ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases

Guillermo Sanchez ROSENBERG [1], Martin MAGNÉLI [2], Niklas BARLE [2], Michael G KONTAKIS [3], Andreas Marc MÜLLER [1], Matthias WITTAUER [1], Max GORDON [2], and Cyrus BRODÉN [3]

[1] Department of Orthopedic and Trauma Surgery, University Hospital Basel, Switzerland; [2] Karolinska Institute, Department of Clinical Sciences at Danderyd Hospital, Stockholm; Sweden [3] Department of Surgical Sciences, Orthopedics, Uppsala University Hospital, Uppsala, Sweden
Correspondence: cyrus.broden@akademiska.se
Submitted 2023-11-01. Accepted 2024-01-28.

Background and purpose — Large language models like ChatGPT-4 have emerged. They hold the potential to reduce the administrative burden by generating everyday clinical documents, thus allowing the physician to spend more time with the patient. We aimed to assess both the quality and efficiency of discharge documents generated by ChatGPT-4 in comparison with those produced by physicians.

Patients and methods — To emulate real-world situations, the health records of 6 fictional orthopedic cases were created. Discharge documents for each case were generated by a junior attending orthopedic surgeon and an advanced orthopedic resident. ChatGPT-4 was then prompted to generate the discharge documents using the same health record information. The quality assessment was performed by an expert panel (n = 15) blinded to the source of the documents. As secondary outcome, the time required to generate the documents was compared, logging the duration of the creation of the discharge documents by the physician and by ChatGPT-4.

Results — Overall, both ChatGPT-4 and physician-generated notes were comparable in quality. Notably, ChatGPT-4 generated discharge documents 10 times faster than the traditional method. 4 events of hallucinations were found in the ChatGPT-4-generated content, compared with 6 events in the human/physician produced notes.

Conclusion — ChatGPT-4 creates orthopedic discharge notes faster than physicians, with comparable quality. This shows it has great potential for making these documents more efficient in orthopedic care. ChatGPT-4 has the potential to significantly reduce the administrative burden on healthcare professionals.

Administrative demands in healthcare significantly contribute to surging healthcare expenses, low job satisfaction, and ultimately physician burnout [1]. Evidence suggests that doctors spend an equivalent amount of time on administrative tasks, such as drafting progress notes and discharge documents, requesting labs and analyzing results, as they do in patient interactions [2]. Administrative costs in the United States account for roughly 31% of total healthcare expenditures, translating to $569 billion when adjusted for current medical inflation [3,4].

The recent introduction of large language models (LLMs) such as ChatGPT (OpenAI, 2022) have already had an impact on healthcare in medical education and clinical decision-making [5], medical licensing examinations [6-8], addressing medical inquiries from patients [9-11], and even crafting scientific articles [12,13].

These LLM models could potentially significantly reduce administrative tasks in healthcare by producing patient-specific texts tailored to the physician documentation requirements. The aim of our study is to explore the capabilities of LLM, specifically ChatGPT-4, by evaluating their quality and efficiency in generating discharge documents compared with physicians.

Methods

Study setting

6 fictional orthopedic and trauma cases were written by 2 orthopedic surgery residents from two university hospitals from different European countries (3 cases from Sweden and

DOI 10.2340/17453674.2024.40182

3 from Switzerland). All health information relevant to the case, including past medical history, radiology reports, laboratory values, progress notes, and interdisciplinary consultation notes were included (see Supplementary data). Although fictional, the cases were thoroughly devised and reviewed by senior orthopedic surgeons to accurately represent a real-life clinical scenario.

The study is reported according to the Consort-AI extension guidelines [14].

Human-generated discharge documents

A junior attending orthopedic surgeon and a senior orthopedic surgery resident independently created 2 discharge documents for 3 cases each: a discharge summary and a discharge letter, setting the gold standard. The discharge summary is aimed at medical professionals, generally family physicians or general practitioners, who are responsible for the post-discharge care. Using technical medical language, it provides a comprehensive account of the patient's hospitalization, diagnosis, history of the present illness, procedures undertaken, clinical summary, and the postoperative and discharge guidelines. In contrast, the discharge letter is aimed for the patient; it seeks to translate the information in the discharge summary into layman's language, so that it can be understood by the patient. In our institutions, clinicians use voice recording for generating discharge summaries. Medical secretaries transcribe these recordings. Upon transcription, clinicians receive the notes for review and approval. In contrast, the discharge letter is written directly by the clinician to ensure the patient obtains the letter without any delays. We diligently logged the duration in the creation of both the discharge summary and letter. Both discharge documents were generated following the standard format of each hospital.

ChatGPT-4 generated discharge documents

Subsequently, the junior attending orthopedic surgeon and the orthopedic surgery resident who initially generated the discharge documents used a specific prompt for ChatGPT-4 to generate the AI-generated discharge documents. Both physicians were unfamiliar with ChatGPT-4 at the time. This prompt was formulated by 2 authors who are proficient with ChatGPT-4. An additional fictional case, which was excluded from the analysis of the study, was created for this purpose. This prompt was applied uniformly across all cases, with minor modifications to accommodate the distinct formats of Swiss and Swedish discharge notes (see Supplementary data). The time required to generate both the discharge summary and the letter using ChatGPT-4 was documented. The version of ChatGPT-4 used in this study was last updated in September 2023.

Evaluation of discharge documents

Both human and GPT-4-generated documents were assessed by an expert panel of orthopedic residents and surgeons, blinded to the author of the document. The expert panel comprised 15 orthopedic surgeons (3 senior residents, 6 consultants, and 6 senior consultants). 6 orthopedic surgeons were from Switzerland (2 consultants and 4 senior consultants) and 9 from Sweden (3 residents, 4 consultants, 2 senior consultants). Swiss surgeons assessed cases from Switzerland, whereas Swedish surgeons evaluated cases from Sweden. The panelists independently evaluated the documents using specifically devised quality assessment criteria focusing on the accuracy of the medical information (diagnosis, history of present illness, hospital course, and discharge plan), the suitability of the language for the intended reader (medical professional or patient), conciseness of the text, the presence of factually incorrect, inconsistent, or nonsensical text (hallucinations), and the suitability for clinical use. Additionally, each panelist awarded a subjective quality score from 1–100 to each document (see Supplementary data). The Quality Assessment criteria were developed by our group by adapting an already established framework previously published by Singhal et al. for assessing medical LLM responses to open-ended questions [15]. 4 answer options were provided for each question. The responses were then translated into scores on an ordinal scale of 100. A higher score indicates better performance. The time required for human and GPT-4 document generation was then compared.

Statistics

Due to the small sample size, we employed descriptive statistics and 95% confidence intervals (CI) to evaluate and compare the quality assessment scores and the time required to generate the human-generated against the GPT-4-generated discharge documents.

Ethics, registration, data sharing plan, funding, and disclosures

Considering the study's exploratory nature and the use of fictitious data without real patient information, ethics committee approval was not sought. Additionally, this research project did not receive any financial support. The authors do not have any conflicts of interest to declare. Complete disclosure of interest forms according to ICMJE are available on the article page, doi: 10.2340/17453674.2024.40182

Results

Our study revealed that, overall, discharge documents generated by ChatGPT-4 are of similar quality to those written by physicians. For the Swedish discharge documents, ChatGPT-4 showed a minor edge in quality compared with the physician-generated versions. However, for the Swiss letters, the physician-crafted versions displayed a minor advantage in quality over those generated by ChatGPT-4 (Figure 1). The expert panel conducted a total of 126 evaluations on these discharge documents. 6 of the evaluations could not be used due

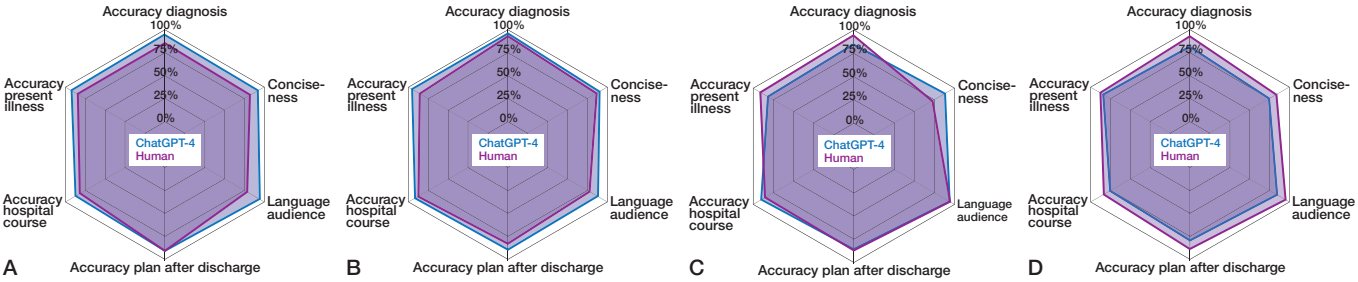

Figure 1. Quality assessment of discharge summaries and letters: A. Swedish summary; B. Swedish letter; C. Swiss summary; D. Swiss letter.

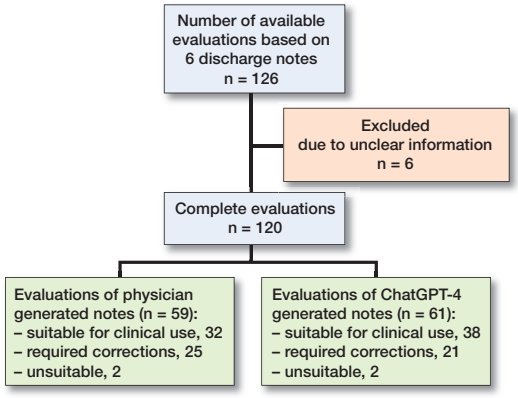

Figure 2. Flowchart of evaluations of discharge notes by the expert panel.

Time in minutes for physician and ChatGPT-4 to generate discharge notes

Case number	Physician-generated notes	ChatGPT-4-generated notes
Swedish		
Case 1	29.2	3.8
Case 2	33.4	2.9
Case 3	30.7	3.2
Swiss		
Case 1	27.5	3.0
Case 2	22.0	2.4
Case 3	24.0	2.1

to unclear information or incomplete information in the evaluation form (Figure 2). 59 evaluations were performed on the physician-generated notes and 61 evaluations on ChatGPT-4 generated notes. Of the documents produced by ChatGPT-4, most (38 assessments) were deemed suitable for clinical use. Some required corrections (21 assessments), while a small number (2 assessments) were deemed unsuitable for clinical use. In contrast, human-generated notes were evaluated 32 times as suitable for clinical use, received 25 indications for corrections, and were considered unsuitable for clinical use on 2 occasions.

Of the 120 evaluations, the panel identified 4 instances of hallucinations in ChatGPT-4-crafted content, in contrast to 6 in human-generated documents.

Regarding subjective quality scores, ChatGPT-4 generated discharge summaries were awarded an average score of 80.7 (CI 76.2–85.2) while those human-generated scored an average 74.3 (CI 68.0–80.7). For discharge letters, ChatGPT-4-generated documents averaged 77.8 (CI 71.1–84.6), while human-generated letters scored 77 (CI 69.3–84.8).

The mean differences in subjective scores between ChatGPT 4 and the physician-produced discharge notes was 6.5 (CI –8.9 to 21.9), and the difference for the discharge letters was 2.9 (CI–14.8 to 20.6)

Discharge documents produced by ChatGPT-4 were generated, on average, 10 times faster than those created by physicians (Table).

Discussion

This represents the first study to explore ChatGPT-4's capability in generating discharge documents in the field of orthopedic surgery. We showed that ChatGPT-4 creates orthopedic discharge notes faster than physicians, with comparable quality.

Our findings are in accordance with a recently published study that showed similar results, but focusing on the possibility to generate clinical letters to patients with a high overall correctness and humanness score with ChatGPT [16].

Our findings suggest that the discharge documents produced by ChatGPT-4 overall align closely with the quality of those authored by the orthopedic surgery physicians. A notable difference in quality was observed in Swedish case 2, where ChatGPT-4 appeared to outperform the physician in generating the discharge summary and letter. However, in Swiss case 3, the physician-generated discharge letters obtained higher scores than those produced by ChatGPT-4. Wimsett et al. delineated the fundamental components of effective medical discharge documents, encompassing diagnosis, administered treatments, test results, and discharge guidelines [17]. Nonetheless, several studies indicate that physician-authored documents frequently omit or misrepresent these essential elements [17-18]. The quality of discharge documents is crucial as it influences ongoing patient care and safety and even reimbursement schemes [19].

The subtle variation of human-generated discharge documents performance between the Swiss and Swedish cases might be due to differences in standards and expertise of the physicians creating the discharge notes, but also on average a slightly more senior expert panel evaluating the Swiss cases.

The elevated subjective satisfaction associated with ChatGPT-4-generated compared with the human-generated notes might be attributable not only to the quality of the content but also to its clear and structured linguistic presentation.

Of the documents produced by ChatGPT-4, the majority were deemed suitable for clinical use; some required corrections, while a small number were deemed unsuitable for clinical use. Notably, the ChatGPT-4-generated notes were subjected to evaluation without prior human review in this study. This emphasizes the need for review by a physician after the note is generated by Chat-GPT-4 before use in the medical context.

Within the clinical context of our study, we discerned 4 instances of hallucinations in the ChatGPT-4-generated content, in contrast to 6 such instances in the physician-generated notes. ChatGPT-4 has previously encountered scrutiny due to its potential for generating misleading or fabricated content, termed "hallucinations," particularly in the context of generating false references in academic articles [20]. Our findings suggest that ChatGPT-4 may, in fact, in the clinical context yield fewer inaccuracies than traditional methods.

ChatGPT-4 generated discharge documents roughly 10 times faster than human writers. This efficiency could free up clinicians to focus on other clinical tasks and may reduce physician burnout [2,21-23]. However, it is important to mention that the review and approval of ChatGPT-4-generated notes may be more time-consuming than assessing notes written by physicians. This can be attributed to clinicians' active engagement in composition during the note creation process using traditional methods. This dimension was not examined in our current study.

Creating the right prompts for ChatGPT-4 usually requires repeated adjustments. Working closely with researchers familiar with these LLMs can help improve the results. As this area of study grows, we are seeing new specialties emerge, such as "prompt engineering." Upcoming versions of ChatGPT-4 may come with ready-to-use prompts or models better adjusted to medical terms, which could speed up document creation. Also, if ChatGPT-4 is trained using specific medical data, it might reduce the need for custom prompts, making the process even more efficient [24]. Our study used cases from Sweden and Switzerland to determine whether ChatGPT-4 can adjust to different styles of discharge notes in various countries. However, as these notes were written in English, it remains unclear how well ChatGPT-4 would perform with healthcare documents in other languages.

To fully leverage the benefits of LLMs in clinical settings, we envision the use of a local LLM capable of running in the hospital's existing IT environment. This approach would possibly ensure that LLMs can effectively and safely interact with the hospital's electronic health records.

Limitations

First, we worked with a limited number of fictional orthopedic cases, as this was an initial exploration of ChatGPT-4 capabilities. Second, we deliberately excluded discharge medication from our study due to variations in prescription practices and formats across medical centers. We advocate for a detailed human review of discharge medication. Third, the gold standard was established by the discharge notes generated by 2 orthopedic physicians. Thus, the diverse writing styles, standards, and competences found within a broader group of clinicians were not fully captured. Additionally, it is worth noting that the discharge notes were generated in English, which is not the resident's native language. However, the evaluation of ChatGPT-4 went beyond the mere examination of written notes by physicians; it also involved a comparison with specific quality criteria. Most of the notes generated by ChatGPT-4 were found to be appropriate and of sufficient quality for use in clinical settings. Another limitation of our study was that evaluators were required to choose from a given Likert scale or answer "yes" or "no" to specific questions. While providing detailed explanations was optional, this approach resulted in a lack of further analysis of types of misinformation and hallucination instances. Nonetheless, the comments received did highlight issues with language use and some misleading phrases in the patient letters generated by GPT-4. Evaluators pointed out a few times that the language in the discharge plans could be confusing. This confusion encompassed areas like unclear directives for initiating anticoagulants, an erroneous distinction between intravenous and oral antibiotics, and 2 instances of vague instructions for referrals to other healthcare services. Consequently, a comprehensive review of AI-generated discharge notes is essential before considering their use in clinical settings. This evaluation could initially be conducted by a physician and, in the long term, potentially by another AI.

Finally, for clinicians already acquainted with ChatGPT-4, potential biases might arise if they were able to distinguish between AI-generated and physician-created notes by the structure and grammar of the note. This recognition could influence our results, as the participants would then not be fully blinded in the study.

Conclusion

Our pilot study indicates that ChatGPT-4 creates orthopedic discharge papers faster than physicians, with comparable quality, and might assist in reducing the administrative load on orthopedic surgeons by generating clinically suitable discharge documents.

In perspective, while ChatGPT-4 shows promise in producing discharge documents, its use requires careful handling. Entering patient data into the chat might violate Patient Health

Information Protection regulations, resulting in substantial fines for unauthorized disclosures [25]. Currently, ChatGPT-4 is not ready for immediate clinical application using patient data because it presents legal and technical challenges that must first be addressed. Further research is needed using patient data from electronic health records. Additionally, exploring the model's fine-tuning and simplify the prompting process would be beneficial.

Supplementary data

Cases, evaluation form, prompt, and generated notes are available as supplementary data on the article page, doi: 10.2340/17453674.2024.40182

GSR, MM, CB, and MG were responsible for the study conception and designed the study. NB and MW created the cases. MK and MW wrote the discharge notes and prompted ChatGPT-4 to generate the AI documents. GSR and CB created the prompt. GSR and CB wrote the manuscript with continuous feedback from MG, MM, MW, MK, and AMM. MG and CB supervised the study.

The authors would like to express their sincere gratitude to Kenneth Jonsson, Mona Mili, Viktor Schmidt, Sara Lindmark, Fredrik Peyronson, Hannah Coudé Adam, Marc-Antoine Burch, Birgit Oberreiter, Franziska Eckers, Sebastian Müller, Cornelia Baum, and Petros Ismailidis for their assistance in evaluating these cases.

Handling co-editor: Taco Gosens
Acta thanks Olivier Groot and Walter van der Weegen for help with peer review of this manuscript.

1. **Patel R S, Bachu R, Adikey A, Malik M, Shah M.** Factors related to physician burnout and its consequences: a review. Behav Sci Basel Switz 2018; 8(11): 98. doi: 10.3390/bs8110098.
2. **Wright A A, Katz I T.** Beyond burnout: redesigning care to restore meaning and sanity for physicians. N Engl J Med 2018; 378(4): 309-11. doi: 10.1056/NEJMp1716845.
3. **Woolhandler S, Campbell T, Himmelstein D U.** Costs of health care administration in the United States and Canada. N Engl J Med 2003; 349(8): 768-75. doi: 10.1056/NEJMsa022033.
4. **Nadeau S, Cusick J, Shepherd M.** Excess administrative costs burden the U.S. health care system. [Published online November 2, 2021.] Available from: https://www.americanprogress.org/article/excess-administrative-costs-burden-u-s-health-care-system.
5. **Kung T H, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al.** Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2(2): e0000198. doi: 10.1371/journal.pdig.0000198.
6. **Gilson A, Safranek C W, Huang T, Socrates V, Chi L, Taylor R A, et al.** How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023; 9: e45312. doi: 10.2196/45312.
7. **Giannos P, Delardas O.** Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ 2023; 9: e47737. doi: 10.2196/47737
8. **Massey P A, Montgomery C, Zhang A S.** Comparison of ChatGPT-3.5, ChatGPT-4, and orthopedic resident performance on orthopedic assessment examinations. J Am Acad Orthop Surg 2023; 31(23): 1173-79. doi: 10.5435/JAAOS-D-23-00396.
9. **Hurley E T, Crook B S, Lorentz S G, Danilkowicz R M, Lau B C, Taylor D C, et al.** Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. Arthroscopy 2023:S0749-8063(23)00642-4. doi: 10.1016/j.arthro.2023.07.048.
10. **Mika A P, Martin J R, Engstrom S M, Polkowski G G, Wilson J M.** Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am 2023; 105(19): 1519-26. doi: 10.2106/JBJS.23.00209.
11. **Ayers J W, Poliak A, Dredze M, Leas E C, Zhu Z, Kelley J B, et al.** Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023; 183(6): 589-96. doi: 10.1001/jamainternmed.2023.1838.
12. **O'Connor S.** Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract 2023; 66:103537. doi: 10.1016/j.nepr.2022.103537.
13. **Ollivier M, Pareek A, Dahmen J, Kayaalp M E, Winkler P W, Hirschmann M T, et al.** A deeper dive into ChatGPT: history, use and future perspectives for orthopedic research. Knee Surg Sports Traumatol Arthrosc 2023; 31(4): 1190-2. doi: 10.1007/s00167-023-07372-5.
14. **Liu X, Rivera S C, Moher D, Calvert M J, Denniston A K, SPIRIT-AI and CONSORT-AI Working Group.** Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020; 370: m3164. doi: 10.1136/bmj.m3164.
15. **Singhal K, Azizi S, Tu T, Mahdavi S S, Wei J, Chung H W, et al.** Large language models encode clinical knowledge. Nature 2023; 620(7972): 172-180. doi: 10.1038/s41586-023-06291-2.
16. **Ali S R, Dobbs T D, Hutchings H A, Whitaker I S.** Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023; 5(4): e179-e181. doi: 10.1016/S2589-7500(23)00048-1.
17. **Wimsett J, Harper A, Jones P.** Review article: Components of a good quality discharge summary: a systematic review. Emerg Med Australas 2014; 26(5): 430-8. doi: 10.1111/1742-6723.12285.
18. **Greer R C, Liu Y, Crews D C, Jaar B G, Rabb H, Boulware L E.** Hospital discharge communications during care transitions for patients with acute kidney injury: a cross-sectional study. BMC Health Serv Res 2016; 16(1): 449. doi: 10.1186/s12913-016-1697-7.
19. **Kripalani S, LeFevre F, Phillips C O, Williams M V, Basaviah P, Baker D W.** Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA 2007; 297(8): 831-41. doi: 10.1001/jama.297.8.831.
20. **Brameier D T, Alnasser AA, Carnino J M, Bhashyam A R, von Keudell A G, Weaver M J.** Artificial intelligence in orthopedic surgery: can a large language model "write" a believable orthopedic journal article? J Bone Joint Surg Am 2023; 105(17): 1388-92. doi: 10.2106/JBJS.23.00473.
21. **West C P, Dyrbye L N, Shanafelt T D.** Physician burnout: contributors, consequences and solutions. J Intern Med 2018; 283(6): 516-29. doi: 10.1111/joim.12752.
22. **Rotenstein L S, Torre M, Ramos M A, Rosales R C, Guille C. Sen S, et al.** Prevalence of burnout among physicians: a systematic review. JAMA 2018; 320(11): 1131-1150. doi: 10.1001/jama.2018.12777.
23. **Panagioti M, Panagopoulou E, Bower P, Lewith G, Kontopantelis E, Chew-Graham C, et al.** Controlled interventions to reduce burnout in physicians: a systematic review and meta-analysis. JAMA Intern Med 2017; 177(2): 195-205. doi: 10.1001/jamainternmed.2016.7674.
24. **Patel S B, Lam K.** ChatGPT: the future of discharge summaries? Lancet Digit Health 2023; 5(3): e107-e108. doi: 10.1016/S2589-7500(23)00021-3.
25. **Kanter G P, Packel E A.** Health care privacy risks of AI chatbots. JAMA 2023; 330(4): 311-12. doi: 10.1001/jama.2023.9618.