technologies

Article

Leveraging Artificial Intelligence for Personalized Rehabilitation Programs for Head and Neck Surgery Patients

Gianluca Marcaccini [1,2], Ishith Seth [1,3,4,*], Jennifer Novo [5], Vicki McClure [1], Brett Sacks [1], Kaiyang Lim [1], Sally Kiu-Huen Ng [4], Roberto Cuomo [2] and Warren M. Rozen [1,3]

[1] Department of Plastic and Reconstructive Surgery, Peninsula Health, Frankston, VIC 3199, Australia
[2] Department of Plastic and Reconstructive Surgery, University of Siena, 53100 Siena, Italy
[3] Faculty of Medicine and Surgery, Central Clinical School, Monash University, Clayton, VIC 3004, Australia
[4] Department of Plastic and Reconstructive Surgery, Austin Health, Heidelberg, VIC 3084, Australia
[5] Faculty of Medicine and Surgery, The University of Notre Dame, Sydney, NSW 2008, Australia
* Correspondence: iseth@phcn.vic.gov.au

Abstract: Background: Artificial intelligence (AI) and large language models (LLMs) are increasingly used in healthcare, with applications in clinical decision-making and workflow optimization. In head and neck surgery, postoperative rehabilitation is a complex, multidisciplinary process requiring personalized care. This study evaluates the feasibility of using LLMs to generate tailored rehabilitation programs for patients undergoing major head and neck surgical procedures. Methods: Ten hypothetical head and neck surgical clinical scenarios were developed, representing oncologic resections with complex reconstructions. Four LLMs, ChatGPT-4o, DeepSeek V3, Gemini 2, and Copilot, were prompted with identical queries to generate rehabilitation plans. Three senior clinicians independently assessed their quality, accuracy, and clinical relevance using a five-point Likert scale. Readability and quality metrics, including the DISCERN score, Flesch Reading Ease, Flesch–Kincaid Grade Level, and Coleman–Liau Index, were applied. Results: ChatGPT-4o achieved the highest clinical relevance (Likert mean of 4.90 $\pm$ 0.32), followed by DeepSeek V3 (4.00 $\pm$ 0.82) and Gemini 2 (3.90 $\pm$ 0.74), while Copilot underperformed (2.70 $\pm$ 0.82). Gemini 2 produced the most readable content. A statistical analysis confirmed significant differences across the models ($p < 0.001$). Conclusions: LLMs can generate rehabilitation programs with varying quality and readability. ChatGPT-4o produced the most clinically relevant plans, while Gemini 2 generated more readable content. AI-generated rehabilitation plans may complement existing protocols, but further clinical validation is necessary to assess their impact on patient outcomes.

Keywords: artificial intelligence; large language models; head and neck surgery; postoperative care; ChatGPT; DeepSeek

Academic Editors: Tamás Haidegger and Daniele Giansanti

Received: 17 February 2025
Revised: 24 March 2025
Accepted: 2 April 2025
Published: 4 April 2025

Citation: Marcaccini, G.; Seth, I.; Novo, J.; McClure, V.; Sacks, B.; Lim, K.; Ng, S.K.-H.; Cuomo, R.; Rozen, W.M. Leveraging Artificial Intelligence for Personalized Rehabilitation Programs for Head and Neck Surgery Patients. *Technologies* **2025**, *13*, 142. https://doi.org/10.3390/technologies13040142

1. Introduction

Artificial intelligence (AI) is increasingly transforming healthcare, with applications ranging from diagnostic assistance to perioperative management and rehabilitation planning. Recent studies have demonstrated the ability of AI-driven models to improve decision-making in surgical and postoperative care, particularly through large language models (LLMs) and natural language processing (NLP) techniques [1,2]. For instance, AI has been successfully applied in thoracic surgery to enhance diagnostic precision and predict postoperative complications, underscoring its expanding role in surgical specialties. Given the growing body of evidence supporting AI's integration into clinical workflows, this

study aims to assess the feasibility of using LLMs for generating personalized rehabilitation programs for patients undergoing major head and neck surgery [1–3]. Among AI-driven tools, large language models (LLMs), such as ChatGPT, DeepSeek, Gemini 2, and Copilot, have gained significant attention for their ability to generate medical information and answer clinical questions based on extensive datasets [4,5].

LLMs in medicine are trained on large repositories of medical literature, clinical guidelines, and anonymized patient case information, allowing them to synthesize complex information and generate sophisticated and structured responses that align with evidence-based practices [4,5]. While LLMs are already being used to streamline administrative tasks and automate clinical documentation, there is growing interest in their potential beyond these functions, particularly in developing tailored patient rehabilitation programs for patients recovering from complex surgeries [6].

Head and neck surgery (HNS) often involves oncological patients who require extensive surgical intervention and complex reconstructions, particularly in cases of squamous cell carcinoma (SCC), basal cell carcinoma (BCC), and other aggressive cutaneous or mucosal malignancies. These procedures range from simple local excisions to complex reconstructions involving microvascular free flaps, mandibulectomies, and radical neck dissections. While these surgeries are often lifesaving, they are also associated with significant morbidity, frequently resulting in functional impairments that affect speech, swallowing, mastication, airway patency, and facial nerve function [7]. The severity of these postoperative complications depends on patient factors, the extent of tissue resection, and the need for adjuvant radiotherapy or chemotherapy treatment, which can further impact function and prolong recovery [8]. Microvascular free flap reconstructions, while providing vascularized soft tissue and bony support, are not without risk, with higher complication rates in previously irradiated fields [8,9]. Additionally, surgical site infections, wound dehiscence, hematomas, and percutaneous fistulae are known postoperative challenges following extensive resections, often requiring prolonged wound management and staged interventions [7,10].

Given this, rehabilitation following HNS must be highly individualized and often requires a multidisciplinary approach with input from surgeons, speech therapists, dietitians, physiotherapists, and psychologists to optimize functional and aesthetic outcomes [11]. Current rehabilitation protocols for HNS patients are primarily guided by standard clinical pathways and expert consensus [9,10]. However, these approaches may not fully account for patient-specific variables, such as age, comorbidities, tumor location, and the extent of resection. This is an area where AI and LLMs could offer a novel solution by utilizing large sets of data to predict patient rehabilitation trajectories, recommend structured post-operative care instructions, and provide real-time clinical support to surgeons, ultimately enhancing patient outcomes.

This study aims to assess the feasibility of using LLMs to generate personalized rehabilitation programs for patients undergoing HNS by comparing the recommendations of multiple LLMs against expert-reviewed standards. By doing so, this study seeks to determine whether AI-generated rehabilitation programs can complement traditional approaches and improve postoperative outcomes. Additionally, this study will assess the readability and quality of LLM-generated responses using validated tools, such as the DISCERN score and Flesch Reading Ease Score, to provide insights into their applicability in patient education and clinical decision-making.

2. Materials and Methods

This experimental study was designed to assess the capacity of LLMs to generate personalized rehabilitation programs for patients undergoing head and neck surgery. Given

that the study utilized hypothetical clinical scenarios rather than real patient data, formal ethical approval was not required. However, all research activities adhered to the ethical principles of the Declaration of Helsinki.

Ten hypothetical clinical scenarios were developed to encompass a range of head and neck surgical procedures. These scenarios were designed by three senior clinicians with expertise in head and neck surgery, ensuring clinical relevance and representativeness. Each scenario was based on real-world cases commonly encountered in surgical practice, incorporating key variables such as patient demographics, comorbidities, tumor location, extent of resection, and reconstructive techniques. The scenarios were iteratively refined through expert review to ensure they reflected a broad spectrum of complexity and post-operative rehabilitation needs. To standardize AI model evaluation, a consistent prompt structure was used for all scenarios, focusing on the recommended rehabilitation program and necessary supportive care measures.

The following prompt was added post-hypothetical scenario: "What rehabilitation program and services are recommended for this patient to optimize their postoperative recovery, considering their specific clinical condition, surgical procedure, and potential risk factors?". These scenarios were formulated by senior clinicians with expertise in head and neck surgery, ensuring their relevance and clinical validity.

2.1. Generation of Rehabilitation Programs

Four state-of-the-art LLMs, ChatGPT-4o, DeepSeek V3, Gemini 2, and Copilot, were selected based on their relevance and accessibility for medical applications. The choice of these models aimed to ensure a balanced comparison between well-established AI tools and emerging alternatives, allowing for a comprehensive evaluation of their suitability in generating rehabilitation programs. ChatGPT-4o was included due to its demonstrated accuracy in medical contexts and its strong performance in previous studies assessing AI-generated clinical recommendations [1–5]. DeepSeek V3 was selected as an emerging model designed for complex reasoning, offering a distinct approach to AI-generated text with an emphasis on scientific and technical content. Gemini 2, developed by Google, was incorporated for its strong capabilities in language understanding and its optimization for producing highly readable outputs, which could be beneficial in patient-centered rehabilitation programs. Copilot, integrated within Microsoft's ecosystem, was chosen to evaluate how general-purpose AI models perform in specialized medical tasks. By including these four LLMs, this study aims to capture a broad spectrum of AI capabilities, from established models optimized for clinical reasoning to newer competitors designed for diverse applications. To ensure consistency in evaluation, all models were prompted simultaneously on 10 February 2025, using identical clinical scenarios.

2.2. Evaluation of LLM-Generated Rehabilitation Programs

Three senior clinicians (WMR, SN, and RC), with over 40 years of cumulative experience, independently assessed the rehabilitation programs suggested by the LLMs for accuracy, clinical relevance, and appropriateness. The clinicians were selected based on their expertise in head and neck surgery and postoperative rehabilitation, each with over ten years of experience in tertiary academic centers. Their selection aimed to ensure a high level of clinical judgment and familiarity with current rehabilitation protocols. To minimize bias, evaluations were conducted independently, and the inter-rater agreement was analyzed to assess consistency in their assessments. Similar studies evaluating AI-generated medical recommendations have also relied on small expert panels for initial validation [9]. Future research may benefit from expanding the number of reviewers to enhance generalizability and account for inter-rater variability. The assessments were

conducted using a five-point Likert scale (1 = Poor, 5 = Excellent) to measure the quality of the responses quantitatively.

To further assess the readability, clarity, and reliability of the generated rehabilitation plans, standardized text analysis metrics were applied, including the following:

- DISCERN Score—evaluating the reliability and quality of health information.
- Flesch Reading Ease Score—measuring the ease of comprehension.
- Flesch–Kincaid Grade Level—indicating the educational level required to understand the text.
- Coleman–Liau Index—assessing the complexity of the generated text.

2.3. Data Analysis

The Likert scale ratings assigned by the two senior clinicians were analyzed for inter-rater reliability using Cohen's kappa coefficient to assess consistency in evaluation. A descriptive statistical analysis was performed to compare the mean scores of each LLM, identifying trends in accuracy and clinical utility. A comparative study of the readability and quality assessment metrics was also conducted, examining model variations and their suitability for patient and clinician use.

3. Results

Quantitative Analysis

To evaluate the performance of four AI models (ChatGPT-4o, DeepSeek V3, Gemini 2, and Copilot) on surgical case responses, we conducted both descriptive and inferential analyses across several metrics. Descriptive statistics revealed that ChatGPT-4o achieved the highest performance in subjective evaluation, with a Likert scale mean of 4.90 ± 0.32, followed by DeepSeek V3 (4.00 ± 0.82) and Gemini 2 (3.90 ± 0.74), while Copilot's responses scored markedly lower, with a mean of 2.70 ± 0.82. In terms of the DISCERN score, which assesses the quality and reliability of the provided information, ChatGPT-4o, DeepSeek V3, and Gemini 2 produced comparable results (ranging from 46.90 ± 2.60 to 47.20 ± 2.53), whereas Copilot consistently scored 40.00.

The readability metrics further differentiated the models. The Flesch Reading Ease Score was notably higher for Gemini 2 (12.25 ± 7.22), suggesting that its texts are more straightforward to comprehend. At the same time, the Flesch–Kincaid Grade Level for Gemini 2 (16.60 ± 1.40) was lower than the other models, indicating reduced complexity. The Coleman–Liau Index, which also reflects text complexity, showed less pronounced differences among the models, with mean values ranging from 18.23 ± 0.79 for Copilot to 19.68 ± 1.32 for ChatGPT-4o.

A one-way variance analysis (ANOVA) was performed for each metric to determine whether these differences were statistically significant. The analysis revealed significant differences among the models for the Likert scale ($F(3, 36) = 16.50$, p value < 0.001), the DISCERN score ($F(3, 36) = 24.63$, p value < 0.001), the Flesch Reading Ease ($F(3, 36) = 4.35$, p value ≈ 0.01), and the Flesch–Kincaid Grade Level ($F(3, 36) = 11.82$, p value < 0.001). In contrast, the result for the Coleman–Liau Index was marginally non-significant ($F(3, 36) = 2.84$, p value ≈ 0.06). These findings indicate that the observed differences in performance and text quality across the AI models are unlikely due to chance, supporting the validity of the differences noted in the descriptive analysis and justifying further post hoc testing to precisely identify which pairs of models differ significantly in performance (Table 1).

Table 1. Mean and median values of AI output across various evaluation metrics.

AI Model	N	Likert Mean	Likert SD	DISCERN Mean	DISCERN SD	Flesch Reading Ease Mean	Flesch Reading Ease SD	Flesch–Kincaid Grade Level Mean	Flesch–Kincaid Grade Level SD	Coleman–Liau Index Mean	Coleman–Liau Index SD
ChatGPT-4o	10	4.90	0.32	46.90	2.60	5.02	8.09	18.44	1.42	19.68	1.32
DeepSeek V3	10	4.00	0.82	46.90	2.60	3.05	4.78	20.15	1.53	19.29	1.39
Gemini 2	10	3.90	0.74	47.20	2.53	12.25	7.22	16.60	1.40	18.50	1.46
Copilot	10	2.70	0.82	40.00	0.00	4.99	3.29	19.23	1.18	18.23	0.79

Statistical analyses were performed using Python (version 3.9) and its associated libraries, including SciPy and Statsmodels. We calculated descriptive statistics means and standard deviations for each metric and conducted a one-way analysis of variance to evaluate differences among the AI models. Prior to testing, we assessed assumptions of normality and homogeneity of variances, and all p values were computed with a significance level set at 0.05. To ensure the accuracy of our analysis, we performed an independent verification of the results using ChatGPT as a supplementary tool, confirming the validity of our findings (Table 2).

Table 2. One-way ANOVA analysis between four AI models across various evaluation metrics.

Metric	F-Value	Degrees of Freedom Between	Degrees of Freedom Within	p Value	Test
Likert Scale	16.50	3	36	<0.001	One-way ANOVA
DISCERN Score	24.63	3	36	<0.001	One-way ANOVA
Flesch Reading Ease	4.35	3	36	~0.01	One-way ANOVA
Flesch–Kincaid Grade Level	11.82	3	36	<0.001	One-way ANOVA
Coleman–Liau Index	2.84	3	36	~0.06	One-way ANOVA

Supplementary Table S1 shows responses from large language models to the following clinical scenario prompt: "What rehabilitation program and services are recommended for this patient to optimize their postoperative recovery, considering their specific clinical condition, surgical procedure, and potential risk factors?"

4. Discussion

This study explores the potential role of large language models (LLMs) in generating rehabilitation programs for patients undergoing head and neck surgery. While AI has been increasingly applied in clinical decision support and administrative tasks, its use in postoperative rehabilitation planning remains an emerging area. By comparing multiple LLMs (ChatGPT-4o, DeepSeek V3, Gemini 2, and Copilot) using standardized clinical scenarios and expert evaluation, this study provides insights into the strengths and limitations of AI-generated rehabilitation recommendations. The dual assessment of clinical relevance and readability offers a structured approach to understanding how these models perform in generating patient-centered rehabilitation plans. These findings contribute to the ongoing discussion on integrating AI into multidisciplinary care, highlighting potential applications and areas for further refinement. The integration of AI-generated rehabilitation plans into clinical practice could enhance multidisciplinary decision-making by providing structured, evidence-based recommendations tailored to individual patient needs. In particular, these tools could be valuable in settings with limited access to specialized rehabilitation providers, where AI models may help bridge the gap in expertise. Moreover, by standardizing rehabilitation protocols, AI has the potential to reduce variability in care while promoting

adherence to best practices. However, for these technologies to be effectively implemented, further research is needed to assess their adaptability to real-world clinical scenarios, their ability to account for patient-specific factors, and their integration into electronic health record (EHR) systems to streamline their use in daily practice.

The key findings reveal that ChatGPT-4o achieved the highest subjective performance, as evidenced by a Likert scale mean of 4.90 ± 0.32. At the same time, DeepSeek V3 and Gemini 2 recorded moderate scores (4.00 ± 0.82 and 3.90 ± 0.74, respectively), and Copilot performed significantly lower with a mean of 2.70 ± 0.82. In parallel, the DISCERN score, a measure of the quality and reliability of clinical information, was similar for ChatGPT-4o, DeepSeek V3, and Gemini 2 (ranging from 46.90 ± 2.60 to 47.20 ± 2.53), whereas Copilot consistently scored 40.0. Readability metrics further differentiated the models: Gemini 2 produced outputs with a notably higher Flesch Reading Ease (12.25 ± 7.22) and a lower Flesch–Kincaid Grade Level (16.60 ± 1.40), suggesting its texts are easier to understand. Meanwhile, the Coleman–Liau Index demonstrated smaller variations among the models, ranging from 18.23 ± 0.79 for Copilot to 19.68 ± 1.32 for ChatGPT-4o. A one-way analysis of variance confirmed that these differences were statistically significant for the Likert scale ($F(3, 36) = 16.50$, p value < 0.001), DISCERN score ($F(3, 36) = 24.63$, p value < 0.001), Flesch Reading Ease ($F(3, 36) = 4.35$, p value ≈ 0.01), and Flesch–Kincaid Grade Level ($F(3, 36) = 11.82$, p value < 0.001), while the Coleman–Liau Index result was marginally non-significant ($F(3, 36) = 2.84$, p value ≈ 0.06). Therefore, ChatGPT-4o's superior performance, coupled with the enhanced readability of Gemini 2, underscores the nuanced strengths of current AI technologies.

Regarding the subjective evaluation, the superior Likert score of ChatGPT-4o suggests that its outputs are perceived as highly acceptable and coherent by clinical experts. This observation is consistent with the emerging literature that emphasizes the growing capability of large language models to solve complex clinical problems intelligently and even outperform traditional clinical decision-making in certain scenarios. The high rating may reflect ChatGPT-4o's advanced natural language processing abilities, which enable it to generate nuanced and contextually appropriate recommendations [12–15]. While this study involved a limited number of evaluators, all were senior specialists in head and neck surgery, ensuring a high level of clinical expertise. Expanding the panel of reviewers in future studies could help refine the assessment by incorporating a broader range of clinical perspectives. However, given the structured evaluation criteria used, the key performance trends observed across different LLMs are likely to remain consistent.

In contrast, the moderate Likert scores for DeepSeek V3 and Gemini 2 indicate that while these models are competent, they might not capture the same level of subtlety or contextual integration as ChatGPT-4o. However, Copilot's significantly lower rating likely stems from its design as a general-purpose AI rather than a model optimized for medical applications. Unlike ChatGPT-4o and Gemini 2, which are trained using extensive medical literature, Copilot's responses were more generic and lacked clinical depth. Additionally, its outputs were often less structured and detailed, making them less useful for rehabilitation planning. These findings highlight the importance of using AI models specifically trained for medical contexts to ensure clinically relevant recommendations. Such discrepancies highlight the importance of selecting the appropriate AI model based on specific clinical applications and demonstrate that not all AI algorithms are equally effective, a point corroborated by previous research [16].

The analysis of the DISCERN scores further supports these interpretations. The comparable scores for ChatGPT-4o, DeepSeek V3, and Gemini 2 suggest that these models can produce reliable and evidence-based information essential for formulating safe rehabilitation protocols. In contrast, Copilot's consistently lower score may indicate deficiencies

in its content generation process, raising concerns about its reliability for clinical decision support. These findings align with the broader trend of exploring AI's potential in perioperative care while emphasizing the need for rigorous quality control.

Readability is a critical factor in the practical application of rehabilitation protocols. Gemini 2's higher Flesch Reading Ease Score and lower Flesch–Kincaid Grade Level imply that its outputs are more accessible, which could facilitate better understanding among clinicians and patients. However, while readability metrics such as DISCERN, Flesch Reading Ease, Flesch–Kincaid Grade Level, and Coleman–Liau Index provide objective measures of text complexity, they also have inherent limitations when applied to medical texts. These indices primarily evaluate structural aspects of language, such as sentence length and word difficulty, but do not assess whether the content is clinically accurate, contextually appropriate, or effectively communicates complex medical concepts. For example, a text with a high readability score may be oversimplified and omit essential clinical details, while a more complex text with a lower score may contain crucial information needed for precise rehabilitation planning. Furthermore, these metrics were originally developed for general educational materials rather than specialized medical content, making their direct applicability to AI-generated rehabilitation programs somewhat limited. Therefore, while readability scores provide useful insights into the accessibility of AI-generated content, they should be interpreted alongside qualitative expert assessments to ensure that rehabilitation plans are both comprehensible and clinically sound.

Using standardized readability measures in our study allows for a valid comparison with the existing literature and reinforces that clarity is paramount in clinical documentation. Although the Coleman–Liau Index did not reveal stark differences among the models, its relatively stable values suggest that while basic text complexity may be similar, other aspects of readability, such as sentence structure and vocabulary, play a more pivotal role in ensuring effective communication.

The statistical significance established through the one-way analysis of variance lends robust support to these findings. The highly significant p values (all below 0.01, except for the borderline result of the Coleman–Liau Index) confirm that the observed differences in performance, quality, and readability among the models are not due to random variation. This reinforces the validity of our conclusions and suggests that each model's distinct strengths and weaknesses are inherent to its design and training data.

It is vital for all who use LLMs to understand that their output is not static. The same prompt entered today will likely not return the same result if entered a year later. These results have broader implications within the context of AI integration in clinical practice. The ability of large language models to generate structured rehabilitation protocols aligns with previous research on AI-assisted clinical decision-making. While prior studies have explored the role of LLMs in generating medical guidelines and decision support tools [12,13], their application in postoperative rehabilitation planning remains less examined. This study builds upon the existing research by evaluating AI-generated rehabilitation recommendations specifically in head and neck surgery, an area where patient-specific variability often limits standardized protocols. However, as with other AI applications in healthcare, the integration of LLMs in rehabilitation requires further validation to ensure their clinical reliability and adaptability [16–18]. For instance, while AI algorithms have shown superiority in specific decision-making tasks, translating this potential into a reliable clinical tool requires adherence to principles such as TURBO (Testable, Usable, Reliable, Beneficial, Operable) to ensure consistent performance [15].

Furthermore, the dynamic nature of AI outputs, wherein the same input may yield different results over time due to continuous updates in training data, necessitates that these tools be used under strict clinical supervision [18]. This study, like others in the field,

acknowledges that while AI can provide valuable insights and augment clinical practice, its output must be validated through prospective studies, randomized controlled trials, and longitudinal research. Additionally, integrating AI with established technologies such as Virtual Surgical Planning, a cornerstone in head and neck surgery, could further refine patient-specific rehabilitation strategies and optimize surgical outcomes.

This study has several limitations that should be considered when interpreting the findings. Firstly, the rehabilitation programs generated by LLMs were evaluated using hypothetical clinical scenarios rather than real patient data, which may limit the direct applicability of the results to clinical practice. Secondly, the number of expert reviewers was limited to three senior clinicians, which, although providing a high level of expertise, may not fully capture the diversity of clinical perspectives. Furthermore, clinician assessments are inherently subjective and may be influenced by individual biases, prior experiences, or expectations, introducing potential variability in the evaluation of LLM outputs. Expanding the panel and incorporating blinded or standardized assessment protocols in future studies could help mitigate these biases and enhance reliability. Thirdly, while readability metrics were used to assess text complexity, they do not fully capture the nuances of patient comprehension or the effectiveness of AI-generated recommendations in real-world rehabilitation settings. Finally, as LLMs continue to evolve, their outputs are subject to ongoing updates, meaning that future versions of these models may yield different results. Prospective studies involving real patient cases and longitudinal follow-up are necessary to validate the clinical utility of AI-generated rehabilitation programs.

This comprehensive analysis demonstrates significant differences in the performance, quality, and readability of rehabilitation protocols generated by various AI models [17–19]. The superior performance of ChatGPT-4o, coupled with the enhanced readability of Gemini 2, underscores the nuanced strengths of current AI technologies while highlighting areas needing improvement [20]. These findings contribute to the evolving landscape of AI in perioperative care by providing critical insights that can inform future research and clinical practice [21]. Moving forward, multidisciplinary collaborations and rigorous clinical validations must guide the integration of AI into healthcare, ensuring that these advanced tools ultimately translate into improved patient outcomes and safer, more effective clinical interventions [22].

5. Conclusions

This study demonstrates that large language models can generate clinically relevant rehabilitation programs for patients undergoing head and neck surgery, with significant variation in quality and readability among models. ChatGPT-4o outperformed others in clinical relevance, as assessed by expert reviewers, while Gemini 2 produced more readable responses. Copilot consistently scored lower across all metrics, highlighting variability in AI-generated outputs. While LLMs show promise in supporting postoperative care, their outputs require rigorous validation before clinical implementation. The findings underscore the need for multidisciplinary oversight and ongoing refinement of AI tools to ensure their reliability in patient rehabilitation. Future research should focus on real-world clinical trials to assess the impact of AI-driven rehabilitation programs on patient outcomes. Integrating AI with existing multidisciplinary approaches could enhance personalized rehabilitation strategies, ultimately improving the functional recovery and quality of life for head and neck surgery patients.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/technologies13040142/s1, Supplementary Figure S1: Representative image. Supplementary Table S1: Responses from large language models to the clinical scenario prompt: "What rehabilitation program and services are recommended for this patient to optimize

their postoperative recovery, considering their specific clinical condition, surgical procedure, and potential risk factors?".

Author Contributions: Conceptualization, I.S., G.M., J.N., and V.M.; methodology, I.S., G.M., and J.N.; software, G.M.; validation, S.K.-H.N., R.C., and W.M.R.; formal analysis, I.S., G.M., J.N., V.M., and B.S.; investigation, I.S.; resources, G.M.; data curation, G.M.; writing—original draft preparation, I.S., G.M., J.N., K.L., V.M., and B.S.; writing—review and editing, all authors; visualization, B.S. and G.M.; supervision, R.C., S.K.-H.N., and W.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable, as this study did not involve human participants.

Data Availability Statement: The authors confirm that the data supporting this study's findings are available within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
HNS	Head and neck surgery
LLMs	Large language models
SCC	Squamous cell carcinoma
BCC	Basal cell carcinoma
ANOVA	One-way analysis of variance

References

1. DiDonna, N.; Shetty, P.N.; Khan, K.; Damitz, L. Unveiling the Potential of AI in Plastic Surgery Education: A Comparative Study of Leading AI Platforms' Performance on In-training Examinations. *Plast. Reconstr. Surg. Glob. Open* **2024**, *12*, e5929. [CrossRef]
2. Aleem, M.U.; Khan, J.A.; Younes, A.; Sabbah, B.N.; Saleh, W.; Migliore, M. Enhancing Thoracic Surgery with AI: A Review of Current Practices and Emerging Trends. *Curr. Oncol.* **2024**, *31*, 6232–6244. [CrossRef] [PubMed]
3. Bajwa, J.; Munir, U.; Nori, A.; Williams, B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc. J.* **2021**, *8*, e188–e194. [CrossRef] [PubMed]
4. Busch, F.; Hoffmann, L.; Rueger, C.; van Dijk, E.H.; Kader, R.; Ortiz-Prado, E.; Makowski, M.R.; Saba, L.; Hadamitzky, M.; Kather, J.N. Current applications and challenges in large language models for patient care: A systematic review. *Commun. Med.* **2025**, *5*, 26.
5. Clusmann, J.; Kolbinger, F.R.; Muti, H.S.; Carrero, Z.I.; Eckardt, J.-N.; Laleh, N.G.; Löffler, C.M.L.; Schwarzkopf, S.-C.; Unger, M.; Veldhuizen, G.P.; et al. The future landscape of large language models in medicine. *Commun. Med.* **2023**, *3*, 141. [CrossRef]
6. Sumner, J.; Lim, H.W.; Chong, L.S.; Bundele, A.; Mukhopadhyay, A.; Kayambu, G. Artificial intelligence in physical rehabilitation: A systematic review. *Artif. Intell. Med.* **2023**, *146*, 102693. [CrossRef]
7. Bhattacharyya, N.; Fried, M.P. Benchmarks for mortality, morbidity, and length of stay for head and neck surgical procedures. *Arch. Otolaryngol. Head Neck Surg.* **2001**, *127*, 127–132. [CrossRef] [PubMed]
8. Tan, B.K.; Por, Y.C.; Chen, H.C. Complications of head and neck reconstruction and their treatment. *Semin. Plast. Surg.* **2010**, *24*, 288–298. [CrossRef]
9. Walia, A.; Lee, J.J.; Jackson, R.S.; Hardi, A.C.; Bollig, C.A.; Graboyes, E.M.; Zenga, J.; Puram, S.V.; Pipkorn, P. Management of flap failure after head and neck reconstruction: A systematic review and meta-analysis. *Otolaryngol. Head Neck Surg.* **2022**, *167*, 224–235. [CrossRef]
10. Genden, E.M.; Rinaldo, A.; Suárez, C.; Wei, W.I.; Bradley, P.J.; Ferlito, A. Complications of free flap transfers for head and neck reconstruction following cancer resection. *Oral Oncol.* **2004**, *40*, 979–984. [CrossRef]
11. Seth, I.; Marcaccini, G.; Lim, K.; Castrechini, M.; Cuomo, R.; Ng, S.K.-H.; Ross, R.J.; Rozen, W.M. Management of Dupuytren's Disease: A Multi-Centric Comparative Analysis Between Experienced Hand Surgeons Versus Artificial Intelligence. *Diagnostics* **2025**, *15*, 587. [CrossRef] [PubMed]

12. List, M.A.; Knackstedt, M.; Liu, L.; Kasabali, A.; Mansour, J.; Pang, J.; Asarkar, A.A.; Nathan, C.A. Enhanced recovery after surgery, current, and future considerations in head and neck cancer. *Laryngoscope Investig. Otolaryngol.* **2023**, *8*, 1240–1256. [CrossRef]
13. Prasad, A.; Chorath, K.; Barrette, L.X.; Go, B.; Deng, J.; Moreira, A.; Rajasekaran, K. Implementation of an enhanced recovery after surgery protocol for head and neck cancer patients: Considerations and best practices. *World J. Otorhinolaryngol. Head Neck Surg.* **2022**, *8*, 91–95. [CrossRef] [PubMed]
14. Loftus, T.J.; Tighe, P.J.; Filiberto, A.C.; Efron, P.A.; Brakenridge, S.C.; Mohr, A.M.; Rashidi, P.; Upchurch, G.R.; Bihorac, A. Artificial intelligence and surgical decision-making. *JAMA Surg.* **2020**, *155*, 148–158. [CrossRef] [PubMed]
15. Palenzuela, D.L.; Mullen, J.T.; Phitayakorn, R. AI versus MD: Evaluating the surgical decision-making accuracy of ChatGPT-4. *Surgery* **2024**, *176*, 241–245. [CrossRef]
16. Hashimoto, D.A.; Rosman, G.; Rus, D.; Meireles, O.R. Artificial intelligence in surgery: Promises and perils. *Ann. Surg.* **2018**, *268*, 70–76. [CrossRef]
17. Seth, I.; Lim, B.; Phan, R.; Xie, Y.; Kenney, P.S.; Bukret, W.E.; Thomsen, J.B.; Cuomo, R.; Ross, R.J.; Ng, S.K.; et al. Perforator selection with computed tomography angiography for unilateral breast reconstruction: A clinical multicentre analysis. *Medicina* **2024**, *60*, 1500. [CrossRef]
18. Abi-Rafeh, J.; Xu, H.H.; Kazan, R.; Tevlin, R.; Furnas, H. Large language models and artificial intelligence: A primer for plastic surgeons on the demonstrated and potential applications, promises, and limitations of ChatGPT. *Aesthetic Surg. J.* **2024**, *44*, 329–343. [CrossRef]
19. Roberts, R.H.; Ali, S.R.; Dobbs, T.D.; Whitaker, I.S. Can large language models generate outpatient clinic letters at first consultation that incorporate complication profiles from UK and USA aesthetic plastic surgery associations? In *Aesthetic Surgery Journal Open Forum*; Oxford University Press: Oxford, UK, 2024; Volume 6, p. ojad109. [CrossRef]
20. Mohapatra, D.P.; Thiruvoth, F.M.; Tripathy, S.; Rajan, S.; Vathulya, M.; Lakshmi, P.; Singh, V.K.; Haq, A.U. Leveraging Large Language Models (LLM) for the plastic surgery resident training: Do they have a role? *Indian J. Plast. Surg.* **2023**, *56*, 413–420. [CrossRef]
21. Kwon, D.Y.; Wang, A.; Mejia, M.R.; Saturno, M.P.; Oleru, O.; Seyidova, N.; Taub, P.J. Adherence of a Large Language Model to Clinical Guidelines for Craniofacial Plastic and Reconstructive Surgeries. *Ann. Plast. Surg.* **2024**, *92*, 261–262. [CrossRef]
22. Seth, I.; Xie, Y.; Rodwell, A.; Gracias, D.; Bulloch, G.; Hunter-Smith, D.J.; Rozen, W.M. Exploring the role of a large language model on carpal tunnel syndrome management: An observation study of ChatGPT. *J. Hand Surg.* **2023**, *48*, 1025–1033. [CrossRef] [PubMed]