HHS Public Access

Author manuscript

Mayo Clin Proc Digit Health. Author manuscript; available in PMC 2025 March 24.

Published in final edited form as:

Mayo Clin Proc Digit Health. 2025 March ; 3(1): . doi:10.1016/j.mcpdig.2025.100198.

Retrospective Comparative Analysis of Prostate Cancer In-Basket Messages: Responses From Closed-Domain Large Language Models Versus Clinical Teams

Yuexing Hao, MS,
Jason Holmes, PhD,
Jared Hobson, MD,
Alexandra Bennett, MD,
Elizabeth L. McKone, MD,
Daniel K. Ebner, MD,
David M. Routman, MD,
Satomi Shiraishi, MD,
Samir H. Patel, MD,
Nathan Y. Yu, MD,
Chris L. Hallemeier, MD,
Brooke E. Ball, MSN,
Mark Waddle, MD,
Wei Liu, PhD

Department of Radiation Oncology, Mayo Clinic, Phoenix, AZ (Y.H., J.H., S.H.P., N.Y.Y., W.L.); Cornell University, Ithaca, NY (Y.H.); Department of Electric Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA (Y.H.); and Department of Radiation Oncology, Mayo Clinic, Rochester, MN (A.B., E.L.M., D.K.E., D.M.R., S.S., C.L.H., B.E.B., M.W.).

Abstract

Objective: To evaluate the effectiveness of RadOnc-generative pretrained transformer (GPT), a GPT-4 based large language model, in assisting with in-basket message response generation for prostate cancer treatment, with the goal of reducing the workload and time on clinical care teams while maintaining response quality.

Patients and Methods: RadOnc-GPT was integrated with electronic health records from both Mayo Clinic-wide databases and a radiation-oncology-specific database. The model was evaluated

Correspondence: Address to Mark Waddle, MD, Department of Radiation Oncology, Mayo Clinic, 200 1st St SW, Rochester, MN 55905 (Waddle.Mark@mayo.edu); or Wei Liu, PhD, Department of Radiation Oncology, Mayo Clinic, 5777 E. Mayo Blvd, Phoenix, AZ, 85054 (liu.wei@mayo.edu; Twitter: @YuexingHao).

POTENTIAL COMPETING INTERESTS

The authors report no competing interests.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data

on 158 previously recorded in-basket message interactions, selected from 90 patients with nonmetastatic prostate cancer from the Mayo Clinic Department of Radiation Oncology in-basket message database in the calendar years 2022–2024. Quantitative natural language processing analysis and 2 grading studies, conducted by 5 clinicians and 4 nurses, were used to assess RadOnc-GPT's responses. Three primary clinicians independently graded all messages, whereas a fourth senior clinician reviewed 41 responses with relevant discrepancies, and a fifth senior clinician evaluated 2 additional responses. The grading focused on 5 key areas: completeness, correctness, clarity, empathy, and editing time. The grading study was performed from July 20, 2024 to December 15, 2024.

Results: The RadOnc-GPT slightly outperformed the clinical care team in empathy, whereas achieving comparable scores with the clinical care team in completeness, correctness, and clarity. Five clinician graders identified key limitations in RadOnc-GPT's responses, such as lack of context, insufficient domain-specific knowledge, inability to perform essential meta-tasks, and hallucination. It was estimated that RadOnc-GPT could save an average of 5.2 minutes per message for nurses and 2.4 minutes for clinicians, from reading the inquiry to sending the response.

Conclusion: RadOnc-GPT has the potential to considerably reduce the workload of clinical care teams by generating high-quality, timely responses for in-basket message interactions. This could lead to improved efficiency in health care workflows and reduced costs while maintaining or enhancing the quality of communication between patients and health care providers.Abbreviations and AcronymsAI; artificial intelligence; LLM; large language model; NLP; natural language processing; RadOnc-GPT; radiation oncology generative pretrained transformer

In-basket is the online portal messaging system integrated within Epic applications functioning similarly to email for communication between patients and their clinical care team. The in-basket messaging system is often used to exchange messages regarding patient concerns, appointments, and follow-up care, particularly when real-time communication is not possible. During or after treatment, patients may not always receive immediate support from their care team. Patients with limited clinical literacy and understanding still need to communicate with health care professionals for various needs, such as disease monitoring, medication, appointments, and billing or insurance issues. In this context, the in-basket serves as a vital tool to bridge the communication gap between patients and clinical professionals.[1]

However, clinical care teams struggle to draft responses on time due to the increasing complexity of patients' supportive care needs.[2] Several studies have shown that an increased workload from responding to in-basket messages can negatively impact clinicians' burnout rates and overall well-being.[2–5] Furthermore, patient messaging volumes increased by more than 50% after COVID-19, placing an undue burden on clinical teams.[6–8] Although these added avenues of communication are beneficial, generally responding to these messages is nonreimbursable as well.[9,10]

Because in-basket messages often contain important real-world concerns from patients, the text-based in-basket message dataset is valuable for reporting patient-centered interactions. We propose using large language models (LLMs), which are connected with the electronic

health record (EHR) system, to provide timely and layman-friendly responses to various categories of in-basket message inquiries.[11–16] The LLMs have already shown strong technical capabilities in clinical context learning, summarization, response generation, decision support, and Q&A.[17–25] Here, we aimed to evaluate the performance of LLM and clinical care teams in 3 key areas, as follows: (1) capturing and interpreting all sources of data; (2) generating personalized and prompt responses; and (3) upholding high clinical standards in terms of completeness, correctness, clarity, and empathy.

Rather than applying LLMs across all types of disease sites, we focused on patients with prostate cancer who received treatment at the Mayo Clinic's Radiation Oncology department. Specializing in a specific field allows the LLM to generate more accurate and relevant responses. We developed RadOnc-GPT, an OpenAI GPT-4o-powered LLM, which is integrated with EHR.[26] Because many in-basket messages require external context for proper understanding and interpretation, RadOnc-GPT can generate more personalized responses with greater details in a zero-shot without training. This approach helps save time for the clinical care team by reducing the need to consult multiple sources of information to draft a response.

METHOD

This retrospective study was approved by the institutional review board of the Mayo Clinic. The need for written informed consent was waived, because our study only involved previously collected data and did not require participant contact. Our study included patients with prostate cancer who were managed at Mayo Clinic in Rochester, Minnesota in the calendar years 2022–2024. RadOnc-GPT is a retrieval-augmented generation system that connects with both the hospital-wide electronic medical record database, Epic, developed by Epic Systems, and the radiation oncology-specific database, Aria, developed by the Varian Medical Systems. The data RadOnc-GPT may access include clinical notes, radiology notes, pathology notes, urology notes, radiology reports, radiation treatment details, diagnosis details, patient details (demographic characteristics), in-basket messages, and more. RadOnc-GPT is able to retrieve data by way of specifying the patient ID and which dataset to retrieve to the backend system. Once retrieved, the data are inserted into the conversation history.

Subject demographic characteristic information retrieved from the EHR system included sex, age, race, ethnicity, preferred language, and the attending physician's name. Information collected from Aria included demographic characteristic information (sex, age, race, ethnicity, preferred language, and the attending physician's name), prostate cancer treatment-specific information (course description, plan intent, treatment orientation, radiation type, radiation oncology machine type, number of fractions, dose prescription, dose delivered, radiation technique, and treatment duration), and diagnosis details (cancer stage, International Classification of Diseases diagnosis code and code type, and onset date). Information collected from Epic included clinical notes, ordered by date. For RadOnc-GPT, the information retrieval order starts with patient demographic characteristic details, followed by treatment details, diagnosis details, and finally, clinical notes.

To ensure every patient inquiry was consistent and under the same GPT generation environment, we developed a graphical user interface for RadOnc-GPT that was reinitialized for each test. This approach ensured that RadOnc-GPT did not generate biased responses from its memory of the previous patient's pair of inquiries and responses. Our study's evaluation was divided into 2 main components: natural language processing (NLP) quantitative assessments and clinical professional grading. The visualization of the pipeline is presented in Figure 1.

From the in-basket messages dataset, we extracted inquiries on patients with prostate cancer and their corresponding care team responses. RadOnc-GPT, integrated with patients' EHR profiles, generates responses to these inquiries. To facilitate evaluation, a randomized dataset was created, containing both RadOnc-GPT responses and clinical care team responses. This dataset is used for NLP-based quantitative evaluation and single-blinded grading, in which clinicians are not informed whether a response originates from RadOnc-GPT or a clinical professional. The responses are distributed randomly, ensuring they do not consistently appear together. Clinician and nurse graders can optionally review and match specific responses to patient EHRs using the patient ID.

For NLP evaluation, we performed 4 types of measurements[27,28]: natural language understanding, reasoning, context readability, and natural language generation.

For the grading study, we focused on the following 6 dimensions of evaluation[29–31]:

- completeness (ranging from 1–5, the higher the better),
- correctness (ranging from 1–5, the higher the better),
- clarity (ranging from 1–5, the higher the better),
- empathy (ranging from 1–5, the higher the better),
- estimated time to respond (in minutes),
- extensive editing required (no use, major editing, minor editing, and no editing needed),
- (optional) text comments section.

We enlisted 4 medical doctors from the Department of Radiation Oncology, all with considerable in-basket response experience, with a mean years of experience of 4.6. Of these, 3 medical residents (C1-C3) independently graded all 158 messages. A chief resident (C4) reviewed discrepancies when conflicts arose, and a board-certified radiation oncologist specializing in prostate cancer (C5) provided the final grading decision if disagreements persisted. Given that nurses typically initiate responses to in-basket messages, we also recruited 4 nurses from the same department to evaluate their capability (whether they can answer the questions or not) and estimate the time in minutes required to answer 158 patient inquiries (mean years of experience=5.25). The nurses provided anonymized estimates of the time in minutes spent responding to and redirecting these in-basket messages to other advanced practice providers or clinicians. (Table)

RESULTS

In-Basket Message Dataset

In-basket message interactions can often be disorganized. Without a standardized format for patient inquiries, under one subject, patients may send multiple messages for a single issue or combine several unrelated questions into one message. This makes it difficult to categorize the messages, as they frequently span multiple categories. In addition, a single thread may include several conversation pairs, in which a pair is defined as one or more patient inquiries followed by one or more care team responses in a time-sequenced manner. For inclusion in our evaluation dataset, each patient-clinician conversation pair must have consisted of 1 patient inquiry and 1 care team response.

We selected 90 patients with nonmetastatic prostate cancer from the Mayo Clinic Department of Radiation Oncology in-basket message database. After filtering patient inquiries that are not relevant to medical advice seeking or receiving no or unrelated care team replies, we finally selected 158 patient inquiries, with each of them containing a clinical care team's reply. We only selected the message type under the Epic category of patient medical advice request. We then pulled 158 patient inquiries' human care team's responses and utilized the patient inquiries to generate 158 RadOnc-GPT responses. We randomized the 316 responses and did not disclose the graders' response source.

We manually summarized the 158 patient inquiries into 9 main categories: test results, adverse effects, medication questions, radiation treatment questions, medical oncology questions, surgical oncology questions, care coordination/logistics, laboratory/radiology/pathology reports, and care journey questions (Figure 2). The 3 most common patient inquiries are adverse effects, medication questions, and radiation treatment questions.

NLP Analysis

To understand the sentiment differences between the human care team and the RadOnc-GPT, we conducted TextBlob and Valence Aware Dictionary and sEntiment Reasoner (VADER) analysis. In the TextBlob sentiment distribution (Figure 3A), RadOnc-GPT responses are observed to skew toward a more positive sentiment, with most responses clustering around a sentiment score of 0.25. In contrast, human care team responses present a more evenly distributed sentiment profile, with a considerable concentration around the neutral score of 0 (grey line in Figure 3A). RadOnc-GPT responses tend to be more positive, whereas care team responses consist of a broader spectrum of sentiments, including neutral and negative tones. The VADER sentiment distribution (Figure 3B) provides further insight into these differences. The box plot reveals that RadOnc-GPT responses exhibit a high median sentiment score, nearing 1.0, indicative of a predominantly positive sentiment. However, there are notable outliers reflecting occasional negative sentiment. Clinical care team responses, by comparison, display a wider range of sentiment scores, with a lower median, indicating a more varied and contextually nuanced sentiment expression. Our sentiment analysis collectively suggests that although RadOnc-GPT responses are generally more positive, care team responses offer a more balanced sentiment distribution, reflecting a greater sensitivity to the contextual nuances of the input data.

To understand how the human care team and RadOnc-GPT responses' inferences with the patients' inquiries, we performed an natural language inference analysis.[32] RadOnc-GPT responses were predominantly neutral, with 92.41% of responses in this category, suggesting a tendency toward generalized statements. In contrast, clinician responses were more varied, with 70.25% neutral and 29.11% entailment, indicating greater relevance and specificity. Both response types found low contradiction rates, though RadOnc-GPT responses had a slightly higher rate at 3.16%, which may point to occasional inconsistencies. The NLI label distribution comparison is shown in Figure 3C.

Comparing the semantic similarity[33] between RadOnc-GPT and human care team responses provided additional context, showing a mean similarity score of 0.85 between RadOnc-GPT and human care team responses. This high score indicated a strong alignment in content, although RadOnc-GPT responses are generally more neutral. The findings suggested that although RadOnc-GPT responses may lack the specificity found in human care team responses, they still captured the core semantic contents, reflecting contextually relevant information. Figure 3D shows the distribution of the semantic similarity scores.

We also compared the relationship between word counts in RadOnc-GPT and human care team responses and their corresponding total mean scores across 4 categories. The results are shown in Supplemental Figure 1 (available online at https://www.mcpdigitalhealth.org/).

Clinician Grading Study.—In the single-blinded grader study, 3 clinician graders first graded all 316 responses (158 human care team responses and 158 RadOnc-GPT responses). The average of 3 graders' results found that GPT consistently performs better in empathy, whereas human responses show higher averages in completeness, correctness, and clarity. The grading rubrics are displayed in Supplementary Figure 2 (available online at https://www.mcpdigitalhealth.org/) and the sample patient inquiry and responses from clinical care team and RadOnc-GPT are in Supplemental Table 1 (available online at https://www.mcpdigitalhealth.org/).

We calculated a total score by summing the scores across the 4 categories (completeness, correctness, clarity, and empathy), each rated on a 0–5 scale, resulting in a combined score ranging 0–20. This total score was used to evaluate performance across 9 categories. The mean scores for the clinical care team responses ranged 16.00–19.25, with the highest score observed in surgical oncology questions (19.25). In comparison, RadOnc-GPT scores ranged 16.72–19.21, with the highest score recorded in test results (19.21). Notably, RadOnc-GPT outperformed the clinical care team in test results, care coordination/logistics, and care journey questions, as illustrated in Figure 4 (right).

Detailed statistics are provided in Supplemental Tables 2 and 3 (available online at https://www.mcpdigitalhealth.org/), and bias comparisons between graders' scores across the 4 categories are presented in Supplemental Figure 3 (available online at https://www.mcpdigitalhealth.org/).

The nurse graders study focused solely on 2 criteria: "*Can you answer this patient inquiry?*" and "*Estimated time to answer this patient inquiry*." We compared the clinician graders'

estimated times to those of the nurses. On average, clinicians took 3.60 minutes (SD 1.44) to respond to an in-basket message, compared with the nurses' 6.39 minutes (SD 4.05). The evaluation of response times between clinicians and nurses is displayed in Figure 5. Although both clinician graders were able to answer all 158 messages, nurses indicated "No" for 90 inquiries, requiring referral to clinicians, and "Yes" for 68 inquiries. For the inquiries marked "Yes," the average response time was 5.54 minutes, and for those marked "No," the average time was 8.83 minutes. Even when nurses could not handle an inquiry directly, they still had to review patient information and gather sufficient details before deciding if the message needed to be escalated to a clinician.

Analyzing Challenges With RadOnc-GPT

After completing all the gradings, we asked all 5 clinician graders to review responses labeled as "*would not use this message*" and identified several key error categories as follows: (1) lack of context; (2) insufficient domain-specific knowledge; (3) hallucination'; and (4) inability to perform meta-tasks.

The first issue, lack of context, arose when a patient inquired about a rash and itching after receiving a hormone deprivation shot and radiation. C3 noted, "*we might want a photo of this rash, it seems unlikely to be related to their ADT (androgen deprivation therapy), and a full body rash generally requires evaluation*" (C3). Because RadOnc-GPT currently only handles text or hyperlinks, it is unable to process images or videos, which can be essential for providing accurate medical insights in such cases. This limitation restricts RadOnc-GPT's ability to fully interpret diverse patient information.

Another challenge is the insufficient domain-specific knowledge of RadOnc-GPT. In prostate cancer care, whereas clinicians generally follow NCCN guidelines, the specifics of care can vary between clinics and even individual providers. For example, when a patient inquired about difficulty urinating, RadOnc-GPT provided "*generic recommendations about urination, while the patient is specifically at risk of urinary retention, a red-flag symptom that requires them to go to the ED for a catheter. They should at least be made aware that this is a possibility*" (C2). RadOnc-GPT struggled to account for these subtle yet important nuances in medical practice, often leading to less accurate or overly generalized responses to patient inquiries. Moreover, it sometimes fails to distinguish the boundaries of prostate cancer care. Some issues, as C1 noted, may require referral to other specialists: "*More specifically, this patient likely needs a workup with their primary care provider, not in radiation oncology*" (C1). This further highlights RadOnc-GPT's limitations in handling domain-specific complexities.

In addition, RadOnc-GPT's inability to perform meta-tasks creates another barrier. Clinicians, when responding to patient messages, often need to perform other tasks such as updating EHR records, adjusting appointments, or managing pharmacy logistics. Although RadOnc-GPT can respond to patient messages and provide some basic task suggestions, it cannot carry out these additional responsibilities, which are integral to comprehensive patient care.

Finally, hallucination is a relevant issue, where RadOnc-GPT generates inaccurate or contradictory information. For example, C2 pointed out that in response to a patient experiencing recurrent bladder spasms, RadOnc-GPT recommended continuing antibiotics, although the patient had already tried them with no lasting success. C2 commented, "*I don't recommend continuing on antibiotics when clearly it has come back after he tried it, would scrap the message and start over*" (C2). This illustrates how RadOnc-GPT can sometimes provide misleading or incorrect advice, especially when dealing with complex clinical scenarios that require more nuanced understanding. The word cloud of qualitative comments from clinician graders is displayed in Supplemental Figure 4 (available online at https://www.mcpdigitalhealth.org/).

DISCUSSION

RadOnc-GPT was well able to provide medical advice to individualized patient in-basket messages in this retrospective comparison study to both trained radiation oncologists and radiation-oncology-specific nurses. Although RadOnc-GPT responses are human-like and generally similar to responses generated by the original human care teams in many aspects, caution is still needed before deploying its messages without human oversight in real-world health care settings. Because the human care team may still need to confirm the evidence in responses by pulling out the imaging or laboratory/examination results to avoid hallucination, RadOnc-GPT may be able to accelerate the response turn-out rate and alleviate the human care team's response pressure.

Our study found that human care team responses often focused on addressing immediate action items, providing clear instructions to guide patients on the next steps. These responses tended to include less extensive patient education, detailed clinical explanations, or broader context about the patient's condition. RadOnc-GPT, by contrast, complemented these efforts by offering more comprehensive background information, helping patients gain a deeper understanding of their basic health conditions. Sample inquiries, along with corresponding care team and RadOnc-GPT responses, are included in the Supplemental Table 1 for reference. In addition, RadOnc-GPT supported clinicians by assisting in drafting in-basket messages, allowing them to focus more on refining message delivery. This approach reduced drafting time and streamlined responses to repetitive inquiries.

We considered prompts to be one of the key factors determining the quality of RadOnc-GPT responses. For the final prompts, we provided instructions in (1) steps of retrieving information to ensure responsibility; (2) acting as the attending physician and provider; (3) step-by-step reasoning from patient health profiles to address patient's inquiries; (4) handling the medications (prioritizing over-the-counter medications); (5) determining the clarity of patient's inquiry and asking for more information if needed; (6) patient's health literacy; and (7) providing the original patient's inquiry. The full prompts were presented in the Supplementary material (available online at https://www.mcpdigitalhealth.org/).

In addition, there is a lack of standardized scales or metrics to evaluate the GPT-generated messages. A few studies have included clear evaluation methods and scoring rubrics for grading. However, the studies in the medical domain are quite specific, and researchers

found it challenging to generalize the grading across all types of medical domains or diseases. Also, in the grading study, which included human evaluators, the subjective grading could potentially introduce bias from years of experience, practicing domain, clinical roles, and clinics.

Among these 158 patient inquiries, the average wait time for clinical care team response is 22.42 hours (SD=32.83, median=11.73 hours), as shown in Figure 4. The purpose of using RadOnc-GPT to generate in-basket message responses was not to replace the human care team's role in managing prostate cancer patients' inquiries. Instead, RadOnc-GPT was intended to streamline the response process and save time for the care team. Typically, responses to in-basket messages were handled sequentially, starting with nurses, then progressing to nurse practitioners or advanced practice providers, and finally to clinicians.

On the basis of our estimates, when using RadOnc-GPT to assist with in-basket message generation, the average word count for patient inquiries was 88.89 (SD: 64.93), resulting in an estimated reading time of 0.51 minutes per message (SD: 0.37 minutes) for an average English reader (175 words per minute). RadOnc-GPT responses averaged 119.55 words (SD: 49.72), with an estimated reading time of 0.68 minutes (SD: 0.28 minutes). The time required for clinical professionals to review each message was ~1.19 minutes. We acknowledged the potential time and efforts per message to account for the time clinical professionals spend validating statistics or evidence presented in RadOnc-GPT-generated content, navigating between systems, and fact-checking information in the patient's EHR or related data sources. On the basis of the clinicians and nurses estimation, RadOnc-GPT could save ~5.2 minutes per message for nurses and 2.41 minutes for clinicians, from reading the patient inquiry message to drafting and sending the response. With Mayo Clinic receiving around 5000 in-basket messages daily[34] and assuming that one-fifth of these are requests for medical advice (which is 1000 messages), the potential time savings for nurses alone would amount to 5200 minutes (or 86.67 hours) per day. On the basis of the NIH salary table, this equates to an annual savings of at least $2.28 million in nurse time ($72 per hour).[35]

Limitations

The retrospective study feature limited our study because we could not ask the patients to add more information or reply to the RadOnc-GPT-generated responses. We only compared a pair of interactions under one subject, which consists of a patient inquiry and a response message from either RadOnc-GPT or the clinical care team. It might deviate from the real-world interaction because sometimes either the clinical care team or patients send out multiple messages under one subject to explain their health concerns. Furthermore, our sample dataset included 158 in-basket message pairs, a relatively small sample that could miss potential corner cases. Because RadOnc-GPT only handles text and cannot process images or files, its applicability may be limited in scenarios requiring more complex inputs—though no such materials were part of our in-basket message grading. The human care team's responses also varied because of factors such as different clinical roles, heavy workloads, and departmental norms; hence, they may not represent the best possible standard of care.

Second, we made efforts to ensure fairness in grading by conducting a single-blinded study, but we recognize that clinician graders might have inferred the source of a response—whether from clinical professionals or RadOnc-GPT—based on tone and context. Although the single-blinded design may not have completely eliminated this possibility, we believe it provided a practical approach to encourage objective grading among all 3 graders. Because our primary goal was to enhance the overall performance of in-basket message responses rather than replicate existing responses from clinical professionals, the single-blinded design aligned well with the aims of our study.

Another limitation of our work is that we used GPT-4o as the sole backend LLM for RadOnc-GPT to generate responses. We did not compare GPT-4o with other LLMs, such as LLaMA 3, Gemini, GPT-4, or GPT-3.5. As a result, the performance observed with GPT-4o may not generalize to other LLMs. Further research is needed to evaluate the performance of various LLMs and measure potential deviations, as GPT-4o represents just one of many high-performing models available.

CONCLUSION

In this single-blinded comparison study, we evaluated 158 in-basket message interactions between RadOnc-GPT and clinical care teams. The results reported RadOnc-GPT's ability to answer patient inquiries, though we observed limitations in its capacity to capture the nuanced information that clinical professionals provide. Utilizing RadOnc-GPT as a foundational tool for generating in-basket message responses helps clinical professionals save time on addressing patient inquiries, allowing them to focus more on the health care delivery process. This approach not only saves time and improves workflow efficiency but also enables clinicians to be more comprehensive in their responses and to focus more on direct patient interaction care. Future studies should further explore the limitations of LLMs in assisting with in-basket message generation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank the nurse practitioners Derek S. Remme, D.N.P., and Jonathan Moonen, D.N.P., as well as the nurses who contributed to this study: Brittainy Johnson, R.N., Shyanne Dobbs, R.N., Brooke Kelly, R.N., and Bailey Krichner, R.N.. This research was supported by National Cancer Institute (NCI) R01CA280134, the Eric Wendy Schmidt Fund for AI Research & Innovation, the Fred C. and Katherine B. Anderson Foundation, and the Kemper Marley Foundation. The authors also acknowledged support from Paul Calabresi program in Clinical/Translational Research at the Mayo Clinic Comprehensive Cancer Center K12CA090628.

REFERENCES

1. Han HR, Gleason KT, Sun CA, et al. Using patient portals to improve patient outcomes: systematic review. JMIR Hum Factors. 2019;6(4):e15038. 10.2196/15038. [PubMed: 31855187]
2. Sandford LM, Fouayzi H, Sundaresan D, et al. Tracking health care team response to electronic health record asynchronous alerts: role of in-basket message burden. J Patient Centered Res Rev. 2016;3(3):201–202. 10.17294/2330-0698.1348.

3. Baxter SL, Saseendrakumar BR, Cheung M, et al. Association of electronic health record inbasket message characteristics with physician burnout. JAMA Netw Open. 2022;5(11):e2244363–e2244363. 10.1001/jamanetworkopen.2022.44363.. [PubMed: 36449288]
4. Overhage JM, McCallie D Jr. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. Ann Intern Med. 2020;172(3):169–174. 10.7326/M18-3684. [PubMed: 31931523]
5. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood). 2019;38(7):1073–1078. 10.1377/hlthaff.2018.05509. [PubMed: 31260371]
6. Nath B, Williams B, Jeffery MM, et al. Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in New England. JAMA Netw Open. 2021;4(10):e2131490. 10.1001/jamanetworkopen.2021.31490. [PubMed: 34636917]
7. Holmgren AJ, Byron ME, Grouse CK, Adler-Milstein J. Association between billing patient portal messages as e-visits and patient messaging volume. JAMA. 2023;329(4):339–342. 10.1001/jama.2022.24710. [PubMed: 36607621]
8. Lieu TA, Altschuler A, Weiner JZ, et al. Primary care physicians' experiences with and strategies for managing electronic messages. JAMA Netw Open. 2019;2(12):e1918287. 10.1001/jamanetworkopen.2019.18287. [PubMed: 31880798]
9. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc. 2020;27(4): 531–538. 10.1093/jamia/ocz220. [PubMed: 32016375]
10. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–596. 10.1001/jamainternmed.2023.1838. [PubMed: 37115527]
11. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. Preprint. Posted Online March 4, 2024. arXiv230308774. 10.48550/arXiv.2303.08774.
12. Matulis J, McCoy R. Relief in sight? Chatbots, in-baskets, and the overwhelmed primary care clinician. J Gen Intern Med. 2023; 38(12):2808–2815. 10.1007/s11606-023-08271-8. [PubMed: 37369892]
13. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. Lancet Digit Health. 2024;6(6):e379–e381. 10.1016/S2589-7500(24)00060-8. [PubMed: 38664108]
14. Gandhi TK, Classen D, Sinsky CA, et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? JAMIA Open. 2023;6(3):ooad079. 10.1093/jamiaopen/ooad079.
15. Baxter SL, Longhurst CA, Millen M, Sitapati AM, Tai-Seale M. Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned. JAMIA Open. 2024;7(2):ooae028. 10.1093/jamiaopen/ooae028.
16. Small WR, Wiesenfeld B, Brandfield-Harvey B, et al. Large language model–based responses to patients' in-basket messages. JAMA Netw Open. 2024;7(7):e2422399–e2422399. 10.1001/jamanetworkopen.2024.22399. [PubMed: 39012633]
17. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI. 2023;1(1):AIp2300031. 10.1056/AIp2300031.
18. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. Preprint. Posted online April 12, 2023. arXiv230313375. 10.48550/arXiv.2303.13375.
19. Hao Y, Liu Z, Riter RN, Kalantari S. Advancing patient-centered shared decision-making with AI systems for older adult cancer patients. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. CHI '24. ACM; 2024:1–20. 10.1145/3613904.3642353.
20. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol. 2023;13:1219326. 10.3389/fonc.2023.1219326. [PubMed: 37529688]

21. Garcia P, Ma SP, Shah S, et al. Artificial intelligence–generated draft replies to patient inbox messages. JAMA Netw Open. 2024; 7(3):e243201. 10.1001/jamanetworkopen.2024.3201. [PubMed: 38506805]
22. Rezayi S, Dai H, Liu Z, et al. ClinicalRadioBERT: Knowledge-infused few shot learning for clinical notes named entity recognition. In: Lian C, Cao X, Rekik I, Xu X, Cui Z, eds. Machine Learning in Medical Imaging. Switzerland: Springer Nature; 2022:269–278. 10.1007/978-3-031-21014-3_28.
23. Liu Z, Zhong A, Li Y, et al. Tailoring large language models to radiology: a preliminary approach to LLM adaptation for a highly specialized domain. In: Cao X, Xu X, Rekik I, Cui Z, Ouyang X, eds. Machine Learning in Medical Imaging. Switzerland: Springer Nature; 2024:464–473. 10.1007/978-3-031-45673-2_46.
24. Xiao Z, Chen Y, Yao J, et al. Instruction-ViT: multi-modal prompts for instruction learning in vision transformer. Inf Fusion. 2024;104:102204. 10.1016/j.inffus.2023.102204.
25. Hao Y, Liu Z, Riter RN, Kalantari S. Advancing patient-centered shared decision-making with AI systems for older adult cancer patients. Proceedings of the CHI Conference on Human Factors in Computing Systems 2024:1–20. 10.1145/3613904.3642353.
26. Liu Z, Wang P, Li Y, et al. RadOnc-GPT: a large language model for radiation oncology. Preprint. Published online November 6, 2023: 10160. arXiv:2309. 10.48550/arXiv.2309.10160.
27. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. ACM Trans Intell Syst Technol. 2024;15(3):1–45. 10.1145/3641289.
28. Iroju OG, Olaleke JO. A systematic review of natural language processing in healthcare. Int J Inf Technol Comput Sci. 2015;8(8): 44–50. 10.5815/ijitcs.2015.08.07.
29. Liu L, Yang X, Li F, et al. Towards automatic evaluation for LLMs' clinical capabilities: metric, data, and algorithm. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM; 2024:5466–5475. 10.1145/3637528.3671575.
30. Abeysinghe B, Circi R. The challenges of evaluating LLM applications: an analysis of automated, human, and LLM-based approaches. Preprint. Posted online June. 2024;13: arXiv240603339. 10.48550/arXiv.2406.03339.
31. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. J Biomed Inform. 2024;151:104620. 10.1016/j.jbi.2024.104620. [PubMed: 38462064]
32. MacCartney B Natural Language Inference. Stanford University; 2009.
33. Miller GA, Charles WG. Contextual correlates of semantic similarity. Lang Cogn Process. 1991;6(1):1–28. 10.1080/01690969108406936.
34. Cognetta-Rieke C Mayo Clinic department of nursing leveraging artificial intelligence for automating draft patient message responses. Mayo AI Summit; 2024. https://attendesource.com/accounts/register123/metroconnection/mayo/events/mayo-aisummit-web-0724/Lightning%20Talk%20Presenters.pdf. Accessed September 25, 2024.
35. Pay Guide. May 7, 2018: https://hr.nih.gov/benefits/pay/pay-guide. Accessed December 17, 2024.

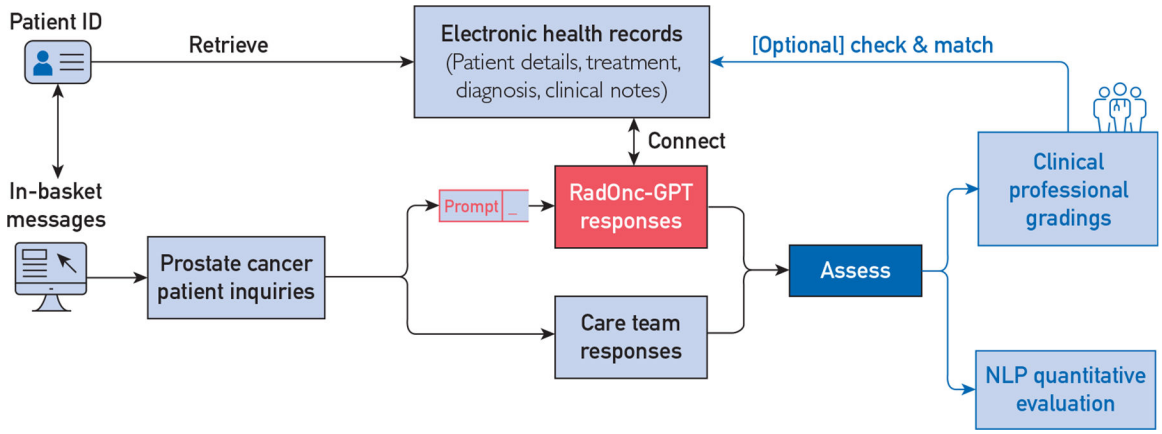

FIGURE 1.
In-basket comparison study workflow overview.

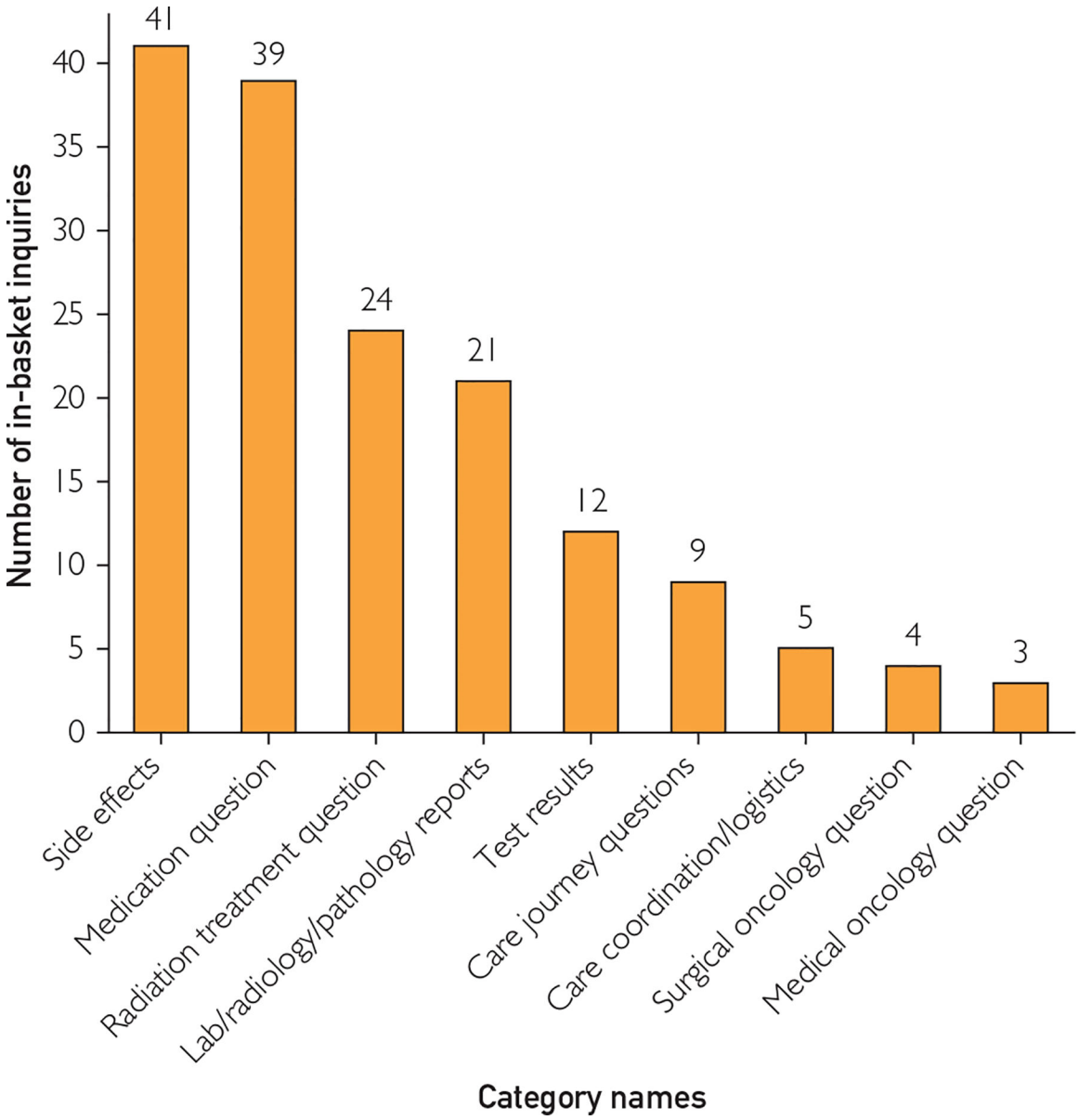

FIGURE 2.
Nine categories of in-basket message patients' inquiries.

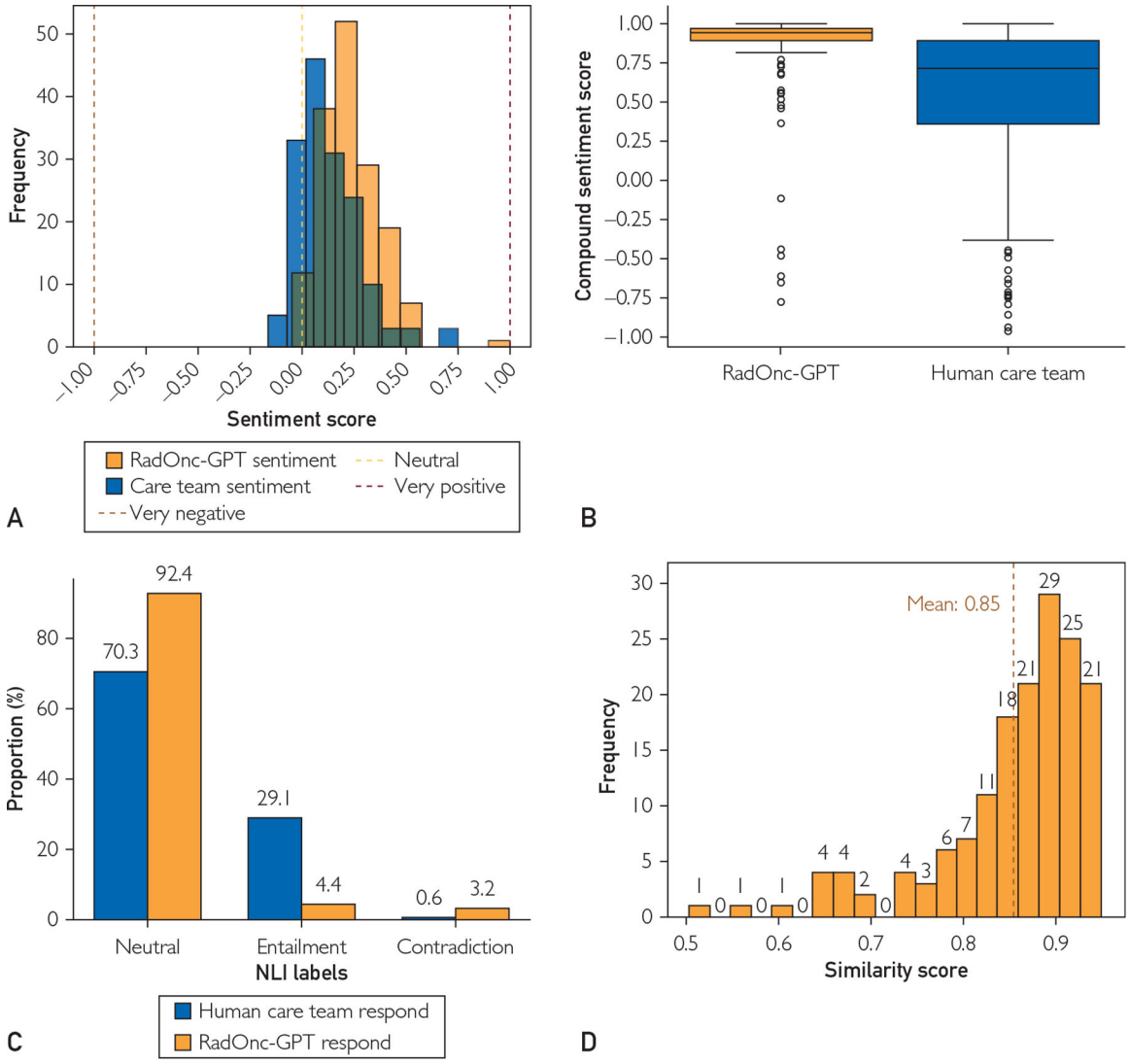

FIGURE 3.

Sentiment analysis. (A) TextBlob sentiment distribution; (B) VADER sentiment distribution; (C) Natural language inference distributions between GPT and care team responses; (D) semantic similarity scores.

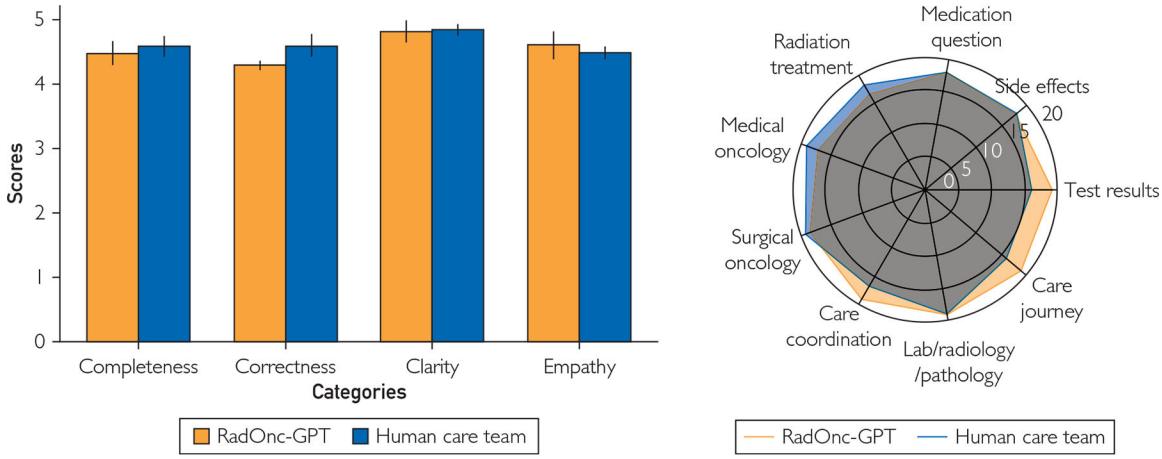

FIGURE 4.
Average score across all 4 categories.

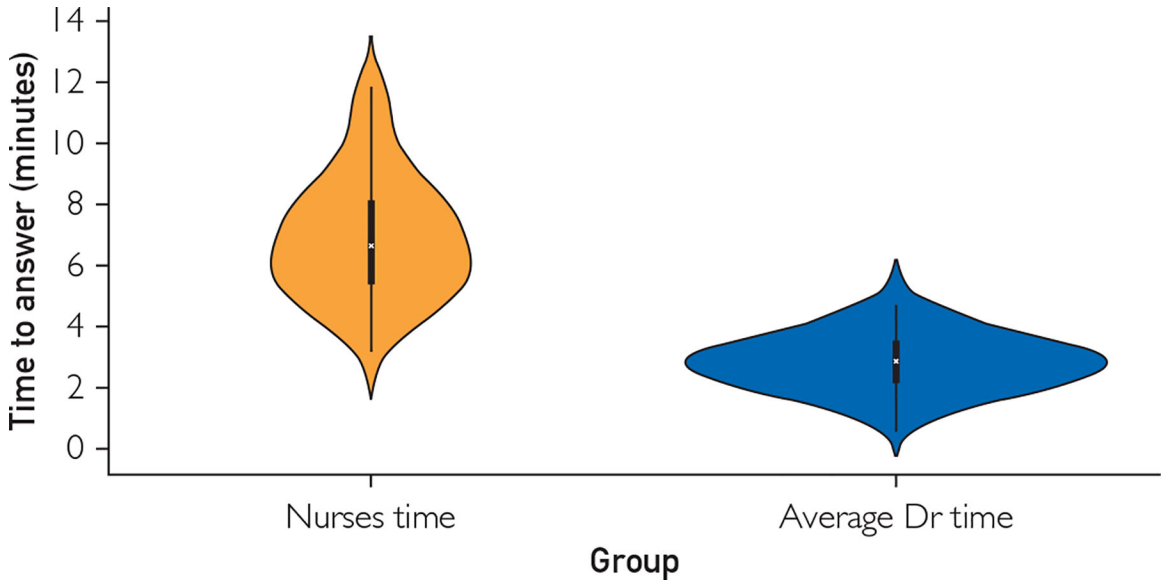

FIGURE 5.
Comparative evaluation of editing effort and response times in in-basket message management. (A) comparisons of 3 graders' average human care team and RadOncGPT editing time; (B) comparative analysis of clinicians and nurses average time in responding to in-basket messages.

TABLE.

Clinician and Nurse Grader Profiles

Clinician	Clinical Domain	Gender	YoE	Nurse	Cancer Domain	Nurse Gender	YoE
C1	Radiation oncology	Male	3 y	N1	Prostate and breast cancer	Female	13 y
C2	Radiation oncology	Female	3 y	N2	Prostate cancer	Female	4
C3	Radiation oncology	Female	3 y	N3	Prostate cancer	Female	2 y
C4	Radiation oncology	Male	5 y	N4	Prostate cancer	Female	2 y
C5	Radiation oncology	Male	9 y				

Abbreviations: YoE, years of experience.

Mayo Clin Proc Digit Health. Author manuscript; available in PMC 2025 March 24.