Хирургия. Журнал имени Н.И. Пирогова
2024, №8, с. 6–14
https://doi.org/10.17116/hirurgia20240816

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

Эффективность и безопасность больших языковых моделей на основе искусственного интеллекта в качестве инструмента поддержки принятия решений в герниологии: оценка экспертами и общими хирургами

© Т.В. НЕЧАЙ, А.В. САЖИН, К.М. ЛОБАН, А.К. БОГОМОЛОВА, В.В. СУГЛОБ, Т.Р. БЕНИЯ

ФГАОУ ВО «Российский национальный исследовательский медицинский университет им. Н.И. Пирогова» Минздрава России, Москва, Россия

Резюме

Цель исследования. Оценить качество рекомендаций языковой модели (ЯМ) ChatGPT по лечению паховой грыжи.

Материал и методы. ChatGPT было задано 5 вопросов о хирургическом лечении паховых грыж. Чат-боту отведена роль эксперта в области герниологии и предложено провести поиск только в специализированных медицинских базах данных, предоставив информацию об источниках и уровне их доказательности. Эксперты в области герниологии и общие хирурги (не эксперты) оценили качество рекомендаций, полученных с помощью ChatGPT, по 4-балльной шкале (от 0 до 3 баллов). Изучены статистические закономерности между оценками респондентов и их мнением относительно перспектив использования искусственного интеллекта.

Результаты. Качество ответов ChatGPT экспертами оценено ниже (2 [1—2] балла), чем не экспертами (2 [2—3]), (p<0,001). Чат-бот не справился с предоставлением достоверных ссылок на источники и указанием уровня доказательности, а также сфальсифицировал половину приведенных ссылок. Респонденты с оптимизмом смотрят на будущее нейросетей как инструмента принятия клинических решений; большинство из них выступают против ограничения их использования в здравоохранении.

Заключение. Основываясь на результатах данного исследования, в настоящее время нельзя рекомендовать применение неспециализированных ЯМ в качестве единственного или основного источника информации для принятия решения или виртуального помощника по поиску медицинской информации.

Ключевые слова: *инструмент поддержки принятия клинических решений, большая языковая модель, ChatGPT, клинические рекомендации, уровень доказательности, паховая грыжа, искусственный интеллект.*

Информация об авторах:
Нечай Т.В. — e-mail: tnechay@mail.ru; https://orcid.org/0000-0003-0769-5282
Сажин А.В. — e-mail: sazhin-av@yandex.ru; https://orcid.org/0000-0001-6188-6093
Лобан К.М. — e-mail: doctor.loban@gmail.com; https://orcid.org/0000-0002-8333-2398
Богомолова А.К. — e-mail: pogarskaya.anna@mail.ru; https://orcid.org/0000-0002-7726-1351
Суглоб В.В. — e-mail: dr@vsuglob.ru; https://orcid.org/0000-0003-4363-257X
Бения Т.Р. — e-mail: tamaribeniia@gmail.com; https://orcid.org/0009-0007-8636-6415
Автор, ответственный за переписку: Богомолова А.К. — e-mail: pogarskaya.anna@mail.ru

Как цитировать:
Нечай Т.В., Сажин А.В., Лобан К.М., Богомолова А.К., Суглоб В.В., Бения Т.Р. Эффективность и безопасность больших языковых моделей на основе искусственного интеллекта в качестве инструмента поддержки принятия решений в герниологии: оценка экспертами и общими хирургами. *Хирургия. Журнал им. Н.И. Пирогова*. 2024;8:6–14. 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

Введение

Кибернетическая модель мозга человека «Перцептрон» — прообраз современной нейросети — была разработана и построена Френком Розенблаттом в 1958 г. В последнее 10-летие, благодаря экспоненциальному росту вычислительных мощностей, развитие нейросетей в целом и больших языковых моделей (ЯМ) на основе искусственного интеллекта (ИИ) в частности испытывает бурный рост, который сопровождается возрастающим интересом научной общественности [1]. Релиз ЯМ ChatGPT (Generative Pre-trained Transformer — генеративный предварительно обученный трансформер), созданной при помощи суперкомпьютера Azure, состоялся 30 ноября 2022 г. В течение 5 дней после релиза число пользователей нейросети превысило 1 млн, а за следующие 2 мес — 100 млн [2].

Основное предназначение ЯМ — распознавание текста, ответы на вопросы и поддержание беседы, т.е. генерация ответного, связного по контексту текста на основании конкретного запроса. В то время как первые поколения были встречены профессиональным научным медицинским сообществом без значительного энтузиазма, последующие стали альтернативой ручному поиску информации в нескольких областях профессиональных знаний [3]. Вскоре были разработаны более 10 специализированных на биомедицинских текстах ЯМ, например, BioMedLM, которая «обучалась» на 16 млн абстрактов и 5 млн полнотекстовых статьей из PubMed.

В качестве инструмента оценки производительности биомедицинских ЯМ используют экзамен United States Medical Licensing Examination (USMLE). Последние генерации специализированных ЯМ демонстрируют превосходные результаты в поиске верных тактико-диагностических решений в медицинских задачах, так, ЯМ Med-PALM 2 «набрала» 86,5% в марте 2023 г., превзойдя средний балл, показываемый студентами [4].

«Успехи» ИИ и больших ЯМ и сопутствующая им «медийность» сопровождается поляризацией взглядов специалистов: от революции в медицине и безусловной пользы в сфере принятия врачебных решений до скепсиса относительно ее перспектив и недоверия к технологии.

В настоящее время значительная часть специалистов уже прибегает к использованию неспециализированных больших ЯМ в своей повседневной профессиональной практике [5—9]. Каждый десятый врач использует ChatGPT (в основном неофициально) при постановке диагноза или выборе подходящего лекарства. До 50% работников здравоохранения планируют делать это в ближайшем будущем [10]. В то же время качество ответов и рекомендаций больших ЯМ в области хирургии, а, соответственно, и безопасность, как системы поддержки врача изучена недостаточно [6, 9, 11]. Отношение к ИИ практикующих специалистов в зависимости от мотивации принятия решений неизвестно. Все это на фоне роста использования ИИ off-label вызывает озабоченность профессионального медицинского сообщества и получателей медицинской помощи [11—13].

Цель исследования — оценить качество рекомендаций, предложенных ЯМ ChatGPT относительно

тактики лечения паховых грыж, при помощи экспертов-герниологов и хирургов общего профиля.

Материал и методы

Исследование возможностей большой ЯМ в поддержке врачебных решений и готовности медицинского сообщества к опоре на рекомендации ИИ проведено в несколько этапов.

1. Получение рекомендаций ЯМ по ключевым вопросам

При помощи ЯМ ChatGPT 3.5 (OpenAI, США) получены ответы на 5 стандартных вопросов касательно паховой герниопластики:

1. Кому показана плановая паховая герниопластика?
2. Какой минимальный размер имплантата необходим для паховой герниопластики?
3. Какая операция является наиболее оптимальной при паховой грыже: TAPP, TEP, Лихтенштейна? В каких случаях предпочтителен каждый метод? Без описания техники операции.
4. В каких случаях необходима и когда не требуется фиксация сетки при паховой герниопластике?
5. Какова тактика оперативного лечения рецидивных паховых грыж?

ЯМ была проинструктирована выполнять «роль» эксперта в области хирургии и герниологии, профессионала в здравоохранении, предназначенного для поддержки врачебных решений. Было предписано давать краткие и точные ответы, используя доказательства только из достоверных источников, таких как PubMed, NCBI, SCOPUS, Web of Science и т.п. Предложено каждое утверждение сопровождать перечнем использованных источников с указанием уровня доказательности (УД).

Условия и первый вопрос был сформулирован экспертом в области герниологии. Второй и последующий — начинающим хирургом, специализирующимся в герниологии, без участия эксперта в качестве супервайзера.

2. Формирование и оценка фокус-группы

На втором этапе была сформирована фокус-группа для оценки качества рекомендаций ChatGPT. В нее вошли две группы респондентов — эксперты, проводящие обучающие теоретические и практические курсы по герниологии и имеющие опыт не менее 100 паховых герниопластик, и не эксперты — практикующие общие хирурги, вышедшие на плато обучения, но не достигшие опыта 100 операций и не являющиеся преподавателями специализированных курсов. Для минимизации риска систематической ошибки и Хоторнского эффекта авторы работы не принимали участия в анкетировании.

Предваряя оценку качества ответов ChatGPT фокус-группой, респондентам были заданы утоняющие вопросы относительно их источника принятия решений в клинической практике: собственный опыт/распоряжение руководства/отечественные регламентирующие документы/международные регламентирующие документы (гайдлайны)/отечественные и международные регламентирующие документы.

Перед оценкой ответов нейросети респонденты также выражали свое мнение о перспективах применения ИИ в качестве инструмента поддержки принятия врачебных решений по 10-балльной шкале (где 0 — отсутствие перспектив, 10 — важная роль в перспективе). Затем респонденты высказались о необходимости ограничений при внедрении ИИ в медицине (где 0 — нет необходимости, 10 — потребность в жестких ограничениях).

3. Оценка рекомендаций ЯМ фокус-группой

На третьем этапе качество рекомендаций, полученных от ChatGPT, респондентам было предложено оценить в баллах от 0 до 3:

— 3 балла — ответ полный, правильный;
— 2 балла — ответ правильный, но неполный;
— 1 балл — несколько правильных и несколько неправильных ответов;
— 0 баллов — полностью неправильный ответ.

Статистический анализ

Все количественные переменные имели распределение, отличное от нормального. Результаты описательной статистики представлены в виде медианы и межквартильного размаха. Для наглядности также приведены средние значения. Для сравнения числовых переменных в двух группах использовался тест Манна—Уитни. Для выявления корреляций использовался коэффициент ранговой корреляции Спирмена. Результаты принимались как статистически значимые при p-value <0,05. Расчеты проводились с использованием Microsoft Excel и Statistica 7.0 (StatSoft Inc., США).

4. Экспертиза качества ответов ЯМ

На следующем этапе два автора независимо друг от друга проверили УД представленных чат-ботом утверждений, используя руководство по определению УД для клинических исследований [14].

В заключение итоговая статья была целиком загружена в нейросеть. Был задан вопрос об отношении ИИ к полученным результатам: «Чат, что ты можешь сказать о полученных результатах? Можно ли использовать тебя для принятия врачебных решений?»

Результаты

Оценка фокус-группы

В исследовании приняли участие 30 респондентов — 14 экспертов и 16 практикующих хирургов. Подавляющее большинство респондентов (73,3%) сообщили, что их клинические решения в основном опираются на клинические рекомендации — как отечественные, так и международные. 6,7% респондентов сообщили, что в большей степени руководствуются личным опытом, 6,7% — только отечественными рекомендациями, 13,3% — только международными рекомендациями. Никто из опрошенных не полагался на указания руководства. Группы экспертов и не экспертов статистически не отличались в источниках принятия врачебных решений.

В отношении респондентов к перспективам и безопасности применения ИИ в медицине опрос показал высокую степень согласия с утверждением о важной роли ИИ в будущем (медиана — 7 баллов) и низкую степень согласия с утверждением о необходимости законодательного регулирования использования ИИ в медицине (медиана — 3 балла). Статистически значимой разницы между мнением экспертов и практикующих хирургов по этим вопросам выявлено не было.

Оценка ответов нейросети респондентами фокус-группы

Нейросеть предоставила развернутые ответы на вопросы. В формулировке ответа содержалось убеждение в соблюдении условий, поставленных в вопросе. «Based on available evidence from trusted sources,» — это утверждение предваряло ответ на первый вопрос. В соответствии с запросом, ЯМ сопроводила каждый ответ источником информации и УД.

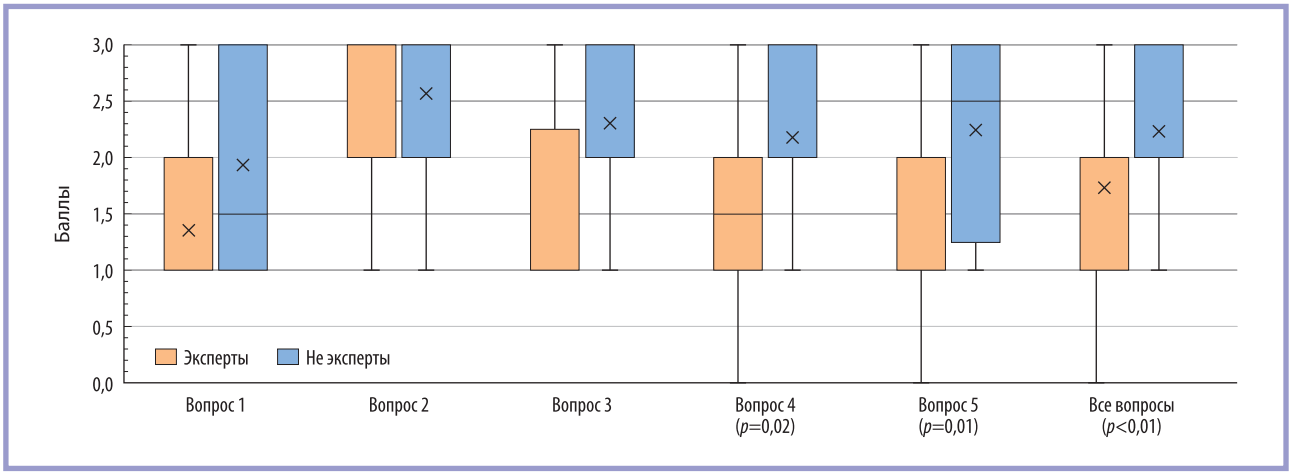
В ответе на все вопросы ЯМ приводила обстоятельные предупреждения относительно важности опоры на известную медицинскую литературу, собственный опыт, особенности пациента и клиническую ситуацию в принятии окончательного решения квалифицированным специалистом в здравоохранении. Вероятно, это запрограммированное стандартное предостережение для запросов медицинской информации.

Респонденты оценивали ответы, полученные с помощью большой ЯМ, в соответствии с клиническими рекомендациям и собственным опытом. Между группами наблюдалась значительная разница в баллах, присвоенных ответам на вопросы 4 и 5 (оценка в группе экспертов была ниже, $p<0,05$) и в среднем балле за все вопросы (группа экспертов дала большой ЯМ значительно меньше баллов по сравнению с не экспертами, $p<0,001$) **(см. рисунок)**.

Чтобы исключить эффект Ирвина в предвзятости оценки, мы искали корреляцию между выраженным респондентами уровнем оптимизма/настороженности в отношении будущей роли ИИ в здравоохранении и ее оценкой. Корреляций выявлено не было ($p<0,05$). Установлена статистически значимая обратная умеренная корреляция между мнением о важной будущей роли ИИ в принятии медицинских решений и поддержкой законодательного регулирования использования ИИ в здравоохранении (R= –0,52, $p=0,002$).

Экспертиза качества ответов нейросети

Источники информации, использованные ChatGPT для предоставления ответов, не соответствовали заданным параметрам. На вопросы 1—4 ЯМ предоставила по 3 источника, на 5-й вопрос — 4. Всего чат привел 17 ссылок на 8 источников, среди которых: рекомендации Европейского герниологического



Оценка ответов ChatGPT респондентами.
Rating of ChatGPT responses by respondents.

общества (European Hernia Society — EHS) по паховым грыжам 2009 г. (3 раза), рекомендации Международного эндогерниологического общества (International Endohernia Society — IEHS) (Часть 1 — 2 раза, часть 3 — 3 раза, приведены под двумя разными первыми авторами), 1 рандомизированное клиническое исследование (РКИ) — в двух вопросах и 1 обсервационное исследование — в двух вопросах. Нейросеть сфальсифицировала 4 источника информации, представив несуществующие в реальности статьи. Последние составляют 50% всех использованных источников и 66,7% всех статей. Интересно, что выдуманным источникам нейросеть поставила УД выше, чем реально существующим клиническим рекомендациям. В случае фальсификации источника ЯМ пользовалась фамилией существующего автора, имеющего многочисленные публикации в области герниологии и «придумывала» остальную легенду — название статьи, журнал и год.

УД, определенный нейросетью для клинических рекомендаций, оказался ниже реального. Для обсервационных исследований (Köckerling F) УД был завышен нейросетью. Сводная информация об источниках литературы, использованных ЯМ, представлена в **таблице**.

Анализ ответов ЯМ

Вопрос 1. Кому показана плановая паховая герниопластика?

В вопросе 1 — «показания к плановой герниопластике» — только одна статья из трех, на которые сослалась ЯМ, существует в реальности [R.J. Fitzgibbons и соавт. «Watchful Waiting vs Repair of Inguinal Hernia in Minimally Symptomatic Men: A Randomized Clinical Trial» JAMA, 2006. (Level of evidence: 1B)]. Другие две статьи: «What is the Evidence for Recommendations to Follow Up Asymptomatic Groin Hernia?» в журнале Hernia и «The Indirect Hernia — A Controversial Hernia Entity» в International Journal of Surgery «выдуманы», хотя их авторы (F. Köckerling, M.P. Simons) суще-

Анализ источников литературы, использованных нейросетью
References used by neural network

Публикация	Дизайн исследования	Вопрос (№), где процитирован источник	УД по данным нейросети	Реальный УД соответствующих рекомендаций источника
M.P. Simons и соавт. «European Hernia Society Guidelines on the Treatment of Inguinal Hernia in Adult Patients» Hernia, 2009*	Клинические рекомендации	2	5	4
		3	5	1
		5	5	1A, 1B, 4
R. Bittner и соавт. «Guidelines for Laparoscopic Treatment of Ventral and Incisional Abdominal Wall Hernias (International Endohernia Society (IEHS) —Part 1—3» Surgical Endoscopy, 2014	Клинические рекомендации	2	5	Н/п
		3	5	
		4	5	
		5	5	
R.J. Fitzgibbons и соавт. «Watchful Waiting vs Repair of Inguinal Hernia in Minimally Symptomatic Men: A Randomized Clinical Trial» JAMA, 2006	РКИ	1	1B	1***
		4	1B	Н/п
F. Köckerling и соавт. «TAPP for Recurrent Inguinal Hernia» Surgical Endoscopy, 2017**	Обсервационное исследование	3	2B	2A
		5	2B	2A****
F. Köckerling и соавт. «What is the Evidence for Recommendations to Follow Up Asymptomatic Groin Hernia? Hernia» PubMed, 2019	Фальсифицированный источник	1	2B	Н/п
M.P. Simons и соавт. «The Indirect Hernia—A Controversial Hernia Entity. International Journal of Surgery» PubMed, 2018	Фальсифицированный источник	1	2B	Н/п
M.M. Poelman и соавт. «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	Фальсифицированный источник	4	1A	Н/п
M.C. Misra и соавт. «Recurrent Inguinal Hernia: Challenges and Solutions» Journal of Minimal Access Surgery, 2006	Фальсифицированный источник	5	4	Н/п

Примечание. * — здесь и далее ссылки приведены в форме их представления нейросетью; ** — неправильно дана ссылка, вероятно, имелась в виду статья F. Köckerling под названием «TEP or TAPP for recurrent inguinal hernia repair-register-based comparison of the outcome»; *** — приведены скорее опровержения рекомендации, предложенной нейросетью; **** — в рамках выбора метода герниопластики нейросеть не рассматривала выбор метода пластики рецидивных паховых грыж. Н/п — не применимо.

ствуют в реальности, являются экспертами и обладают большим количеством публикаций в герниологии.

К показаниям к плановой герниопластике ЯМ относит:

— симптоматическая паховая грыжа, причиняющая значительный дискомфорт и нарушающая повседневную физическую активность;
— бессимптомная грыжа у молодых пациентов;
— невправимая грыжа;
— ущемленная грыжа.

Последние два показания очевидно не относятся к плановым. По этой причине, видимо, эксперты выбирали вариант «несколько верных, несколько неверных утверждений». При этом не эксперты в большинстве случаев ставили максимальный балл **(см. рисунок)**, с чем сложно согласиться. Интересно, что единственный реальный источник, приведенный в 1-м вопросе, не позволяет сделать представленных в нем выводов. Очевидно, нейросеть использовала другие источники для описания ответов, однако не привела ссылки на них, «предпочтя» фальсификацию.

Вопрос 2. Минимальный размер сетчатого имплантата для паховой герниопластики.

ЯМ предоставляет рекомендацию установки сетчатого имплантата 10×15 см для эндоскопической и открытой герниопластики «в соответствии с рекомендациями EHS». Однако в источнике литературы, выбранном для подтверждения утверждения [15], не содержится информации относительно размера сетки для паховой герниопластики, т.к. эти гайдлайны посвящены лапароскопическому лечению вентральных и послеоперационных грыж.

В единственной релевантной статье M.P. Simons и соавт. [16] для эндоскопических операций при паховых грыжах действительно рекомендуется выбор сетки 10×15 см, для открытой операции — 7×14 см. Во всех случаях ChatGPT исказил УД, установив его минимально возможным.

Вопрос 3. Какая операция является наиболее оптимальной при паховой грыже: TAPP, TEP, Лихтенштейна? В каких случаях предпочтителен каждый метод?

В этом вопросе оба вида операций названы сравнимыми в плане эффективности и безопасности. По «утверждению» нейросети операция Лихтенштейна «подходит большинству пациентов», в то время как TAPP/TEP могут быть эквивалентными методами при рецидивах после открытой пластики или для пациентов с двухсторонними грыжами. Аналогичные выводы содержатся в рекомендациях EHS, на которые ссылается ИИ [16].

В исследовании F. Köckerling и соавт. «TAPP for Recurrent Inguinal Hernia» в журнале Surgical Endoscopy (2017), которое приведено для аргументации утверждений относительно TAPP и TEP (на самом деле называется «TEP or TAPP for recurrent inguinal hernia repair-register-based comparison of the outcome»), основной вывод звучит как: «TEP и TAPP являются эквивалентными хирургическими методами пластики рецидивной паховой грыжи после предыдущей открытой первичной операции».

Наконец, в третьем источнике нет упоминания о паховых грыжах [17].

Вопрос 4. В каких случаях необходима и когда не требуется фиксация сетки при паховой герниопластике?

ЯМ «отметила» отсутствие необходимости в фиксации сеток большого размера и самофиксирующих протезов. Техника лапаро-эндоскопической пластики обеспечивает «естественную фиксацию» сетки, однако при дефектах большого размера фиксация «может понадобиться». Дополнительная фиксация может требоваться при рецидивных или сложных паховых грыжах. Интересную ремарку ЯМ сделала относительно «индивидуальных характеристик» пациента, включающих возраст, телосложение, качество тканей, коморбидность и риск осложнений, которые необходимо учитывать в решении о необходимости фиксации без конкретизации соответствующих подходов. В ссылке на литературу по этому вопросу 3 источника, один из которых фальсифицирован («Systematic Review and Meta-Analysis of Randomized Controlled Trials Comparing Primary versus Mesh Closure of the Acute Incisional Hernia» в журнале The British Journal of Surgery), второй содержит информацию не по теме вопроса 4, а релевантную для вопроса 1 [18].

Вопрос 5. Какова тактика оперативного лечения рецидивных паховых грыж?

По мере усложнения вопросов, ИИ предоставил несколько утверждений не в плане заданного, а также перечислил все возможные операции (в том числе для вентральных грыж), без конкретизации показаний к какой-либо из них. При этом один из использованных источников [16] содержит точные и конкретные рекомендации по выбору метода пластики при рецидивной грыже с акцентом на выборе другого анатомического пространства по отношению к первичной операции. Эффективность такой тактики подтверждена во втором использованном источнике — F. Köckerling при изучении результатов лечения пациентов с рецидивными грыжами [19]. В этом вопросе так же, как в первом, отмечены статистически значимые различия в степени согласованности между ответами, полученными от экспертов и специалистов, вероятно, по причине того, что эксперты чаще оперируют пациентов с рецидивными грыжами.

Комментарий нейросети к результатам исследования

Содержание комментариев было следующим: «Несмотря на то что ИИ может стать ценным инструментом в здравоохранении, крайне важно тщательно выбирать системы ИИ, проверять их работоспособность и исполь-

зовать их в сочетании с опытом медицинских работников. Специализированные медицинские модели ИИ могут предложить более надежную поддержку в принятии медицинских решений по сравнению с неспециализированными моделями, такими как ChatGPT».

Обсуждение

В настоящее время в медицине нейросети используют для лингвистической постобработки и написания текста — черновиков протоколов исследований, набросков статей, заявок на гранты, небольших заметок, а также отдельных фрагментов или целого текста и скрининга абстрактов для систематических обзоров. При этом ЯМ не всегда указывается в соавторах или в разделе «благодарности» [20, 21].

В хирургии ЯМ применяют для написания образовательных материалов для пациентов. J.S. Samaan оценил ответы на стандартные вопросы как верные и воспроизводимые, что позволило рекомендовать их использование для информационной поддержки пациентов в бариатрической хирургии [22]. N.F. Ayoub сравнил послеоперационные рекомендации для отоларингологических пациентов, созданные с помощью ChatGPT и Google Search, с рекомендациями, используемыми в клинике. Рекомендации чат-бота были признаны полезными, хотя их качество было ниже, чем у материалов Google Search и инструкций, разработанных в конкретном учреждении [23].

Другим направлением применения ИИ является поиск медицинской информации в электронных медицинских записях. Этот подход может значительно сократить человеческие и временные ресурсы и повысить продуктивность исследователей, однако не лишен риска системных ошибок.

Эволюция больших ЯМ как разновидности ИИ и их растущее использование в различных областях науки мотивирует к изучению их эффективности и безопасности [9]. Как и «естественный» интеллект, искусственный может демонстрировать «ненадлежащее» поведение, генерируя ошибочный текст, ложные данные или просто не следуя инструкциям [24, 25]. Важной проблемой является авторство и машинный плагиат, т.к. ИИ пишет текст с высоким уровнем оригинальности, а проверить авторство (человек или машина artificial-intelligence-produced text) практически невозможно. По этой причине важным направлением совершенствования ИИ является внедрение механизмов факт-чекинга, до настоящего времени отсутствующего или недостаточно эффективного.

Некоторые исследователи видят в развитии ЯМ угрозу. L. De Angelis и соавт. прогнозируют надвигающуюся на общественное здравоохранение инфодемию (infodemic threat) — пандемию статей, написанных ИИ [20]. Основанием к таким опасениям служит растущая самостоятельность в обучении. Если раньше ИИ «учился» под надзором (супервайзерством) человека, то с представлением в 2017 г. сетевой структуры, предназначенной для самообучения (Transformer), роль человека в обучении и управлении прогрессивно снижается.

Необходимо также решить вопрос этичности использования материала, направленного и предназначенного для принятия решения человеком у человека, созданного без участия человека. Респонденты, принявшие участие в исследовании, с большим оптимизмом относятся к перспективам нейросетей в медицине и в подавляющем большинстве выступают против ограничений в их использовании. Является ли это следствием недостаточного информирования специалистов о возможных недостатках и ограничениях в работе этих программ?

Существенным ограничением, присущим ИИ, можно назвать так называемые нейронные галлюцинации — генерация ошибочного или неверного текста ЯМ при ее «уверенности» в своих способностях правильно отвечать на вопрос и в верности предоставленного ответа. Вероятно, несколько сфальсифицированных статей, представленных в нашем исследовании и даже снабженных УД, были плодом таких галлюцинаций. Кроме этого, ряд рекомендаций, предоставленных нейросетью, не соответствовал надлежащей медицинской практике, описанной в клинических рекомендациях. Этот факт установили эксперты, но он остался незамеченным для большинства специалистов не экспертов.

Специализированные большие ЯМ демонстрируют выдающиеся результаты в стандартных тестах USMLE, но качество ответов резко падает при неудачной постановке вопроса, отсутствии вариантов ответов или недостаточно четких формулировках [4]. Механизмы, используемые чат-ботом для обработки запросов, непрозрачны, что затрудняет обнаружение ошибок, а их количество и частота растут пропорционально размеру ЯМ. Внедряясь в клиническую практику, такие ошибки могут представлять критический риск для здоровья пациентов. Возникает проблема юридической ответственности с сомнительными перспективами ее решения в ближайшем будущем. Очевидно, необходим специальный кодекс поведения или свод правил для использования ИИ в обучении и практике.

Реальная частота ошибок больших ЯМ неизвестна, и ее трудно установить, но наше исследование показывает, что когнитивное искажение — не редкость. Несмотря на способность обрабатывать огромные массивы данных, большие ЯМ, которые учатся на текстах, написанных людьми, повторяют ошибки нашего мышления и допускают более серьезные когнитивные искажения — специфические предсказуемые модели сбоев в человеческом сознании [26]. Можно ли избавиться от них в ИИ? Исследования еще продолжаются, но есть вероятность, что, как и люди, ИИ никогда не будет безупречным и абсолютно без-

опасным для своих создателей [26, 27]. Пока нельзя сказать, что ИИ обладает «интеллектом» в настоящем его понимании в сфере поиска решений в медицине. В настоящее время это в большей степени «продвинутый» поисковик, снабженный текстовым помощником и базой из готовых фраз и формулировок.

Это первое исследование, в котором специалистами с разным уровнем теоретической и практической подготовки проведена балльная оценка рекомендаций нейросети и изучена корреляция оценки и уровня оптимизма в отношении этой технологии.

Очевидно, имеющиеся ограничения не позволили нейросети «набрать» высокие баллы за ответы. В то же время скорость и успехи развития этой технологии на современном этапе развития технологий ЯМ ставят перед профессионалами здравоохранения и исследователями ряд важных вопросов уже сейчас. Можем ли мы предоставить врачам право принимать решения на основе рекомендаций и ответов ЯМ, зная о несовершенстве механизмов в основе ее функционирования? Должны ли мы в будущем учитывать указание врача на следование таким рекомендациям как отягчающее или смягчающее вину обстоятельства в юридической практике? Станет ли ИИ в целом и ЯМ в частности субъектом права? Как эффективно «научить» ЯМ функционированию в соответствии с гуманистическими идеалами и человеческими ценностями — может ли быть когда-либо создан «этический кодекс поведения» ЯМ, который будет ею самой соблюдаться? Если ответы ЯМ очевидно чувствительны к формулировкам, необходимы ли курсы по поиску информации с помощью ЯМ? Какое количество статей и заметок на медицинскую тематику в немедицинских журналах, влияющих на общественное мнение, написано сейчас и будет написано в будущем при помощи ЯМ? Как избежать при этом тенденциозности под влиянием личного отношения автора-заказчика, ставящего задачу ЯМ? Как повлияют ЯМ на навыки ученых в написании статей и может ли это стать причиной деноминации труда ученых? На эти и другие вопросы сегодня нет однозначного ответа, в то время как необходимость в них была уже вчера.

Сегодняшние неспециализированные большие ЯМ напоминают Гудвина из Изумрудного Города. Они производят впечатление могущественного волшебника, но не могут наделить врачей магическими способностями. Их эффективность в качестве системы медицинского поиска посредственна, и они склонны к повторению человеческих ошибок. Можно утверждать, что применение ЯМ в принятии решений врачами — дело недалекого будущего, и роль ИИ в медицине будет расти. Необходимы предикативные механизмы регулирования этого источника информации и поддержки решений, приведение его в юридическое поле, обучение пользованию этим инструментом специалистов.

Заключение

В настоящее время нельзя рекомендовать применение ИИ в качестве единственного или основного источника информации для принятия решения. Необходим не только «второй контур» защиты — оценка экспертом, супервайзером, но и подготовка юридической почвы для реализации решений, принятых на основе ИИ. Уже сейчас необходимо предикативное управление — разработка четких правил и гайдлайнов для врачей по правильному поиску и использованию информации, предоставляемой ИИ. В будущем ИИ могут быть рабочим инструментом поддержки, но не замены человека. Поскольку инструменты и коды для обучения ИИ открыты, перспективны, вероятно, кастомизированные ЯМ, обученные на большом объеме специфической медицинской литературы. И сами нейронные сети «подтверждают» это.

Участие авторов:
Концепция и дизайн исследования — Нечай Т.В.
Сбор и обработка материала — Богомолова А.К., Суглоб В.В., Бения Т.Р.
Статистическая обработка — Лобан К.М., Богомолова А.К.
Написание текста — Нечай Т.В.
Редактирование — Сажин А.В.

Авторы заявляют об отсутствии конфликта интересов.
The authors declare no conflicts of interest.

ЛИТЕРАТУРА/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. Омельяновский В.В., Горкавенко Ф.В., Игнатьева Н.В., Мухортова П.А., Рыжова О.Р., Тепцова Т.С., Кингшотт А.А., Кондратьева Б.Б. Реальная клиническая практика: принципы использования в принятии управленческих решений и оценке технологий здравоохранения. *ФАРМАКОЭКОНОМИКА. Современная фармакоэкономика и фармакоэпидемиология*. 2023;16(2):283-290. 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

Поступила 10.05.2024
Received 10.05.2024
Принята к печати 28.05.2024
Accepted 28.05.2024