Artificial intelligence in clinical practice: a cross-sectional survey of paediatric surgery residents' perspectives

Francesca Gigola ![ORCID],[1] Tommaso Amato,[1] Marco Del Riccio,[2] Alessandro Raffaele,[3,4] Antonino Morabito,[5,6] Riccardo Coletta[5,6]

¹School of Pediatric Surgery, University of Florence, Florence, Italy
²Department of Health Sciences, University of Florence, Florence, Italy
³Department of Maternal and Child Health, IRCCS Policlinico San Matteo Pavia, Pavia, Italy
⁴Department of Clinical-Surgical, Diagnostic and Pediatric Science, University of Pavia, Pavia, Italy
⁵Department of Neurosciences Psychology, University of Florence, Florence, Italy
⁶Department of Paediatric Surgery, Meyer Children's Hospital IRCCS, Florence, Italy

Correspondence to
Francesca Gigola;
francesca.gigola@unifi.it

ABSTRACT

Objectives The aim of this study was to compare the performances of residents and ChatGPT in answering validated questions and assess paediatric surgery residents' acceptance, perceptions and readiness to integrate artificial intelligence (AI) into clinical practice.

Methods We conducted a cross-sectional study using randomly selected questions and clinical cases on paediatric surgery topics. We examined residents' acceptance of AI before and after comparing their results to ChatGPT's results using the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) model. Data analysis was performed using Jamovi V.2.4.12.0.

Results 30 residents participated. ChatGPT-4.0's median score was 13.75, while ChatGPT-3.5's was 8.75. The median score among residents was 8.13. Differences appeared statistically significant. ChatGPT outperformed residents specifically in definition questions (ChatGPT-4.0 vs residents, p<0.0001; ChatGPT-3.5 vs residents, p=0.03). In the UTAUT2 Questionnaire, respondents expressed a more positive evaluation of ChatGPT with higher mean values for each construct and lower fear of technology after learning about test scores.

Discussion ChatGPT performed better than residents in knowledge-based questions and simple clinical cases. The accuracy of ChatGPT declined when confronted with more complex questions. The UTAUT questionnaire results showed that learning about the potential of ChatGPT could lead to a shift in perception, resulting in a more positive attitude towards AI.

Conclusion Our study reveals residents' positive receptivity towards AI, especially after being confronted with its efficacy. These results highlight the importance of integrating AI-related topics into medical curricula and residency to help future physicians and surgeons better understand the advantages and limitations of AI.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI) is becoming more and more integrated in clinical practice and could provide support to clinicians in their work. Residents and medical students are positioned at the confluence of tradition and innovation. Knowing the perspectives of residents on AI integration is fundamental to understanding the future of surgical care and defining the necessary steps to help prepare the next generation of surgeons.

WHAT THIS STUDY ADDS

⇒ Very few reports try to analyse the perception of residents on this topic. Our results show that ChatGPT-3.5 and ChatGPT-4.0 both outperform residents in answering clinical questions. When confronted with the results of ChatGPT compared with their own, residents showed a change in perception towards technology, expressing a more positive evaluation with reduced fear of technology.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our outcomes should not lead to the assumption that AI can outperform healthcare professionals in everyday clinical practice, but they can indicate that residents and doctors could rely on this tool for guidance and support, while being mindful of its limitations. Our results suggest that participants' initial beliefs regarding AI were challenged by the efficacy demonstrated by ChatGPT in answering the test questions and its performance compared with the residents. We believe that AI-themed topics should be introduced in the training of surgical residents to teach them how to use this tool to their advantage and help them understand its pitfalls. AI could be used as an aid during surgical training as well to improve residents' learning curves.

INTRODUCTION

The advent of artificial intelligence (AI) in medicine promises to redefine clinical practices, enhance diagnostic accuracy and personalise patient care. In paediatric surgery, where precision and adaptability are paramount, AI's potential to revolutionise the field is met with enthusiasm and caution.[1] Recent studies have showcased AI's capability to augment clinical decision-making, improve diagnostic processes and enhance surgical outcomes, particularly in paediatric populations where unique anatomical and physiological considerations exist.[2 3]

ChatGPT-4.0 is a cutting-edge AI language model that can present opportunities in the clinical field. Compared with other AI

systems, ChatGPT is easily accessible and user-friendly.[4 5] Even if the medical community appears optimistic about implementing AI in clinical practice, some clinicians still appear hesitant, fearing that treatment plans created by AI would be too generalised and not patient specific and some AI-generated answers could be wrong.[6 7]

Today, residents and medical students are positioned at the confluence of tradition and innovation. Knowing their perspectives on AI integration is fundamental to understanding the future of surgical care and defining the necessary steps to help prepare the next generation of surgeons to navigate a new way of working in healthcare. However, even if there are some literature reports focusing on the comparison between AI and residents' abilities in answering closed questions, very few try to analyse the perception of residents on this topic.[6 8–10] This study aims to assess the readiness, acceptability and perceived challenges towards AI among Italian residents in paediatric surgery, providing a foundation for addressing educational gaps and fostering a conducive environment for the seamless incorporation of AI technologies in paediatric surgery.

MATERIALS AND METHODS

We conducted a comparative longitudinal observational study with pre–post acceptance assessment to critically examine the proficiency of residents in paediatric surgery in contrast to the capabilities of ChatGPT-3.5 and ChatGPT-4. At the time of the study, the knowledge cut-off date for ChatGPT-3.5 (ie, the point in time when the data feeding the AI model was last updated) was January 2022, while for the new version of ChatGPT-4, it was April 2023. Additionally, we examined residents' acceptance of AI before and after comparing their results to ChatGPT results. Our methodology was segmented into three phases: AI acceptance assessment, comparative performance analysis and postexposure acceptance re-evaluation. The study was conducted through a survey which was digitised and disseminated to the residents through QualtricsXM to facilitate seamless participation, providing a comprehensive, unified form. Residents from all Italian Pediatric Surgery residency programmes were contacted through a national informal webchat with 74 participants and enrolled voluntarily. Participants were accorded a 30-day timeframe to submit their responses. Recruitment and data collection were performed in January and February 2024, and subsequent analyses were executed via Jamovi (V.2.4.12.0), ensuring rigorous statistical scrutiny.

Phase I: AI acceptance assessment

Using the validated Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) framework, this phase aimed to gauge participants' perception towards AI.[11–13]

Employing a 5-point Likert scale (ranging from 1=strongly disagree to 5=strongly agree), participants responded to a series of 38 questions proposed by a previously published study[14] to probe into 10 critical constructs of the UTAUT2 model: performance expectancy (eg, 'I believe that medical AI can help me grasp key information during the diagnosis and treatment process'), effort expectancy (eg, 'The process of learning how to use medical AI is simple'), social influence (eg, 'People who influence my behaviour think I should use medical AI'), facilitating conditions (eg, 'I have the necessary knowledge to use medical AI'), price value (eg, 'I believe that the pricing of medical AI is reasonable'), hedonic motivation (eg, 'Using medical AI is a pleasurable experience'), technology trust (eg, 'Medical AI can achieve the functionalities advertised'), technology fear (eg, 'I'm worried that relying too much on medical AI might lead to a decline in my disease diagnosis and treatment skills'), behavioural intention (eg, 'I will try using medical AI in my daily disease diagnosis and treatment processes') and user behaviour (eg, 'Using medical AI has become a habit for me'). The outcomes are presented as median and IQR.

Phase II: comparative performance analysis

In this phase of the study, we deployed a multiple-choice questionnaire encompassing 20 questions randomly extracted from a comprehensive pool of 81 potential queries. These questions were sourced from renowned databases, including the 'European Board of Pediatric Surgery 2017', 'EPSITE' and a compilation of '300+Top Pediatric Surgery Objective Questions and Answers'. Each query was engineered to have a singular correct response among five options (A–E), deliberately excluding image-based questions to accommodate the limitations of ChatGPT-3.5. The same set of questions was administered to the residents and ChatGPT-3.5 and ChatGPT-4, respectively, with scoring criteria meticulously outlined. Each correct answer earned 1 point, while each incorrect answer incurred a penalty of −0.25 points, and unanswered questions received 0 points. On ChatGPT-4.0 and ChatGPT-3.5, the multiple-choice questions were submitted into the chat by copying and pasting. Each time, the questions were asked following the same prompting, and before each attempt, the previous chat was deleted, the internet page was closed, the browser's chronology and cookies were deleted, and the cache was emptied.

Phase III: postexposure acceptance re-evaluation

Following the comparative analysis, participants were exposed to the mean scores and accuracy percentages derived from the residents and ChatGPT entities. Subsequently, they were prompted to repeat the initial UTAUT2 Questionnaire, enabling a comparative insight into any shifts in AI acceptance post exposure. To evaluate residents' responses to the UTAUT2 Questionnaire at two time points (T1=before the multiple-choice question test and T2=after the multiple-choice question test), a comprehensive Cronbach's alpha reliability analysis was initially conducted to assess the consistency and effectiveness of

each measure of the questionnaire's constructs. Subsequently, the individual items were combined for each construct to generate aggregate measures of the specific construct. This process was carried out for UTAUT2 at both T1 and T2. A paired-sample Wilcoxon test was employed to compare median values for each construct at T1 versus T2.

RESULTS
In total, 72 residents received the survey, and 30 (42%) completed the study. Among them, 9 were in the first year of residency (30%), 2 in the second year (7%), 12 in the third year (40%), 4 in the fourth year (13%) and 3 in the fifth and last year (10%).

Multiple-choice question test
Regarding the multiple-choice question test, ChatGPT-4.0's median score was 13.75 (IQR: 12.50–13.75, 74.5% accuracy, 17.5 peaks), conversely ChatGPT-3.5 obtained a score of 8.75 (8.44–10.00, 55.8% accuracy, 11.25 peak). The score among the residents was 8.13 (3.50–10.00, 46.7% accuracy, 12.75 peak). The Wilcoxon test revealed that the performance of both ChatGPT-3.5 and ChatGPT-4.0 was significantly better than that of residents in providing accurate responses (residents vs ChatGPT-3.5, $p=0.02$; residents vs ChatGPT-4.0, $p<0.001$).

More specifically, regarding the clinical reasoning questions, ChatGPT-4.0 and ChatGPT-3.5 have higher median accuracy than residents (ChatGPT-4.0=57.1%, IQR: 57.1–57.1; ChatGPT-3.5=57.1%, IQR: 42.9–57.1; residents=42.9%, IQR: 14.3–57.1; ChatGPT-4.0 vs residents, $p=0.005$; ChatGPT-3.5 vs residents, $p=0.1$), but the most important difference is in the definition questions in which the results of ChatGPT-3.5 but especially ChatGPT-4.0 outperform the results of the residents (ChatGPT-4.0=84.6%, IQR: 76.7–84.6; ChatGPT-3.5=61.5%, IQR: 53.8–61.5; residents=53.8%, IQR: 38.5–61.5; ChatGPT-4.0 vs residents, $p<0.0001$; ChatGPT-3.5 vs residents, $p=0.03$).

UTAUT2 Questionnaire (pre and post multiple-choice question test)
The Cronbach's alpha values for each construct indicated satisfactory to excellent internal consistency across different scales, with values ranging from 0.714 to 0.868, demonstrating the reliability of the measures utilised in our assessment.

At T1, the performance expectancy (PE) and hedonic motivation (HM) exhibited the highest value (PE $M_{t1}=3.83$, 3.58–4.08; HM $M_{t1}=3.83$, 3.33–4.00), while the technology fear displayed the lowest value ($M_{t1}=2.64$, 2.29–3.00). At T2, the PE and HM remained the highest value (PE $M_{t2}=4.00$, 3.92–4.00; HM $M_{t2}=4.00$, 3.58–4.08), while the technology fear was the lowest ($M_{t2}=2.29$, 2.14–3.14).

Specifically, paediatric surgery residents rated the price value of 2.67 (2.33–3.08) before the test and 3.00

(2.58–3.33) after being informed about the performance of ChatGPT, with a statistically significant difference ($p=0.007$).

Respondents also expressed higher and statistically significant ratings at T2 regarding user behaviour ($M_{t1}=3.00$, 2.50–3.31 vs $M_{t2}=3.25$, 2.75–3.50, $p=0.03$), PE ($M_{t1}=3.83$, 3.58–4.08 vs $M_{t2}=4.00$, 3.92–4.00, $p=0.03$), effort expectancy ($M_{t1}=3.00$, 2.75–3.56 vs $M_{t2}=3.25$, 3.00–4.00, $p=0.05$) and social influence ($M_{t1}=3.01$, 2.33–3.67 vs $M_{t2}=3.33$, 2.92–4.00, $p=0.02$).

Furthermore, even if not statistically significant, we can observe a more positive attitude of respondents towards ChatGPT in terms of technology trust ($M_{t1}=3.33$, 3.00–4.00 vs $M_{t2}=3.67$, 3.25–4.00, $p=0.07$), HM ($M_{t1}=3.83$, 3.33–4.00 vs $M_{t2}=4.00$, 3.58–4.08, $p=0.32$) and facilitating condition ($M_{t1}=2.83$, 2.33–3.33 vs $M_{t2}=3.00$, 2.67–3.67, $p=0.66$). Behavioural intention median score remained the same at T1 and at T2 ($M_{t1}=3.67$, 3.33–4.00 vs $M_{t2}=3.67$, 3.25–4.00, $p=0.76$).

By contrast, the technology fear value decreased at T2, even if the results were not statistically significant ($M_{t1}=2.64$, 2.29–3.00 vs $M_{t2}=2.29$, 2.14–3.14, $p=0.67$). Results are shown in table 1.

DISCUSSION
AI is becoming more and more involved in everyday lives and in clinical practice. It is fundamental to teach healthcare professionals how to deal with this tool, understand its advantages and risks and learn how to integrate it into their everyday work.

Applications of AI and chatbots, such as ChatGPT, range from clinical to surgical to patient counselling. Recent studies show that ChatGPT could help draft responses to patients' questions that physicians could edit. It has been shown that a chatbot could generate responses to patient questions in an online forum that were considered more empathetic than physicians' responses.[15] ChatGPT could also help give patients dietary advice, even if a professional check should always be considered, because in complex situations with the need for a tailored approach the efficacy appears reduced.[16] Recent studies also showed that AI could be a useful tool in surgery as a guide to identify anatomical structures during laparoscopic procedures, which has the potential to help reduce adverse events during surgery.[17]

Our study compared the performance of paediatric surgery residents versus ChatGPT in answering multiple-choice questions and then analysing the residents' perception of AI before and after being confronted with the test results. We observed a statistically significant difference between the results obtained by surgical residents and both versions of ChatGPT and between ChatGPT-3.5 and ChatGPT-4.0, with the latter outperforming the former. The accuracy seemed to decline when confronted with more complex or specialised questions, particularly those involving clinical cases. In fact, most ChatGPT errors occur when there are multiple data to consider and

Table 1 Mean results for each construct of the UTAUT2 Questionnaire at T1 and T2

	Items	Cronbach's alpha (at T1)	Cronbach's alpha (at T2)	Median† (at T1)	IQR (at T1)	Median† (at T2)	IQR (at T2)	P value
PE	3	0.740	0.726	3.83	3.58–4.08	4.00	3.92–4.00	0.03
EE	4	0.754	0.776	3.00	2.75–3.56	3.25	3.00–4.00	0.05
SI	3	0.797	0.867	3.01	2.33–3.67	3.33	2.92–4.00	0.02
FC	3	0.714	0.771	2.83	2.33–3.33	3.00	2.67–3.67	0.66*
PV	3	0.793	0.863	2.67	2.33–3.08	3.00	2.58–3.33	0.01
HM	3	0.868	0.835	3.83	3.33–4.00	4.00	3.58–4.08	0.32*
HT	3	0.743	0.764	3.00	2.50–3.31	3.25	2.75–3.50	0.03
TR	3	0.867	0.793	3.33	3.00–4.00	3.67	3.25–4.00	0.07*
BI	3	0.788	0.769	3.67	3.33–4.00	3.67	3.25–4.00	0.76*
TF	7	0.781	0.852	2.64	2.29–3.00	2.29	2.14–3.14	0.67*

*p=not significant (>0.05).
†Median of 1–5 Likert scale; T1=before multiple-choice test and T2 after multiple-choice test.
BI, behavioural intention; EE, effort expectancy; FC, facilitating condition; HM, hedonic motivation; HT, user behaviour; PE, performance expectancy; PV, price value; SI, social influence; TF, technology fear; TR, technology trust; UTAUT2, Unified Theory of Acceptance and Use of Technology 2.

especially when clinical scenarios are complex. However, residents showed lower accuracy in responses to these questions as well and were outperformed by ChatGPT. ChatGPT-3.5 showed lower accuracy than ChatGPT-4.0 when answering questions related to medical definitions, while both versions performed better than residents. Our results align with other findings: one study showed that AI demonstrated high accuracy in different medical specialities, achieving 97% in multiple-choice questions when tested with The New England Journal of Medicine quiz. In another study, ChatGPT outperformed residents nationally on Plastic Surgery In-Service Examinations.[18] Furthermore, it has been shown that when ChatGPT is provided with specific knowledge, its performance improves, approaching human-level accuracy.[19]

The limited number of participants in our study makes it difficult to draw definitive conclusions, and this outcome should not lead us to assume that AI can outperform healthcare professionals in everyday clinical practice. It is essential to recognise that the randomly selected questions aim to primarily assess participants' punctual knowledge: this method may not accurately esteem AI's performance in real-life scenarios, but it indicates that residents and doctors could rely on this tool for guidance and support, while being mindful of its limitations. It must also be acknowledged that, to account for the limitations of ChatGPT-3.5, image-based questions were excluded. This further restricted the ability to simulate real-life scenarios, limiting the accuracy of assessing ChatGPT's efficacy. We must remember that ChatGPT was not explicitly designed for this purpose and was trained using different sources, some of which may not always be reliable. Consequently, the information provided could be inaccurate or even misleading, and the system needs continuous updating. It must also be kept in mind that

AI has the potential to reproduce biases in the training data.[20]

In our study, participants were also asked to fill in the UTAUT2 Questionnaire both before and after being confronted with their results on the test and those of ChatGPT. The UTAUT2 Questionnaire is an acceptance model designed by Venkatesh et al[11 12] to understand attitudes towards technology. It is based on evaluating several aspects that can define acceptance of technology. Interestingly, all items of the UTAUT2 Questionnaire showed higher ratings at T2 than T1, except for fear of technology, which decreased. Among the included residents, the predominant aspect that emerged was their high perception of the potential of using AI: PE was the highest mean value at both T1 and T2. PE refers to the degree to which technology will benefit users in performing certain activities. Its rating was notably higher at T2, almost reaching 4 out of 5 points. Participants perceived AI as simple to use, as reflected by the high value of effort expectancy, which represents the degree of ease associated with consumers' use of technology. Interestingly, its value increased from 3.15 to 3.42 at T2, indicating that residents found AI even easier to use after being presented with the test results. Price value, which shows how consumers balance the benefits with the cost, also increased significantly from T1 to T2: its rating reaches a mean of 2.63/5 at T1 and a mean of 3.06 at T2 with a difference that appears to be statistically significant. Furthermore, technology fear also appears lower at T2, although not statistically significant. These results suggest that participants' initial beliefs regarding the high price of AI compared with its benefits were challenged by the efficacy demonstrated by ChatGPT in answering the test questions and its performance compared with the residents. Interestingly, this knowledge also decreased the fear of AI at T2.

The limitations of this study include a low response rate among residents (42%), which may introduce selection bias by predominantly including individuals with a more favourable opinion of AI. The relatively small sample size of the cohort must also be acknowledged, as it may reduce the statistical power of the findings. However, despite its limited size, our cohort represents nearly half of all Italian Pediatric Surgery residents, making it a strong representation of this small population. Additionally, the residents who participated were from various residency years, making training level a variable that could influence their education and ability to respond to the questions. Unfortunately, the limited number of participants prevented a subgroup analysis based on residency year. It would be interesting in the future to propose this study aiming to include a larger number of residents and even specialists to compare performances and results. Moreover, although the proposed questions were primarily drawn from the European Board of Pediatric Surgery 2017 and EPSITE tests, they were randomly selected from larger question pools without validation for difficulty level. This may raise concerns about the consistency and reliability of the assessment.

The results of this survey highlight that learning about the capabilities of ChatGPT and the potential associated with integrating AI in clinical care could lead to a shift in perception, resulting in a more positive attitude toward AI. The statistically significant modification in the residents' perceptions of AI post exposure to comparative performance data with ChatGPT highlights the fundamental role of direct experience and evidence in shaping attitudes towards technology. Integrating AI into paediatric surgery residency programmes could significantly impact residents' training and preparedness for future clinical practice: residents can gain valuable experience in utilising advanced technologies for patient care, decision-making and research.

AI can also serve as an educational tool and allow objective evaluation of residents' skills and performance, particularly in minimally invasive surgery training through video-based assessments and as an aid in laparoscopic and robotic surgery or in virtual reality simulation.[21 22] By incorporating motion-tracking systems, AI can analyse factors such as dissection speed and precision, as well as study biometric data to assess surgical performance.[23–26] Additionally, some studies have explored artificial systems designed to provide feedback and tutoring to surgical residents, while others have examined the use of AI to clarify and break down the steps of specific surgical procedures for training purposes.[27–30] AI systems such as ChatGPT can also serve as valuable support tools for students and residents preparing for examinations. They can provide quick insights into specific questions, reducing the time needed for extensive research while enhancing understanding of complex topics.[18] Moreover, AI could serve as a valuable tool in various disciplines and contexts, including the development of educational programmes for healthcare workers and patients to address vaccine hesitancy, a critical issue that must be addressed, aiming to improve awareness and attitudes.[31] A recent review highlights the role of chatbots in fostering a vaccine-literate environment by combating misinformation and enhancing communication with healthcare professionals.[32] The integration of AI in the educational field is even more justified by the fact that it is estimated that the doubling time of medical knowledge went from 50 years in the 1950s to 3.5 years in the 2010s. This expansion of knowledge will force medical schools and residency programmes to redefine the essential core of what students must learn and how they have to do it.[33] This integration of AI in different medical fields should also be associated with creating guidelines regarding ethical aspects, privacy, data security and patient autonomy. Professionals such as clinicians, computer scientists, and ethicists must develop AI tools that prove ethically safe and scientifically accurate. Paediatric surgery residents' inputs and ideas will be invaluable in creating a future where AI enhances, rather than supplants, the work of paediatric surgeons.

CONCLUSIONS

Our study reveals residents' positive receptivity towards AI, especially after being confronted with its efficacy. These results highlight the importance of integrating AI-related topics into medical curricula and residency to help future physicians and surgeons better understand the advantages of AI, while also keeping in mind ethical considerations and remembering its limitations. With the evolution of AI and its possible integration in the paediatric surgery field, we will need ongoing education, transparency and multidisciplinary collaboration to improve paediatric healthcare.

ORCID iD
Francesca Gigola http://orcid.org/0000-0001-6732-4243

REFERENCES

1. Aoki T, Yamada A, Aoyama K, *et al*. Clinical usefulness of a deep learning-based system as the first screening on small-bowel capsule endoscopy reading. *Dig Endosc* 2020;32:585–91.
2. Bien N, Rajpurkar P, Ball RL, *et al*. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med* 2018;15:e1002699.
3. Steiner DF, MacDonald R, Liu Y, *et al*. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am J Surg Pathol* 2018;42:1636–46.
4. Xiao D, Meyers P, Upperman JS, *et al*. Revolutionizing Healthcare with ChatGPT: An Early Exploration of an AI Language Model's Impact on Medicine at Large and its Role in Pediatric Surgery. *J Pediatr Surg* 2023;58:2410–5.
5. Cascella M, Montomoli J, Bellini V, *et al*. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst* 2023;47:33.
6. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for Clinical Practice and Medical Education: Cross-Sectional Survey of Medical Students' and Physicians' Perceptions. *JMIR Med Educ* 2023;9:e50658.
7. Pedro AR, Dias MB, Laranjo L, *et al*. Artificial intelligence in medicine: A comprehensive survey of medical doctor's perspectives in Portugal. *PLoS One* 2023;18:e0290613.
8. Guerra GA, Hofmann H, Sobhani S, *et al*. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. *World Neurosurg* 2023;179:e160–5.
9. Lum ZC. Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. *Clin Orthop Relat Res* 2023;481:1623–30.
10. St John A, Cooper L, Kavic SM. The Role of Artificial Intelligence in Surgery: What do General Surgery Residents Think? *Am Surg* 2024;90:541–9.
11. Venkatesh V, Morris MG, Davis GB, *et al*. User Acceptance of Information Technology: Toward a Unified View. *MIS Q* 2003;27:425.
12. Venkatesh V, Thong JYL, Xu X. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Q* 2012;36:157.
13. Oshlyansky L, Cairns P, Thimbleby H. Validating the unified theory of acceptance and use of technology (UTAUT) tool cross-culturally. Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK; 2007
14. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. *BMC Med Educ* 2023;23:852.
15. Ayers JW, Poliak A, Dredze M, *et al*. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023;183:589–96.
16. Ponzo V, Goitre I, Favaro E, *et al*. Is ChatGPT an Effective Tool for Providing Dietary Advice? *Nutrients* 2024;16:469.
17. Madani A, Namazi B, Altieri MS, *et al*. Artificial Intelligence for Intraoperative Guidance: Using Semantic Segmentation to Identify Surgical Anatomy During Laparoscopic Cholecystectomy. *Ann Surg* 2022;276:363–9.
18. Hubany SS, Scala FD, Hashemi K, *et al*. ChatGPT-4 Surpasses Residents: A Study of Artificial Intelligence Competency in Plastic Surgery In-service Examinations and Its Advancements from ChatGPT-3.5. *Plast Reconstr Surg Glob Open* 2024;12:e6136.
19. Russe MF, Fink A, Ngo H, *et al*. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. *Sci Rep* 2023;13:14215.
20. Braga AVNM, Nunes NC, Santos EN, *et al*. Use of ChatGPT in Urology and its Relevance in Clinical Practice: Is it useful? *Int Braz J Urol* 2024;50:192–8.
21. Kirubarajan A, Young D, Khan S, *et al*. Artificial Intelligence and Surgical Education: A Systematic Scoping Review of Interventions. *J Surg Educ* 2022;79:500–15.
22. Guerrero DT, Asaad M, Rajesh A, *et al*. Advancing Surgical Education: The Use of Artificial Intelligence in Surgical Training. *Am Surg* 2023;89:49–54.
23. Uemura M, Tomikawa M, Miao T, *et al*. Feasibility of an AI-Based Measure of the Hand Motions of Expert and Novice Surgeons. *Comput Math Methods Med* 2018;2018:9873273.
24. Kowalewski K-F, Garrow CR, Schmidt MW, *et al*. Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. *Surg Endosc* 2019;33:3732–40.
25. Oquendo YA, Riddle EW, Hiller D, *et al*. Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surg Endosc* 2018;32:1840–57.
26. Gao Y, Yan P, Kruger U, *et al*. Functional Brain Imaging Reliably Predicts Bimanual Motor Skill Performance in a Standardized Surgical Task. *IEEE Trans Biomed Eng* 2021;68:2058–66.
27. DiPietro R, Ahmidi N, Malpani A, *et al*. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. *Int J CARS* 2019;14:2005–20.
28. Ahmidi N, Poddar P, Jones JD, *et al*. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *Int J CARS* 2015;10:981–91.
29. Despinoy F, Bouget D, Forestier G, *et al*. Unsupervised Trajectory Segmentation for Surgical Gesture Recognition in Robotic Training. *IEEE Trans Biomed Eng* 2016;63:1280–91.
30. Guzmán-García C, Gómez-Tome M, Sánchez-González P, *et al*. Speech-Based Surgical Phase Recognition for Non-Intrusive Surgical Skills' Assessment in Educational Contexts. *Sensors (Basel)* 2021;21:1330.
31. Lanza TE, Paladini A, Marziali E, *et al*. Training needs assessment of European frontline health care workers on vaccinology and vaccine acceptance: a systematic review. *Eur J Public Health* 2023;33:591–5.
32. Cosma C, Radi A, Cattano R, *et al*. Exploring Chatbot contributions to enhancing vaccine literacy and uptake: A scoping review of the literature. *Vaccine (Auckl)* 2025;44:126559.
33. Densen P. Challenges and opportunities facing medical education. *Trans Am Clin Climatol Assoc* 2011;122:48–58.