JMIR Publications
The Leading eHealth Publisher

Artificial intelligence in dental radiology: improving patient communication with ChatGPT – a comparative study

Daniel Stephan, Annika Bertsch, Sophia Schumacher, Behrus Puladi, Mathias Burwinkel, Bilal Al-Nawas, Peer W Kämmerer, Daniel GE Thiem

Submitted to: Journal of Medical Internet Research
on: March 02, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint
JMIR Publications

Artificial intelligence in dental radiology: improving patient communication with ChatGPT – a comparative study

Daniel Stephan[1] Dr med; Annika Bertsch[1]; Sophia Schumacher[1]; Behrus Puladi[2] Dr med, Dr med dent; Mathias Burwinkel[1] Dr med dent; Bilal Al-Nawas[1] Prof Dr Med, Dr med dent; Peer W Kämmerer[1] Prof Dr Med, Dr med dent, MA; Daniel GE Thiem[1] PD, Dr med, Dr med dent, MHBA

[1] Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz Mainz DE
[2] Department of Oral and Maxillofacial Surgery University Hospital RWTH Aachen Aachen DE

Corresponding Author:
Daniel Stephan Dr med

Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz
Augustusplatz 2
Mainz
DE

Abstract

Objectives: Artificial intelligence (AI) has emerged as a transformative tool in healthcare, particularly in improving medical documentation and communication. This study investigates the potential of ChatGPT, an AI language model, to enhance patient comprehension of radiology reports through simplification.
Materials and

Methods: Three versions of radiology reports were evaluated: original AI-generated, simplified, and further simplified versions. Patient feedback of 300 patients was collected via structured questionnaires assessing clarity, tone, structure, and patient engagement.

Results: The results demonstrated that both simplified versions significantly outperformed the original AI-generated reports across all dimensions, with the further simplified version receiving the highest ratings. Patients consistently reported greater clarity, improved tone, and enhanced engagement with the simplified texts. Readability analysis confirmed these findings, as simplified reports achieved higher Flesch Reading Ease scores and lower LIX indices (readability index), indicating improved accessibility without compromising clinical relevance.

Conclusion: Overall, simplified reports effectively empower patients to understand their medical conditions and actively engage in healthcare decisions.
Clinical Relevance: This study therefore highlights the transformative potential of ChatGPT in advancing patient-centered communication while emphasizing the need for a cautious and collaborative approach to AI integration in clinical workflows.

(JMIR Preprints 02/03/2025:73337)
DOI: https://doi.org/10.2196/preprints.73337

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone.

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Artificial intelligence in dental radiology: improving patient communication with ChatGPT – a comparative study

Stephan D.[1], Bertsch A.[1], Schumacher S.[1], Puladi B.[2], Burwinkel M.[1], Al-Nawas B.[1], Kämmerer P.W.[1], Thiem D.G.E.[1]

Corresponding author:

Dr. med. Daniel Stephan, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Augustusplatz 2, 55131 Mainz, Germany. Phone: +49-6131-17-3080, Fax: +49-6131-17-8468, Email: stephand@uni-mainz.de

Authors:

Annika Bertsch, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Sophia Schumacher, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre of the Johannes Gutenberg-University Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Dr. med. Dr. med. dent. Behrus Puladi, [2]Department of Oral and Maxillofacial Surgery University Hospital RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany

Dr. med. dent. Matthias Burwinkel, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Prof. Dr. med. Dr. med. dent. Bilal Al-Nawas, [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Prof. Dr. med. Dr. med. dent. Peer W. Kämmerer, MA [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany.

Priv.-Doz. Dr. med. Dr. med. dent. Daniel G.E. Thiem, MHBA [1]Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery, University Medical Centre Mainz, Augustusplatz 2, 55131 Mainz, Germany. Phone: +49-6131-17-3080, Fax: +49-6131-17-8468, Email: daniel.thiem@uni-mainz.de

Preprint JMIR Publications

ABSTRACT

Objectives: Artificial intelligence (AI) has emerged as a transformative tool in healthcare, particularly in improving medical documentation and communication. This study investigates the potential of ChatGPT, an AI language model, to enhance patient comprehension of radiology reports through simplification.

Materials and Methods: Three versions of radiology reports were evaluated: original AI-generated, simplified, and further simplified versions. Patient feedback of 300 patients was collected via structured questionnaires assessing clarity, tone, structure, and patient engagement.

Results: The results demonstrated that both simplified versions significantly outperformed the original AI-generated reports across all dimensions, with the further simplified version receiving the highest ratings. Patients consistently reported greater clarity, improved tone, and enhanced engagement with the simplified texts. Readability analysis confirmed these findings, as simplified reports achieved higher Flesch Reading Ease scores and lower LIX indices (readability index), indicating improved accessibility without compromising clinical relevance.

Conclusion: Overall, simplified reports effectively empower patients to understand their medical conditions and actively engage in healthcare decisions.

Clinical Relevance: This study therefore highlights the transformative potential of ChatGPT in advancing patient-centered communication while emphasizing the need for a cautious and collaborative approach to AI integration in clinical workflows.

Key words: artificial intelligence, ChatGPT, radiology report, patient communication, patient-centered care, natural language processing, patient education, medical accessibility

Preprint JMIR Publications

Introduction

The recording and communication of diagnostic findings, treatment plans and patient progress are fundamental to high-quality patient care [1]. Structured and standardized medical documentation should support two essential objectives in healthcare: it should enable effective communication among medical professionals while also improving therapy involvement of patients. However, the complexity of medical documentation often possesses significant barriers, particularly for patients, due to technical language and specialized terminology limiting comprehension. Addressing these challenges is essential to advance patient-centered communication, which builds trust, enhances patient satisfaction, and strengthens adherence to treatment plans. Clear and comprehensible medical information enables patients to better understand their conditions and actively participate in shared decision-making, highlighting the need for innovative approaches to improve accessibility and understanding [2, 3].

Despite its critical role in healthcare, traditional manual documentation presents numerous challenges. Writing detailed medical reports is both time-consuming and error-prone, particularly when multiple versions are required to address different audiences, including patients and healthcare professionals. Additionally, variability in style and content can hinder interpretation, especially when simplified language compromises technical accuracy. Healthcare providers further cope with extensive documentation requirements, driven by legal, regulatory, and insurance standards, contributing to documentation overload [4]. These demands not only impair clinical workflow efficiency but also demonstrate the importance of solutions balancing accuracy, clarity, and usability in fast-paced medical environments.

Artificial intelligence (AI) has emerged as a transformative tool in medicine, expanding its applications beyond disease diagnosis and pharmaceutical drug development [5]. AI technologies have demonstrated remarkable potential across various medical disciplines, particularly in radiology, where advanced algorithms facilitate medical image interpretation. These innovations enable early disease detection, improve diagnostic precision, and support clinical decision-making [6, 7]. Furthermore, AI-driven methodologies are increasingly employed to optimize and streamline healthcare workflows including documentation practices. Due to automating repetitive tasks and translating complex medical nomenclature into comprehensible language, AI addresses persistent inefficiencies [8]. One major benefit of using AI in clinical practice is that it can streamline the creation of clear, easy-to-understand medical documents—without adding extra work for healthcare professionals. This capability is particularly crucial in radiology, as reports are highly technical and predominantly intended for medical professionals.

Patients are often provided with their radiology reports digitally or as printed document, frequently without sufficient explanation. This lack of information can result in misinterpretation and confusion. Unfamiliar terms such as "lesion" or "opacity” may cause patients to overestimate the severity of findings, leading to unnecessary anxiety. Conversely, the urgency of critical recommendations might be overlooked if findings are not communicated effectively,

Preprint JMIR Publications

resulting in delayed or missed follow-ups. Additionally, the absence of clear communication in radiology reports may compromise informed consent, leaving patients unable to fully understand the implications of their findings or treatment options. The AI-driven translation of technical medical language into easily comprehensible terms offers the potential to bridge the gap between complex medical findings and patient understanding.

Although traditional methods of simplifying medical language often impose a significantly increased workload, advancements in AI offer a more efficient and comprehensive opportunity. Large language models (LLMs), such as OpenAI's generative pre-training transformer (GPT), which was first introduced in 2018 [9], have gained substantial attention in recent years as transformative tools in natural language processing. GPT has undergone consecutive refinement through extensive training, resulting in progressively enhanced capabilities in language understanding and generation, including medical texts [10, 11]. Notably, ChatGPT, an implementation of GPT, excels in generating coherent, human-like text by leveraging advanced deep learning techniques, including neural networks, to process and produce language with remarkable precision. ChatGPT's potential in medical applications is underscored by its strong performance in standardized examinations [12, 13], demonstrating its ability to process complex medical content at a level comparable to human expertise. This capability has imparted its integration into clinical practice and is employed to automate routine documentation tasks, including drafting examination findings, composing operative notes and generating radiology reports [14, 15]. Beyond its utility for healthcare professionals, ChatGPT presents an opportunity to address challenges in patient communication by simplifying complex medical information. This potential to enhance patient communication by generating accessible and comprehensible radiology reports represents an emerging area of interest [8].

While advancements in AI have demonstrated its utility in generating technically accurate and readable medical texts, research on their impact on patient communication remains limited. Most existing studies emphasize the evaluation of technical attributes of AI-generated texts, such as accuracy, grammar, and readability, without adequately addressing their effects on patient understanding or engagement [16]. Building on prior research, this study therefore aimed to investigate the potential of ChatGPT in generating and simplifying radiology reports by primarily analyzing patient evaluations regarding text comprehensibility, clarity, and communication quality of both AI-generated and AI-simplified radiology reports with the null hypothesis stating no differences among the three sets of reports. Secondary outcomes, including text quality and readability, were assessed to identify opportunities for improvement in AI-driven radiology reporting.

Material and Methods

The present study aimed to evaluate the efficacy of utilizing AI language models, specifically ChatGPT, in simplifying and enhancing the readability of radiology reports to improve patient communication. Radiology reports were generated in two ways: First, dental students provided written reports based on a panoramic radiography. Second, checkbox lists of diagnoses filled out by students were used to generate radiology reports using ChatGPT. Additionally, ChatGPT was used to create one simplified and a second further simplified version of these AI-generated reports to assess the impact of simplification on readability and comprehension. The primary evaluation focused on patient responses collected through a questionnaire, while secondary parameters included text quality and readability differences between student-written, original AI-generated, and simplified AI-generated reports. Notably, the evaluation of student-written reports by patients was not part of the study, as the focus was exclusively on AI-generated content.

Text generation:

The analysis of dental panoramic radiographs and the generation of radiology reports were not part of this study. Instead, the student-written reports and the initial AI-generated reports (based on checkbox lists) were obtained from a prior study. The report generation process, previously described in detail by our group [16], is summarized briefly as follows: Two panoramic radiographs were independently analyzed by 100 dental students. For the first radiograph, students produced a written radiology report, whereas for the second radiograph, they completed a structured checkbox list of diagnoses. These checkbox lists were transcribed into an Excel datasheet, with individual spreadsheets created for each diagnostic category to ensure systematic data organization. Subsequently, radiology reports were generated using ChatGPT 4.0 (OpenAI, San Francisco, CA, USA), a state-of-the-art AI language model with following settings: a temperature of 0.7, a maximum token limit of 1500, a frequency penalty of 0.0, and a presence penalty of 0.6. Each report was generated in a new session of the ChatGPT web interface to maintain consistency and standardization throughout the process using the following prompt:

“Formulate a structured X-ray report in the sense of an X-ray report of an OPG based on the following checkbox list of the entire Excel table, and do not omit any columns. Please mention only those statements for which a box is marked with an X in the X-ray report. The statements not marked with an X should not be included in the report. The figures given should be interpreted in the sense of an odontogram. Analyze each spreadsheet in the Excel file in the order given. The column with the markings (X) is marked with the term "checkbox". The report should be written in continuous text from the perspective of the treating dentist. Formulate a continuous text without subheadings.”

Simplification of Radiology Reports

The AI-generated reports have been analyzed in detail in prior studies. Based on these reports, two additional simplified versions were created for each. Each of the 100 AI-generated texts was reformulated into two simplified versions using ChatGPT 4.0 (OpenAI, San Francisco, CA, USA) under the following settings: a temperature of 0.7 (controlling the randomness of responses), a maximum token limit of 1500 (restricting output length), a frequency penalty of 0.0 (preventing repetitive word usage), and a presence penalty of 0.6 (encouraging the inclusion of new content), resulting in a total of 300 texts. Reports were generated via the ChatGPT web interface in individual sessions conducted between April 6th and May 8th, 2024. Potential variability in performance due to server load fluctuations and biases in output quality were reduced by random distribution across different weekdays, hence ensuring consistency and reliability.

The following prompts were used to create the simplified versions of the reports:

1. *“Rewrite the radiology report to make it easier for a patient to understand. Do not leave out any information or content.”*
2. *“Rewrite the radiology report to make it understandable for patients of all educational backgrounds. Do not leave any information or content.”*

The resulting texts were analyzed for readability and text quality. To minimize bias in the patient evaluation phase, a single text was randomly selected from each group (original AI-generated, first simplified, and second simplified). The selected texts were assessed for readability (Flesch reading ease score and LIX readability index) to ensure that they represent the mean of their respective groups. This approach avoided the inclusion of outliers. These three representative texts were subsequently used for patient evaluation.

Study Setting and Participants

The study was conducted in the Department of Oral and Maxillofacial Surgery, Facial Plastic Surgery at the University Medical Centre Mainz, Germany. Patients were randomly selected and invited to participate. Those who agreed to participate were provided with one randomly assigned text and the corresponding questionnaire. Patients with prior medical education or knowledge were excluded from the study to ensure the data reflected the perspectives of a general audience. All data were collected anonymously, with no personal information recorded. Participants were informed of their right to withdraw consent at any time without consequences. A total of 300 patients participated in the study, with 100 patients assigned to evaluate each text version.

Text Analysis

To compare AI-generated with student-written radiology reports descriptive text analysis was performed. Metrics included word count, sentence count and proportion of long words (defined as words with more than six characters). Since AI-generated texts are inherently free of spelling and grammatical errors, no evaluation of errors or error rates was carried out.

Preprints JMIR Publications

Readability Indices

The readability and complexity of the texts were assessed using the Flesch Reading Ease (FRE) score [17] and the LIX readability index, as described in detail in our prior publication. These metrics evaluate text comprehensibility based on linguistic features such as sentence length and word complexity.

The FRE score calculates readability based on the average sentence length (ASL) and the average number of syllables per word (ASW) [18] using the following formula:

$$Flesch\,Reading\,Ease = 180 - ASL - 58{,}5 * ASW$$

As the established metric, higher FRE scores indicate more readable texts, while lower scores reflect greater complexity [19, 20]. The LIX-Index considers the average sentence length and the prevalence of long words with more than six letters to assess text readability using the following formula:

$$LIX = \frac{Total\,number\,of words}{Total\,number\,of\,sentences} + \frac{\left(Number\,of\,long\,words * 100\right)}{Total\,number\,of\,words}$$

Higher LIX scores denote increased complexity, while lower scores suggest simpler texts [21, 22]. Both indices are well-established tools validated in medical and non-medical contexts and provide a robust measure for comparing text readability.

Experimental Design

Patients were randomly assigned to one of three groups, each receiving one text (text 1, text 2, or text 3) along with the corresponding questionnaire. After completing the questionnaire, the responses were collected and analyzed. There were no additional restrictions or limitations imposed in this study and participants were not subjected to a time limit for completing the questionnaire.

Questionnaire:

The questionnaire consisted of 11 questions divided into five thematic categories:

1. Clarity of the Report
    - I found the radiology report easy to understand. (Question number 1)
    - The medical terms in the radiology report were clearly explained. (Question number 2)
    - I was able to understand the findings of the radiology report without additional help. (Question number 3)
2. Structure and Information Content
    - The structure of the radiology report helped me to easily understand the information. (Question number 4)
    - The information in the radiology report was detailed enough to answer my

questions. (Question number 5)

3. Empathy and Tone
    - The tone of the radiology report was respectful and took my situation as a patient into account. (Question number 6)
    - The radiology report conveyed an empathetic attitude towards my level of understanding. (Question number 7)
4. Guidance and Motivation for Action
    - The radiology report helps me ask the right questions to my physician/dentist. (Question number 8)
    - After reading the radiology report, I can discuss my diagnosis with my physician/dentist and have a discussion on equal terms. (Question number 9)
    - The radiology report motivates me to focus more on oral hygiene and health in the future. (Question number 10)
5. Future Perspective
    - I would like all my medical reports to be as understandable as this radiology report. (Question number 11)

Questions were answered on a 5-point Likert scale ranging from 1 (“do not agree at all”) to 5 (“fully agree”), with intermediate options: “rather do not agree” (2), “neither agree nor disagree” (3) and “rather agree” (4).

Ethical Approval

This study was conducted in strict adherence to ethical guidelines, with informed consent obtained from all participants verbally prior to their involvement. Confidentiality and data privacy were rigorously upheld throughout the research process to ensure the participant's well-being and privacy. The local ethics committee (Ethics Committee of the State Medical Association of Rhineland-Palatinate, Mainz, Germany) confirmed that an ethics approval was not required as the study primarily constituted a quality assurance measure, and the survey was conducted anonymously (Application Number: 2024-17526). Furthermore, all data were analyzed anonymously.

Power Analysis and Sample Size

The sample size for this study was determined through an a priori power analysis based on the primary outcome: patient responses to the questionnaire. A recent study evaluating a patient information leaflet for liver cirrhosis, patients demonstrated a significant improvement in pre- and post-test knowledge scores [23]. Specifically, the study reported an improvement of 4 points (pre-test: median = 12, range: 6–20; post-test: median = 16, range: 12–22; $p < 0.05$), corresponding to a moderate effect size (Cohen’s $d = 0.4$). Assuming a similar effect size, to achieve a power of 80% and maintain a significance level of 5%, a minimum of 53 samples per group (159 participants in total) is required. To ensure robustness and account for potential

dropouts or incomplete data, the study included 100 participants per group, resulting in a total of 300 participants.

Statistical Analysis

Statistical analyses were conducted using the following software packages: GraphPad Prism 9.0 (GRAPHPAD SOFTWARE, LLC, Boston, USA), G*Power 3.1 (Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany), Excel 16.76 (Microsoft Corporation, Redmond, USA) and SPSS Statistics 29 (IBM Deutschland GmbH, Böblingen, Germany). ChatGPT 4.0 was used for proofreading. Differences in text quality and readability between student-written and AI-generated texts were analyzed using a one-way repeated measures ANOVA. Post hoc comparisons were performed with Tukey's multiple comparisons test. All numerical data are presented as mean ± standard deviation (SD), with statistical significance set at $p < 0.05$. Analysis of the patient questionnaire was conducted using an ordinary one-way ANOVA with Tukey's multiple comparisons test for post hoc comparison. Each graph corresponds to a specific question displayed above the respective graph and data are presented as box-and-whisker plots, illustrating medians and interquartile ranges.

Preprint JMIR Publications

Results

Text quality, readability, and comprehensibility of AI-generated and AI-simplified radiology reports were evaluated. Additionally, patient feedback was integrated to assess the effectiveness of simplified reports in enhancing understanding and engagement with medical information. The findings demonstrate that simplified versions significantly improved readability and accessibility, aligning with the primary objective of enhancing patient communication.

Simplified radiology reports possess higher sentence and word count with reduced use of long words

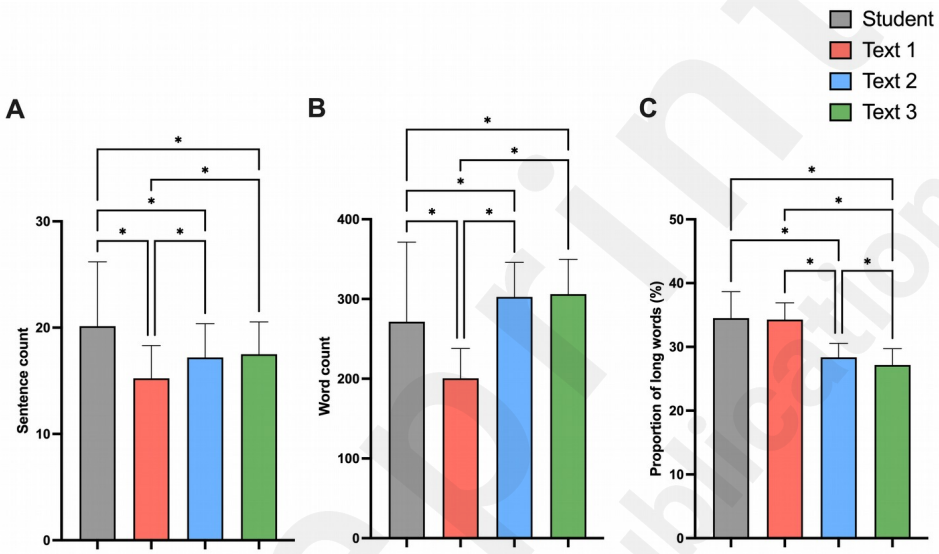

Figure 1: Analysis of sentence count(A), word count (B) and proportion of long words (C) of AI-simplified radiology reports compared to student-written reports. Data represent mean +/- standard deviation. Sample size: n = 100; *p<0.05.

The one-way repeated measures ANOVA revealed a significant difference in sentence count among the four text groups, including the student-written text, the AI-generated text, and its simplified versions (*Student:* 20.1 ± 6.0 vs. *text 1*: 15.2 ± 3.1 vs *text 2*: 17.2 ± 3.2 vs. *text 3*: 17.5 ± 3.0; $p < 0.0001$; $F (1.657, 164.0) = 30.26$; $df = 3$). In particular, all AI-generated texts contained fewer sentences than the student-written texts as presented in figure 1A, while both simplified versions had a higher sentence count compared to the original AI-generated text. A similar reduction could be observed in word count, with significant differences across groups (*Student:* 271.6 ± 99.7 vs. *text 1*: 200.6 ± 37.3 vs *text 2*: 302.6 ± 43.3 vs. *text 3*: 306.1 ± 43.8; $p < 0.0001$; $F (1.419, 140.5) = 73.20$; $df = 3$). Text 1 contained significantly fewer words than the student-written reports, while both simplified versions (text 2 and text 3) included significantly more words than text 1 and 2, with no significant difference in word count between text 2 and text 3. Those differences were also reflected in the proportion of long words, which showed a significant difference across all groups ($p < 0.0001$; $F (2.213, 219.1) = 179.6$; $df = 3$) as presented in figure 1C. Whereas no difference between student-written and AI-generated reports could be shown, post hoc analysis revealed a significant reduction in the proportion of long words in the simplified text versions compared to both the student-written and original AI-

generated texts.

Simplified text versions demonstrated improved readability

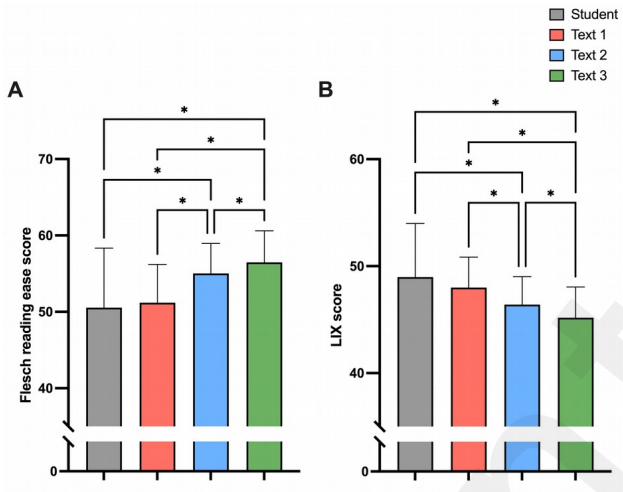

Figure 2: Metric evaluation of readability of AI-simplified reports and student-written radiology reports assessed with the Flesch Reading Ease score (A) and LIX score (B). Data represent mean +/- standard deviation. Sample size: n = 100; *p<0.05.

Significant differences in readability were observed across the four text groups, as identified by one-way repeated measures ANOVA (Flesch reading ease score (figure 2A): $p < 0.0001$; $F(2.090, 207.0) = 29.41$; $df = 3$; LIX score (figure 2B): $p < 0.0001$; $F(2.058, 203.8) = 23.60$; $df = 3$). No difference in readability was found between both student-written and AI-generated texts (text 1). However, both showed significantly lower Flesch Reading Ease scores compared to the simplified texts (text 2 and text 3), indicating improved readability due to AI-simplification. Text 3 displayed a significantly higher score than text 2. Similar results could be observed in the LIX scores as depicted in figure 2B. The student-written reports and text 1 showed comparable LIX scores, which were significantly higher than those of text 2 and 3, indicating greater complexity. In contrast, both simplified versions (text 2 and text 3) exhibited significantly lower LIX scores, with text 3 being the simplest and most readable text compared to the other groups.

Descriptive statistics of questionnaire responses: comparison of patients' perceptions of AI-generated and simplified texts

Table 1 provides a summary of descriptive statistics for patient responses to text 1 (AI-generated radiology reports) across 11 survey questions. Median ratings ranged from 2 to 4, with question 2 receiving the highest mean score (3.2 ± 1.3) and question 6 the lowest (2.1 ± 1.1). The overall range of responses spanned from 1 (minimum) to 5 (maximum), reflecting variability in patient perceptions of clarity, tone, and informativeness. Coefficients of variation were generally moderate, indicating consistent responses across most questions.

Table 1: Descriptive statistics for questionnaire responses to text 1 (AI-generated radiology report): Summary of questionnaire responses for text 1 (student-written radiology reports). Data represent mean ± standard deviation, minimum, maximum, percentiles (25%, 50% [median], and 75%), range, standard error of the mean, and 95% confidence intervals (lower and upper bounds). Sample size: n = 98-100.

Table 2 summarizes descriptive statistics for patient responses to text 2 (simplified AI-generated radiology reports) for all 11 survey questions. Median ratings generally ranged from 1 to 2, with lower mean scores compared to text 1, such as 1.6 ± 0.8 for question 1 and 1.7 ± 1.0 for question 11. The overall response range remained consistent from 1 to 5, indicating variability in patient perceptions. Standard deviations were slightly lower than those of text 1, suggesting more consistent responses. The coefficients of variation highlight moderate variability across questions, with question 7 showing one of the highest at 57.04%.

Table 2: Descriptive statistics for questionnaire responses to text 2 (AI-simplified radiology report): Summary of questionnaire responses for text 2 (AI-generated radiology reports based on a checkbox table). Data represent mean ± standard deviation, minimum, maximum, percentiles (25%, 50% [median], and 75%), range, standard error of the mean, and 95% confidence intervals (lower and upper bounds). Sample size: n = 99-100.

Table number 3 provides an overview of descriptive statistics for patient responses to text 3 (further simplified AI-generated radiology reports). Median ratings were consistently lower, ranging from 1 to 2, with mean scores slightly reduced compared to text 2 (e.g., question 1: 1.5 ± 0.7; question 11: 1.6 ± 0.8). The response range remained between 1 and 5, showing variability in patient perceptions. Standard deviations and coefficients of variation were generally similar to Text 2, with the highest variability observed in question 7 (59.20%).

Preprint JMIR Publications

Table 3: Descriptive statistics for questionnaire responses to text 2 (AI-simplified radiology report): Summary of questionnaire responses for text 3 (AI-generated simplified radiology reports). Data represent mean ± standard deviation, minimum, maximum, percentiles (25%, 50% [median], and 75%), range, standard error of the mean, and 95% confidence intervals (lower and upper bounds). Sample size: n = 100.

Patient survey responses regarding the comprehensibility of radiology reports

Text clarity, structure and information content, empathy and tone, motivation for action and future perspective of AI-generated radiology reports and their AI-simplified versions were evaluated through patient survey responses. Three text groups were assessed: text 1 (AI-generated), text 2 (simplified AI-generated), and text 3 (further simplified AI-generated). Each participant completed a questionnaire referring to only one text version and multiple participations were not allowed. Patients with prior medical knowledge or education were excluded from the survey to ensure responses reflected the perceptions of a general audience.

Questions were answered on a 5-point Likert scale ranging from 1 ("do not agree at all") to 5 ("fully agree"), with intermediate options: "rather do not agree," "neither agree nor disagree," and "rather agree." Each graph corresponds to a specific question displayed above the respective graph, with data being presented as box-and-whisker plots, showing medians and interquartile ranges. Both simplified report versions (text 2 and text 3) received significantly better ratings across all evaluated aspects, highlighting a preference for more accessible medical documents in the future.

Clarity of the Report

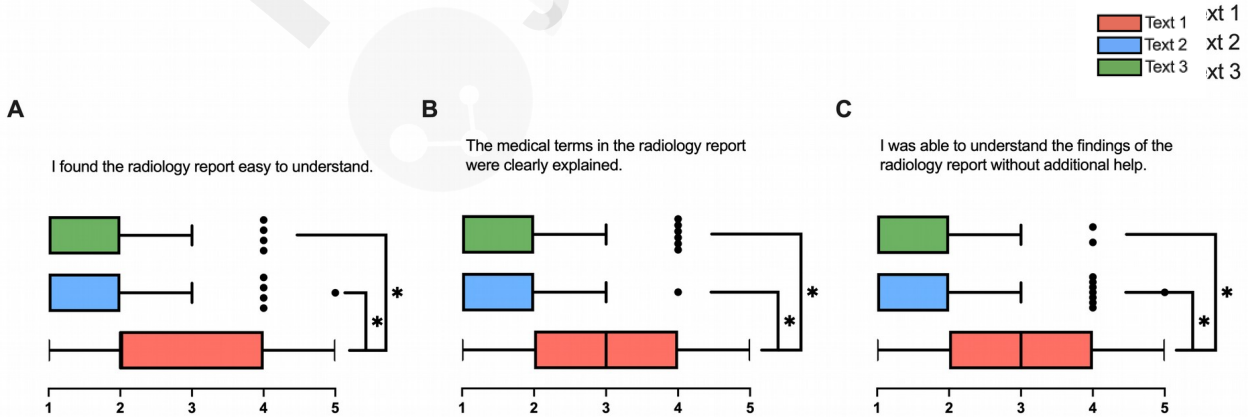

Figure 3: Patient ratings of their comprehension of radiology reports for text 1 (AI-generated), text 2 (simplified AI-generated), and text 3 (further simplified AI-generated). Responses were assessed for (A) ease of understanding the radiology report, (B) clarity of the medical terms, and (C) ability to understand the findings without additional help. Data are presented as box-and-whisker plots with medians and interquartile ranges. Sample size: n = 100; *p < 0.05.

As presented in figure 3 significant differences were identified among the three text versions. The simplified text versions significantly improved patient's ability to understand the information (Figure 3A: $F(2, 296) = 51.61$, $df = 2$, $p < 0.0001$), the clarity of medical terms (Figure 3B: $F(2, 297) = 84.34$, $df = 2$, $p < 0.0001$) and the ability to comprehend the findings of the report without additional help (Figure 3C: $F(2, 297) = 70.68$, $df = 2$, $p < 0.0001$). Throughout all questions regarding clarity and comprehensibility, text 3 received the highest ratings, followed by text 2, with both being significantly higher than text 1.

Structure and Information Content

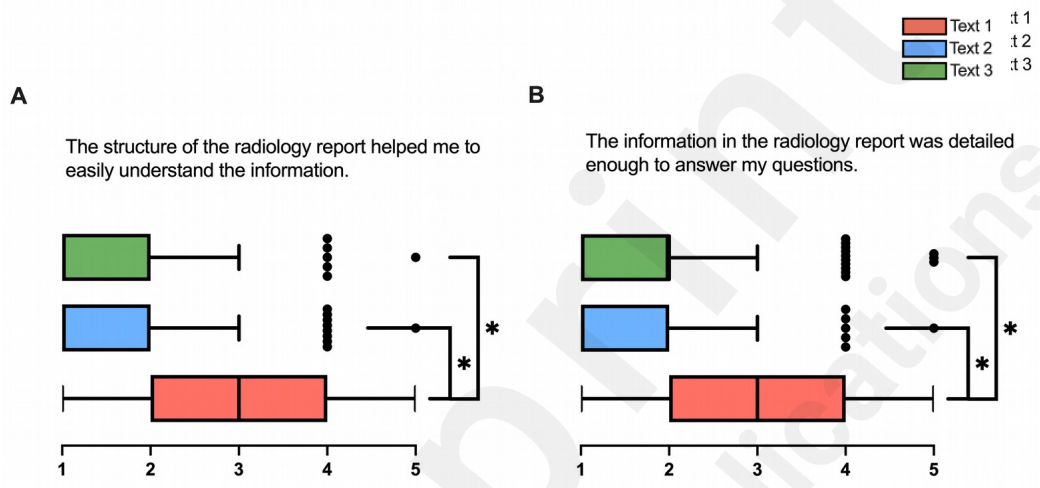

Figure 4: Patient ratings comparing text 1 (AI-generated), text 2 (simplified AI-generated), and text 3 (further simplified AI-generated) radiology reports. Ratings were assessed for (A) how the structure of the report helped in understanding the information and (B) whether the information in the report was detailed enough to answer patients' questions. Data are presented as box-and-whisker plots, showing medians and interquartile ranges. Sample size: $n = 100$; $^*p < 0.05$.

The three text versions demonstrated significant differences in how patients perceived the structure and information content of the radiology reports. Patients rated the structure of the report as being most helpful for understanding the information in the simplified versions (Figure 4A: $F(2, 297) = 43.63$, $df = 2$, $p < 0.0001$). Similarly, the simplified texts were evaluated as containing sufficient detail to adequately answer patient questions (Figure 4B: $F(2, 296) = 31.53$, $df = 2$, $p < 0.0001$). Text 3 demonstrated the highest ratings across both measures, followed by text 2, with text 1 showing lower medians. All those results are significant.

Empathy and Tone

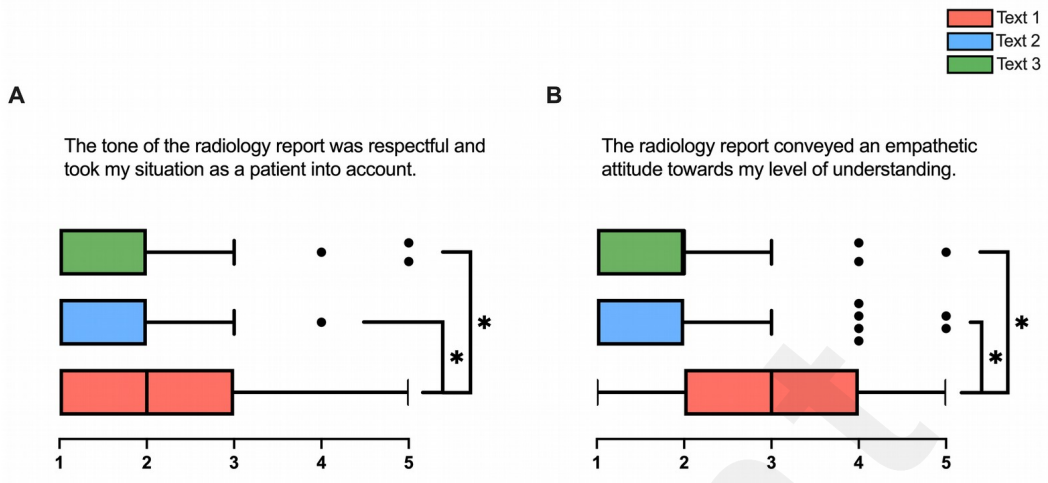

Figure 5: Patient ratings comparing text 1 (AI-generated), text 2 (simplified AI-generated), and text 3 (further simplified AI-generated) radiology reports. Ratings were assessed for (A) whether the tone of the report was respectful and took the patient's situation into account, and (B) whether the report conveyed an empathetic attitude towards the patient's level of understanding. Data are presented as box-and-whisker plots, showing medians and interquartile ranges. Sample size: n = 100; *p < 0.05.

Significant differences were observed in patient responses to the tone and empathy of the radiology reports across the three text groups. (Figure 5A: F (2, 297) = 17.72, df = 2, p < 0.0001; Figure 5B: F (2, 295) = 49.57, df = 2, p < 0.0001). Patients rated the tone of the radiology report as significantly more respectful and considerate of their situation in the simplified text versions (text 2 and text 3) compared to the original AI-generated radiology report. Similarly, differences were identified in how the reports conveyed an empathetic attitude towards the patient's level of understanding with text 2 and text 3 receiving significantly higher ratings than text 1.

Guidance and Motivation for Action

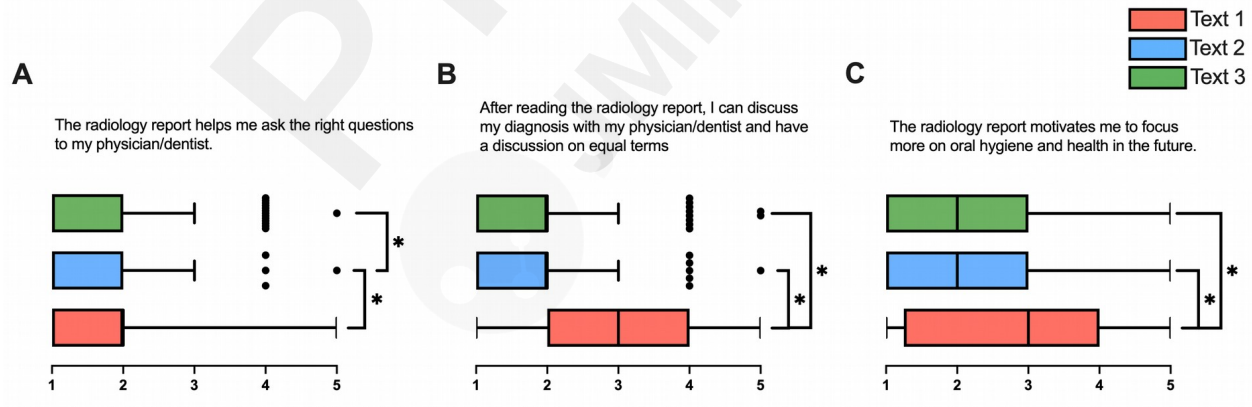

Figure 6: Patient ratings comparing text 1 (AI-generated), text 2 (simplified AI-generated), and text 3 (further simplified AI-generated) radiology reports. Ratings were assessed for (A) whether the report helps patients ask the right questions to their physician or dentist, (B) whether the report enables a discussion on equal terms with their physician or dentist, and (C) whether the report motivates patients to focus more on oral hygiene and health in the future. Data are presented as box-and-whisker plots, showing medians and interquartile ranges. Sample size: n = 100; *p < 0.05.

Patient responses showed significant differences across the three text versions regarding the guidance and motivation provided by the radiology reports. Patients were asked if the reports helped them formulate the right questions for their physician or dentist (Figure 6A: F (2, 295) =

9.387, df = 2, p< 0.0001), supported discussions with their healthcare provider on equal terms (Figure 6B: F (2, 297) = 29.03, df = 2, p < 0.0001), and motivated them to prioritize oral hygiene and health in the future (Figure 6C: F (2, 297) = 8.873, df = 2, p < 0.0001). Median values differed across significantly across text groups, with text 2 and text 3 receiving significantly higher ratings than text 1 for facilitating discussions with healthcare providers (Figure 6B) and motivating patients to focus on oral hygiene and health (Figure 6C).

Future Perspective

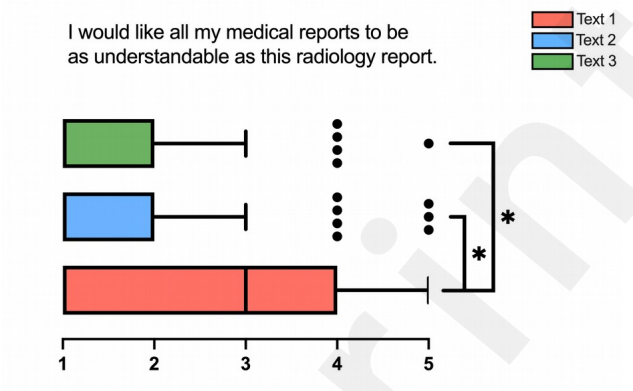

Figure 7: Patient ratings comparing text 1 (AI-generated), text 2 (simplified AI-generated), and text 3 (further simplified AI-generated) radiology reports. Ratings reflect the preference for all medical reports to be as understandable as the radiology report presented. Data are shown as box-and-whisker plots, displaying medians and interquartile ranges. Sample size: n = 100; *p < 0.05.

Among all three text versions, the one-way ANOVA revealed a significant difference regarding the preference for the understandability of medical reports as presented in figure 7 (F (2, 297) = 33.75, df = 2, p < 0.0001). Text 2 and text 3 were rated significantly higher than text 1, indicating a strong preference for the simplified versions.

Discussion

Integrating AI into clinical workflows has attracted significant interest due to its potential to enhance efficiency. Our study, therefore, aimed to investigate the potential of ChatGPT in improving the clarity and comprehensibility of radiology reports for patients. Simplified versions of AI-generated reports with higher readability scores were consistently rated higher by patients than the original AI-generated reports across all evaluated dimensions, including clarity, structure, tone, and patient engagement. The most simplified version demonstrated the greatest improvements, indicating that simplification enhances accessibility without compromising the quality of the content. Patients found the simplified reports more empathetic and respectful, which positively influenced their ability to engage in meaningful healthcare discussions and focus on maintaining health.

AI in clinical documentation and communication

The use of medical documents must account for the diverse communication needs within healthcare settings. Healthcare professionals require detailed and comprehensive reports to ensure accurate clinical decision-making and effective inter-professional communication. Contrary, patients benefit from simplified and accessible documents enabling them to understand their medical conditions and actively participate in treatment discussions [2, 3]. In the United States for example, it is recommended that patient-facing health literature is written at or below a sixth-grade reading level (age 11–12 years) to accommodate the general population's literacy levels [24]. Health literacy (HL), defined as an individual’s ability to seek, comprehend, and use health information to make appropriate health-related decisions, is a critical factor in patient communication. Therefore, clinical letters or individual patient reports serve as communication tools to support health literacy and should hence be written in a manner that can be easily understood [25]. However, many patients still receive medical documents intended for professional communication, which are often too complex to comprehend for individuals without medical education. A recent meta-analysis by Armache et al., evaluated the readability of 1,124 head and neck cancer patient education materials (PEM) produced by professional societies, hospitals, and other organizations intended to inform patients, promote their engagement and their adherence to treatment plans. The mean Flesch-Kincaid Grade Level of these materials ranged from 8.8 to 14.8, with none of them meeting the recommended sixth-grade readability standard. Additionally, eight studies included in the analysis reported inadequate HL prevalence ranging from 11.9% to 47.0% [26]. Concomitantly, another study found that none of the 11 publications by the American Society of Metabolic and Bariatric Surgery met the readability criteria for patient education materials [27]. Similar results were reported for the assessment of online PEMs for graft-versus-host disease with an average Flesch Reading Ease Score of 46.4 underlining the persistent issue of insufficiently accessible medical information [28]. These findings therefore highlight the responsibility of healthcare

professionals to diminish preventable literacy barriers by providing patient-centered education materials.

To further address these communication challenges, the integration of AI in healthcare documentation provides the opportunity to meet the needs of both healthcare professionals and patients. Whereas detailed reports support clinical decision-making and inter-professional communication, simplified, patient-friendly texts enable patients to better understand their medical conditions and engage actively in healthcare decisions. Large language models (LLMs), such as GPT-3.5 and GPT-4, have demonstrated significant utility in summarizing complex data and generating comprehensible information tailored to different audiences [8, 29]. This adaptability emphasizes ChatGPT's potential to enhance communication across various medical contexts, as demonstrated by its ability to provide templates, refine the content of clinical consultation letters, and generate patient letters with a humane tone [30]. Furthermore, ChatGPT has been proven efficient in various methods documenting standardized patient histories, including typing and dictation, though remaining constrained to human-written content regarding readability and accessibility [31]. Our findings align with these observations, particularly regarding text structure and readability. AI-generated radiology reports contained fewer sentences and words compared to student-written reports, indicating that AI communicates information concisely. Simplified AI-generated texts included more sentences and words than the original AI-generated versions, reflecting efforts to enhance clarity through paraphrasing and additional explanations. The readability analysis further revealed higher Flesch reading ease scores, a validated measure for the assessment of text difficulty [17, 19, 20], for simplified AI texts (56.4 for Text 3 and 55.0 for Text 2) compared to the original AI-generated text (51.1) and student-written reports (50.5). While the latter two did not differ significantly, the scores for simplified texts approached the range expected for sixth-grade readability (around 60), whereas scores below 50 are typically associated with college-level content. These results highlight ChatGPT's ability to effectively simplify medical information and enhance accessibility. Consequently, ChatGPT’s ability to meet the technical precision demanded by healthcare professionals, as demonstrated in prior studies [16], while simultaneously addressing patients' need for accessible information, underscores its transformative potential in medical communication.

Patient evaluation vs. professional evaluation

Most existing LLM outputs prioritize technical quality metrics, such as accuracy and grammar, rather than patient comprehension [29]. Several recent studies have demonstrated the potential of ChatGPT and other LLMs in medical applications but have largely excluded patient input. For example, a recent study reported that chatbot-generated responses to patient questions on social media were rated higher in quality and empathy compared to physician-generated responses [32]. However, this evaluation relied solely on assessments by healthcare

professionals, with no involvement of patients. Similarly, another study assessed the factual correctness and humanness of ChatGPT-generated clinical letters using clinician feedback alone, while another group acknowledged the promise of ChatGPT for providing medical information but emphasized limitations in quality without considering patient perspectives [14, 33]. Further Studies demonstrated ChatGPT's capacity to generate understandable and evidence-based responses, such as frequently asked questions about total hip arthroplasty and cirrhosis-related inquiries, but these investigations also relied exclusively on expert grading, thereby limiting their clinical applicability to patient communication [34, 35]. In contrast to evaluations conducted by professionals [8], the findings of this study integrate patient feedback through structured questionnaires and revealed that simplified AI-generated reports were consistently rated significantly higher across key dimensions, including clarity, tone, and patient engagement, compared to the original AI-generated version. Notably, these simplified reports significantly enhanced patients' ability to understand radiological findings, empowering them to engage more effectively in healthcare discussions. This patient-focused approach not only addresses a key limitation in existing research but also demonstrates the potential of AI-driven solutions to transform medical communication by making complex information accessible to diverse audiences representing a critical advancement in patient-centered care.

Prompt design in AI performance

Previous research validated ChatGPT's ability to generate high-quality radiology reports from structured checkbox information, producing error-free texts with strong similarity to reference materials and excellent readability using a precise prompt [16]. These findings highlighted the importance of carefully crafted instructions in guiding AI outputs effectively, aligning with the established understanding of prompt design as a pivotal factor influencing AI performance [36, 37]. Two distinct prompts were employed: the first instructed, "*Rewrite the radiology report to make it easier for a patient to understand. Do not leave out any information or content*," while the second specified, "*Rewrite the radiology report to make it understandable for patients of all educational backgrounds. Do not leave any information or content*." These prompts were designed to ensure content completeness and accessibility, targeting a broad range of health literacy levels. By adhering to these directives, ChatGPT successfully modified its outputs for patient-facing communication without compromising the integrity of the medical information. This targeted approach underscores the importance of precision in prompt formulation, which directly impacts the relevance and clarity of AI-generated texts [36]. These findings align with prior studies that emphasize the role of precise and detailed prompts in mitigating omissions, inaccuracies, and potential misinformation [37, 38]. One study, for instance, highlighted those errors of omission accounted for 86% of inaccuracies in AI-generated subjective, objective, assessment and plan (SOAP) notes, underscoring the necessity of refining input frameworks to enhance data completeness and consistency [39]. While ChatGPT demonstrates strong

adherence to clinical guidelines, as seen in findings indicating 97% compliance, its ability to provide tailored medical advice or address nuanced clinical scenarios remains limited [40]. These limitations, coupled with the variability observed in outputs across repeated cases, highlight the importance of ongoing refinement in AI input structures and ethical considerations. By incorporating structured protocols and patient feedback, as demonstrated in this study, AI systems can better address the dual needs of healthcare professionals and patients, advancing communication and accessibility in clinical settings.

Limitations and future directions:

While this study highlights the potential of ChatGPT in generating accessible and comprehensible radiology reports, certain limitations and risks need to be considered. The simplified texts were not systematically evaluated for content completeness, as the study prioritized assessing patient comprehension and feedback over engineering the perfect prompt or achieving absolute precision. Nonetheless, the simplifications successfully produced accessible and understandable radiology reports, as evidenced by the readability indices. Considering the focus on evaluating ChatGPT’s potential for enhancing patient communication, there was no need to further refine or evaluate the content and prompt design. Future research could, however, explore how to balance simplification with clinical precision to further optimize AI-generated texts. Ethical and legal considerations present critical challenges to the integration of AI-generated content in clinical workflows. Accountability, particularly in cases of errors in AI-generated reports, remains unresolved and requires thorough discussion at both institutional and regulatory levels. These concerns are particularly significant within the sensitive doctor-patient relationship, where trust is essential. Overreliance on AI-generated materials for patient communication risks diminishing healthcare professionals' direct communication skills, potentially undermining this trust. Consequently, it is crucial to position AI as a complementary tool that supports, rather than replaces, human interaction in clinical practice. The reliability of transferred information remains a key limitation in the use of AI, as healthcare providers are ultimately responsible for the accuracy of the texts they distribute. Thorough proofreading of AI-generated content is therefore essential. Prior research has highlighted that LLMs like ChatGPT often face challenges in tailoring outputs to individual patients' specific needs and circumstances, with nuances in phrasing and contextual accuracy significantly influencing patient understanding and trust [29, 41]. Despite these general limitations, this study demonstrated ChatGPT’s strong adherence to instructions, emphasizing the critical role of precise prompt design in mitigating errors. Importantly, no hallucinations — plausible-sounding but fabricated information [42-44] —were observed in the AI-generated texts. ChatGPT consistently adhered to its directive to rephrase and restructure existing content without inventing or interpreting additional information, thereby minimizing biases and ensuring reliability.

Future research should investigate whether AI-generated text simplifications maintain their

clinical accuracy while enhancing readability and accessibility. These findings underscore the need for collaborative efforts between clinicians, AI developers, and policymakers to establish ethical and legal frameworks that address accountability, accuracy, and trustworthiness in AI-generated content. As ChatGPT continues to evolve, its integration into healthcare workflows holds significant promise for enhancing both professional and patient communication. However, a balanced and cautious approach remains essential to fully realize its transformative potential while mitigating associated risks.

Conclusion

In conclusion, this study highlights the significant potential of Large Language Models like ChatGPT in generating simplified and patient-friendly radiology reports, demonstrating its ability to enhance patient comprehension and engagement. By integrating patient feedback, the findings emphasize the importance of tailoring AI-generated content to meet diverse communication needs within healthcare. While ethical, legal, and practical limitations persist, particularly regarding accountability and content reliability, the study underscores the transformative role of AI as a complementary tool in clinical workflows. Future research should focus on refining AI prompts and balancing simplification with clinical precision to further optimize its application in patient-centered care.

Preprint JMIR Publications

Conflicts of Interest

The authors declare no conflict of interest.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Annotation:

This study was conducted as part of Annika Bertsch’s doctoral thesis.

Preprint JMIR Publications

References

1. Ebbers, T., et al., *The Impact of Structured and Standardized Documentation on Documentation Quality; a Multicenter, Retrospective Study.* J Med Syst, 2022. **46**(7): p. 46.
2. Sharkiya, S.H., *Quality communication can improve patient-centred health outcomes among older patients: a rapid review.* BMC Health Services Research, 2023. **23**(1): p. 886.
3. Pinto, R.Z., et al., *Patient-centred communication is associated with positive therapeutic alliance: a systematic review.* J Physiother, 2012. **58**(2): p. 77-87.
4. Lorkowski, J., I. Maciejowska-Wilcock, and M. Pokorski, *Overload of Medical Documentation: A Disincentive for Healthcare Professionals.* Adv Exp Med Biol, 2021. **1324**: p. 1-10.
5. Vemula, D., et al., *CADD, AI and ML in drug discovery: A comprehensive review.* Eur J Pharm Sci, 2023. **181**: p. 106324.
6. Weisberg, E.M. and E.K. Fishman, *The future of radiology and radiologists: AI is pivotal but not the only change afoot.* J Med Imaging Radiat Sci, 2024.
7. Lee, J.H., et al., *Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study.* Korean J Radiol, 2022. **23**(5): p. 505-516.
8. Jeblick, K., et al., *ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports.* Eur Radiol, 2023.
9. Floridi, L. and M. Chiriatti, *GPT-3: Its Nature, Scope, Limits, and Consequences.* Minds and Machines, 2020. **30**(4): p. 681-694.
10. Hwang, T., et al., *Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts.* PLoS One, 2024. **19**(2): p. e0297701.
11. Dave, T., S.A. Athaluri, and S. Singh, *ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations.* Front Artif Intell, 2023. **6**: p. 1169595.
12. Kung, T.H., et al., *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.* PLOS Digit Health, 2023. **2**(2): p. e0000198.
13. Yaneva, V., et al., *Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment.* Acad Med, 2024. **99**(2): p. 192-197.
14. Ali, S.R., et al., *Using ChatGPT to write patient clinic letters.* Lancet Digit Health, 2023. **5**(4): p. e179-e181.
15. Waisberg, E., et al., *GPT-4 and Ophthalmology Operative Notes.* Ann Biomed Eng, 2023. **51**(11): p. 2353-2355.
16. Stephan, D., et al., *AI in Dental Radiology-Improving the Efficiency of Reporting With ChatGPT: Comparative Study.* J Med Internet Res, 2024. **26**: p. e60684.
17. Flesch, R., *A new readability yardstick.* Journal of Applied Psychology, 1948. **32**(3): p. 221-233.
18. Amstad, T., *Wie verständlich sind unsere Zeitungen?* 1978: Studenten-Schreib-Service.
19. Gajjar, A.A., et al., *Readability of cerebrovascular diseases online educational material from major cerebrovascular organizations.* J Neurointerv Surg, 2024.
20. Friedman, D.B. and L. Hoffman-Goetz, *A systematic review of readability and comprehension instruments used for print and web-based cancer information.* Health Educ Behav, 2006. **33**(3): p. 352-73.
21. Skrzypczak, T. and M. Mamak, *Assessing the Readability of Online Health Information for Colonoscopy - Analysis of Articles in 22 European Languages.* J Cancer Educ, 2023. **38**(6): p.

1865-1870.
22. Anderson, J., *Lix and Rix: Variations on a Little-known Readability Index*. Journal of Reading, 1983. **26**(6): p. 490-496.
23. Rajpurohit, S., et al., *Development and evaluation of patient information leaflet for liver cirrhosis patients*. Clinical Epidemiology and Global Health, 2023. **24**: p. 101436.
24. Health, U.S.D.o., O.o.D.P. Human Services, and P. Health, *National Action Plan to Improve Health Literacy*. 2010, Washington, DC: Author.
25. Nutbeam, D. and J.E. Lloyd, *Understanding and Responding to Health Literacy as a Social Determinant of Health*. Annu Rev Public Health, 2021. **42**: p. 159-173.
26. Armache, M., et al., *Readability of Patient Education Materials in Head and Neck Cancer: A Systematic Review*. JAMA Otolaryngol Head Neck Surg, 2024. **150**(8): p. 713-724.
27. Massie, P.L., S.A. Arshad, and E.D. Auyang, *Readability of American Society of Metabolic Surgery's Patient Information Publications*. J Surg Res, 2024. **293**: p. 727-732.
28. Kaundinya, T., S. El-Behaedi, and J.N. Choi, *Readability of Online Patient Education Materials for Graft-Versus-Host Disease*. J Cancer Educ, 2023. **38**(4): p. 1363-1366.
29. Thirunavukarasu, A.J., et al., *Large language models in medicine*. Nat Med, 2023. **29**(8): p. 1930-1940.
30. Cheng, S.W., et al., *The now and future of ChatGPT and GPT in psychiatry*. Psychiatry Clin Neurosci, 2023. **77**(11): p. 592-596.
31. Baker, H.P., et al., *ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial*. J Am Acad Orthop Surg, 2024. **32**(3): p. 123-129.
32. Ayers, J.W., 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): p. 589-596.
33. Walker, H.L., et al., *Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument*. J Med Internet Res, 2023. **25**: p. e47479.
34. Mika, A.P., et al., *Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty*. J Bone Joint Surg Am, 2023. **105**(19): p. 1519-1526.
35. Yeo, Y.H., et al., *Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma*. Clin Mol Hepatol, 2023. **29**(3): p. 721-732.
36. Nazary, F., Y. Deldjoo, and T. Di Noia. *ChatGPT-HealthPrompt. Harnessing the Power of XAI in Prompt-Based Healthcare Decision Support using ChatGPT*. 2024. Cham: Springer Nature Switzerland.
37. White, J., et al., *Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design*. arXiv preprint arXiv:2303.07839, 2023.
38. Hu, X., et al., *Opportunities and challenges of ChatGPT for design knowledge management*. arXiv preprint arXiv:2304.02796, 2023.
39. Kernberg, A., J.A. Gold, and V. Mohan, *Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study*. J Med Internet Res, 2024. **26**: p. e54419.
40. Nastasi, A.J., et al., *A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts*. Sci Rep, 2023. **13**(1): p. 17885.
41. Shah, Y.B., et al., *Comparison of ChatGPT and Traditional Patient Education Materials for Men's Health*. Urol Pract, 2024. **11**(1): p. 87-94.
42. Ji, Z., et al., *Survey of Hallucination in Natural Language Generation*. ACM Comput. Surv., 2023. **55**(12): p. Article 248.

43. Alkaissi, H. and S.I. McFarlane, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*. Cureus, 2023. **15**(2): p. e35179.
44. Huang, L., et al., *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. arXiv preprint arXiv:2311.05232, 2023.

Preprint
JMIR Publications