European Journal of Cancer 205 (2024) 114100

Contents lists available at ScienceDirect

European Journal of Cancer

journal homepage: www.ejcancer.com

Original research

ChatGPT's Gastrointestinal Tumor Board Tango: A limping dance partner?

Ughur Aghamaliyev [a], Javad Karimbayli [b], Clemens Giessen-Jung [c], Ilmer Matthias [a,d], Kristian Unger [d,e,f], Dorian Andrade [a], Felix O. Hofmann [a,d], Maximilian Weniger [a], Martin K. Angele [a], C. Benedikt Westphalen [c,d], Jens Werner [a], Bernhard W. Renz [a,d,*]

[a] Department of General, Visceral and Transplantation Surgery, LMU University Hospital, LMU Munich, Germany
[b] Division of Molecular Oncology, Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, National Cancer Institute, Aviano, Italy
[c] Comprehensive Cancer Center Munich & Department of Medicine III, LMU University Hospital, LMU Munich, Germany
[d] German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany
[e] Department of Radiation Oncology, University Hospital, LMU Munich, 81377
[f] Bavarian Cancer Research Center (BZKF), Munich, Germany

ARTICLE INFO

Keywords:
ChatGPT
AI
MTB
Concordance
Gastrointestinal cancers
PDAC
CCC
HCC

ABSTRACT

Objectives: This study aimed to assess the consistency and replicability of treatment recommendations provided by ChatGPT 3.5 compared to gastrointestinal tumor cases presented at multidisciplinary tumor boards (MTBs). It also aimed to distinguish between general and case-specific responses and investigated the precision of ChatGPT's recommendations in replicating exact treatment plans, particularly regarding chemotherapy regimens and follow-up protocols.

Material and methods: A retrospective study was carried out on 115 cases of gastrointestinal malignancies, selected from 448 patients reviewed in MTB meetings. A senior resident fed patient data into ChatGPT 3.5 to produce treatment recommendations, which were then evaluated against the tumor board's decisions by senior oncology fellows.

Results: Among the examined cases, ChatGPT 3.5 provided general information about the malignancy without considering individual patient characteristics in 19% of cases. However, only in 81% of cases, ChatGPT generated responses that were specific to the individual clinical scenarios. In the subset of case-specific responses, 83% of recommendations exhibited overall treatment strategy concordance between ChatGPT and MTB. However, the exact treatment concordance dropped to 65%, notably lower in recommending specific chemotherapy regimens. Cases recommended for surgery showed the highest concordance rates, while those involving chemotherapy recommendations faced challenges in precision.

Conclusions: ChatGPT 3.5 demonstrates potential in aligning conceptual approaches to treatment strategies with MTB guidelines. However, it falls short in accurately duplicating specific treatment plans, especially concerning chemotherapy regimens and follow-up procedures. Ethical concerns and challenges in achieving exact replication necessitate prudence when considering ChatGPT 3.5 for direct clinical decision-making in MTBs.

1. Introduction

Following its launch in November 2022, ChatGPT 3.5 emerged as a major highlight of 2023. This publicly available artificial intelligence (AI) model swiftly gained massive popularity, drawing in more than a million users in just the first week of its release [1]. Notably, the scientific community rapidly adopted ChatGPT 3.5, leveraging its features across various disciplines [2]. Its significant impact was particularly felt in scientific writing, where ChatGPT 3.5 was extensively utilized [3]. The AI demonstrated its prowess by producing abstracts that passed plagiarism checks and were often mistaken for genuine work by human evaluators, with around 32% of abstracts considered authentic [2]. A landmark moment was reached when ChatGPT was named the primary author in a scientific paper, sparking both fascination and controversy [4].

This unconventional acknowledgment of authorship attracted

* Correspondence to: Department of General, Visceral, and Transplantation Surgery Hospital of the University of Munich, Ludwig-Maximilians-University Munich, Marchioninistr. 15, 81377 München, Germany.
E-mail address: bernhard.renz@med.uni-muenchen.de (B.W. Renz).

https://doi.org/10.1016/j.ejca.2024.114100
Received 12 April 2024; Accepted 23 April 2024
Available online 7 May 2024

criticism from parts of the research community [5]. The ethical considerations of assigning authorship to an AI like ChatGPT in scholarly writing became a hot topic of debate [6]. However, the AI's ability to formulate cogent written arguments led to further exploration of its potential in academic testing.

A significant investigation by Gilson et al. assessed ChatGPT's abilities in answering questions similar to those in the United States Medical Licensing Examination Steps 1 and 2, where it performed on par with a third-year medical student. This finding ignited further interest and research into ChatGPT's capabilities [7].

The journey of ChatGPT into more rigorous scrutiny began here. Researchers globally started to explore how it could be used in clinical settings [8]. This exploration was inspired by Watson for Oncology [9], which provides expert recommendations in multidisciplinary tumor boards (MTBs) [10] [11]. Some ventured into examining ChatGPT's potential in clinical decision-making [1]. A notable study involving 157 colorectal cancer cases compared ChatGPT's input with that of MTB, finding a high concordance in postoperative recommendations [12]. Interestingly, discrepancies arose when surgery was advised by the board, while ChatGPT suggested neoadjuvant chemotherapy [12]. Despite the growing research, the reproducibility of ChatGPT's advice has yet to be fully assessed.

Therefore, this study seeks to determine the reliability and replicability of ChatGPT's recommendations compared to those of a MTB for gastrointestinal tumor cases, alongside examining the reproducibility of these recommendations.

2. Material and methods

2.1. Patient population

A retrospective study was carried out using data from 448 patients who were discussed at our MTB from August to November 2022. Out of these, 115 cases met the selection criteria, including first-time diagnoses of pancreatic ductal adenocarcinoma (PDAC), cancers of the stomach (GC) and esophagus (EC), hepatocellular carcinoma (HCC), cholangiocarcinoma (CCC), or cases presented post-surgery at the MTB. Colorectal cancer cases were excluded because local colon cancer cases were not systematically presented in MTB meetings preoperatively at our facility. We also excluded patients with complicated cancer histories and cases where the treatment recommendations were informed by guidelines or research published after September 2021. The comprehensive patient selection process is outlined in Fig. 1.

2.2. Study design

Patient information, including age, gender, ASA classification, primary diagnosis, details of oncological treatments, and if relevant, surgical procedures, were compiled from medical records and translated into English. Subsequently, all patient information was permanently anonymized following data collection. The creation of this database received approval from the local ethics committee [reference number: 23–0175].

A senior resident, unaware of the MTB's final recommendations, manually entered the selected cases into ChatGPT 3.5. The queries were framed to ask for the first-line treatment option for a given patient based on the most recent German guidelines for specific cancers, up to the latest update. The case descriptions included age, gender, the diagnosis of local or locally advanced cancer, and the status of lymph node/distant metastasis. Where applicable, details of endoscopic or computed tomographic staging and previous cancer treatments were also provided. Each case was entered into the chat interface individually, and ChatGPT's recommendations were iterated at least three times without further discussions or analysis of the responses. These recommendations were later compared with the MTB's decisions by two senior oncology fellows who regularly chair tumor board meetings. The comparison focused on treatment strategy (e.g., neoadjuvant therapy, primary surgery, adjuvant therapy, follow-up, systemic therapy) and specific treatment details (type of chemotherapy or intervention).

For this study, ChatGPT's therapy recommendations were categorized as follows: responses providing general disease information without considering individual patient characteristics were tagged as "general information," while those directly addressing the clinical case, starting with phrases like "in this particular case", “for a patient like the one described”, or "for this patient", were marked as "case-specific". If at least one out of three responses significantly diverged from the others, it was considered "non-reproducible" and automatically classified as "not recommended". Responses that matched or closely resembled the MTB's recommendations were labeled as “recommended” and “in consideration”, indicating concordance. Conversely, recommendations from ChatGPT that did not align with these standards were identified as non-concordant.

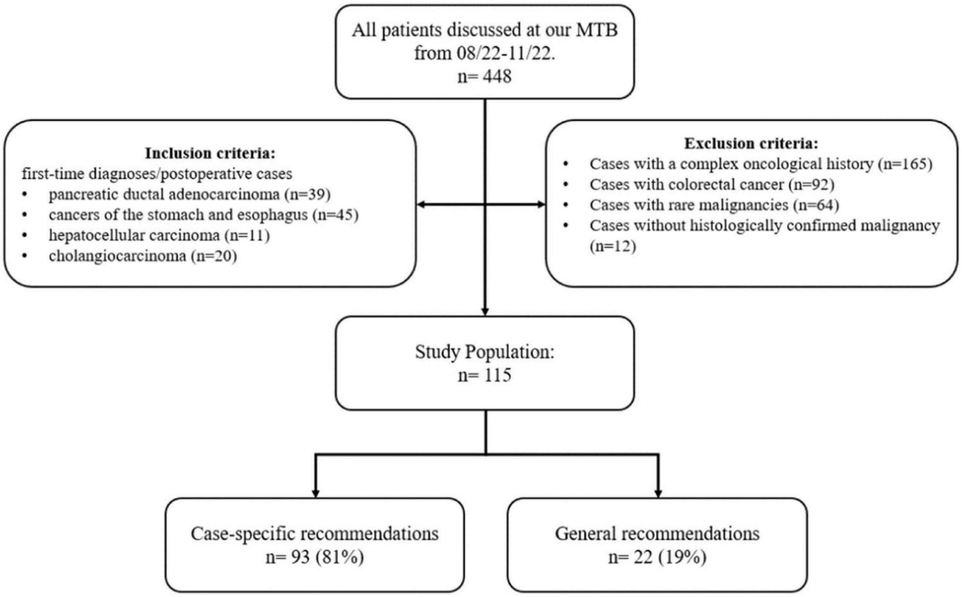
All patients discussed at our MTB
from 08/22-11/22.
n= 448
Inclusion criteria:
first-time diagnoses/postoperative cases
• pancreatic ductal adenocarcinoma (n=39)
• cancers of the stomach and esophagus (n=45)
• hepatocellular carcinoma (n=11)
• cholangiocarcinoma (n=20)
Exclusion criteria:
• Cases with a complex oncological history (n=165)
• Cases with colorectal cancer (n=92)
• Cases with rare malignancies (n=64)
• Cases without histologically confirmed malignancy (n=12)
Study Population:
n= 115
Case-specific recommendations
n= 93 (81%)
General recommendations
n= 22 (19%)

Fig. 1. Study flow.

2.3. Statistics

Categorical variables were compared using the chi-squared test and Fisher’s exact test. Graphical illustrations were conducted using R-Software.

3. Results

3.1. Patient characteristics

In our analysis of 115 cases of gastrointestinal malignancies, we found that ChatGPT provided general information in 22 patients (19.1%) (Table 1). Importantly, there were no significant differences in age, sex, and ECOG status between the group receiving personalized responses and the one provided with general advice. Moreover, the variety of diagnoses between these groups was similar. Remarkably, among cases with prior therapies, case-specific recommendations were provided in 93% of instances. Moreover, among the 22 cases receiving general advice, 19 (86%) had not been treated previously.

3.2. Treatment strategy concordance between ChatGPT and MTB stratified by type of malignancy

Excluding the 22 cases with "general recommendations," we analyzed a refined group of 93 cases for treatment strategy agreement following our predefined approach. This process highlighted an overall concordance rate of 83% (77 out of 93 cases), with the highest agreement observed in gastric cancer (GC), pancreatic ductal adenocarcinoma (PDAC), esophageal cancer (EC), and cholangiocarcinoma (CCC) at 87.5%, 84%, 84%, and 82%, respectively. Hepatocellular carcinoma (HCC) cases showed a lower concordance rate at 70%. The distribution of concordance across various cancer types is detailed in Fig. 2.

Table 1
Characteristics of cases stratified by case-specific response and general recommendations.

Variables	Case-specific response (n = 93)	General recommendations (n = 22)	p-value
Age			0.2
< 80	79 (83.2%)	16 (16.8%)	
> 80	14 (70.0%)	6 (30.0%)	
Sex			0.1
male	52 (81.5%)	12 (18.5%)	
female	41 (80.5%)	10 (19.5%)	
ECOG			0.4
0-1	87 (81.3%)	20 (18.7)	
> 1	6 (75)	2 (25)	
Previous therapies			**0.02**
yes	38 (92.7)	3 (7.3)	
no	55 (74.3)	19 (25.7)	
Diagnosis			0.8
PDAC	31 (79.5)	8 (20.5)	
GC	16 (76.2)	5 (23.8)	
EC	19 (79.2)	5 (20.8)	
HCC	10 (90.9)	1 (9.1)	
CCC	17 (85)	3 (15)	
Therapy Recommendation of multidisciplinary tumor board			0.5
Neoadjuvant	20 (71.4)	8 (28.6)	
Surgery	21 (80.8)	5 (19.2)	
Adjuvant	23 (88.5)	3 (11.5)	
System therapy 18 (78.3)	5 (21.7)		
Intervention	6 (85.7)	1 (14.3)	
Follow-up	0	5 (100)	

ECOG: Eastern Cooperative Oncology Group; PDAC: pancreatic ductal adenocarcinoma; GC: gastric cancer; EC: esophageal cancer; HCC: hepatocellular carcinoma; CCC: cholangiocarcinoma

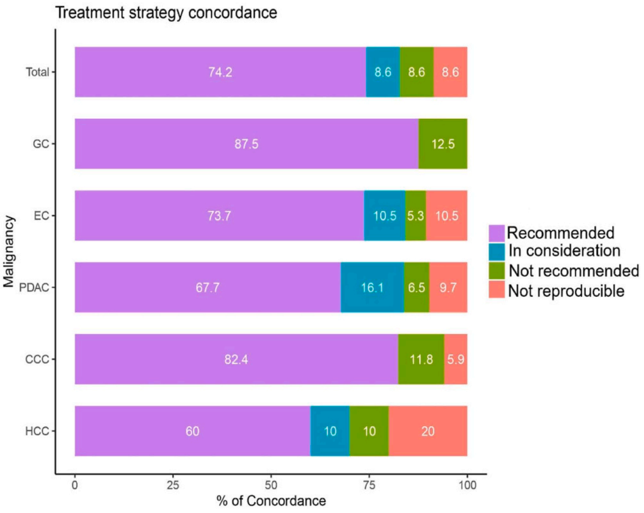

Fig. 2. Treatment Strategy Concordance between ChatGPT and MTB stratified by type of malignancy. PDAC: pancreatic ductal adenocarcinoma; GC: gastric cancer; EC: esophageal cancer; HCC: hepatocellular carcinoma; CCC: cholangiocarcinoma. Purple: recommended; blue: in consideration; green: not recommended; red: not reproducible.

3.3. Treatment strategy concordance between ChatGPT and MTB stratified by MTB recommendations

Further, we investigated the relationship between the treatment strategy agreement and the MTB’s recommendations. Fig. 3 shows that cases recommended for neoadjuvant therapy by the MTB had a 95% agreement rate (20 out of 21 cases), with one instance being non-reproducible. Following neoadjuvant therapy, both surgery and adjuvant therapy had concordance rates of 90% and 85%, respectively. In contrast, other therapeutic strategies like systemic therapy, locoregional interventions, and follow-up had lower conceptual concordance rates.

3.4. Exact treatment concordance between ChatGPT and MTB recommendation stratified by type of malignancy

After observing a high rate of overall conceptual agreement, we examined ChatGPT’s precision in offering exact treatment

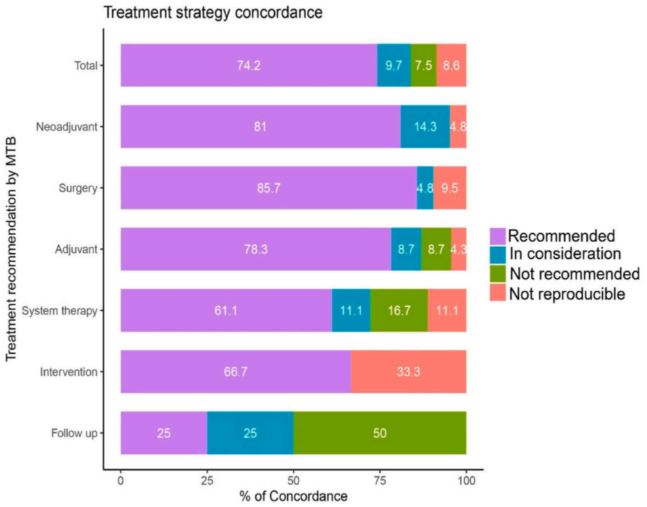

Fig. 3. Treatment Strategy Concordance between ChatGPT and MTB stratified by MTB recommendations. Purple: recommended; blue: in consideration; green: not recommended; red: not reproducible.

recommendations, categorizing cases by cancer type. Unlike the broad conceptual agreement, the overall exact treatment concordance was notably lower at 65% (61 out of 93 cases) (Fig. 4). The greatest precision was seen in GC cases at 75% (12 out of 16 cases), with PDAC and EC following closely. Like the conceptual agreement, HCC cases had the lowest precision rate at 60% (6 out of 10 cases).

3.5. Exact treatment concordance between ChatGPT and MTB stratified by MTB recommendations

To identify the impact of MTB treatment recommendations on the decreased precision of therapy concordance, we categorized cases by the specific treatment recommendations made by the MTB (Fig. 5). Interestingly, cases receiving MTB recommendations for surgery exhibited the highest concordance at 90% (19 out of 21 cases). Notably, two cases demonstrated unreproducible responses, paralleling observations in conceptual treatment concordance. Intriguingly, there was a substantial decline in exact treatment concordance rates among cases recommended for neoadjuvant, adjuvant chemotherapy, and systemic therapy, as illustrated in Fig. 5. In these instances, the specific type of chemotherapy either lacked recommendation or exhibited unreproducible responses in non-concordant cases. The lowest concordance rate was observed in the context of follow-up (25%).

4. Discussion

To the best of our knowledge, this study for the first time evaluates the concordance between recommendations made by ChatGPT and those from an MTB in various gastrointestinal cancers. This research stands out also as the first to differentiate between generic responses and tailored recommendations, thoroughly assessing their consistency and reproducibility.

During our analysis, we found that ChatGPT offered generic treatment information in about 19% of cases. A similar observation of generic advice aligning with MTB recommendations was made in the study by Lukac et al., although they did not differentiate between generic and personalized advice due to their study's limited size of ten patients [13]. Intriguingly, in our larger sample, ChatGPT issued tailored recommendations in 93% of cases that had undergone previous treatments like neoadjuvant therapy or surgery, a figure that fell to 73% among patients without previous treatments. Several factors could account for this variation. It's possible that ChatGPT provides more accurate responses when presented with more comprehensive data. Moreover, individuals who have received prior treatments may have a more defined treatment pathway for ChatGPT to enhance. On the other hand, patients needing a new treatment strategy might pose a greater challenge for ChatGPT to generate specific recommendations.

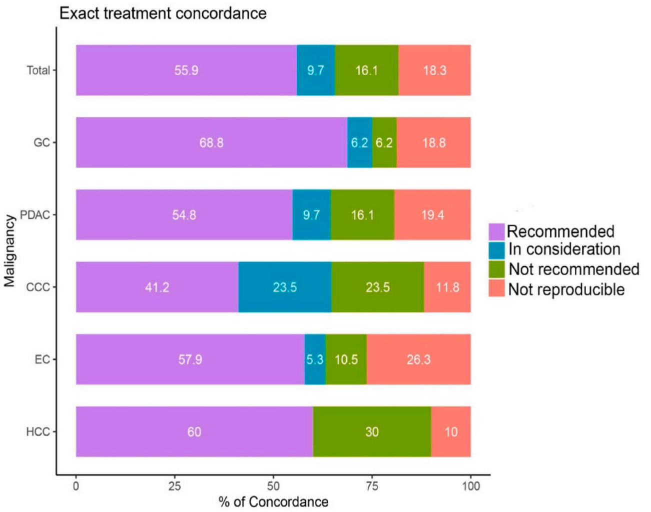

Fig. 4. Exact Treatment Concordance between ChatGPT and MTB stratified by type of malignancy. PDAC: pancreatic ductal adenocarcinoma; GC: gastric cancer; EC: esophageal cancer; HCC: hepatocellular carcinoma; CCC: cholangiocarcinoma. Purple: recommended; blue: in consideration; green: not recommended; red: not reproducible.

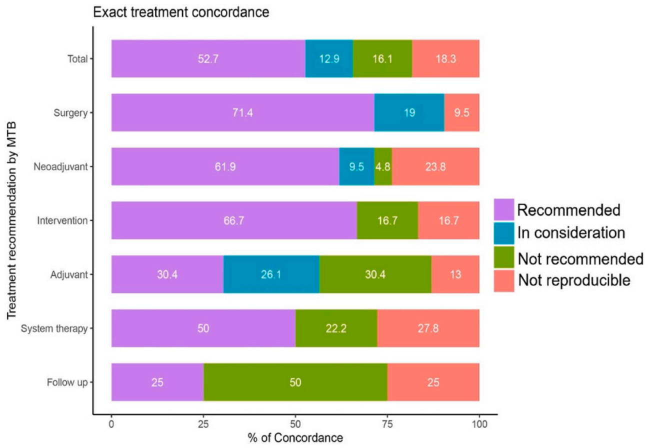

Fig. 5. Exact Treatment Concordance between ChatGPT and MTB stratified by MTB recommendations. Purple: recommended; blue: in consideration; green: not recommended; red: not reproducible.

After excluding responses containing general information, we identified an 83% overall treatment strategy concordance rate among ChatGPT's case-specific responses (77 out of 93 cases). Notably, the lowest concordance rate was found in HCC at 70%. This aligns with a recent study where two transplant hepatologists evaluated ChatGPT's responses to questions about HCC, demonstrating a 74% accuracy rate [14]. On the other hand, cases involving MTB recommendations for neoadjuvant therapy, surgery, and adjuvant therapy showed the highest levels of concordance. This notably high concordance rate is promising as it signifies a strong alignment between ChatGPT's suggestions and established clinical guidelines. However, there was a considerable drop in the concordance rate for aspects like follow-up, systemic therapy, or loco-regional interventions. This suggests specific areas that require further fine-tuning and validation of AI-generated recommendations.

When focusing specifically on the concordance rate for exact treatment recommendations, we observed a noticeable decrease to 65%. This finding aligns with earlier studies on breast cancer, where ChatGPT achieved concordance rates between 60% and 70% [15,16]. Particularly, cases receiving surgical recommendations from the MTB showed the highest level of agreement and reproducibility in ChatGPT's answers. However, when chemotherapy was recommended by the MTB, ChatGPT struggled, often correctly identifying chemotherapy as the appropriate strategy but failing to recommend the specific regimen needed. Notably, in more than 20% of instances, ChatGPT was unable to accurately replicate the advised chemotherapy regimen, highlighting its current limitations in offering detailed guidance on chemotherapy treatments.

In conclusion, while ChatGPT shows potential in formulating conceptual treatment strategies that align with MTB recommendations, its inability to accurately reproduce specific treatment plans, particularly regarding chemotherapy protocols and follow-up procedures, even in cases with straightforward medical histories, makes it unsuitable for direct application in clinical decision-making within tumor boards. Ethical concerns regarding patient safety and the quality of care necessitate further development and caution before integrating AI tools like ChatGPT into these sensitive healthcare contexts.

This study represents a pioneering exploration of the concordance between treatment strategies and exact treatment recommendations made by ChatGPT compared to those of MTBs. However, it is not without limitations. The research focused on a cohort of 115 cases that lacked complex medical histories, failing to capture the full diversity of gastrointestinal cancers. A broader sample encompassing more varied and complex cases, including different interventions or chemotherapy histories, might offer a deeper insight into ChatGPT's capabilities across a wider array of situations, potentially leading to more conclusive findings. Moreover, the approach of measuring reproducibility—relying solely on the "regenerate" function—may not fully capture the nuances of reproducibility. Investigating reproducibility through different accounts, over various times, or across multiple locations could shed additional light on that matter. Future studies that incorporate these broader considerations could enrich our understanding of the reliability and utility of AI-generated recommendations like those from ChatGPT 3.5.

Funding/Support

The authors declare that no financial support was received for this study.

CRediT authorship contribution statement

Dorian Andrade: Data curation, Formal analysis, Investigation, Methodology. **Maximilian Weniger:** Conceptualization, Data curation, Formal analysis, Investigation. **Martin K Angele:** Formal analysis, Supervision, Validation. **C. Benedikt Westphalen:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – review & editing. **Ughur Aghamaliyev:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. **Jens Werner:** Conceptualization, Supervision, Writing – review & editing. **Javad Karimbayli:** Formal analysis, Methodology, Visualization, Writing – original draft. **Clemens Giessen-Jung:** Conceptualization, Investigation, Methodology, Supervision. **Matthias Ilmer:** Supervision, Writing – review & editing, Conceptualization, Data curation, Investigation. **Kristian Unger:** Conceptualization, Writing – review & editing, Formal analysis, Methodology, Supervision. **Felix O. Hofmann:** Data curation, Investigation. **Bernhard W Renz:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, UA utilized ChatGPT 3.5 to refine language usage, correct grammar and punctuation errors to improve overall writing quality. After using ChatGPT3.5, the authors (UA, BR) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? Nature 2022.

[2] Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613(7944)):423.

[3] Graham F. Daily briefing: Will ChatGPT kill the essay assignment? Nature 2022.

[4] Zhavoronkov A, ChatGPT Generative Pre-trained Transformer. Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. Oncoscience 2022;9:82–4.

[5] Lee JY. Can an artificial intelligence chatbot be the author of a scholarly article? J Educ Eval Health Prof 2023;20:6.

[6] Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023;613(7945):620–1.

[7] Gilson A, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023;9:e45312.

[8] Allahqoli L, et al. Diagnostic and Management Performance of ChatGPT in Obstetrics and Gynecology. Gynecol Obstet Invest 2023;88(5):310–3.

[9] Zhao X, et al. Concordance between treatment recommendations provided by IBM Watson for Oncology and a multidisciplinary tumor board for breast cancer in China. Jpn J Clin Oncol 2020;50(8):852–8.

[10] Grunebaum A, et al. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 2023;228(6):696–705.

[11] Rajjoub R, et al. ChatGPT and its Role in the Decision-Making for the Diagnosis and Treatment of Lumbar Spinal Stenosis: A Comparative Analysis and Narrative Review. Glob Spine J 2023:21925682231195783.

[12] Vela Ulloa J, et al. Artificial intelligence-based decision-making: can ChatGPT replace a multidisciplinary tumour board? Br J Surg 2023;110(11):1543–4.

[13] Lukac S, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. Arch Gynecol Obstet 2023;308(6):1831–44.

[14] Yeo YH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29(3):721–32.

[15] Griewing S, et al. Challenging ChatGPT 3.5 in Senology-An Assessment of Concordance with Breast Cancer Tumor Board Decision Making. J Pers Med 2023;13(10).

[16] Sorin V, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023;9(1):44.