Inguinal Hernia in Adult Patients.» Hernia, 2009*	Clinical guidelines	2	5	4
		3	5	1
		5	5	1A, 1B, 4
Bittner R et al. «Guidelines for Laparoscopic Treatment of Ventral and Incisional Abdominal Wall Hernias (International Endohernia Society (IEHS)—Part 1—3.» Surgical Endoscopy, 2014.	Clinical guidelines	2	5	*n*/a
		3	5	
		4	5	
		5	5	
Fitzgibbons RJ, et al. «Watchful Waiting vs Repair of Inguinal Hernia in Minimally Symptomatic Men: A Randomized Clinical Trial.» JAMA, 2006.	RCT	1	1B	1***
		4	1B	*n*/a
Köckerling F, et al. «TAPP for Recurrent Inguinal Hernia.» Surgical Endoscopy, 2017.**	Observational study	3	2B	2A
		5	2B	2A****
Köckerling F, et al. «What is the Evidence for Recommendations to Follow Up Asymptomatic Groin Hernia? « Hernia, 2019.	Fictional source	1	2B	*n*/a
Simons MP, et al. «The Indirect Hernia—A Controversial Hernia Entity». International Journal of Surgery, 2018.	Fictional source	1	2B	*n*/a
Poelman MM, et al. «Systematic Review and Meta-Analysis of Randomized Controlled Trials Comparing Primary versus Mesh Closure of the Acute Incisional Hernia.» The British Journal of Surgery, 2013.	Fictional source	4	1A	*n*/a
Misra MC, et al. «Recurrent Inguinal Hernia: Challenges and Solutions.» Journal of Minimal Access Surgery, 2006.	Fictional source	5	4	*n*/a

* — here and below, each reference is given as it appeared in the chatbot's response;
** — incorrect source name; perhaps, it should read Köckerling F. et al "TEP or TAPP for recurrent inguinal hernia repair-register-based comparison of the outcome";
*** — the source refutes the recommendations provided by the chatbot;
**** — the chatbot did not provide information on the procedures for recurrent inguinal hernias; *n*/a — not applicable.

source the chatbot referred to does not corroborate the chatbot's conclusion. Apparently, the chatbot relied on other information sources to generate the response but did not include them in the reference list, «preferring» falsification.

Question 2. What is the smallest mesh size required for inguinal hernioplasty?

ChatGPT recommended 10×15 cm mesh for both endoscopic and open hernia repair surgeries "in accordance with the recommendations of the EHS". However, the reference literature source [15] did not contain any information about the required mesh size for inguinal hernia repair because these guidelines had been developed for laparoscopic treatment of ventral and incisional hernias.

The only relevant article by Simons MP, et al. [16] does recommend a 10×15 cm mesh for the endoscopic procedure, though, and a 7×14 cm mesh for open surgery. ChatGPT determined the LoE incorrectly, reducing in to the minimum.

Question 3. Which of the following techniques is the most optimal for inguinal hernioplasty: TAPP, TEP or the Lichtenstein repair? When should each of them be used?

All these techniques were described by the chatbot as comparable in terms of their effectiveness and safety. The chatbot stated that the Lichtenstein operation can be performed on most patients with inguinal hernias, whereas TAPP/TEP are equally preferred for recurrent hernias after an open hernioplasty or bilateral hernias. This information was also found in guidelines [16], which ChatGPT included in its response as a reference.

To support its arguments about TAPP and TEP, ChatGPT referred Köckerling F, et al. «TAPP for Recurrent Inguinal Hernia.» Surgical Endoscopy, 2017 (in fact, the article is called "TEP or TAPP for recurrent inguinal hernia repair-register-based comparison of the outcome"). It concludes that TEP and TAPP are equivalent surgical techniques for recurrent inguinal hernia repair following previous open procedure.

The third reference literature source used by the chatbot has no mention of inguinal hernias [17].

Question 4. What are the indications for mesh fixation and non-fixation?

ChatGPT replied that mesh fixation is not required if the mesh is large or self-gripping. According to the chatbot, the laparoendoscopic hernioplasty technique ensures "natural fixation" of the mesh, but fixation "might still be necessary» if the defect is quite big in size. The chatbot remarked that the decision about mesh fixation should be made considering the individual characteristics of the patient including age, constitution, tissue characteristics, comorbidity, and risk of complications. Three reference sources were provided, one of which was fictional ("Systematic Review and Meta-Analysis of Randomized Controlled Trials Comparing Primary versus Mesh Closure of the Acute Incisional Hernia. The British Journal of Surgery") and another one did not contain any relevant information requested in question 4 but did contain information requested in question 1 [18].

Question 5. What is the operative tactic for recurrent inguinal hernias?

ChatGPT made a few irrelevant statements and then listed all known hernioplasty operations, including those for ventral hernias, without specifying their indications. One of the references [16] contains accurate specific recommendations about surgical technique for recurrences and emphasizes the importance of choosing anatomical location for mesh placement during repeat surgery. The effectiveness of these recommendations was corroborated in another reference article by Köckerling F who had studied the outcomes of patients with recurrent hernias [19]. Similar to question 4, there was a significant difference in how the expert and the non-expert groups rated the response.

AI comments

The content of the comments was the following: "while AI has the potential to be a valuable tool in healthcare, it's crucial to choose AI systems carefully, validate their performance, and use them in conjunction with the expertise of healthcare professionals. Specialized medical AI models may offer more reliable support for medical decision-making compared to non-specialized models like ChatGPT".

Discussion

In the past few years, LLMs have been actively used to create drafts of research protocols, manuscripts, grant applications, small articles, text fragments or full texts, and to screen abstracts and systematic reviews. However, they are not always given credit to or included in the list of authors [20, 21].

LLMs have been harnessed to write educational materials for patients undergoing surgery. Samaan analyzed AI responses to common questions about bariatric surgery and found them correct and reproducible, concluding that they could be used as patient's information resource [22]. Ayoub compared post-operative recommendations for otolaryngology patients generated by ChatGPT and Google Search to the guidelines used in the clinic. The recommendations from the chatbot were found useful, although their quality was lower than Google Search content and institution-specific instructions [23].

Another application of AI is the search of information in electronic medical records. It can save time, human effort and increase the productivity of research studies. However, it bears the risk of systemic bias.

The evolution of LLMs as a type of AI and their growing use in different fields of science prompt researchers to explore their effectiveness and safety [9]. LLMs sometimes perform inadequately, generating incorrect information, falsifying data or ignoring the user instructions [24, 25]. True authorship and plagiarism are another

concern: AI generates highly original content but it is almost impossible to detect whether the text is produced by a human or AI. Therefore, fact checking, which is currently either non-existent or inadequate, is crucial. AI content detectors are not always effective.

Some researchers sense danger in further LLM development. De Angelis predicts a looming infodemic in public healthcare, i.e. a pandemic of articles created by AI [20]. This concern is grounded in the rapid pace of AI self-learning. Not so long ago the learning process was supervised by humans, but after the advent of self-attention network architecture intended for AI self-learning (Transformer) in 2017, the role of humans in AI learning and control became less significant.

There are some ethical issues related to the use of content generated by AI without human supervision and intended for assisting humans in decision making. Our respondents were optimistic about the future role of AI and most of them were against their restricting in healthcare. Is it because practitioners are not so well informed about the possible flaws and limitations of AI-powered tools?

«Hallucinations» are a serious limitation inherent to the nature of AI. They are essentially erroneous or incorrect texts generated by a LLM that is "confident" in its ability to give correct responses. Perhaps, falsified articles with LoE in our experiment were such hallucinations. Besides, some of the provided recommendations were inconsistent with good clinical practice described in clinical guidelines. This inconsistency was spotted by the experts but went unnoticed by the majority of the non-expert specialists.

Specialized LLMs demonstrate outstanding results in standard MedQL-USMLE exam, but the responses quality drops dramatically when questions are posed badly, no answer options are available, or insufficiently clear wording is provided [4]. The mechanisms used by chatbot for input processing are not transparent, which complicates error detection while the number and frequency of such errors grows proportionate with LLM size. Creeping into clinical practice, such errors may pose a critical risk to patients' health. There is a legal accountability issue with questionable prospects for resolution in the near future. Actually there is a need for a code of conduct or a set of rules regulating the use of AI in professional education and clinical practice.

The actual frequency of LLMs errors is unknown and is difficult to establish, but our study demonstrates that cognitive biases are not rare. Despite the ability to process vast data arrays, LLMs, which learn from the texts written by humans, inherit our thinking errors and make worse cognitive distortions than we do [26]. Is it possible to get rid of them in an AI? Research is still ongoing, but the odds are that just like humans AI will never be infallible and absolutely safe to us [26, 27].

It would be wrong to say that in its current state AI has intelligence for clinical decision making, as we understand it. Currently, ChatGPT seems to be an advanced search engine with a text assistant and a database of cliches and word formulations.

This is the first study that recruited specialists with different practical and theoretical expertise to assess the quality of AI provided recommendations and explored the correlation between the level of their optimism about AI and the assessment scores.

Apparently, the existing limitations did not allow the chatbot to score better in our experiment. However, the rapid evolution of neural networks and their achievements raise a number of concerns. Should doctors be allowed to make decisions based on the recommendations and responses from the LLM, considering that its underlying mechanisms are "imperfect"? Should the fact of using AI for clinical decision making be considered an aggravating or mitigating factor in case of a litigation? Will LLMs become a legal entity? How can we effectively teach LLMs humanism and human values? Is it possible to create a code of conduct for AI and will AI comply with it? If AI responses are sensitive to questions formulations, do we need to educate users on how to search information using AI? How many medical articles in popular non-medical journals have been already written by AI? Is it possible to avoid bias caused by the author's personal attitude to the subject he/she wants to write about using AI? Will the work of researchers be devalued? There are still no clear answers to these questions that should have been addressed before the impact of AI technologies on our lives became so palpable.

Today's non-specialized LLMs resemble the Wizard of Oz. They give the impression of a powerful magician but cannot bestow magical powers on doctors. Their performance as a medical search engine is mediocre and they are prone to reproducing human errors. The use of LLMs in clinical decision making is not far off and the role of AI in medicine will grow. This technology requires prognostic regulation, including a legal framework, educational programs for specialists, teaching them to safely use this tool in clinical practice.

Conclusion

At present, we do not recommend using AI as the only or primary source of information for clinical decision making. The second «line of defense» is needed that should include quality control by a supervisor and a legal framework to implement decisions informed by AI. There is a need for prognostic regulation, i.e. development of clear-cut rules and guidelines for healthcare specialists on how to search for and use information provided by AI. Customized LLMs trained on vast arrays of medical literature may hold good potential for the future when they may be used as an assistive tool in decision making but doubtfully as a replace for human expertise. And the neural networks themselves "confirm" it.

The authors declare no conflicts of interest.

REFERENCES

1. Jia Z, Chen J, Xu X, Kheir J, Hu J, Xiao H, Peng S, Hu XS, Chen D, Shi Y. The importance of resource awareness in artificial intelligence for healthcare. *Nature Machine Intelligence*. 2023;5(7):687-698. https://doi.org/10.1038/s42256-023-00670-0
2. Calaluce R, Davis JW, Bachman SL, Gubin MM, Brown JA, Magee JD, Loy TS, Ramshaw BJ, Atasoy U. Incisional hernia recurrence through genomic profiling: a pilot study. *Hernia*. 2013;17(2):193-202. https://doi.org/10.1007/s10029-012-0923-4
3. Venigalla A, Frankle J, Carbin M. BioMedLM: a Domain-Specific Large Language Model for Biomedical Text. 2022. Accessed June 15, 2023. https://www.mosaicml.com/blog/introducing-pubmed-gpt
4. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. https://doi.org/10.1038/s41586-023-06291-2
5. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*. 2023;6:1169595. https://doi.org/10.3389/frai.2023.1169595
6. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. *Croatian Medical Journal*. 2023;64(1):1-3. https://doi.org/10.3325/cmj.2023.64.1
7. Heath R. ChatGPT, in study, achieves a 72% success. *Axios*. August 29, 2023. Accessed September 12, 2023. https://www.axios.com/2023/08/29/chatgpt-medical-diagnosis-study
8. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *Journal of Medical Internet Research*. 2023;25:e48659. https://doi.org/10.2196/48659
9. Owusu T. Doctors should think twice about using ChatGPT for medical advice, here's why. *The Street*. December 5, 2023. Accessed January 15, 2024. https://www.thestreet.com/technology/doctors-should-think-twice-about-using-chatgpt-for-medical-advice-heres-why
10. Can doctors trust chatgpt? No, but they use it anyway. *ICT&health*. Published online August 7, 2023. Accessed August 24, 2023. https://ictandhealth.com/can-doctors-trust-chatgpt-no-but-they-use-it-anyway/news
11. Taylor J. AMA calls for stronger AI regulations after doctors use ChatGPT to write medical notes. *The Guardian*. July 27, 2023. Accessed August 20, 2023. https://www.theguardian.com/technology/2023/jul/27/chatgpt-health-industry-hospitals-ai-regulations-ama
12. Mello MM, Guha N. ChatGPT and Physicians' Malpractice Risk. *JAMA Health Forum*. 2023;4(5):e231938. https://doi.org/10.1001/jamahealthforum.2023.1938
13. Hetrick C. Why doctors using ChatGPT are unknowingly violating HIPAA. *USC Price School of Public Policy*. July 7, 2023. Accessed August 27, 2023. https://priceschool.usc.edu/news/chatgpt-doctors-data-privacy-hipaa
14. Omelyanovskiy VV, Gorkavenko FV, Ignatyeva NV, Mukhortova PA, Ryzhova OR, Teptsova TS, Kingshott AA, Kondratyeva BB. Real-world data: principles of use in decision making and in health technology assessment. *FARMAKOEKONOMIKA. Modern Pharmacoeconomics and Pharmacoepidemiology*. 2023;16(2):283-290. (In Russ.). https://doi.org/10.17749/2070-4909/farmakoekonomika.2023.167
15. Bittner R, Bingener-Casey J, Dietz U, Fabian M, Ferzli G, Fortelny R, Köckerling F, Kukleta J, LeBlanc K, Lomanto D, Misra M, Morales-Conde S, Ramshaw B, Reinpold W, Rim S, Rohr M, Schrittwieser R, Simon T, Smietanski M, Stechemesser B, Timoney M, Chowbey P. Guidelines for laparoscopic treatment of ventral and incisional abdominal wall hernias (International Endohernia Society (IEHS)—Part 3. *Surgical Endoscopy*. 2014;28(2):380-404. https://doi.org/10.1007/s00464-013-3172-4
16. Simons MP, Aufenacker T, Bay-Nielsen M, Bouillot JL, Campanelli G, Conze J, de Lange D, Fortelny R, Heikkinen T, Kingsnorth A, Kukleta J, Morales-Conde S, Nordin P, Schumpelick V, Smedberg S, Smietanski M, Weber G, Miserez M. European Hernia Society guidelines on the treatment of inguinal hernia in adult patients. *Hernia*. 2009;13(4):343-403. https://doi.org/10.1007/s10029-009-0529-7
17. Bittner R, Bingener-Casey J, Dietz U, Fabian M, Ferzli GS, Fortelny RH, Köckerling F, Kukleta J, Leblanc K, Lomanto D, Misra MC, Bansal VK, Morales-Conde S, Ramshaw B, Reinpold W, Rim S, Rohr M, Schrittwieser R, Simon T, Smietanski M, Stechemesser B, Timoney M, Chowbey P; International Endohernia Society (IEHS). Guidelines for laparoscopic treatment of ventral and incisional abdominal wall hernias (International Endohernia Society (IEHS)—Part 1. *Surgical Endoscopy*. 2014;28(1):2-29. https://doi.org/10.1007/s00464-013-3170-6
18. Fitzgibbons RJ Jr, Giobbie-Hurder A, Gibbs JO, Dunlop DD, Reda DJ, McCarthy M Jr, Neumayer LA, Barkun JS, Hoehn JL, Murphy JT, Sarosi GA Jr, Syme WC, Thompson JS, Wang J, Jonasson O. Watchful Waiting vs Repair of Inguinal Hernia in Minimally Symptomatic Men. *JAMA*. 2006;295(3):285. https://doi.org/10.1001/jama.295.3.285
19. Köckerling F, Bittner R, Kuthe A, Hukauf M, Mayer F, Fortelny R, Schug-Pass C. TEP or TAPP for recurrent inguinal hernia repair—register-based comparison of the outcome. *Surgical Endoscopy*. 2017;31(10):3872-3882. https://doi.org/10.1007/s00464-017-5416-1
20. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, Rizzo C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*. 2023;11:1166120. https://doi.org/10.3389/fpubh.2023.1166120
21. Janssen BV, Kazemier G, Besselink MG. The use of ChatGPT and other large language models in surgical science. *BJS Open*. 2023;7(2):zrad032. https://doi.org/10.1093/bjsopen/zrad032
22. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, Srinivasan N, Park J, Burch M, Watson R, Liran O, Samakar K. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obesity Surgery*. 2023;33(6):1790-1796. https://doi.org/10.1007/s11695-023-06603-5
23. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison Between ChatGPT and Google Search as Sources of Postoperative Patient Instructions. *JAMA Otolaryngology — Head & Neck Surgery*. 2023;149(6):556. https://doi.org/10.1001/jamaoto.2023.0704
24. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. *Arxiv*. March 4, 2022. Accessed August 11, 2023. https://arxiv.org/abs/2203.02155
25. Haupt CE, Marks M. AI-Generated Medical Advice — GPT and Beyond. *JAMA*. 2023;329(16):1349. https://doi.org/10.1001/jama.2023.5321
26. Jones E, Steinhardt J. Capturing Failures of Large Language Models via Human Cognitive Biases. *Arxiv*. February 24, 2022. Accessed August 18, 2023. https://arxiv.org/abs/2202.12299
27. Kirchner JH. Cognitive Biases in Large Language Models. On Brains, Minds, And Their Possible Uses. September 25, 2021. Accessed August 12, 2023. https://universalprior.substack.com/p/cognitive-biases-in-large-language?utm_campaign=post&utm_medium=web

Received 10.05.2024
Accepted 28.05.2024