JAMA Network Open

Original Investigation | Health Informatics

Large Language Model–Based Responses to Patients' In-Basket Messages

William R. Small, MD, MBA; Batia Wiesenfeld, PhD; Beatrix Brandfield-Harvey, BS; Zoe Jonassen, PhD; Soumik Mandal, PhD; Elizabeth R. Stevens, PhD; Vincent J. Major, PhD; Erin Lostraglio, BA; Adam Szerencsy, DO; Simon Jones, PhD; Yindalon Aphinyanaphongs, MD, PhD; Stephen B. Johnson, PhD; Oded Nov, PhD; Devin Mann, MD

Abstract

IMPORTANCE Virtual patient-physician communications have increased since 2020 and negatively impacted primary care physician (PCP) well-being. Generative artificial intelligence (GenAI) drafts of patient messages could potentially reduce health care professional (HCP) workload and improve communication quality, but only if the drafts are considered useful.

OBJECTIVES To assess PCPs' perceptions of GenAI drafts and to examine linguistic characteristics associated with equity and perceived empathy.

DESIGN, SETTING, AND PARTICIPANTS This cross-sectional quality improvement study tested the hypothesis that PCPs' ratings of GenAI drafts (created using the electronic health record [EHR] standard prompts) would be equivalent to HCP-generated responses on 3 dimensions. The study was conducted at NYU Langone Health using private patient-HCP communications at 3 internal medicine practices piloting GenAI.

EXPOSURES Randomly assigned patient messages coupled with either an HCP message or the draft GenAI response.

MAIN OUTCOMES AND MEASURES PCPs rated responses' information content quality (eg, relevance), using a Likert scale, communication quality (eg, verbosity), using a Likert scale, and whether they would use the draft or start anew (usable vs unusable). Branching logic further probed for empathy, personalization, and professionalism of responses. Computational linguistics methods assessed content differences in HCP vs GenAI responses, focusing on equity and empathy.

RESULTS A total of 16 PCPs (8 [50.0%] female) reviewed 344 messages (175 GenAI drafted; 169 HCP drafted). Both GenAI and HCP responses were rated favorably. GenAI responses were rated higher for communication style than HCP responses (mean [SD], 3.70 [1.15] vs 3.38 [1.20]; $P = .01$, $U = 12\,568.5$) but were similar to HCPs on information content (mean [SD], 3.53 [1.26] vs 3.41 [1.27]; $P = .37$; $U = 13\,981.0$) and usable draft proportion (mean [SD], 0.69 [0.48] vs 0.65 [0.47], $P = .49$, $t = -0.6842$). Usable GenAI responses were considered more empathetic than usable HCP responses (32 of 86 [37.2%] vs 13 of 79 [16.5%]; difference, 125.5%), possibly attributable to more subjective (mean [SD], 0.54 [0.16] vs 0.31 [0.23]; $P < .001$; difference, 74.2%) and positive (mean [SD] polarity, 0.21 [0.14] vs 0.13 [0.25]; $P = .02$; difference, 61.5%) language; they were also numerically longer (mean [SD] word count, 90.5 [32.0] vs 65.4 [62.6]; difference, 38.4%), but the difference was not statistically significant ($P = .07$) and more linguistically complex (mean [SD] score, 125.2 [47.8] vs 95.4 [58.8]; $P = .002$; difference, 31.2%).

CONCLUSIONS In this cross-sectional study of PCP perceptions of an EHR-integrated GenAI chatbot, GenAI was found to communicate information better and with more empathy than HCPs,

(continued)

Key Points

Question Can generative artificial intelligence (GenAI) chatbots aid patient–health care professional (HCP) communication by creating high-quality draft responses to patient requests?

Findings In this cross-sectional study of 16 primary care physicians' opinions on the quality of GenAI- and HCP-drafted responses to patient messages, GenAI responses were rated higher than HCPs' for communication style and empathy. GenAI responses were longer, more linguistically complex, and less readable than HCP responses; they were also rated as more empathetic and contained more subjective and positive language.

Meaning In this study, primary care physicians perceived that GenAI chatbots produced responses to patient messages that were comparable in quality with those of HCPs, but due to GenAI responses' use of complex language, these responses could cause problems for patients with lower health or English literacy.

+ Multimedia

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

highlighting its potential to enhance patient-HCP communication. However, GenAI drafts were less readable than HCPs', a significant concern for patients with low health or English literacy.

JAMA Network Open. 2024;7(7):e2422399. doi:10.1001/jamanetworkopen.2024.22399

Introduction

The surge in patient-health care professional (HCP) messaging due to COVID-19 has increased electronic health record (EHR) inbox management burden,[1-6] particularly for primary care physicians (PCPs), contributing to burnout (especially among female and Hispanic/Latino physicians).[3,7-11] Each additional message adds more than 2 minutes of EHR time, encompassing message drafting, information searching, order placing, and documentation.[1,5,10] Proposed relief strategies include EHR window switching reduction (improving user interface design), upskilling support staff, and billing for messages,[1,7,10,11] a practice now permitted by the Centers for Medicare & Medicaid Services, which evidence suggests reduces messaging burden.[12] Using EHR-integrated generative artificial intelligence (GenAI) chatbots to automate drafting responses to patient messages could streamline workflows and thus alleviate burnout.

GenAI chatbots are large language models (LLMs)[13] that synthesize massive text volumes, including medical literature, and have potential in many health care applications, including clinical note generation and medical text simplification.[14-17] Substantial implementation challenges include processing needs, model biases, privacy concerns, and absent evaluation benchmarks.[18-22] Addressing these challenges will enhance understanding of this technology's benefits and limitations.[18] Successful adoption depends on understanding HCPs' and patients' perceptions of GenAI outputs.[13,16,18,20,21,23,24]

Studies investigating output quality, evaluation methods, and benchmarking are burgeoning as institutions pilot GenAI for in-basket messaging.[14,16,18,20,22,25-34] Studies of physicians' perceptions of GenAI response efficacy on various dimensions often find them equivalent to HCP-drafted responses, especially in empathy.[14,25-28] Despite advocacy for including private information in AI models[20] and the institutional push for EHR implementation,[35] few studies have assessed chatbots' use of private patient information to answer their messages.[32-34]

Our study addresses this gap by using private patient-HCP message-response pairs to investigate PCPs' perceptions of GenAI drafts and explore underlying linguistic factors associated with equity and perceived empathy. We hypothesize that GenAI draft quality, assessed by PCPs on information content quality, communication style, and usability, will be equivalent to HCP-generated responses.

Methods

This blinded cross-sectional quality improvement study evaluates PCP perceptions of GenAI responses to patient messages compared with HCP-generated responses. Subgroup analyses evaluated whether response quality varied with HCP type (physicians and nonphysicians) and patient message classification (laboratory results, medication refill requests, paperwork, and general medical advice; determined by the EHR's proprietary message classification LLM [Epic]). Computational linguistics analyses compare response content to elucidate potential equity concerns and why drafts were considered empathetic. As part of an operational pilot program to implement and curate GenAI in-basket drafts most acceptable for end-users, this study met NYU criteria for quality improvement work and did not undergo institutional review board review. All study procedures complied with institutional ethical standards and those set by the Declaration of Helsinki

and are reported using the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines for cross-sectional studies.

Study Setting and Participants

A convenience sample of 16 PCPs were recruited from a large urban academic health system via a listserv email to 1189 internal medicine physician email addresses. PCPs affiliated with NYU Grossman School of Medicine were eligible. Participants provided consent by accepting the request to complete the survey and could opt out at any time. Surveys were collected between September 23 and November 16, 2023.

Survey

Surveys were conducted in REDCap.[36] Participants were provided a random selection of message-response pairs over 2 surveys, masked to whether the response was generated by GenAI or an HCP. The first contained 5 to 8 message-response pairs but no branching logic, while the second survey contained 15 to 20 message-response pairs and branching logic (eAppendix 1 in Supplement 1).

In both surveys, participants assessed the quality of response information content and communication style using 5-point Likert scale questions (scale 1-5, with 1 indicating strongly disagree and 5 indicating strongly agree), then answered whether it was preferable to starting from a blank page (usable vs unusable). Branching logic followed negative responses to Likert questions (bottom 2 box) to explore PCPs' rationale, assessing for aspects like relevance and empathy. Regardless of whether a draft was considered usable, respondents selected from a list of items (eAppendix 1 in Supplement 1).

To construct the first survey, 200 random in-basket messages were extracted on September 12, 2023, including the corresponding HCP and AI-generated response. A total of 112 patient messages were reviewed, and 53 were excluded because they needed outside context (eg, laboratory values or medication names) for adequate evaluation of the response by participants, leaving 59 patient messages paired with both HCP and GenAI responses (52.7%). For the second survey, 500 random patient messages were extracted from the data warehouse on October 12, 2023, of which 464 were reviewed, and 146 were excluded due to the need for external context to properly evaluate the response, leaving 318 patient messages paired with both HCP and GenAI responses (68.5%) from which respondents' questions were randomly assigned. Not all extracted messages were reviewed because the desired sample size (determined by the effort required of our participants) was achieved beforehand. Message-response pairs were randomly sampled (with replacement) for review, yielding some pairs being reviewed by different reviewers.

Message-Response Sample Selection

The survey used in-basket message-response pairs from outpatient internal medicine departments participating in the pilot study of Generated Draft Replies (Epic), which generated responses using GPT-4 (OpenAI) through an EHR-integrated, vendor-prepared system. Pairs were randomly selected during the system's silent validation, where drafts were being generated using Epic's standard prompts but not seen by HCPs. The patient message subcategory (laboratory results, medication refill requests, paperwork, and general medical advice) determined which prompt (utilizing unique instructions and patient-specific details) generated the response (eg, laboratory results messages auto-populate recent test results, while medication refill requests include the active medication list). Evaluating standard prompts allows for benchmarking future prompt engineering efforts.

Inclusion criteria dictated that the first patient-initiated message between the patient and their HCP was chosen. If multiple HCP messages were sent in response, they were combined to minimize artificially incomplete responses. Responses from physicians, nurses, and frontline staff were included to reflect how patient requests are answered at many institutions.

Statistical Analysis

Statistical analysis was conducted in Python version 3.9.16 (Python Software Foundation) in May 2024. We used a priori levels of significance of $P < .05$ for 2-sided tests of the null hypothesis that GenAI drafts would be equal to HCP responses on our 3 main survey questions. Mann-Whitney tests, robust to outliers and nonnormal distributions,[14,37,38] evaluated differences between GenAI and HCP responses for the 2 main Likert questions and the 2-way paperwork messages subgroup comparison. Kruskal-Wallis tests compared the Likert scale means of physicians, nonphysicians, and GenAI across the 4 message subcategories.[14] Independent sample *t* tests were used to compare differences in the proportion of GenAI vs HCP responses considered usable and all computational linguistics measures. One-way analysis of variance was used to compare the proportions of drafts considered usable by physicians, nonphysicians, and GenAI across 3 of 4 message subcategories. *P* values for all secondary analyses underwent a Sidak correction[39] to account for multiple comparisons.

Because our data are ordinal and pairs were randomly assigned, the 1-way intraclass correlation coefficient (ICC) was used to estimate interrater reliability from the double-reviewed questions.[40] Linear mixed models with random effects for individual reviewer variation and fixed effects for patient message subcategory and HCP subcategory were built (eAppendix 3 in Supplement 1) to assess how these factors affected survey results.

Computational linguistics methods analyzed responses' length, complexity, and sentiment as well as the prevalence of specific content dimensions, such as positive emotion words. Such measures characterize writing styles and can anticipate readers' attitudes and behavior toward the content, including their perception of its usefulness.[41] Analysis was performed in Python with the pandas package (version 2.1.1) used to calculate word counts. Lexical diversity, or the variety of words used in a text, was assessed with the measure of textual, lexical diversity, calculated using the lexical_diversity package (version 0.1.1) and chosen due to its insensitivity to text length.[42-45] Lexical diversity reflects language proficiency; highly diverse text indicates the author is using a broad range of vocabulary to express their thoughts and ideas.[46] The textstat package (version 0.7.3) calculated the Flesch-Kincaid grade level, which is calculated from the average syllables per word and average words per sentence, and describes an English passage's comprehensibility.[47,48]

Content analysis of the main response groups and empathetic subgroups utilized the latest Linguistic Inquiry and Word Count (LIWC) application, LIWC-22, the preferred application for automated text analysis in social science research.[49] LIWC-22 matches response words to various dictionaries and subdictionaries that represent themes like positive and negative emotion and reports metrics as a percentage of words in the text that exist in a given theme's dictionary.[50-53] The textblob package (version 0.17.1) facilitated sentiment analysis, a common method used to assess subjectivity (range 0 to 1, higher indicating more subjectivity) and polarity, or the overall positive/negative tone of a text (range −1 to 1, higher is more positive).[54-56]

Results

Of 1189 email addresses on the internal medicine listserv, 16 PCP participants (1.3%) volunteered. All were outpatient faculty, 8 (50.0%) were female, 7 (43.8%) worked primarily at NYU Langone Health, 5 (31.2%) at Bellevue Hospital, and 2 (12.5%) at the Manhattan Veteran's Affairs hospital.

Of 344 evaluated survey message-response pairs (175 GenAI drafted; 169 HCP drafted), there were 157 single-reviewed, 73 double-reviewed, 11 triple-reviewed, and 2 quadruple-reviewed message-response pairs, resulting in 117 unique HCP and 126 unique GenAI message-response pairs. Branching logic only occurred during the second survey and was available for 207 unique questions (85.2%).

Survey Results

Participant evaluations were generally positive for GenAI and HCP responses (**Figure 1** and **Table 1**). ICC (range −1 to 1) was 0.11 for information content quality, 0.094 for communication style, and 0.012

for draft usability. Low ICC results were not explained by message source (HCP vs GenAI) (eAppendix 5 in Supplement 1). To address ICC concerns, analyses treated multireviewed questions as independent observations rather than average their scores.

The information content quality (accuracy, completeness, and relevance) of GenAI and HCP responses did not differ statistically (mean [SD], 3.53 [1.26] vs 3.41 [1.27]; $P = .37$; $U = 13\,981.0$), a finding that persisted when controlling for individual reviewer variance, HCP subcategory (physician vs nonphysician) as a random effect, and patient message subcategory as a fixed effect (eAppendix 3 in the Supplement). For responses with inadequate information content quality (Likert score <3), HCP responses, compared with GenAI responses, were more often incomplete (24 of 39 [61.5%] vs 11 of 33 [33.3%]) while GenAI responses, compared with HCP responses, were more often irrelevant (10 [30.3%] vs 5 [12.8%]) (**Figure 2**). The other category for HCP responses (9 [23.1%]) received comments on unresponsiveness or even rudeness (eAppendix 4 in Supplement 1), while GenAI received comments (6 [18.2%]) about insensitivity to clinical urgency.

GenAI responses significantly outperformed HCP responses in communication style (understandability, tone, verbosity; mean [SD], 3.70 [1.15] vs 3.38 [1.20]; $P = .01$; $U = 12\,568.5$). Subgroup analyses (eAppendix 2 in Supplement 1) and linear mixed models (eAppendix 3 in Supplement 1) revealed that physician underperformance was associated with this discrepancy. Low-scoring responses (Likert score <3) criticized HCPs more than GenAI for inappropriate tone (18 of 40 [45.0%] vs 3 of 27 [11.1%]) and criticized GenAI more than HCPs for verbosity (13 [48.1%] vs 7 [17.5%]). Free-text comments (HCP, 11 [30.0%]; GenAI, 12 [40.7%]) highlighted HCPs' use of jargon and GenAI's extraneous information.

GenAI and HCP responses were considered usable in similar proportions (mean [SD] proportion usable, 0.69 [0.48] vs 0.65 [0.47]; $P = .49$; $t = -0.6842$). Overall, 50% or less of the physician and laboratory results responses were considered usable (eAppendix 2 in Supplement 1). For unusable drafts, criticisms were evenly spread for HCPs, while GenAI was more often deemed unhelpful (28 of 38 [73.7%]). Insufficient professionalism occurred almost 4 times as frequently in the HCP responses vs GenAI responses (9 [19.6%] vs 2 [5.3%]). HCPs' and GenAI's free-text comments (1 [2.2%] vs 4

Figure 1. Distribution of Health Care Professional (HCP) and Generative Artificial Intelligence (GenAI) Responses to Each Main Survey Question

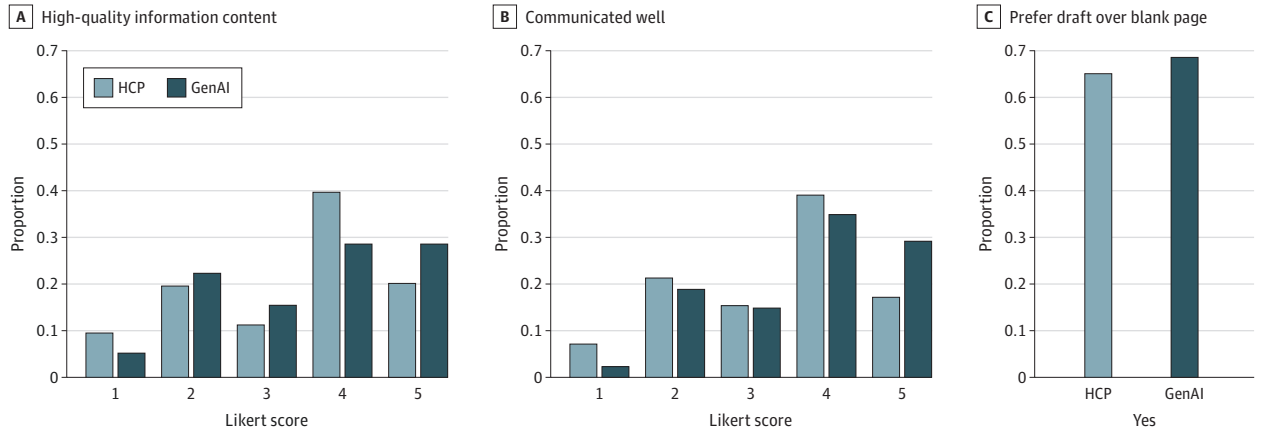


Table 1. Primary Care Physicians' Ratings of HCP and GenAI Responses for Each of the Main Survey Questions

Message drafter	No.	Information content quality (5-point Likert), mean (SD)	*P* value	Communication quality (5-point Likert), mean (SD)	*P* value	Proportion of responses preferred to a blank page, mean (SD)	*P* value
HCP	169	3.41 (1.27)	.37	3.38 (1.20)	.01	0.65 (0.47)	.49
GenAI	175	3.53 (1.26)		3.70 (1.15)		0.69 (0.48)	

Abbreviations: GenAI, generative artificial intelligence; HCP, health care professional.

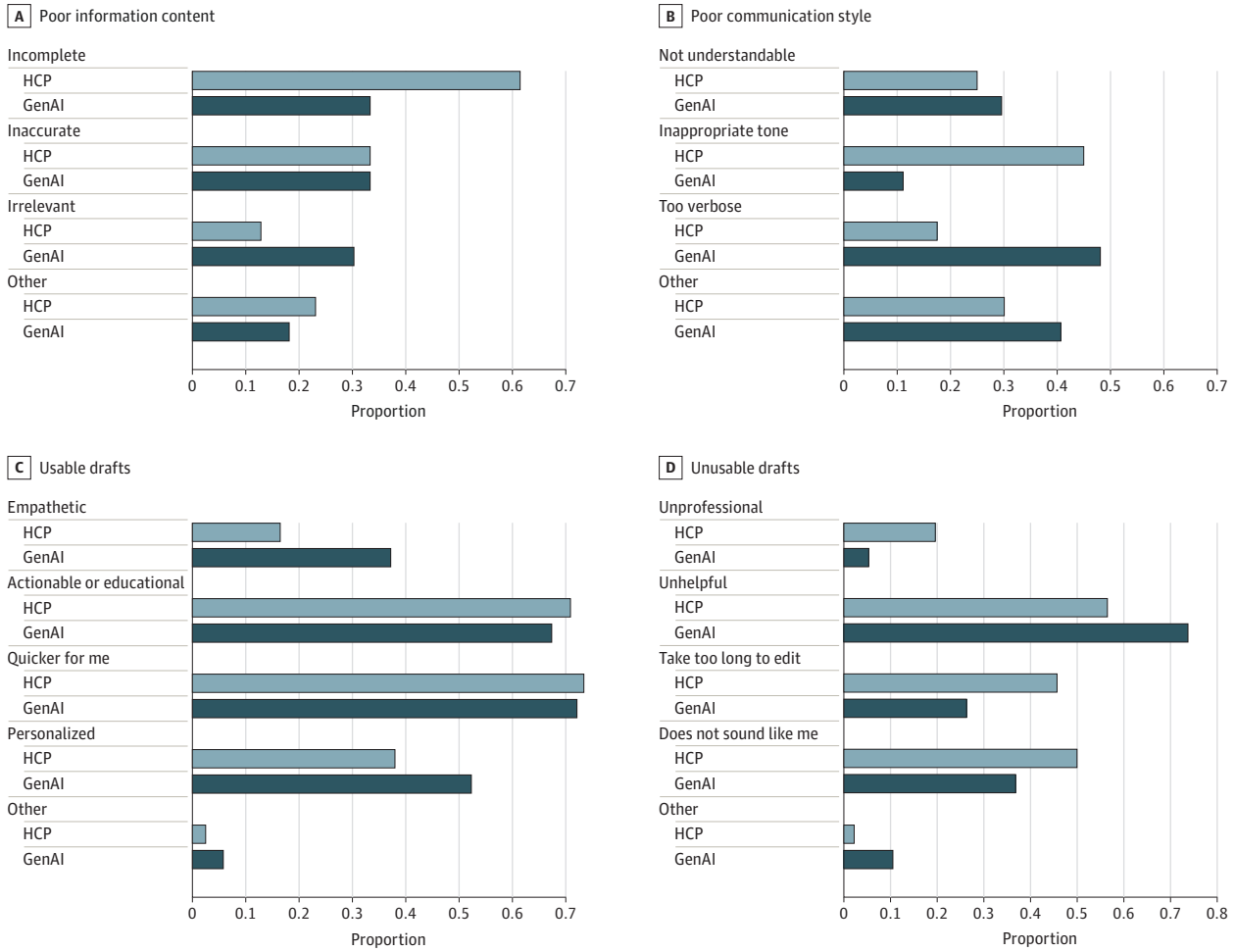
[10.5%]) both mentioned inadequate concern with a message's clinical urgency, and GenAI responses were criticized for insufficient reasoning for proposed actions.

PCPs noted drafts were usable mainly because they would have been quicker to edit than start anew (58 of 79 HCP drafts [73.4%]; 62 of 86 GenAI drafts) and were more actionable or educational (56 HCP drafts [70.9%]; 58 GenAI drafts [67.4%]). GenAI responses, compared with HCP responses, were more often perceived as personalized (45 [52.3%] vs 30 [38.0%]) and empathetic (32 [37.2%] vs 13 [16.5%]; difference, 125.5%). Comments on HCP (2 [2.5%]) and GenAI (5 [5.8%]) responses critiqued clarity, and some GenAI responses were identified as computer-generated.

Computational Linguistics Results

GenAI responses were 38% longer than HCPs' (imposing a burden on readers' time), but the difference was not statistically significant (mean [SD] word count, 90.5 [32.0] vs 65.4 [62.6]; *P* = .07; difference, 38.4%); had greater lexical diversity (requiring a wider vocabulary for readers to comprehend) (mean [SD] score, 125.2 [47.8] vs 95.4 [58.8]; *P* = .002; difference, 31.2%), and required higher levels of education to understand (mean [SD] Flesch-Kincaid grade level, 8.1 [1.6] vs 6.5 [2.3]; *P* < .001) (**Table 2**). GenAI responses utilized more polarity (positivity) (mean [SD], 0.21 [0.14] vs 0.13 [0.25]; *P* = .02; difference, 61.5%) and subjectivity (mean [SD], 0.54 [0.16] vs 0.31 [0.23]; *P* < .001; difference, 74.2%), a pattern maintained in empathetic responses, which also

Figure 2. Branching Logic Results for Inadequate Information Content, Communication Style, and Both Usable and Unusable Draft Responses



GenAI indicates generative artificial intelligence; HCP, health care professional.

contained a significantly higher proportion of (particularly positive) emotion words and a greater use of affiliative language (**Table 3**).

Discussion

This analysis suggests that GenAI draft responses to patient requests, rated similarly to HCPs responses, could help mitigate the growing burden of in-basket messages, a known contributor to physician burnout.[1-4,57] According to the PCP respondents, and consistent with prior studies,[25-28] GenAI drafts outperformed HCPs' responses on communication quality. Despite poor interrater reliability, the sensitivity analysis revealed consistent patterns of findings even after incorporating random effects for reviewers. Subsequently including fixed effects for HCP and patient message subcategories revealed that physicians were responsible for HCP responses underperforming GenAI on communication quality. This may be because physicians responded to more challenging messages than their nonphysician colleagues.[11]

GenAI responses matched HCP responses in information quality, indicating effective use of health care–related training data[20] and patient health data within the standard prompts. This deviates from Ayers et al,[26] where chatbots had 3.6 times higher quality responses to public patient messages, but still supports chatbots' utility. A crucial caveat is that intentional guardrails restrict the LLM's confidence in providing medical information[22] and are designed to limit hallucinations and automation bias,[14,18] but may explain why PCPs found GenAI responses more often unhelpful and irrelevant.

GenAI's poor performance on certain subgroups, especially laboratory results, likely results from the differing prompts by message type, reinforcing the need for benchmarking and thoughtful prompt engineering.[18-20] Future implementers of GenAI into EHR in-basket messaging should direct resources toward revising prompts related to laboratory results.

The prevalence of affiliation words, positivity, and subjectivity in GenAI drafts may explain why they were perceived as more empathetic than HCPs'. Affiliation content, such as "together" and "us," implies a partnership between the HCP and patient. Although empathy is context-sensitive, responses that PCPs perceived as empathetic contained more positive language, which may convey hopefulness and potentially better outcomes.[58] GenAI could thus improve virtual communications between HCPs (physicians in particular) and patients. HCPs surprisingly did not leverage knowledge of their patients to communicate more empathetically than GenAI. GenAI's more consistent language

Table 2. Basic Computational Linguistics and Lexical Complexity Metrics per Unique Responses in Each Group

Metric	Mean (SD)		*P* value
	HCP (n = 117)	GenAI (n = 126)	
Word count	65.4 (62.6)	90.5 (32.0)	.07
Measure of textual lexical diversity	95.4 (58.8)	125.2 (47.8)	.002
Flesch-Kincaid grade level	6.5 (3.3)	8.1 (1.6)	<.001

Abbreviations: GenAI, generative artificial intelligence; HCP, health care professional.

Table 3. Content and Sentiment Analysis of GenAI vs HCP and Empathetic vs Nonempathetic Responses

Metric	Mean (SD)		*P* value	Mean (SD)		*P* value
	HCP (n = 117)	GenAI (n = 126)		Empathetic (n = 41)	Nonempathetic (n = 166)	
Polarity	0.13 (0.24)	0.21 (0.14)	.02	0.26 (0.21)	0.15 (0.19)	.04
Subjectivity	0.31 (0.23)	0.54 (0.16)	<.001[2]	0.56 (0.18)	0.40 (0.22)	.003
Affiliation	1.92 (2.46)	2.51 (1.97)	.56	3.06 (2.01)	1.98 (2.20)	.08
Emotion	1.21 (2.28)	1.41 (1.42)	.99	2.42 (2.06)	1.12 (1.82)	.002
Positive emotion	0.87 (2.12)	1.20 (1.38)	.95	1.95 (2.14)	0.86 (1.67)	.01
Negative emotion	0.29 (1.05)	0.20 (0.55)	.99	0.41 (0.95)	0.22 (0.87)	.99

Abbreviations: GenAI, generative artificial intelligence; HCP, health care professional.

structure, reflected by smaller SDs for most metrics, and its use of more emotional and affiliative language, suggests PCPs may utilize structured responses that fill a gap in their typical responses.

Despite critiques of GenAI's verbosity and readability, it maintained more appropriate tone and professionalism, potentially reflecting HCPs' time constraints when drafting responses.[1-11] In fact, PCPs cited quicker for me as the main reason drafts were considered usable. Although not the primary audience, PCPs must still perceive GenAI drafts as high quality before utilizing them. Patients are the ultimate recipients of drafts, and future research must assess their perceptions of GenAI responses, whose linguistic complexity may be preferred (or ignored) by physicians but burden those with low health or English literacy. Research must also explore concerns about whether GenAI perpetuates bias and health inequity of various patient demographic characteristics[32-34,59,60] and determine whether communication gains outweigh such risks.

This study addressed GenAI implementation challenges, including benchmarking draft and prompt quality and understanding PCPs' perceptions. A critical finding of our study was the inability of PCPs to agree with each other on what makes a draft high quality, suggesting that successful utilization of drafts by PCPs requires a personalized approach. Future research should investigate the impact of prompt refinement and personalization on end-users' perceptions of draft quality. Computational linguistics may drive more intelligent prompt engineering to enhance outputs' empathy, reduce their linguistic complexity, and improve personalization.

Limitations

This study has limitations. Generalizability may be limited due to this study's single-center focus and small sample size. The evaluated GenAI responses were not used to deliver patient care, which may limit our findings' practical applicability. Low ICC suggests a need to adjust the survey questions or instructions (although variance in reviewer responses did not affect our findings) or conduct follow-up interviews to investigate reasons for disagreement. This study did not evaluate the perceptions of patients and nonphysician HCPs who participate in outpatient messaging. We acknowledge that for some HCPs who answer patient messages, particularly nonphysicians, templates are used to draft responses rather than a blank page; the presence of templates was not assessed, and future studies should treat templated HCP responses as a separate group for comparison. Furthermore, our study did not examine whether response quality varied with patient demographics.

Conclusions

In this study, PCPs' found EHR-integrated GenAI responses to private patient messages similar to HCPs in terms of information content quality, better with respect to communication style, and similar in their usability compared with starting from scratch. While poorly rated GenAI responses lacked relevance, were less helpful, or more verbose, they outperformed HCP responses in completeness, empathy, and professionalism. GenAI drafts acceptable to HCPs may offset the increasing workload (and diminishing well-being) they face from in-basket messages from patients. Future research should focus on optimizing the perceived quality of GenAI responses to end-users', particularly patients', perceptions; quantifying efficiency gains; and mitigating biases and hallucinations.

ARTICLE INFORMATION

Accepted for Publication: May 16, 2024.

Published: July 16, 2024. doi:10.1001/jamanetworkopen.2024.22399

Corresponding Author: William R. Small, MD, MBA, NYU Grossman School of Medicine, 550 1st Ave, New York, NY 10016 (will.small@nyulangone.org).

Author Affiliations: NYU Grossman School of Medicine, New York, New York (Small, Brandfield-Harvey, Jonassen, Mandal, Stevens, Major, Lostraglio, Szerencsy, Jones, Aphinyanaphongs, Johnson, Mann); NYU Stern School of Business, New York, New York (Wiesenfeld); NYU Tandon School of Engineering, New York, New York (Nov).

Author Contributions: Dr Small had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Small, Wiesenfeld, Brandfield-Harvey, Jonassen, Major, Lostraglio, Szerencsy, Aphinyanaphongs, Nov, Mann.

Acquisition, analysis, or interpretation of data: Small, Wiesenfeld, Jonassen, Mandal, Stevens, Major, Szerencsy, Jones, Johnson, Mann.

Drafting of the manuscript: Small, Brandfield-Harvey, Jonassen, Stevens, Major, Lostraglio, Jones, Mann.

Critical review of the manuscript for important intellectual content: Small, Wiesenfeld, Brandfield-Harvey, Jonassen, Mandal, Stevens, Major, Szerencsy, Aphinyanaphongs, Johnson, Nov, Mann.

Statistical analysis: Small, Mandal, Major, Jones, Mann.

Obtained funding: Jonassen, Nov, Mann.

Administrative, technical, or material support: Small, Wiesenfeld, Brandfield-Harvey, Jonassen, Mandal, Stevens, Major, Lostraglio, Szerencsy, Aphinyanaphongs, Nov, Mann.

Supervision: Small, Szerencsy, Johnson, Nov, Mann.

Conflict of Interest Disclosures: Dr Wiesenfeld reported receiving grants from the National Science Foundation (award Nos. 1928614 and 2129076) during the conduct of the study. Dr Jonassen reported receiving grants from the Swiss National Science Foundation General Postdoc Mobility Fellowship (award Nos. P500PS_202955 and P5R5PS_217714) during the conduct of the study. No other disclosures were reported.

Funding/Support: This work was supported by the Medical Center Information Technology Department of Health Informatics, NYU Langone Health.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

Additional Information: Chat GPT (GPT-4; OpenAI) was used to provide suggestions for editing the text, such as reducing word count, and code used to generate figures. GPT-4 was used from October 2023 to April 2024. The authors affirm that the original intent and meaning of the content remain unaltered during editing, and that ChatGPT had no involvement in shaping the intellectual content of this work. The authors assume full responsibility for upholding the integrity of the content presented in this manuscript.

Additional Contributions: We would like to thank our participants and colleagues for generously giving their time and sharing their insights with us. Without their contributions, this research would not have been possible. These include, but are not limited to, Arielle Elmaleh-Sachs, MD (NYU Grossman School of Medicine), Ann Garment, MD (Bellevue Hospital), Andrew Wallach, MD (Bellevue Hospital), Charles Gillihan, MD (NYU Grossman School of Medicine), Craig Tenner, MD (Veterans Affairs New York Harbor Healthcare System), Dany Haddad, MD (NYU Grossman Long Island School of Medicine), Holly Lofton, MD (NYU Grossman School of Medicine), Jessica Allan, MD (NYU Grossman School of Medicine), Jeffrey Friedman, MD (NYU Grossman School of Medicine), Mitchell Adler, MD, MPH (Bellevue Hospital), Margarita Rohr, MD (NYU Grossman School of Medicine), Nathanael Horne, MD (NYU Grossman School of Medicine), Rachael Hayes, MD (NYU Grossman School of Medicine), Xi Chu, MD (NYU Grossman School of Medicine), Sabrina Felson, MD (Veterans Affairs New York Harbor Healthcare System), Paolo Dib, MD (Bellevue Hospital), Jonathan So, MS (OptimizeRx), Nicholas Genes, MD, PhD (NYU Grossman School of Medicine), James Davydov, MSc (NYU Langone Medical Center), Gavriil Ilizarov, DO (NYU Grossman School of Medicine), Nina Singh, MD (UCSF School of Medicine), Anthony Cardillo, MD (NYU Grossman School of Medicine), Kiran Malhotra, MD (NYU Grossman School of Medicine), and Conner Polet, MD, MBA (NYU Grossman School of Medicine). These individuals were not compensated for their time.

REFERENCES

1. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. *J Am Med Inform Assoc*. 2022;29(3):453-460. doi: 10.1093/jamia/ocab268

2. Mandal S, Wiesenfeld BM, Mann DM, Szerencsy AC, Iturrate E, Nov O. Quantifying the impact of telemedicine and patient medical advice request messages on physicians' work-outside-work. *NPJ Digit Med*. 2024;7(1):35. doi:10.1038/s41746-024-01001-2

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. doi:10.1001/jamanetworkopen.2022.44363

4. 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. doi:10.1001/jamanetworkopen.2021.31490

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. doi:10.1377/hlthaff.2018.05509

6. Yan Q, Jiang Z, Harbin Z, Tolbert PH, Davies MG. Exploring the relationship between electronic health records and provider burnout: a systematic review. *J Am Med Inform Assoc*. 2021;28(5):1009-1021. doi:10.1093/jamia/ocab009

7. Akbar F, Mark G, Prausnitz S, et al. Physician stress during electronic health record inbox work: in situ measurement with wearable sensors. *JMIR Med Inform*. 2021;9(4):e24014. doi:10.2196/24014

8. Akbar F, Mark G, Warton EM, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. *J Am Med Inform Assoc*. 2021;28(5):923-930. doi:10.1093/jamia/ocaa229

9. Rittenberg E, Liebman JB, Rexrode KM. Primary care physician gender and electronic health record workload. *J Gen Intern Med*. 2022;37(13):3295-3301. doi:10.1007/s11606-021-07298-z

10. Escribe C, Eisenstat SA, O'Donnell WJ, Levi R. Understanding physicians' work via text analytics on EHR inbox messages. *Am J Manag Care*. 2022;28(1):e24-e30. doi:10.37765/ajmc.2022.88817

11. Escribe C, Eisenstat SA, Palamara K, et al. Understanding physician work and well-being through social network modeling using electronic health record data: a cohort study. *J Gen Intern Med*. 2022;37(15):3789-3796. Published online January 28, 2022. doi:10.1007/s11606-021-07351-x

12. 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. doi:10.1001/jama.2022.24710

13. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv*. Preprint published March 4, 2022.

14. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. Preprint posted online February 28, 2023. doi:10.21203/rs.3.rs-2566942/v1

15. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29(3):721-732. doi:10.3350/cmh.2023.0089

16. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388(13):1201-1208. doi:10.1056/NEJMra2302038

17. Li H, Rotenstein L, Jeffery MM, et al. Quantifying EHR and policy factors associated with the gender productivity gap in ambulatory, general internal medicine. *J Gen Intern Med*. 2024;39(4):557-565. doi:10.1007/s11606-023-08428-5

18. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. *JMIR Med Inform*. 2022;10(2):e32875. doi:10.2196/32875

19. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA*. 2023;330(4):315-316. doi:10.1001/jama.2023.9651

20. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

21. Dash D, Thapa R, Banda JM, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. . arXivPreprint updated May 1, 2023. doi:10.48550/arXiv.2304.13714

22. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. doi:10.1038/s41746-023-00873-0

23. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

24. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for diagnosing rare eye diseases? a pilot study. *Ophthalmol Ther*. 2023;12(6):3395-3402. doi:10.1007/s40123-023-00789-8

25. Liu S, Wright AP, Patterson BL, et al. Assessing the value of chatgpt for clinical decision support optimization. *medRxiv* Preprint posted online February 23, 2023. doi:10.1101/2023.02.21.23286254

26. 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. doi:10.1001/jamainternmed.2023.1838

27. Sorin V, Brin D, Barash Y, et al. Large language models (LLMs) and empathy—a systematic review. *medRxiv*. Preprint posted online August 7, 2023. doi:10.1101/2023.08.07.23293769

28. Nov O, Singh N, Mann D. Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study. *JMIR Med Educ*. 2023;9:e46939. doi:10.2196/46939

29. Copeland-Halperin LR, O'Brien L, Copeland M. Evaluation of Artificial Intelligence-generated Responses to Common Plastic Surgery Questions. *Plast Reconstr Surg Glob Open*. 2023;11(8):e5226. doi:10.1097/GOX.0000000000005226

30. 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. doi:10.1007/s11606-023-08271-8

31. Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial intelligence vs clinician performance in estimating probabilities of diagnoses before and after testing. *JAMA Netw Open*. 2023;6(12):e2347075. doi:10.1001/jamanetworkopen.2023.47075

32. 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. doi:10.1016/S2589-7500(24)00060-8

33. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw Open*. 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201

34. Tai-Seale M, Baxter SL, Vaida F, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw Open*. 2024;7(4):e246565. doi:10.1001/jamanetworkopen.2024.6565

35. Bruce G. Stanford to roll out ChatGPT-like feature for physicians next week. Becker's Health IT. May 8, 2023. Accessed June 14, 2024. https://www.beckershospitalreview.com/innovation/stanford-to-roll-out-chatgpt-like-feature-for-physicians-next-week.html

36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010

37. de Winter JFC, Dodou D. Five-point Likert Items: t test versus Mann-Whitney-Wilcoxon. *Pract Assess, Res Eval*. 2010;15:1-12.

38. Okeh UM. Statistical analysis of the application of Wilcoxon and Mann-Whitney U test in medical research studies. *Biotechnol Mol Biol Rev*. 2009;4(6):128-131.

39. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*. 1967;62(318):626-633. doi:10.2307/2283989

40. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23-34. doi:10.20982/tqmp.08.1.p023

41. Weiss SM, Indurkhya N, Zhang T, Damerau F. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer Science & Business Media; 2010.

42. Herbold S, Hautli-Janisz A, Heuer U, Kikteva Z, Trautsch A. A large-scale comparison of human-written versus ChatGPT-generated essays. *Sci Rep*. 2023;13(1):18617. doi:10.1038/s41598-023-45644-9

43. Fergadiotis G, Wright HH, West TM. Measuring lexical diversity in narrative discourse of people with aphasia. *Am J Speech Lang Pathol*. 2013;22(2):S397-S408. doi:10.1044/1058-0360(2013/12-0083)

44. Weiss Z, Riemenschneider A, Schröter P, Meurers D. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 30–45, doi:10.18653/v1/W19-4404

45. Koizumi R, In'nami Y. Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*. 2012;40(4):554-564. doi:10.1016/j.system.2012.10.012

46. Argamon S, Dhawle S, Koppel M, Pennebaker JW. Lexical predictors of personality type. In: Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America. January 2005.

47. Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of Flesch formula. *Educ Health (Abingdon)*. 2017;30(1):84-88. doi:10.4103/1357-6283.210517

48. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? *Clin Orthop Relat Res*. 2023;481(11):2260-2267. doi:10.1097/CORR.0000000000002668

49. Dudău DP, Sava FA. Performing multilingual analysis with linguistic inquiry and word count 2015 (LIWC2015): an equivalence study of four languages. *Front Psychol*. 2021;12:570568. doi:10.3389/fpsyg.2021.570568

50. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. *The Development and Psychometric Properties of LIWC-22*. University of Texas at Austin; 2022. doi:10.13140/RG.2.2.23890.43205

51. Pennebaker JW, Boyd RL, Booth RJ, Ashokkumar A, Francis ME. (2022). Linguistic Inquiry and Word Count: LIWC-22. Pennebaker Conglomerates. https://www.liwc.app

52. Boyd RL, Schwartz HA. Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. *J Lang Soc Psychol*. 2021;40(1):21-41. doi:10.1177/0261927X20967028

53. Boyd RL, Pasca P, Lanning K. The personality panorama: conceptualizing personality through big behavioural data. *Eur J Pers*. 2020;34(5):599-612. doi:10.1002/per.2254

54. Kanaparthi SD, Patle A, Naik KJ. Prediction and detection of emotional tone in online social media mental disorder groups using regression and recurrent neural networks. *Multimed Tools Appl*. Published online April 25, 2023. doi:10.1007/s11042-023-15316-x

55. He L, Zheng K. How do general-purpose sentiment analyzers perform when applied to health-related online social media data? *Stud Health Technol Inform*. 2019;264:1208-1212. doi:10.3233/SHTI190418

56. Ramya Sri VIS, Niharika C, Maneesh K, Ismail M. sentiment analysis of patients' opinions in healthcare using lexicon-based method. *Int J Eng Adv Technol*. 2019;9(1):6977. doi:10.35940/ijeat.A2141.109119

57. Dyrbye LN, Gordon J, O'Horo J, et al. Relationships between EHR-based audit log data and physician burnout and clinical practice process measures. *Mayo Clin Proc*. 2023;98(3):398-409. doi:10.1016/j.mayocp.2022.10.027

58. Decety J, Fotopoulou A. Why empathy has a beneficial impact on others in medicine: unifying theories. *Front Behav Neurosci*. 2015;8:457. doi:10.3389/fnbeh.2014.00457

59. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12-e22. doi:10.1016/S2589-7500(23)00225-X

60. Singh N, Lawrence K, Richardson S, Mann DM. Centering health equity in large language model deployment. *PLOS Digit Health*. 2023;2(10):e0000367. doi:10.1371/journal.pdig.0000367

SUPPLEMENT 1.

eAppendix 1. Survey Design Flow Diagram
eAppendix 2. Subgroup Analyses Stratified First by HCP Type (Physician vs Nonphysician), Then by Patient Message Subcategory
eAppendix 3. Linear Mixed Models With Random and Fixed Effects to Explore the Extent to Which Reviewer Variance or That Attributed to Patient Message (General Medical Advice, Laboratory Results, Medication Refill Requests, Paperwork) and HCP (Physician, Nonphysician) Subgroups Affected Final Results for the Main 3 Survey Questions
eAppendix 4. Select Free-Text Comments From Each Questions When Other Was Chosen in the Branching Logic
eAppendix 5. Exploration of Intraclass Correlation by Subgroup

SUPPLEMENT 2.

Data Sharing Statement