

Evaluating Quality and Understandability of Radiology Report Summaries Generated by ChatGPT: A Pilot Study

Alexis Sunshine, Grace H Honce, Andrew Callen, David Zander, Jody Tanabe,
Samantha P Petrucci, Chen-Tan Lin, Justin M Honce

Submitted to: JMIR Formative Research
on: April 16, 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 its 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 expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5
Supplementary Files..... 16
 Multimedia Appendixes 17
 Multimedia Appendix 1..... 17



Evaluating Quality and Understandability of Radiology Report Summaries Generated by ChatGPT: A Pilot Study

Alexis Sunshine¹ MD; Grace H Honce² MPH; Andrew Callen¹ MD; David Zander¹ MD; Jody Tanabe¹ MD; Samantha P Petrucci^{1*} MD; Chen-Tan Lin³ MD; Justin M Honce^{4*} MD

¹Department of Radiology University of Colorado Anschutz Medical Campus Aurora US

²Hartway Evaluation Group Denver US

³University of Colorado Health Aurora US

⁴University of Colorado Anschutz Medical Campus Aurora US

*these authors contributed equally

Corresponding Author:

Justin M Honce MD

University of Colorado Anschutz Medical Campus
12700 E. 19th Ave. Mail Stop C278
Aurora
US

Abstract

Background: Radiology reports play a crucial role in conveying medical information to healthcare providers and patients. Unfortunately, they are difficult for patients to comprehend, causing confusion and anxiety, thereby limiting patient engagement in healthcare decisions. Large language models (LLMs) such as ChatGPT have the potential to create simplified, patient-friendly report summaries to increase accessibility. However, there's also a risk of errors being introduced during summarization.

Objective: The objectives of this study were to: (1) evaluate the accuracy and clarity of ChatGPT-generated summaries compared to original radiology reports as assessed by radiologists; (2) assess patients' understanding and satisfaction with the summaries compared to the original reports; and (3) compare the readability of the original reports and summaries using validated readability metrics.

Methods: We anonymized 30 radiology reports created by neuroradiologists at our institution (6 MR brain, 6 CT brain, 6 CTA head/neck, 6 CT neck, 6 CT spine). These anonymized reports were processed by ChatGPT to produce patient-centric summaries. 4 board-certified neuroradiologists evaluated the paired original reports and ChatGPT-generated summaries on quality and accuracy, and 4 patient volunteers evaluated the reports on perceived understandability and satisfaction. Readability was assessed via word count and validated readability scales.

Results: There was substantial improvement in patient confidence of understanding after reading the summary (98% vs. 26%), satisfaction regarding the level of jargon/terminology (91% vs. 8%), and satisfaction regarding the time taken to understand the content (97% vs. 23%). 92% of responses indicated the summary clarified questions patients had about the report, and 98% of responses indicated patients would utilize the summary if available, with 67% of responses indicating they would want to have access to both the report and summary, while 26% of responses indicated only wanting the summary.

83% of radiologist responses indicated the summary represented the original report "extremely" or "very" well, with only 5% of responses indicating it did so "slightly" or "not well at all." 5% of responses indicated there was missing relevant medical information in the summary, 12% reported instances of overemphasis of non-significant findings, and 18% reported instances of underemphasis of significant findings. No fabricated findings were identified. Overall, 83% of responses indicated that the summary would definitely/probably not lead patients to incorrect conclusions about the original report, with 10% of responses indicating the summaries may do so.

Conclusions: These findings demonstrate that ChatGPT-generated summaries can significantly improve perceived comprehension and satisfaction while accurately reflecting most key information from original radiology reports. Instances of minor omissions and under/over emphasis were noted in some summaries, underscoring the need for ongoing validation and oversight. Overall, these patient-centric summaries hold promise for enhancing patient-centered communication in radiology.

(JMIR Preprints 16/04/2025:76097)

DOI: <https://doi.org/10.2196/preprints.76097>

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 visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>

Original Manuscript

Evaluating Quality and Understandability of Radiology Report Summaries Generated by ChatGPT: A Pilot Study

BACKGROUND

Radiology reports play a crucial role in conveying medical information to healthcare providers and patients. Unfortunately, they are difficult for patients to comprehend, causing confusion and anxiety, thereby limiting patient engagement in healthcare decisions. Large language models (LLMs) such as ChatGPT have the potential to create simplified, patient-friendly report summaries to increase accessibility. However, there's also a risk of errors being introduced during summarization.

OBJECTIVE

The objectives of this study were to: (1) evaluate the accuracy and clarity of ChatGPT-generated summaries compared to original radiology reports as assessed by radiologists; (2) assess patients' understanding and satisfaction with the summaries compared to the original reports; and (3) compare the readability of the original reports and summaries using validated readability metrics.

METHODS

We anonymized 30 radiology reports created by neuroradiologists at our institution (6 MR brain, 6 CT brain, 6 CTA head/neck, 6 CT neck, 6 CT spine). These anonymized reports were processed by ChatGPT to produce patient-centric summaries. 4 board-certified neuroradiologists evaluated the paired original reports and ChatGPT-generated summaries on quality and accuracy, and 4 patient volunteers evaluated the reports on perceived understandability and satisfaction. Readability was assessed via word count and validated readability scales.

RESULTS

There was substantial improvement in patient confidence of understanding after reading the summary (98% vs. 26%), satisfaction regarding the level of jargon/terminology (91% vs. 8%), and satisfaction regarding the time taken to understand the content (97% vs. 23%). 92% of responses indicated the summary clarified questions patients had about the report, and 98% of responses indicated patients would utilize the summary if available, with 67% of responses indicating they would want to have access to both the report and summary, while 26% of responses indicated only wanting the summary.

83% of radiologist responses indicated the summary represented the original report "extremely" or "very" well, with only 5% of responses indicating it did so "slightly" or "not well at all." 5% of responses indicated there was missing relevant medical information in the summary, 12% reported instances of overemphasis of non-significant findings, and 18% reported instances of underemphasis of significant findings. No fabricated findings were identified. Overall, 83% of responses indicated that the summary would definitely/probably not lead patients to incorrect conclusions about the original report, with 10% of responses indicating the summaries may do so.

CONCLUSIONS

These findings demonstrate that ChatGPT-generated summaries can significantly improve perceived comprehension and satisfaction while accurately reflecting most key information from original radiology reports. Instances of minor omissions and under/over emphasis were noted in some summaries, underscoring the need for ongoing validation and oversight. Overall, these patient-centric summaries hold promise for enhancing patient-centered communication in radiology.

INTRODUCTION

The 21st Century Cures Act increased patients' access to their medical information.^{1,2} This change reflected a growing desire among patients to engage with their care—a desire that is exaggerated in the field of radiology, where patients want access to reports and images, and access these materials at higher rates than general patient portals or clinical notes.³⁻⁶ Unfortunately, radiology reports, which are primarily written for referring clinicians, use complex medical terminology that is well above the established guidelines for patient friendly material.^{1-4,7-9} Limited comprehension of these reports can leave patients feeling disempowered, anxious, and uncertain about their diagnoses and treatment plans.⁴ This is particularly worrisome given that patient engagement with and comprehension of their healthcare information correlates with improved patient experience, improved outcomes, and decreased costs.^{1,3,8,10-12} This evolving medical

landscape necessitates urgent adaptations within radiology in order to provide easily understandable, patient-friendly reports.¹²

To date, there have been several attempts at improving readability of radiology reports. For example University of Pennsylvania's PORTER platform is a structured reporting system, which provides automated lay-language translation of reports, and UC San Diego has implemented a CHV automated translation, and embedded hyperlinks for defining key terms.^{4,13,14} However, there is an ongoing need to find ways to simplify language and avoid jargon in order to facilitate clear patient communication.^{4,8,9,9,15,16} Radiologists face increasing workloads, making the manual preparation of simplified reports for every patient impractical.¹⁷ Recent improvements in artificial intelligence, in particular, large language models (LLMs) such as ChatGPT, have emerged as a promising solution for generating accessible and context-appropriate textual summaries.² These models show promise at efficiently processing complex medical information and producing coherent summaries that minimize technical jargon while preserving essential clinical content.^{18,19}

The ability of LLMs to generate patient-centric summaries may help address critical barriers to patient engagement—specifically, low health literacy levels and limited healthcare professional time for conversational patient education.^{8,12} Automating this process with an LLM offers a potential solution that could be integrated into clinical workflows while preserving radiologists' time for critical interpretative tasks. However, important concerns persist regarding the completeness and accuracy of LLM-generated outputs, biases in the underlying training data, and the consequences of generating misleading clinical information. Therefore, effective deployment of such technology requires rigorous assessment of and balance between the quality and simplicity of these summaries, as well as their actual benefit to patients.²⁰

This study evaluates ChatGPT's capacity to produce patient-centered radiology report summaries through three primary aims: (1) assessing the technical accuracy and completeness of the summaries by a radiologist, (2) measuring improvements in patient comprehension and satisfaction, and (3) quantifying changes in readability using validated metrics. Through this evaluation, we aimed to provide insights into how LLM-driven text generation could advance patient-centered radiology while maintaining clinical accuracy, ultimately contributing to broader discussions about the role of AI in healthcare communication.

METHODS

This study evaluated the quality and understandability of ChatGPT-generated radiology report summaries through both quantitative and qualitative analyses. The study protocol was approved by our institutional review board (COMIRB), and all participants provided informed consent.

Radiology Report Collection and Anonymization:

The electronic medical record was queried for radiology reports performed at our institution in September 2023. A total of 30 radiology reports were selected for analysis (6 MR Brain, 6 CT Brain, 6 CTA Head/Neck, 6 CT Neck, and 6 CT Lumbar Spine) to reflect a broad range of imaging types and pathologies encountered in our neuroradiology section. Each radiology report was anonymized to remove all identifying information, including dates of comparison studies, if referenced.

Patient Centric Summary Creation:

Each radiology report was processed by ChatGPT 4 within a new Chat dialogue in October 2023 to ensure consistent performance of the model across each report. No custom instructions or memory features were activated, thus the prompt template for each report conversion followed a zero-shot prompt format, which is included in the Online Supplemental Appendix.

Recruitment:

Four patients were recruited to complete survey questions about the reports and summaries from our health system's Patient and Family Advisory Council. Four board-certified radiologists were recruited to complete survey questions on the reports and summaries from a single institution.

Survey Design:

The patient survey was designed to evaluate patient's understanding, satisfaction, and perceived utility of the radiology reports and summaries. The radiologist survey was designed to assess the technical accuracy, clinical completeness, clarity, and potential errors that may have been introduced into the summary. Both surveys included 5-point Likert-scaled questions and open-ended responses. Complete survey instruments are provided in the Multimedia Appendix.

Word Count and Readability Analysis:

Textual characteristics of the original radiology reports and patient-centric summaries were analyzed via word count and validated readability metrics. These included the Flesch-Kincaid Grade Level, Flesch Reading Ease Score, Gunning-Fog Index, and SMOG Index and provide estimates of complexity and accessibility of the reports and summaries.²¹⁻²³

Statistical Analysis:

Differences in word count and readability scores between the reports and summaries were calculated with paired T-tests. For the survey questions, closed-ended responses were rated on Likert Scales, with scores reported as percentages, and were reflected as Top 2 (or Bottom 2) Box scores as appropriate, aggregating from the Likert scale. Each patient and radiologist response for each report-summary pair were considered individual data points.

Data from the open-ended questions were analyzed using a multifaceted qualitative approach. To capture a comprehensive understanding of patient and provider perspectives, the method of analysis was based on the structure of each question. During inductive analysis open-ended responses were coded manually, sorting responses by found themes. To ensure that information was not lost, or assumptions about implied importance were not lost, each response could be coded to a variety of themes.

To determine how well patients understood the original radiology report and summary, their responses to the question “What do you believe to be the most important / main point that the radiologist was trying to convey?” were directly compared for congruence to answers supplied by the radiologist for each report/summary pair. Comparisons were then categorized as congruent or incongruent with the radiologist interpretation, which was considered the gold standard.

Additionally, patient responses after reading the summary were compared to their answers after reading only the report. If the answers from the summary matched the answer from the report, responses were considered congruent (i.e. the summary did not change the patient’s subjective understanding). If the answers did not match, responses were considered as incongruent (i.e. the summary did change the patient’s subjective understanding), with an additional positive or negative change modifier. A positive change modifier was assigned when responses included additional details or demonstrated closer alignment with the radiologist responses. Conversely, a negative change modifier was assigned when responses contained fewer details or diverged further from the radiologist responses.

RESULTS:

Readability:

Patient-centric summaries demonstrated significant improvements in readability compared to original radiology reports for all measured metrics (Table 1). The summaries achieved a mean Flesch-Kincaid Grade Level of 7.2, representing middle school reading level, compared to 12.5 for original reports ($p < 0.001$). Similar improvements were seen in Flesch Reading Ease Score, Gunning-Fog Index, and SMOG Index, consistently indicating enhanced accessibility of the summary content.

Table 1 – Comparison of Readability Metrics Between Original Reports and Summaries

Metric	Original Reports (Mean +/-SD)	Summaries (Mean +/- SD)	p-value
Word count	241.0 ± 130.5	189.0 ± 34.7	p=0.018
Flesch-Kincaid Grade Level	12.5 ± 1.63	7.2 ± 1.14	p<0.0001
Flesch-Kincaid Reading Ease Score	21.5 ± 11.39	69.8 ± 6.2	p<0.0001
Gunning-Fog Index	16.4 ± 2.38	10.3 ± 1.41	p<0.0001
Smog Index	12.7 ± 1.55	10.4 ± 1.05	p<0.0001

Radiologist Survey:

Four board-certified neuroradiologists completed surveys for each of the thirty radiology report/summary pairs, resulting in a total of 120 responses for each question. The neuroradiologists had a mean 11.6 (3.5-25) years of experience in neuroradiology after completing fellowship.

Radiologist assessment of summaries was generally favorable (Table 2). In 83% (100/120) of responses, radiologists indicated the summaries represented the original reports extremely/very well and 90% (108/120) were satisfied with the balance of medical jargon to patient-centric language. Regarding accuracy, 5% (6/120) of responses indicated missing information, 12% (14/120) indicated overemphasis of findings, and 18% (22/120) indicated underemphasis of findings.

Notably, no hallucinations or fabricated findings were identified in any summary. When assessing for the potential for patient misunderstanding their condition due to the summary, 82% (99/120) of radiologist responses indicated that summaries would likely not lead to incorrect conclusions, while 10% (12/120) indicated a perceived risk of misinterpretation.

Table 2 – Radiologist Assessment of Quality and Accuracy of Summaries

Measure	Responses n (%)
Representation of Original Findings	n=120
- Extremely Well / Very Well	100 (83%)
- Neutral	14 (12%)
- Slightly Well / Not Well at All	6 (5%)
Satisfaction with Balance of Medical Jargon and Patient-Centric Language	n=120
- Very Satisfied / Somewhat Satisfied	108 (90%)
- Neutral	5 (4%)
- Somewhat Dissatisfied / Very Dissatisfied	7 (6%)
Error Analysis	n=120
- Missing Relevant Information	6 (5%)
- Overemphasis of Findings	14 (12%)
- Underemphasis of Findings	22 (18%)
Risk of Leading Patients to Incorrect Conclusions	n=120
- Definitely Yes / Probably Yes	12 (10%)
- Neutral	9 (8%)
- Probably Not / Definitely Not	99 (82%)

In their free text responses, neuroradiologists indicated several issues with the summaries. They noted overemphasis of minor findings (e.g. mild sinus disease, white matter changes, renal cysts), and perceived unnecessary follow-up recommendations. Radiologists also indicated disagreement on specific terminology choices in some cases (e.g. “mildly enlarged” to describe a “stable to incrementally” enlarged collection). The radiologists also noted that summaries underemphasized some critical findings—arterial occlusions, perfusion deficits, nerve root compression—by not adequately conveying their clinical significance. Additionally, radiologists expressed concerns that oversimplified vocabulary diminished the importance of certain findings, such as describing a vertebral artery as “a smaller artery in your neck” or referring to a normalized brain volume at the 6th percentile as “a bit below average”.

Qualitative analysis of radiologist feedback revealed several themes regarding desired improvements to the summaries. Fifty four percent of free-text responses indicated no improvements were needed. In 38%, radiologists expressed discomfort with perceived oversimplification of language (e.g. “lump” instead of nodule, “thinning” instead of osteopenia, “squishing” instead of height loss), and generally found the summaries could have been more concise. In 4% of responses, radiologist indicated a preference for more detail including more clearly defined terminology and more explicit localization of relevant findings (e.g. the level of spinal stenoses).

Patient Survey:

Four patient volunteers completed surveys for each of the 30 report-summary pairs, resulting in a total of 120 responses across all 30 reports and summaries for each question. A total of 118 responses were valid, with the remaining two responses left blank. Patient’s reported age ranged between 55-84 : [1 (55-64), 2 (65-74), 1 (75-84)]. Patient’s reported educational attainment included high school graduate (1), college degree (2), and master’s degree (1).

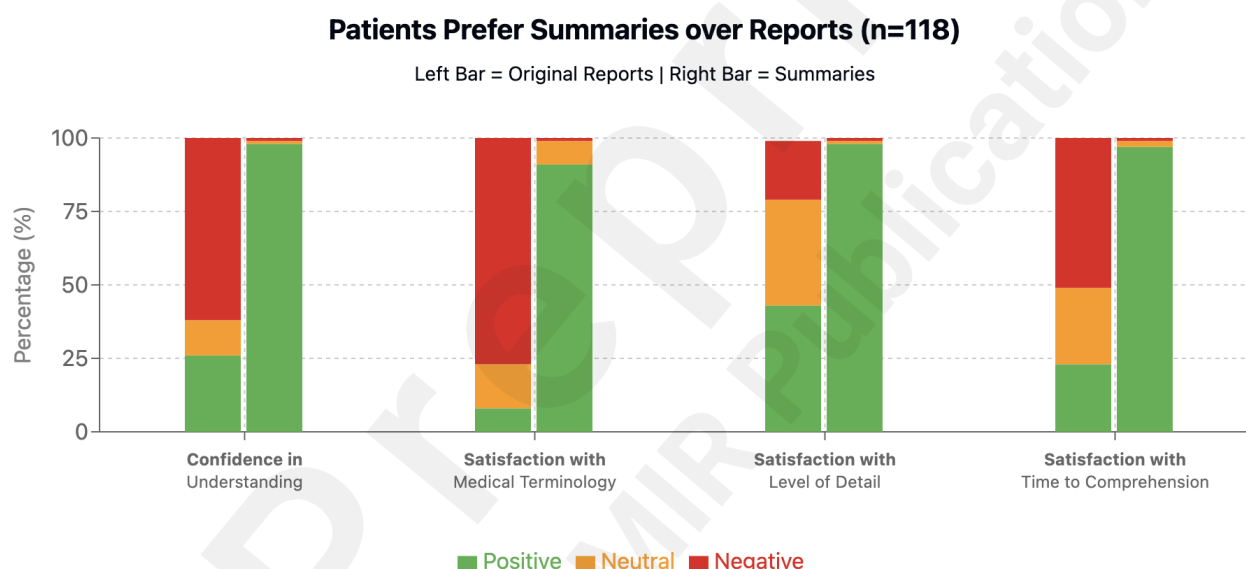
Patient survey responses indicated substantially improved subjective comprehension and satisfaction with summaries compared to the original reports (Table 3). While only 26% (31/118) of responses indicated confidence in understanding the original reports, this increased to 98% (116/118) of responses for summaries. Satisfaction with medical terminology showed similar improvement, rising from 8% (9/118) for original reports to 91% (107/118) for summaries. Time required for comprehension was a significant concern with original reports, with only 23% (27/118) satisfied with the time needed for understanding, compared to 97% (115/118) satisfaction with summaries.

Table 3 – Patient Survey Responses: Comparison of Reports and Summaries

Measure	Original Report	Summaries n (%)
---------	-----------------	--------------------

	n (%)	
Confidence in Understanding	n=118	n=118
- Very Confident / Somewhat Confident	31 (26%)	116 (98%)
- Neutral	14 (12%)	1 (1%)
- Very Unconfident / Somewhat Unconfident	73 (62%)	1 (1%)
Satisfaction with Medical Jargon / Terminology	n=118	n=118
- Very Satisfied / Somewhat Satisfied	9 (8%)	107 (91%)
- Neutral	18 (15%)	9 (8%)
- Very Dissatisfied / Somewhat Dissatisfied	91 (77%)	2 (2%)
Satisfaction with Level of Detail	n=118	n=118
- Very Satisfied / Somewhat Satisfied	51 (43%)	116 (98%)
- Neutral	43 (36%)	1 (1%)
- Very Dissatisfied / Somewhat Dissatisfied	24 (20%)	1 (1%)
Satisfaction with Time to Comprehension	n=118	n=118
- Very Satisfied / Somewhat Satisfied	27 (23%)	115 (97%)
- Neutral	31 (26%)	2 (2%)
- Very Dissatisfied / Somewhat Dissatisfied	60 (51%)	1 (1%)

Figure 1 (alternative presentation style)



The summaries demonstrated high utility, with 92% (108/118) of responses indicating they clarified questions about the original report, and improved patient's understanding of clinical implications. When asked about format preferences, 67% (79/118) of responses indicated a preference to having access to both the summary and original report, while 26% (31/118) of responses indicated they would prefer to have the summary alone. Only 1% preferred the original report exclusively, with the remaining 6% indicating a preference for alternative formats.

Qualitative analysis of patient feedback revealed several themes regarding desired improvements to the summaries. Of the free-text responses, 49.5% indicated no improvements were needed. The remaining responses suggested enhancing professional tone (13%), providing more specific follow-up recommendations (14%), and including additional detail about medical terminology and physiologic processes relevant to the findings (24%).

Effect of the Summary on Patient Understanding:

Patient understanding was assessed by comparing the patient's identification of the main point of the report/summary against the radiologists' response regarding the main point of the report. When reading the original radiologist report, 60% (71/119) of patient responses aligned with radiologist interpretations, while 40% (48/119) were incongruent. Patient understanding improved with the summary, with 74% (88/119) of responses aligning with radiologist interpretations, while 27% (32/119) remained incongruent.

When comparing patient responses regarding the main point the radiologist was trying to convey after reading only the report with their response after reading the summary, answers did not change in 52% (62/119) responses (i.e. responses were congruent). In 29% (34/119), the responses after reading the summary were different (i.e. the summary changed a patient's understanding). Specifically in 23% (27/119), responses after reading the summary contained additional relevant details or demonstrated closer alignment with radiologist responses, suggesting an increase in understanding as a result of the summary. In 6% (7/119), responses after reading the summary contained fewer details or diverged further from radiologist responses, suggesting decreased understanding as a result of the summary.

DISCUSSION:

This study demonstrates that large language models can successfully generate radiology reports which are more understandable to patients while maintaining clinical accuracy. The AI-generated summaries significantly improved readability metrics across all measured scales, reducing the required reading level from college (grade 12.5) to middle school (grade 7.2) while substantially increasing reading ease scores.^{24,25} This improvement in accessibility was achieved while maintaining high accuracy, with radiologists confirming that 83% of summaries represented the original findings "extremely" or "very" well. Similar studies have shown that such summaries maintain high accuracy.²⁶

A unique strength of this study is the inclusion of patients in evaluation of reports and AI-generated summaries, as the vast majority of research to date focuses on evaluation by radiologist and non-radiologist physicians.^{2,27-29} Patient response data revealed a striking contrast between comprehension of traditional reports and AI-generated summaries. While only 26% of patients reported confidence in understanding the original reports, 98% reported confidence in understanding the summaries. Additionally, the high satisfaction rates with the summaries' level of detail (98%) and medical terminology (91%) indicate that the AI successfully struck a balance between accessibility and informativeness.

The dramatic improvement in patient confidence was accompanied by objective improvements in understanding, with congruence between patient and radiologist interpretations increasing from 60% on the original reports to 74% after reading the summaries. Additionally, 23% of responses after reading the summary contained additional relevant details or demonstrated closer alignment with radiologist responses, consistent with an increased understanding as a result of the summary. This suggests that AI-generated summaries not only increase patient confidence but also lead to more accurate understanding of radiological findings. This aligns with prior evidence that patient-friendly summaries can enhance patient understanding, for example, non-physician readers in one study had comprehension scores of 4.69/6 for AI generated summaries versus 2.71/5 for original reports.²⁶

Particularly noteworthy was the strong patient preference for having access to both the original report and the summary (67% of responses), with 98% of responses indicating they would utilize the summary if available. This suggests that patients value the combination of detailed technical information with accessible explanations, rather than seeing the summary as a replacement for the full report. These findings align with the shifting preference among the general public towards greater transparency in healthcare.^{3,4,7,8}

Importantly, no frank hallucinations were observed in the summaries, which is crucial for clinical implementation. However, a small percentage (10%) of radiologist responses indicated that summaries may be misleading. Radiologists noted instances of omission and both over- and under-emphasis of findings (12% and 18% of responses, respectively), particularly regarding incidental findings and critical abnormalities. The concern about misleading summaries was corroborated by the 6% of patient responses that contained fewer details or diverged further from radiologist responses after reading the summary. Misleading information has the potential to lead to a variety of negative consequences for the patient, further emphasizing the need for comprehensive standards surrounding LLM-generated resources.²⁰

Additionally, in some instances, language that was deemed too casual by radiologists (e.g., "squished" instead of "compressed"), highlighting the need for more sophisticated prompt engineering to maintain professional medical communication standards while improving accessibility. An area of improvement identified by the patient-reviewers was highlighted by the 14% of patient responses indicating a desire for more explicit guidance on next steps and implications of findings.

This study has several limitations. Principally, its focus on neuroradiology reports from a single institution potentially limit the generalizability of our findings to other radiologic subspecialties and other institutional reporting patterns. Additionally, the study included a relatively small number of patient and radiologist reviewers and thus we hope that this work encourages other investigators to replicate this study with larger cohorts of reviewers to assess its applicability in larger and more diverse patient populations. Patient volunteers had a relatively high education level, though we anticipate that if anything, this biased the results towards more understanding of the original report, not less. Finally, volunteers were not reading their own reports, which removes the emotional salience, decidedly altering the experience.

Future Research:

AI-generated patient-friendly summaries appear to represent a promising tool for improving patient understanding of radiology reports. The summaries had high accuracy rates and dramatically improved readability and patient comprehension while maintaining clinical precision. The perceived benefit to patients in conjunction with the objective increase in comprehension suggest potential benefit in terms of patient experience and outcomes, though further investigation is needed using patient's own scans and reports (which we are pursuing).^{1,3,8,10-12} Integration of these summaries into existing radiology workflows and patient portals should be studied, with particular attention on the ultimate impact on radiologist workload, patient experience, and patient outcomes. Conscientious integration of this technology is required given the extraction of sensitive patient data and associated privacy concerns.³⁰

Future research should explore options to continue refining clinical accuracy of these reports. Though LLMs lack the ability to independently reason, there are several ways to optimize their output including optimization of prompt engineering, incorporating structured templates for critical findings, and using standardized language for common pathologies.^{2,28,29} Additionally, giving the LLM more clinical context (e.g. recent note from an ordering provider, problem list, patient class (IP/OP/ED)), may improve the quality of the summary and recommendations provided.^{29,31}

REFERENCES

1. Amin K, Khosla P, Doshi R, Chheang S, Forman HP. Artificial Intelligence to Improve Patient Understanding of Radiology Reports. *Yale J Biol Med.* 2023;96(3):407-417. doi:10.59249/NKOY5498
2. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. *Radiology.* 2024;310(3):e231593. doi:10.1148/radiol.231593
3. Alarifi M, Patrick T, Jabour A, Wu M, Luo J. Full Radiology Report through Patient Web Portal: A Literature Review. *Int J Environ Res Public Health.* 2020;17(10):3673. doi:10.3390/ijerph17103673
4. Lee CI, Langlotz CP, Elmore JG. Implications of Direct Patient Online Access to Radiology Reports Through Patient Web Portals. *J Am Coll Radiol JACR.* 2016;13(12 Pt B):1608-1614. doi:10.1016/j.jacr.2016.09.007
5. Halaska C, Sachs P, Sanfilippo K, Lin CT. Patient Attitudes About Viewing Their Radiology Images Online: Preintervention Survey. *J Med Internet Res.* 2019;21(7):e12595. doi:10.2196/12595
6. Kemp J, Short R, Bryant S, Sample L, Befera N. Patient-Friendly Radiology Reporting-Implementation and Outcomes. *J Am Coll Radiol JACR.* 2022;19(2 Pt B):377-383. doi:10.1016/j.jacr.2021.10.008
7. Farmer CI, Bourne AM, O'Connor D, Jarvik JG, Buchbinder R. Enhancing clinician and patient understanding of radiology reports: a scoping review of international guidelines. *Insights Imaging.* 2020;11(1):62. doi:10.1186/s13244-020-00864-9
8. Hansberry DR, Donovan AL, Prabhu AV, Agarwal N, Cox M, Flanders AE. Enhancing the Radiologist-Patient Relationship through Improved Communication: A Quantitative Readability Analysis in Spine Radiology. *AJNR Am J Neuroradiol.* 2017;38(6):1252-1256. doi:10.3174/ajnr.A5151
9. Yi PH, Golden SK, Harringa JB, Kliewer MA. Readability of Lumbar Spine MRI Reports: Will Patients Understand? *AJR Am J Roentgenol.* 2019;212(3):602-606. doi:10.2214/AJR.18.20197
10. Johnson AM, Brimhall AS, Johnson ET, et al. A systematic review of the effectiveness of patient education through patient portals. *JAMIA Open.* 2023;6(1):ooac085. doi:10.1093/jamiaopen/ooac085
11. Alturkistani A, Qavi A, Anyanwu PE, Greenfield G, Greaves F, Costelloe C. Patient Portal Functionalities and Patient Outcomes Among Patients With Diabetes: Systematic Review. *J Med Internet Res.* 2020;22(9):e18976. doi:10.2196/18976
12. Bange M, Huh E, Novin SA, Hui FK, Yi PH. Readability of Patient Education Materials From RadiologyInfo.org: Has There Been Progress Over the Past 5 Years? *AJR Am J Roentgenol.* 2019;213(4):875-879. doi:10.2214/AJR.18.21047
13. Qenam B, Kim TY, Carroll MJ, Hogarth M. Text Simplification Using Consumer Health Vocabulary to Generate Patient-Centered Radiology Reporting: Translation and Evaluation. *J Med Internet Res.* 2017;19(12):e417. doi:10.2196/jmir.8536
14. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. *Radiology.* 2011;260(1):174-181. doi:10.1148/radiol.11101913

15. Norris EC, Halaska C, Sachs PB, Lin CT, Sanfilippo K, Honce JM. Understanding Patient Experiences, Opinions, and Actions Taken After Viewing Their Own Radiology Images Online: Web-Based Survey. *JMIR Form Res*. 2022;6(4):e29496. doi:10.2196/29496
16. McEnteggart GE, Naeem M, Skierkowski D, Baird GL, Ahn SH, Soares G. Readability of Online Patient Education Materials Related to IR. *J Vasc Interv Radiol JVIR*. 2015;26(8):1164-1168. doi:10.1016/j.jvir.2015.03.019
17. Recht MP, Donoso-Bach L, Brkljačić B, Chandarana H, Jankharia B, Mahoney MC. Patient-centered radiology: a roadmap for outpatient imaging. *Eur Radiol*. 2024;34(7):4331-4340. doi:10.1007/s00330-023-10370-3
18. Liebreinz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Health*. Published online February 6, 2023:S2589-7500(23)00019-5. doi:10.1016/S2589-7500(23)00019-5
19. Yeo YH, Samaan JS, Ng WH, et al. *Assessing the Performance of ChatGPT in Answering Questions Regarding Cirrhosis and Hepatocellular Carcinoma*. *Gastroenterology*; 2023. doi:10.1101/2023.02.06.23285449
20. Goodman KE, Yi PH, Morgan DJ. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA*. 2024;331(8):637-638. doi:10.1001/jama.2024.0555
21. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Res Soc Adm Pharm RSAP*. 2013;9(5):503-516. doi:10.1016/j.sapharm.2012.05.009
22. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32(3):221-233. doi:10.1037/h0057532
23. Sabharwal S, Badarudeen S, Unes Kunju S. Readability of online patient education materials from the AAOS web site. *Clin Orthop*. 2008;466(5):1245-1250. doi:10.1007/s11999-008-0193-8
24. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the Use of ChatGPT to Accurately Simplify Patient-centered Information about Breast Cancer Prevention and Screening. *Radiol Imaging Cancer*. 2024;6(2):e230086. doi:10.1148/rycan.230086
25. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit Health*. 2023;9:20552076231221620. doi:10.1177/20552076231221620
26. Park J, Oh K, Han K, Lee YH. Patient-centered radiology reports with generative artificial intelligence: adding value to radiology reporting. *Sci Rep*. 2024;14(1):13218. doi:10.1038/s41598-024-63824-z
27. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. *Skeletal Radiol*. 2024;53(8):1621-1624. doi:10.1007/s00256-024-04599-2
28. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. 2024;34(5):2817-2825. doi:10.1007/s00330-023-10213-1
29. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art*. 2023;6(1):9. doi:10.1186/s42492-023-00136-5
30. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health*. Published

online February 6, 2023:S2589-7500(23)00021-3. doi:10.1016/S2589-7500(23)00021-3

31. Campbell DJ, Estephan LE. ChatGPT for patient education: an evolving investigationResponse to Kleebayoon A, Wiwanitkit V. ChatGPT, obstructive sleep apnea and patient education. *J Clin Sleep Med* . 2023;19(12):2133. doi:10.5664/jcsm.10768. *J Clin Sleep Med*. 2023;19(12):2135-2136. doi:10.5664/jcsm.10808



Supplementary Files

Multimedia Appendixes

Untitled.

URL: <http://asset.jmir.pub/assets/ec50eb324161fe025638cb55bfae27ac.docx>