Original Paper

Exploring the Role of AI in Managing Treatment Recommendations for Lymphedema: International, Multidisciplinary, Multiprofessional Survey Study of Trust, Reliability, and Impact on Decision-Making

Adriano Fabi[1,2*], MD; Caroline E Egli[1,2*], MD; Séverin R Wendelspiess[1,2], MD; Sebastian Griewing[3,4], MD, PhD; Yvonne Haas[1,2], MD; Laura De Pellegrin[5], MD; Dirk J Schaefer[1,2], MD; Shan S Qiu[6], MD; Yves Harder[7,8], MD; Elisabeth A Kappos[1,2,9], MD

[1]Department of Plastic, Reconstructive, Aesthetic and Hand Surgery, University Hospital of Basel, Basel, Switzerland
[2]Faculty of Medicine, University of Basel, Basel, Switzerland
[3]Institute for Digital Medicine, Philipps University of Marburg, Marburg, Germany
[4]Department of Gynecology and Obstetrics, Philipps University of Marburg, Marburg, Germany
[5]Department of Plastic and Hand Surgery, University Hospital of Bern, Bern, Switzerland
[6]Department of Plastic and Reconstructive Surgery, Maastricht University Medical Centre, Maastricht, The Netherlands
[7]Department of Plastic, Reconstructive and Aesthetic Surgery and Hand Surgery, University Hospital of Lausanne (CHUV), Lausanne, Switzerland
[8]Faculty of Biology and Medicine, University of Lausanne (UNIL), Lausanne, Switzerland
[9]Breast Center, University Hospital of Basel, Basel, Switzerland
*these authors contributed equally

Corresponding Author:

Elisabeth A Kappos, MD
Department of Plastic, Reconstructive, Aesthetic and Hand Surgery
University Hospital of Basel
Spitalstrasse 21
Basel 4031
Switzerland
Phone: 41 613286254
Email: Elisabeth.Kappos@usb.ch

Abstract

Background: Upper and lower extremity lymphedema is a chronic, progressive condition that significantly impairs the quality of life of affected patients. Despite the recently established effectiveness of physical therapy and supermicrosurgical interventions, current guidelines frequently lag behind emerging evidence and commonly do not offer stage-specific treatment algorithms. This gap in evidence-based guidance may prompt clinicians with limited experience to seek support from large language models such as ChatGPT.

Objective: Given the potential of artificial intelligence to rapidly integrate emerging research, this study evaluated how clinicians from different professional backgrounds rate the quality and reliability of personalized lymphedema management recommendations generated by ChatGPT.

Methods: In this exploratory cross-sectional study, ChatGPT generated treatment recommendations for 6 standardized lymphedema case scenarios. An international panel of 67 participants (resident doctors, board-certified specialists, physiotherapists, and advanced practice nurses) from 34 institutions across 11 countries assessed the recommendations using a modified DISCERN questionnaire with a 9-point agreement scale ranging from 1 (completely disagree) to 9 (completely agree). Ratings were summarized as pooled means with variability measures and compared across clinician groups (residents vs board-certified physicians vs physiotherapists or advanced practice nurses) using group comparison testing.

Results: ChatGPT was rated most favorably for diagnostic accuracy and treatment relevance, with higher ratings among residents than board-certified physicians. Residents assigned significantly lower scores for source indication, source currency, and communication of uncertainty. Between-group differences were observed across multiple DISCERN items, consistent with

systematically more critical appraisal by experienced specialists. Participants reported moderate to high trust and willingness to consider ChatGPT as a supplementary resource, with more favorable perceptions among younger respondents.

Conclusions: Clinicians perceived ChatGPT as potentially useful for preliminary orientation and educational support in lymphedema management, especially for less experienced users. Despite not being blinded, lower ratings in evidence transparency and uncertainty communication, particularly among experienced specialists, suggest that current artificial intelligence outputs should not be used as stand-alone guidance. Future work should test clinically integrated, citation-grounded workflows in prospective settings and evaluate whether they improve decision quality and efficiency.

JMIR Med Inform 2026;14:e80553; doi: 10.2196/80553

Keywords: artificial intelligence; AI; ChatGPT; decision-making; digital health; large language models; lymphedema; personalized medicine

Introduction

Lymphedema is a chronic condition characterized by the accumulation of interstitial fluid, typically resulting from damage or disruption of lymphatic pathways [1]. In the Western world, lymphedema is commonly caused by oncologic surgery or radiotherapy and significantly reduces quality of life due to extremity swelling, limited mobility, and psychosocial distress [1-5].

Lymphedema treatment is highly individualized, influenced by age, preexisting comorbidities, and prior interventions [1]. Conservative treatment includes complex decongestive therapy; however, recent literature has demonstrated promising long-term outcomes after supermicrosurgical interventions, such as lymphovenous anastomosis (LVA) and vascularized lymph node transfer (VLNT) [6-9]. Due to the personalized nature of lymphedema treatment, following the current guidelines of the AWMF (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, S2k) or the International Society of Lymphology (ISL) may result in difficulties providing universally applicable, stage-specific recommendations [10,11]. Furthermore, the guidelines may include outdated information, primarily due to the time-intensive nature of expert consensus and validation processes [10,11]. Given the rapid advancements in lymphatic surgery, the incorporation of novel treatment approaches into official guidelines may take years, as this process requires a predefined and structured consensus. Consequently, there is a possibility that patients may receive care that is no longer up to date in a clinical context.

Large language models (LLMs) such as ChatGPT (OpenAI) represent a new approach to the processing of knowledge, which also applies to medical information [12,13]. These technologies have the potential to reshape medical decision-making by accelerating the integration of novel research into clinical routines [14-17]. Thus, unlike traditional guidelines, artificial intelligence (AI) is hypothesized to accelerate the transition of new findings into personalized treatment recommendations. However, whether this potential is truly realized in practice strongly depends on data implementation and prompting [18].

Therefore, the goal of this study was to assess ChatGPT's ability to define and generate treatment recommendations for lymphedema by evaluating the quality, accuracy, and perceived utility of its recommendations across an interdisciplinary panel of relevant medical specialties. Additionally, the differences in acceptance were compared between resident doctors in training and board-certified plastic surgeons with experience in either lymphatic surgery or lymphedema management.

Methods

Study Design

The study was designed as a multidisciplinary assessment of ChatGPT-generated treatment recommendations of 6 cases of chronic lymphedema, ranging from ISL stage I to ISL stage 3 of the upper and lower limb (Multimedia Appendix 1). To ensure consistency, all cases were submitted to GPT-4o (GPT-4 Omni) on December 18, 2024, at noon Central European Time by the first authors (AF and CEE). Each case was entered into a distinct, separate chat session to limit session history bias. Only the initial response was considered, and the option to regenerate the response was not used. The responses obtained from GPT-4o were compiled into a PDF and were subsequently evaluated by a panel of resident doctors (considered inexperienced), board-certified doctors (considered experienced), and physiotherapists or advanced practice nurses (APNs) using a 9-point agreement scale [19].

Clinical Vignettes and Treatment Recommendations

The clinical vignettes were designed with the intention of incorporating the underlying etiology of lymphedema through past medical history, current patient symptoms, and findings from clinical examination. Cases were prepared to be as realistic as possible and to reflect real clinical scenarios. Patients were clearly categorized into the 3 ISL stages, with corresponding indocyanine green studies illustrating the lymphatic patterns consistent with the stage described in each clinical vignette [10]. Furthermore, the cases were illustrated with images adapted from Principles and Practice of Lymphedema Surgery by Cheng et al [20]. A standardized prompt was used for each case and is provided in Multimedia Appendix 1.

Evaluation

The evaluators had access to the original prompts, all 6 clinical cases, and the corresponding treatment recommendations generated by ChatGPT (Multimedia Appendix 1). The web-based questionnaire for subsequent evaluation was developed and distributed using the Research Electronic Data Capture (REDCap; Vanderbilt University) platform [21]. The DISCERN tool, a validated questionnaire for assessing the quality of written treatment information, was adapted to a 9-point agreement scale and then used for the quantitative evaluation of each generated response [19].

In accordance with the instructions provided in the DISCERN handbook [19], questions 1 and 2 ("Are the aims clear?" and "Does it achieve its aims?") were excluded from the evaluation as they were found to be irrelevant to the defined research question of this study. Two additional questions were included in the evaluation of each case: one concerning correct staging (question 1) and the other regarding overall final agreement with the treatment option recommended by ChatGPT (question 16). In summary, each response was evaluated using 16 questions rated on a 9-point agreement scale, where 1 represents complete disagreement and 9 represents complete agreement. Finally, 3 concluding questions were added to evaluate the overall agreement regarding the treatment recommendation by ChatGPT and the attitude regarding AI use as a clinical decision tool (Multimedia Appendix 2). In addition, each participant was given the option to offer feedback. These open-text comments were reviewed descriptively by the first authors to provide contextual insight alongside the quantitative results. No formal qualitative analysis or coding approach was applied. All available comments were read in full, and relevant observations are summarized in Multimedia Appendix 3.

Survey Distribution

The survey was conducted from January 7 to March 18, 2025. Resident doctors without prior experience in lymphedema management were recruited internationally through the same broader professional networks as the expert panel, primarily via participating centers in the LYMPH trial, as well as international resident and student networks. Experienced, board-certified doctors specializing in lymphedema treatment were identified and contacted, most of whom had previously participated in the international lymphedema consortium through existing collaborations [22]. A total of 64 board-certified doctors were invited via email to take part in the study, receiving a survey link along with a comprehensive explanation of its objectives and requirements (response rate 24/64, 37.5%). Additionally, 30 physiotherapists and lymphedema APNs specializing in lymphedema treatment were invited (response rate 10/30, 33.3%) through the same consortium and a contact list provided by the Lymphödem Vereinigung Schweiz (a Swiss patient organization) [22,23]. As resident recruitment was conducted via mass email and a preexisting student group, the response rate of invited resident doctors could not be calculated.

Statistical Analysis

Statistical analysis was conducted using R software (version 4.4.3; R Foundation for Statistical Computing) [24]. The primary end point was the scoring of agreement and reliability of ChatGPT-generated treatment recommendations stratified by experience groups. Consequently, participants were categorized into 3 groups: resident doctors, board-certified doctors, and physiotherapists or APNs. For each question, group-wise means and SDs were calculated. The Levene test was used to confirm homogeneity of variances prior to conducting 1-way ANOVA. Depending on the result of the Levene test, either standard ANOVA (for equal variances) or Welch ANOVA (for unequal variances) was used to test for differences between groups. Where applicable, an "overall" group comprising all participants was included for descriptive purposes only and was excluded from inferential statistical testing. A 2-sided P value of ≤.05 was considered statistically significant.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. On the basis of the fictional nature of the cases and the absence of human subject data, the local ethics committee (Ethikkommission Nordwest- und Zentralschweiz) deemed that no formal ethics approval was required. Participation in the web-based survey was voluntary. All participants were informed about the purpose of the study, and informed consent was implied through completion of the survey. Participants had the option to discontinue the survey at any time without any consequences. All data were collected anonymously. Data protection and confidentiality were ensured in accordance with applicable regulations. No compensation was provided. No identifiable individuals are shown in any figures or supplementary materials.

Results

A total of 67 participants from 34 different institutions across 11 countries completed the survey, including 33 (49.3%) resident doctors, 24 (35.8%) board-certified doctors, and 10 (14.9%) lymphedema physiotherapists or APNs. The mean age was 35.6 (SD 14.1) years, with resident doctors averaging 28.0 (SD 6.0) years, board-certified doctors 45.0 (SD 14.0) years, and physiotherapists or APNs 42.0 (SD 19.9) years.

Diagnostic Accuracy and Relevance (Q1 and Q2)

Staging and diagnostic accuracy (question 1; Q1) were generally rated high (mean 7.3, SD 1.5), with board-certified doctors assigning significantly lower scores than resident doctors (P<.001; Table 1). Similarly, the relevance of the recommendations (Q2) was rated highly (mean 7.5, SD 1.4), with board-certified doctors rating significantly lower compared to resident doctors or physiotherapists or APNs (P<.001; Figure 1; Table 1). A detailed overview of scoring per case is provided in Multimedia Appendix 4.

Figure 1. Heat map of pooled mean item ratings across 6 case scenarios by clinician group. Tiles are color-coded by the pooled mean score for each question (Q1-Q16) within each group; numeric values within tiles indicate the corresponding mean. Higher scores indicate more favorable ratings on a 9-point agreement scale (1=completely disagree and 9=completely agree); variability measures (SD or CI) are not shown in this figure. BC: board-certified; R: resident; P/A: physiotherapist/advanced practice nurse.

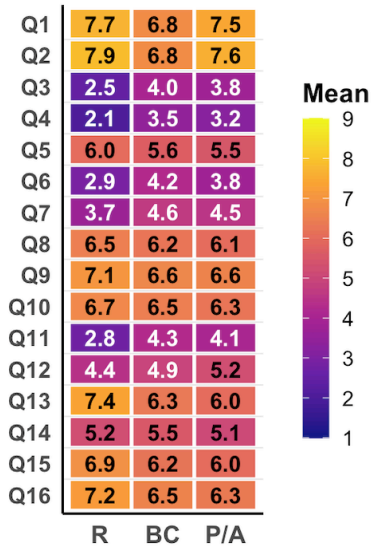

Table 1. Mean scores across all 6 scenarios[a].

Question[b]	Overall (n=67), mean (SD)	Resident doctors (n=33), mean (SD)	Board-certified physicians (n=24), mean (SD)	Physiotherapist or APN[c] (n=10), mean (SD)	*P* value
Q1—diagnostic accuracy	7.3 (1.5)	7.7 (1.2)	6.8 (1.6)	7.5 (1.9)	*<.001*[d]
Q2—relevance	7.5 (1.4)	7.9 (1.0)	6.8 (1.6)	7.6 (1.1)	*<.001*
Q3—indication of sources used	3.2 (2.5)	2.5 (2.1)	4.0 (2.4)	3.8 (2.9)	*<.001*
Q4—currency of sources used	2.7 (2.1)	2.1 (1.6)	3.5 (2.1)	3.2 (2.9)	*<.001*
Q5—balanced and unbiased information	5.8 (1.8)	6.0 (1.7)	5.6 (1.6)	5.5 (2.2)	.07
Q6—additional supporting information	3.5 (2.3)	2.9 (2.1)	4.2 (2.0)	3.8 (2.7)	*<.001*
Q7—communication of areas of uncertainty	4.1 (2.1)	3.7 (2.0)	4.6 (1.8)	4.5 (2.5)	*<.001*
Q8—explanation of how the treatment works	6.3 (1.8)	6.5 (2.0)	6.2 (1.5)	6.1 (2.0)	.23
Q9—benefits of treatment	6.8 (1.6)	7.1 (1.8)	6.6 (1.3)	6.6 (1.7)	*.04*
Q10—risks of treatment	6.6 (1.9)	6.7 (2.0)	6.5 (1.5)	6.3 (2.0)	.22
Q11—consequences of treatment refusal	3.5 (2.2)	2.8 (2.0)	4.3 (1.9)	4.1 (2.7)	*<.001*
Q12—impact on quality of life	4.7 (2.4)	4.4 (2.3)	4.9 (2.1)	5.2 (2.8)	*.03*
Q13—availability of multiple treatment options	6.8 (1.9)	7.4 (1.7)	6.3 (1.8)	6.0 (2.2)	*<.001*
Q14—support for shared decision-making	5.3 (2.3)	5.2 (2.4)	5.5 (2.1)	5.1 (2.3)	.37
Q15—overall quality of the recommendation	6.5 (1.6)	6.9 (1.4)	6.2 (1.8)	6.0 (1.8)	*<.001*
Q16—overall agreement	6.8 (1.8)	7.2 (1.6)	6.5 (1.9)	6.3 (2.3)	*<.001*

[a]Values represent the pooled mean (SD) across the 6 standardized lymphedema case scenarios from an exploratory cross-sectional survey of 67 clinicians (resident doctors, board-certified physicians, and physiotherapists or advanced practice nurses) from 34 institutions in 11 countries (January-March 2025).
[b]Q: question.
[c]APN: advanced practice nurse.
[d]Values in italic demonstrate statistical significance ($P<.05$).

Source Transparency and Objectivity (Q3-Q5)

The clarity of source citation (Q3) and the dating of sources (Q4) were rated very low across all groups, with mean scores of 3.2 (SD 2.5) and 2.7 (SD 2.1), respectively (Figure 2). Interestingly, resident doctors gave both questions significantly lower ratings than the other groups (P<.001). Interestingly, Q5 (objectivity or neutrality of ChatGPT) received an average score of 5.8 (SD 1.8), with no significant differences between groups.

Figure 2. Mean scores by professional group for all cases. Overview of mean scores across 6 lymphedema case scenarios by professional group. Bars represent mean scores. APN: advanced practice nurse; Q: question. Exact values can be found in Multimedia Appendix 4.

Supportive Information and Areas of Uncertainty (Q6 and Q7)

The quality of additional resources (Q6) was rated badly, with a mean score of 3.5 (SD 2.3) across all cohorts. Similarly, the communication of uncertainty (Q7) received moderate scores of 4.1 (SD 2.1), with residents giving the lowest scores across both questions (P<.001).

Treatment Explanation and Options, Risk Communication, and Quality of Life Considerations (Q8-Q13)

ChatGPT's performance in communicating treatment options and possible risks was generally rated positively, with Q8 (how the treatment would work), Q9 (benefits of treatment), and Q10 (potential risks) all receiving favorable scores (mean 6.3, SD 1.8; mean 6.8, SD 1.6; and mean 6.6, SD 1.9, respectively). In contrast, Q11 (consequence of no treatment) consistently scored low (mean 3.5, SD 2.2). Q12 assessed whether the provided information addressed the impact of the treatment on patients' quality of life. Ratings were moderate (mean 4.7, SD 2.4), with residents rating the responses significantly lower than other groups. Q13 (ChatGPT mentioning that there may be more than one treatment choice) was generally rated positively (mean 6.8, SD 1.9).

Shared Decision-Making and Overall Quality (Q14-Q16)

Support for shared decision-making (Q14) was rated moderately (mean 5.3, SD 2.3). The overall impression of the quality of responses (Q15) and overall agreement with treatment recommendations (Q16) both received high ratings (mean 6.5, SD 1.6 and mean 6.8, SD 1.8, respectively), with physiotherapists or APNs rating the answers the lowest (P<.001).

Concluding Questions—Trust and Future Use

Concluding question (CQ) 1 assessed trust in ChatGPT's responses, CQ2 assessed willingness to use ChatGPT in future clinical work, and CQ3 assessed whether participants would recommend it for clinical decision-making. Resident doctors rated all 3 concluding questions (CQ1-CQ3) significantly higher than board-certified doctors and physiotherapists or APNs. CQ2 showed a similar pattern,

with resident doctors at a mean of 6.5 (SD 1.4), board-certified doctors at 5.3 (SD 2.2), and physiotherapists or APNs at 5.3 (SD 2.3). Finally, for CQ3, which assessed whether participants would recommend AI as a clinical decision-making tool, residents again scored highest, followed by board-certified doctors and physiotherapists or APNs (Figure 3; Table 2). All survey comments are listed in Multimedia Appendix 3.

Figure 3. Concluding questions (CQs) by professional groups. Boxplot overview of concluding questions (trust, willingness to use, and recommendation of ChatGPT) by professional group. APN: advanced practice nurse.

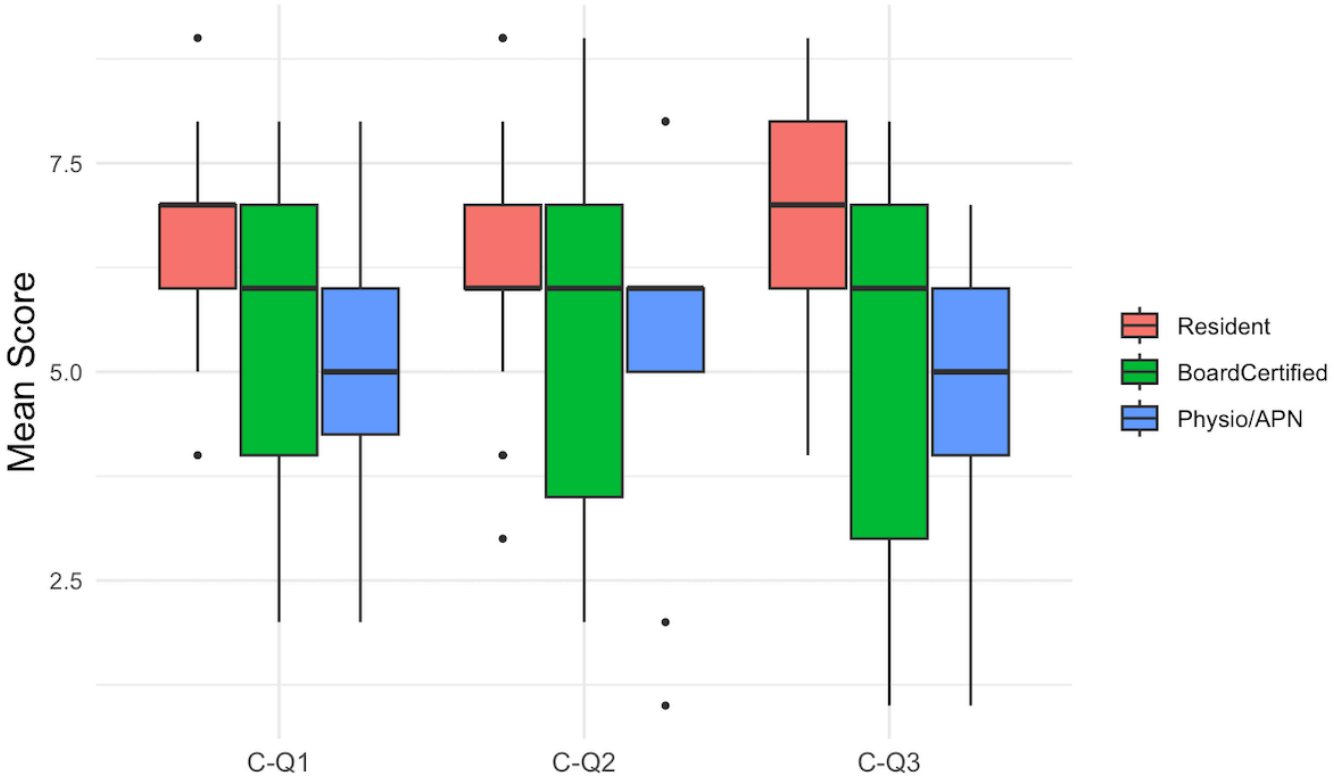

Table 2. Scores by professional groups across all concluding questions (CQs)[a].

Question	Overall (n=67), mean (SD)	Resident doctors (n=33), mean (SD)	Board-certified physicians (n=24), mean (SD)	Physiotherapist or APN[b] (n=10), mean (SD)	*F* test (*df*)	Levene test for equality of variables	ANOVA *P* value
CQ1—trust in AI[c]	5.9 (1.5)	6.5 (1.0)	5.4 (1.9)	5.1 (1.8)	5.21 (2, 20.58)	0.027	*.007*[d]
CQ2—willingness to use AI	5.9 (1.9)	6.5 (1.4)	5.3 (2.2)	5.3 (2.3)	3.20 (2, 63.00)	0.119	*.04*
CQ3—recommendation of AI	6.1 (2.1)	7.0 (1.4)	5.4 (2.4)	4.7 (2.0)	8.13 (2, 21.55)	0.015	*<.001*

[a]Mean scores for the CQS (trust, willingness to use, and recommendation of ChatGPT), stratified by professional group.
[b]APN: advanced practice nurse.
[c]AI: artificial intelligence.
[d]Values in italics were statistically significant ($P<.05$).

Discussion

In this exploratory, cross-sectional study, a total of 67 participants, comprising 33 resident doctors still in training, 24 board-certified doctors, and 10 lymphedema physiotherapists or APNs, from 34 different institutions across 11 countries evaluated ChatGPT-generated treatment recommendations for 6 fictional lymphedema cases. This is the first study to evaluate ChatGPT-generated treatment recommendations for lymphedema.

Notably Higher Acceptance of ChatGPT Among Younger Doctors

The overall high level of agreement with the ChatGPT-generated treatment recommendations (Q16) indicates broad acceptance across all cases and groups (Table 1). Resident doctors expressed a more positive opinion toward ChatGPT, a tendency that is presumably attributable to their lesser clinical experience and greater openness to digital technologies in medical practice, indicating that the evaluation of ChatGPT may also be influenced by generational factors (Figure 3) [25,26]. Importantly, however, this favorable perception was not uncritical. Residents assigned significantly lower scores than board-certified physicians for source transparency (Q3-Q4; Figure 1). The lower ratings may reflect stricter expectations regarding explicit source citation, differences in grading thresholds, or a reduced ability to recognize implicit sources that experienced clinicians may infer based on prior knowledge [27]. As the study design does not allow discrimination between these explanations, no conclusions can be drawn regarding the underlying cause of this difference. Nevertheless, several resident doctors reported that ChatGPT already serves them as a useful orientation tool, particularly in situations with limited prior familiarity with a condition or treatment options. This underscores the tool's potential as a rapid decision support or educational aid, as it only uses a few seconds to generate an extensive response [18]. It also highlights the need to integrate the critical use of AI tools into medical education. As residents showed greater willingness to use ChatGPT, targeted training could help prevent misconceptions and ensure these tools complement, rather than replace, clinical reasoning.

However, board-certified doctors demonstrated a lower level of agreement with treatment recommendations provided by ChatGPT, possibly due to their greater clinical experience and higher expectations regarding therapeutic nuances. This may, in part, stem from the perception that AI systems are not yet capable of providing fully personalized treatment algorithms. This viewpoint was supported in the open comments section (Multimedia Appendix 3), where experienced clinicians repeatedly pointed out that well-established surgical interventions, such as LVA and VLNT, were insufficiently addressed. In this study, LVA and VLNT were only briefly mentioned in cases 2, 3, 4, and 6 and only as potential options if compression therapy and liposuction were deemed insufficient. Reconstructive procedures such as LVA and VLNT are mentioned in both the German AWMF (S2k) guidelines and the ISL consensus document [10,11]. While the AWMF (S2k) guidelines suggest considering surgical options only after 6 months of conservative treatment without any lasting beneficial effect, the ISL guidelines refer to microsurgical interventions without specifying any time frame regarding previous complex decongestive therapy [10,11]. However, recent literature suggests that these procedures may also be used earlier or even as a preventative measure, thus challenging the traditional time threshold [28,29]. Given that key papers and guideline discussions on LVA and VLNT were already available well before October 2023, their limited presence in the responses is unlikely to be explained by the model's knowledge cutoff alone [30]. A more plausible explanation is that ChatGPT defaults to what it has seen most often: conservative management pathways that are widely taught and described, and a generally "play it safe" tendency when asked broad clinical questions [31]. In other words, the model may not be unaware of surgery; rather, it may just not surface it unless the prompt explicitly asks for surgical options or provides more detail about operative candidacy [12]. Nevertheless, newer ideas such as prophylactic LVA may still be underrepresented due to the recency of the evidence. Future work should test whether prompting specifically for surgical options changes the balance and completeness of the recommendations. These findings also emphasize the necessity of incorporating temporal limitations when using AI tools in clinical decision-making processes [7,8,20,28-30,32,33]. Furthermore, the strong emphasis on lymphedema management using conservative treatment options may result from the broader and more established evidence base in this area, which has developed over decades and is therefore more prominently represented in ChatGPT's training data.

Physiotherapists or APNs expressed greater skepticism regarding ChatGPT's treatment recommendations, as highlighted in the open comments section. The recommendations were regarded as being too generalized and lacking personalization. It was emphasized that effective lymphedema care necessitates ongoing, individualized assessment, a component that the AI recommendations did not adequately address.

Prompt Quality and Lack of Sources Limit ChatGPT's Clinical Reliability

One of the key characteristics of ChatGPT is its reliance on user input—responses are generated solely based on the prompt it receives [12]. As a result, the specificity and clarity of the input significantly affect the quality, depth, and relevance of the output [12]. Shorter prompts tend to produce shorter, less detailed responses. Citations were completely absent in ChatGPT's responses and consistently rated poorly by participants (Q3; Table 1), demonstrating that the absence of sources significantly limited the reliability of AI-generated treatment recommendation and, consequently, did not earn the physicians' trust. Furthermore, the absence of information regarding additional sources (Q6), the lack of mention of areas of uncertainty (Q7), and the omission of potential consequences of forgoing treatment (Q11) were also rated negatively (Table 1).

As an LLM, ChatGPT does not inherently verify the information it provides, but instead generates content based on linguistic patterns, which may influence the accuracy of treatment recommendations [12]. Notably, previous studies emphasized the need to explicitly ask for citations, as the model may otherwise omit references or produce so-called "hallucinations," meaning fabricated or incorrect references [34,35]. As our prompt did not explicitly request sources, thus reflecting the expected real-world use of ChatGPT, the absence of citations was expected. However, this highlights 2 key limitations. First, when users lack the

background knowledge to formulate precise prompts, the model's effectiveness is significantly reduced. Second, users currently have no possibilities to verify the origin or quality of the information provided. Previous studies have shown that such uncritical reliance on unverifiable information has the potential to result in inefficient or inappropriate treatment decisions and, consequently, potentially harm the patient [36]. It is also important to emphasize that ChatGPT does not *collect* information but rather generates responses based on statistical patterns in its training data, which further limit its interpretive reliability.

ChatGPT may also pose security risks [12]. Of particular concern is the use of false or deliberately misleading prompts, the aim of which is to manipulate the model into generating inaccurate responses [12]. To ensure patient safety, any clinical implementation would require strict regulatory oversight. A more realistic, and potentially safer, alternative could involve embedding ChatGPT within a clinical decision support system using prevalidated clinical data [37]. In this scenario, ChatGPT would function as a predictive model, assisting in decision-making within a strictly regulated and supervised data environment.

Study Strengths and Limitations

This is the first exploratory, cross-sectional study evaluating the agreement, reliability, and decision-making value of AI using LLMs in generating treatment recommendations for lymphedema management in a global, multidisciplinary, and interprofessional panel. Strengths of this study include the sizeable expert panel of 67 doctors and lymphedema therapists from 34 different institutions across 11 countries. A potential limitation of this study is that evaluators were aware that responses were generated by ChatGPT, which may have introduced either positive bias (novelty or expectation effect) or negative bias (preexisting skepticism toward AI). The results should therefore be interpreted as clinicians' perceptions of AI-generated recommendations rather than a purely objective measure of model performance. The lack of human-generated reference standard represents an important limitation, as it prevents the interpretation of absolute DISCERN scores; these should instead be understood in a relative context across professional groups. Although the 6 vignettes were deliberately designed to reflect common clinical presentations across the ISL stage spectrum across both extremities, they cannot capture the full heterogeneity of lymphedema. Consequently, generalizability to all clinical contexts is limited, and future studies should evaluate a larger and more diverse set of cases. Another limitation of this study is the use of an adapted assessment tool based on the DISCERN criteria. Although the underlying framework is validated, modifications to the item set and scoring scale may have altered the original psychometric properties of the instrument. As resident recruitment was conducted via large email and international student group, the response rate of invited resident doctors could not be calculated. Moreover, all analyses were based on a single, time-stamped response per case and were not controlled for random seed values or multirun variability, as this would have substantially increased survey burden. Furthermore, because the study was conducted using the standard ChatGPT web interface rather than the application programming interface, we were unable to access or control key inference parameters such as temperature, top-p, or system prompts. As a result, these factors were not reported. Finally, the study did not assess how variations in prompt design might affect the consistency or accuracy of the generated recommendations. Additionally, participants did not generate the prompts themselves, which may limit user satisfaction to the generated responses.

Conclusions

ChatGPT has the capacity to provide relevant basic knowledge for lymphedema management and may be particularly useful for less experienced clinicians. However, it has not yet succeeded in replacing expert consultation and offers only limited personalization. Additionally, its constraints regarding source transparency and prompt dependency highlight the need for cautious and context-aware integration into clinical practice.

Acknowledgments

During the revision of this manuscript, the authors used GPT-5.2 to improve readability. The authors subsequently reviewed and edited the content as needed and take full responsibility for the final version of the publication.

Funding

This study was funded by the Freiwillige Akademische Gesellschaft Basel, the Swiss Cancer Research Foundation, Rising Tide Foundation, the Swiss National Science Foundation, and the Theodor und Ida Herzog-Egli Stiftung.

Data Availability

The datasets generated and analyzed during this study are stored on the secure Research Electronic Data Capture platform hosted by the Department of Clinical Research at the University of Basel and can be provided by the corresponding author upon reasonable request.

Authors' Contributions

All authors contributed to the study conception and design. Clinical vignettes were prepared by AF, CEE, and EAK. Data collection was conducted by AF, CEE, and EAK. Statistical analysis and figure preparation were performed by AF and SRW. Data interpretation and drafting of the manuscript were carried out by AF and CEE. The study was supervised by EAK. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt, cases, and ChatGPT treatment recommendations.
[DOCX File (Microsoft Word File), 36 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Research Electronic Data Capture questionnaire with the validated DISCERN tool.
[DOCX File (Microsoft Word File), 18 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Survey comments per subgroup (anonymous).
[DOCX File (Microsoft Word File), 26 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Group comparisons for questionnaire responses divided by case.
[DOCX File (Microsoft Word File), 64 KB-Multimedia Appendix 4]

References

1. Lin WC, Safa B, Buntic RF. Approach to lymphedema management. Semin Plast Surg. Nov 16, 2022;36(4):260-273. [doi: 10.1055/s-0042-1758691] [Medline: 36561430]
2. Grada AA, Phillips TJ. Lymphedema: pathophysiology and clinical manifestations. J Am Acad Dermatol. Dec 2017;77(6):1009-1020. [doi: 10.1016/j.jaad.2017.03.022] [Medline: 29132848]
3. Ridner SH, Bonner CM, Deng J, Sinclair VG. Voices from the shadows: living with lymphedema. Cancer Nurs. 2012;35(1):E18-E26. [doi: 10.1097/NCC.0b013e31821404c0] [Medline: 21558848]
4. Vassard D, Olsen MH, Zinckernagel L, Vibe-Petersen J, Dalton SO, Johansen C. Psychological consequences of lymphoedema associated with breast cancer: a prospective cohort study. Eur J Cancer. Dec 2010;46(18):3211-3218. [doi: 10.1016/j.ejca.2010.07.041] [Medline: 20797846]
5. Bowman C, Piedalue KA, Baydoun M, Carlson LE. The quality of life and psychosocial implications of cancer-related lower-extremity lymphedema: a systematic review of the literature. J Clin Med. Oct 2, 2020;9(10):3200. [doi: 10.3390/jcm9103200] [Medline: 33023211]
6. Chang DW, Dayan J, Greene AK, et al. Surgical treatment of lymphedema: a systematic review and meta-analysis of controlled trials. Results of a consensus conference. Plast Reconstr Surg. Apr 1, 2021;147(4):975-993. [doi: 10.1097/PRS.0000000000007783] [Medline: 33761519]
7. Seidenstuecker K, Fertsch S, Ghazaleh AA, et al. Improving quality of life after breast cancer: a comparison of two microsurgical treatment options for breast cancer-related lymphedema (BCRL). Clin Exp Med. Apr 23, 2024;24(1):82. [doi: 10.1007/s10238-024-01344-w] [Medline: 38653874]
8. Kappos EA, Fabi A, Halbeisen FS, et al. Vascularized lymph node transfer (VLNT) versus lymphaticovenous anastomosis (LVA) for chronic breast cancer-related lymphedema (BCRL): a retrospective cohort study of effectiveness over time. Breast Cancer Res Treat. Apr 2025;210(2):319-327. [doi: 10.1007/s10549-024-07567-5] [Medline: 39653884]
9. Yang JCS, Wang YM, Wu SC, et al. Lymphaticovenous anastomosis for treating secondary lower limb lymphedema in older patients-a retrospective cohort study. J Clin Med. May 30, 2022;11(11):3089. [doi: 10.3390/jcm11113089] [Medline: 35683479]
10. Executive Committee of the International Society of Lymphology. The diagnosis and treatment of peripheral lymphedema: 2020 Consensus Document of the International Society of Lymphology. Lymphology. 2020;53(1):3-19. [Medline: 32521126]
11. Wilting J, Bartkowski R, Baumeister R, Földi E, Stöhr S, Strubel G, et al. S2k Leitlinie: diagnostik und therapie der lymphödeme. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften; 2017. URL: http://register.awmf.org/assets/guidelines/058-001l_S2k_Diagnostik_und_Therapie_der_Lymphoedeme_2019-07-abgelaufen.pdf [Accessed 2026-03-19]
12. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst. 2023;2(2):81-83. [doi: 10.54097/fcis.v2i2.4465]
13. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High Confid Comput. Jun 2024;4(2):100211. [doi: 10.1016/j.hcc.2024.100211]

14. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. Nov 2023;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
15. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in senology-an assessment of concordance with breast cancer tumor board decision making. J Pers Med. Oct 16, 2023;13(10):1502. [doi: 10.3390/jpm13101502] [Medline: 37888113]
16. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. Jun 28, 2023;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
17. Nietsch KS, Shrestha N, Mazudie Ndjonko LC, et al. Can large language models (LLMs) predict the appropriate treatment of acute hip fractures in older adults? Comparing appropriate use criteria with recommendations from ChatGPT. J Am Acad Orthop Surg Glob Res Rev. Aug 9, 2024;8(8):e24.00206. [doi: 10.5435/JAAOSGlobal-D-24-00206] [Medline: 39137403]
18. Trinkley KE, An R, Maw AM, Glasgow RE, Brownson RC. Leveraging artificial intelligence to advance implementation science: potential opportunities and cautions. Implement Sci. Feb 21, 2024;19(1):17. [doi: 10.1186/s13012-024-01346-y] [Medline: 38383393]
19. Charnock D. The DISCERN Handbook: Quality Criteria for Consumer Health Information on Treatment Choices. Radcliffe Medical Press; 1998. ISBN: 1857753100
20. Cheng MH, Chang DW, Patel KM. Principles and Practice of Lymphedema Surgery. 2nd ed. Elsevier Health Sciences; 2021. ISBN: 9780323694186
21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. Apr 2009;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
22. Kappos EA, Haas Y, Schulz A, et al. The LYMPH trial: comparing microsurgical with conservative treatment for chronic breast cancer-associated lymphoedema - study protocol of a pragmatic randomised international multicentre superiority trial. BMJ Open. Feb 17, 2025;15(2):e090662. [doi: 10.1136/bmjopen-2024-090662] [Medline: 39961719]
23. Lymphödem Vereinigung Schweiz. 2025. URL: https://www.lv-schweiz.ch [Accessed 2026-03-19]
24. RStudio: integrated development environment for R. Posit. URL: https://posit.co [Accessed 2026-03-20]
25. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ. Apr 10, 2024;24(1):401. [doi: 10.1186/s12909-024-05400-7] [Medline: 38600457]
26. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? Digit Health. Jan 26, 2023;9:20552076231152167. [doi: 10.1177/20552076231152167] [Medline: 36762024]
27. Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med. Nov 2024;30(11):3098-3100. [doi: 10.1038/s41591-024-03180-7] [Medline: 39054373]
28. Sun JM, Yamamoto T. Primary surgical prevention of lymphedema. J Chin Med Assoc. Jun 1, 2024;87(6):567-571. [doi: 10.1097/JCMA.0000000000001101] [Medline: 38666773]
29. Wan R, Hussain A, Kuruoglu D, Houdek MT, Moran SL. Prophylactic lymphaticovenous anastomosis (LVA) for preventing lymphedema after sarcoma resection in the lower limb: a report of three cases and literature review. Microsurgery. Mar 2023;43(3):273-280. [doi: 10.1002/micr.30975] [Medline: 36226524]
30. OpenAI. GPT-4o system card. arXiv. Preprint posted online on Oct 25, 2024. [doi: 10.48550/arXiv.2410.21276]
31. Saban M, Dubovi I. A comparative vignette study: evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. J Adv Nurs. Nov 2025;81(11):7489-7499. [doi: 10.1111/jan.16101] [Medline: 38366690]
32. Cheng MH, Loh CYY, Lin CY. Outcomes of vascularized lymph node transfer and lymphovenous anastomosis for treatment of primary lymphedema. Plast Reconstr Surg Glob Open. Dec 20, 2018;6(12):e2056. [doi: 10.1097/GOX.0000000000002056] [Medline: 30656125]
33. Scaglioni MF, Fontein DB, Arvanitakis M, Giovanoli P. Systematic review of lymphovenous anastomosis (LVA) for the treatment of lymphedema. Microsurgery. Nov 2017;37(8):947-953. [doi: 10.1002/micr.30246] [Medline: 28972280]
34. Athaluri SA, Manthena SV, Kesapragada VS, Yarlagadda V, Dave T, Duddumpudi RT. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. Apr 11, 2023;15(4):e37432. [doi: 10.7759/cureus.37432] [Medline: 37182055]
35. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res. May 22, 2024;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]

36. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. arXiv. Preprint posted online on Aug 11, 2024. [doi: 10.1056/AIoa2300015]
37. Liao Z, Wang J, Shi Z, Lu L, Tabata H. Revolutionary potential of ChatGPT in constructing intelligent clinical decision support systems. Ann Biomed Eng. Feb 2024;52(2):125-129. [doi: 10.1007/s10439-023-03288-w] [Medline: 37332008]

Abbreviations

AI: artificial intelligence
APN: advanced practice nurse
AWMF: Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften
ISL: International Society of Lymphology
LLM: large language model
LVA: lymphovenous anastomosis
REDCap: Research Electronic Data Capture
VLNT: vascularized lymph node transfer

Edited by Andrew Coristine, Arriel Benis; peer-reviewed by Elizabth Dylke, Nicolas Bievre, Thomas Lefèvre; submitted 12.Jul.2025; final revised version received 02.Feb.2026; accepted 02.Feb.2026; published 08.Apr.2026

Please cite as:
Fabi A, Egli CE, Wendelspiess SR, Griewing S, Haas Y, De Pellegrin L, Schaefer DJ, Qiu SS, Harder Y, Kappos EA
Exploring the Role of AI in Managing Treatment Recommendations for Lymphedema: International, Multidisciplinary, Multiprofessional Survey Study of Trust, Reliability, and Impact on Decision-Making
JMIR Med Inform 2026;14:e80553
URL: https://medinform.jmir.org/2026/1/e80553
doi: 10.2196/80553

© Adriano Fabi, Caroline E Egli, Séverin R Wendelspiess, Sebastian Griewing, Yvonne Haas, Laura De Pellegrin, Dirk J Schaefer, Shan S Qiu, Yves Harder, Elisabeth A Kappos. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 08.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.