

PATIENT COMMUNICATION

AI-drafted replies to patient messages: what the deployments measured

ABSTRACT

Health systems are increasingly deploying language models that draft replies to patient portal messages for a clinician to edit and send. This paper appraises what the evaluations of that feature measured. Of 23 studies in a 2025 systematic review, one was randomized: drafts increased physicians' read time by 21.8% and left reply time unchanged. Log-based deployment studies show no consistent time saving; the reported relief of burden comes from uncontrolled surveys and divides sharply by role; and the comparisons that made the case rewarded longer answers. In a simulation, physicians missed two in three known draft errors. None of the deployments appraised audited the replies actually sent or measured an outcome in patients.

Type Systematic appraisal of the literature **References** 24 **Reading time** 14 min

Last reviewed September 2026

1 An increasingly deployed feature, thinly tested

A patient sends a question through the portal. Before a clinician opens it, a language model has written a reply, which the clinician can edit and send or discard. Health systems are increasingly adopting drafting of this kind.¹ It sits inside an existing workflow and keeps a person between the model and the patient, which makes it an early candidate for deployment. This paper appraises what its evaluations measured under four headings, clinician time, clinician burden, the quality of the reply and the safety of what is sent, and what they left unmeasured.

The evidence base is small and mostly descriptive. The one systematic review located for this paper identified 23 studies, all from the United States and published between 2023 and 2025; 7 were conducted in live electronic health records, 16 in simulated settings, and one used a randomized design.² That study, a randomized waiting-list quality improvement study, found that access to drafts increased physicians' time reading messages by 21.8%, left their time replying unchanged, and made their replies 17.9% longer.³ The searches for this paper, run in September 2026, found no later randomized trial of message drafting.

1 of 23

studies in a 2025 systematic review used a randomized design

+21.8%

physician read time with drafts available, in that randomized study

2.67 of 4

known draft errors missed per physician, on average, in a simulation

Four findings follow. Time savings have not been shown on objective measures: log-based studies find no change, more reading time, or a small decrease that cannot be separated from which messages clinicians chose to draft. The burden result is real but self-reported, uncontrolled and divided by professional role. The comparisons that made the case for the feature rewarded longer answers and did not separate length from empathy. And the safety risk sits in the edits a reviewer does not make: physicians reviewing drafts with known errors missed two in three of them,⁴ and no deployment appraised here audited the replies its clinicians actually sent.

2 The workload the feature was built for

The case for drafting starts from message volume. Holmgren and colleagues analyzed metadata on ambulatory clinicians in 366 health systems using the Epic record from December 2019 to December 2020.⁵ Messages from patients showed the largest increase of any inbox category, rising to 157% of the prepandemic average, and each additional patient message was associated with 2.32 more minutes of EHR time per day ($P < .001$). The 157% is a level, not a change. It describes an increase of 57%, a distinction easily lost when the figure is repeated.

The increase persisted. A research letter using Epic Signal data on 280,712 US ambulatory physicians found that patient medical advice requests rose in March 2020 and remained elevated, with no evidence that patients were substituting messages for telephone calls.⁶ Primary care physicians received a mean of 16 such messages and 24 patient calls a week, and their EHR time rose 6.5%, from 10.6 to 11.3 hours a week.

Patient messages are one stream in the inbox, and before the pandemic they were not the largest. In a multispecialty practice studied by Tai-Seale and colleagues, physicians received an average of 243 inbox messages a week: 114 generated by the EHR system, 53 from colleagues and 30 from patients.⁷ A drafting tool acts on the patient stream only, and on time per message rather than the number of messages.

Volume can be moved by other means, though weakly. When UCSF Health switched to clinician-initiated billing of portal exchanges as e-visits on 14 November 2021, the share of message threads billed rose from 0.26% to 1.40%.⁸ An interrupted time-series analysis found an immediate decrease of 3,388 message threads a week (95% CI $-4,781.7$ to $-1,994.7$; $P < .001$), with no significant level or trend change in scheduled visits or unscheduled telephone calls. The authors' own summary is that adoption was low.

3 The comparisons behind the feature rewarded length

Ayers and colleagues' comparison of physician and chatbot answers is not a study of patient messages. They drew 195 exchanges at random from Reddit's r/AskDocs forum, whose moderators verify responders' credentials, and compared each physician's public answer with a response from ChatGPT generated in December 2022.⁹ Licensed health care professionals from the study team, blinded to source, rated each pair in triplicate and preferred the chatbot in 78.6% of 585 evaluations (95% CI 75.0% to 81.8%), rated its answers good or very good in quality 78.5% of the time against 22.1% for physicians, and empathetic or very empathetic 45.1% of the time against 4.6%.

Three features of the design limit what those numbers mean. Physician answers averaged 52 words and chatbot answers 211, and when the comparison was restricted to physician answers above the 75th percentile of length, preference for the chatbot fell to 62.0% (95% CI 54.0% to 69.3%). The authors state that the additional length could have been erroneously associated with greater empathy, that the evaluators were also coauthors, and that accuracy and fabricated content were not assessed independently. And the exchanges were public answers to strangers, with no chart and no relationship behind them. It is easy to read the result as a patient preference; no patient rated anything in the study.

Studies nearer to practice reproduce the direction at smaller scale. At NYU Langone, 16 primary care physicians rated 344 real patient messages from three practices piloting the feature, each paired either with the clinician's reply or with a draft produced by the record's standard prompts.¹⁰ Drafts scored higher for communication style (3.70 against 3.38; $P = .01$) and no differently on information content or the share judged usable (0.69 against 0.65; $P = .49$). Among usable responses, 37.2% of drafts and 16.5% of clinician replies were rated empathetic. The drafts were numerically longer, 90.5 words against 65.4, though not significantly ($P = .07$), and significantly more linguistically complex ($P = .002$), which the authors flag as a concern for patients with low health or English literacy.

In an evaluation of 50 patient messages with negative sentiment, model drafts averaged 119 words against clinicians' 42, and the authors observed that some drafts could have escalated emotionally charged exchanges.¹¹ In a simulation built on 100 cancer-patient messages, radiation oncologists' unaided replies averaged 34 words, GPT-4 drafts 169, and the same physicians' edited drafts 160 ($p < 0.0001$).¹² The last figure matters more than the first two: the edited reply kept nearly all of the draft's length. In the one randomized deployment, replies sent with drafts available were 17.9% longer (95% CI 10.1% to 26.2%).³

TABLE 1

Reply length in studies comparing model text with clinicians' own replies. Model text was longer in every comparison; only one study tried to separate length from preference.

Study	Material	Clinician reply	Model text	What was rated or found
Ayers 2023 ⁹	195 public forum exchanges	52 words	211 words	Chatbot preferred 78.6%; 62.0% vs longest physician answers
Chen 2024 ¹²	100 simulated oncology messages	34 words	169 words; 160 after editing	Edited replies kept the draft's length
Baxter 2024 ¹¹	50 negative-sentiment messages	42 words	119 words	Some drafts could have escalated the exchange
Small 2024 ¹⁰	344 real portal messages	65.4 words	90.5 words (P = .07)	Empathetic: 37.2% vs 16.5% of usable replies
Tai-Seale 2024 ³	Live replies, randomized access	Baseline	Sent replies +17.9% longer	Reply quality not an outcome

WHAT THE PREFERENCE STUDIES ESTABLISH

They establish that raters, mostly clinicians, prefer model text read in isolation, sometimes by large margins. They do not establish that the preference survives adjustment for length, that it reflects accuracy, or that patients receiving a reply within an ongoing relationship value the same things. Empathy in these studies is a rating of text, not an outcome measured in patients, and the one study that tried to separate it from length, by restriction, found the preference shrank.

4 Time: what the logs show

Eight evaluations of drafting in live use report utilization, time or both (Table 2). Seven name the EHR vendor's integrated feature, Epic's, as the drafting pipeline,^{13,14,15,16,17,18,19} two of them with eligible messages routed to the vendor for classification and drafting.^{13,15} They evaluate a vendor product, and in at least one the prompts were revised while the study ran.¹⁵

How often the draft is used

In US deployments, reported utilization ranged from 12% to about 20%, although the denominator varies: replies sent in some studies, drafts generated in others.^{13,14,15,16,19} At NYU Langone, prompt revisions raised use from 12% to 20%, and drafts were generated for every eligible message, including

roughly 80% (44,454) that were resolved without any reply, each adding to the review burden.¹⁵ Definitions differ enough to prevent pooling: one Dutch hospital counted a draft as used in 58% of replies, of which 43% retained a large part of it,¹⁷ and a second reported adoption that began high and declined to an average of 16.7%.¹⁸

How long a message takes

At Stanford, where 162 clinicians used drafts for five weeks, audit logs showed no change in reply action time, write time or read time between the prepilot and pilot periods.¹³ At UC San Diego, 52 primary care physicians were randomized to immediate or delayed activation, with 70 contemporaneous physicians as controls.³ The hypothesis stated in advance was that drafts would reduce time spent reading and replying. Read time rose 21.8% (95% CI 5.2% to 41.0%; $P = .008$) and reply time changed by -5.9% (95% CI -16.6% to 6.2% ; $P = .33$). In absolute terms the change was small: median read time in the immediate group went from 26 seconds at baseline to 31 seconds at 3 and at 6 weeks, while the control group moved from 21 to 23 seconds.

The Dutch cohort of 919 messages found total response time of 157 seconds with a blank reply and 153 seconds with a draft that matched the sent reply by at least 10% ($p = 0.69$).¹⁷ A second Dutch evaluation, a preprint covering 8,410 drafts over six months, found review and drafting times similar for users with and without the tool.¹⁸

One observational result favors the drafts. In the NYU audit-log study, replies that used a draft had a 6.76% shorter median turnaround, 331 against 355 seconds, while median message open time was 7.27% longer, 59 against 55 seconds.¹⁵ Although adjusted for patient and draft complexity, it compares messages on which clinicians chose to use a draft with those on which they did not, so it cannot fully separate the draft's effect from the choice of message. Self-report runs higher than any log: in a pediatric and obstetric pilot, clinicians reported perceived savings of 1.2 and 0.74 minutes per reply.¹⁹ No log-based measure in Table 2 shows a saving of that size.

TABLE 2

Live deployments of drafted replies: what each measured. A dash means the outcome was not reported in the sources read for this paper.

Study	Design	Users	Draft use	Objective time	Self-report
Garcia 2024 ¹³	Single-group pre-post, 5 weeks	162 clinicians	20%	No change in read, write or reply time	Task load −13.87; exhaustion −0.33
Tai-Seale 2024 ³	Randomized waiting list, with controls	52 randomized; 70 controls	—	Read +21.8%; reply −5.9% (NS)	Value recognized; improvements suggested
English 2024 ¹⁴	Quality improvement; survey at 2 weeks	166 users	12% of 21,323 drafts	—	Net promoter: nurses 58; clinicians −43
Liang 2025 ¹⁹	Prospective pilot	61 clinicians	13.3% and 18.3%	—	Pediatric task load down; burnout unchanged
Mandal 2025 ¹⁵	Retrospective audit logs	75 professionals	19.4%	Turnaround −6.76%; open time +7.27%	—
Crowley 2026 ¹⁶	Pilot with survey	80 users	20.2%	—	66% useful; 46% better quality
Bootsma-Robroeks 2025 ¹⁷	Prospective cohort, 16 weeks	100 physicians	58% of replies	157 s vs 153 s (p = 0.69)	—
Bladder 2026 ¹⁸	Effectiveness—implementation; preprint	237 professionals	16.7%, declining	Similar with and without	Perceived efficiency declined

5 The burden result is self-reported and divided by role

The favorable findings in this literature concern how clinicians feel, and they are real. At Stanford, the 73 clinicians (45.1%) who completed both surveys reported a four-item task-load score falling from 61.31 to 47.26 (paired difference −13.87; 95% CI −17.38 to −9.50) and work exhaustion from 1.95 to 1.62 (−0.33; 95% CI −0.50 to −0.17), both $P < .001$.¹³ The same study found no change in time. It was a single-group, five-week pre-post comparison with no control group, and more than half the clinicians analyzed did not complete both surveys.

Results elsewhere are narrower. In the pediatric and obstetric pilot, pediatric clinicians reported lower task load on the NASA Task Load Index (59.9 to 50.9; $P = .04$), while burnout scores did not change significantly in either group.¹⁹ In the Dutch preprint, perceived clinical efficiency declined significantly between baseline and four months, well-being did not change, and usability and intention-to-use scores fell.¹⁸

The sharpest pattern is by role. At UCHHealth, a survey two weeks after launch found a net promoter score of 58 among nurses, −29 among medical assistants and −43 among physicians and advanced practice clinicians ($P = .004$).¹⁴ Eleven of 12 nurses agreed they could reply more quickly, against 20 of 43 clinicians (46%). Eight of 12 nurses but 12 of 43 clinicians (28%) agreed that the risk of giving incorrect

information was minimal, and 5 of 43 clinicians (12%) agreed that significant edits were rarely needed. In a Pennsylvania pilot, 66% of survey respondents found the drafts useful but 46% thought they improved reply quality, and nurses were more favorable than physicians.¹⁶

“Drafting reduces burden” is therefore not one finding. In the study that compared roles directly, net promoter scores were favorable among nurses and unfavorable among medical assistants and clinicians.¹⁴ Whether that gradient reflects how close each role's replies are to a template was not tested in the studies appraised here.

Institutional communication has run ahead of the measurements. The UC San Diego press release for the randomized study acknowledged that drafts did not reduce response time, then quoted the senior author as saying that longer messages suggest higher quality and that physicians' appreciation of the help lowered cognitive burden.²⁰ The study's main outcomes were read time, reply time, reply length and likelihood to recommend the drafts.³ Neither reply quality nor cognitive burden was among them.

6 The safety risk sits in the edits not made

A draft is as safe as the reviewer's ability to catch what is wrong with it. Two simulation studies tested that ability. Chen and colleagues had six radiation oncologists answer simulated cancer-patient messages, 100 in all and themselves generated by GPT-4, first unaided and then by editing GPT-4 drafts of replies to the same messages.¹² The assessing physicians judged that the drafts posed a risk of severe harm in 11 of 156 assessments (7.1%) and of death in one (0.6%), mostly because the drafts misjudged or misstated the acuity of the scenario and the action it called for. The 7.1% is a share of physician assessments, not a per-draft rate of any harm, and it comes from simulated messages.

The same study found reviewers moving toward the draft. Compared with unaided replies, drafts were less likely to tell patients to present for evaluation, urgently or not, or to describe an action the clinician would take, and more likely to offer education, self-management advice and a contingency plan. Agreement between physicians on the clinical content of replies rose from a mean kappa of 0.10 unaided to 0.52 with drafts. The authors' interpretation is that physicians may adopt the model's assessment rather than use the model to phrase their own.

Biro and colleagues measured detection directly. Twenty practicing primary care physicians handled 18 portal messages in a simulated portal modeled on a commercial record, each with a draft generated by ChatGPT-4.0. An experienced primary care physician had identified errors in four of the drafts in advance, each an objective inaccuracy or a potentially harmful omission.⁴ Participants missed an average of 2.67 of the 4 errors (66.6%). Each erroneous draft was missed by at least 13 of the 20 physicians and sent entirely unedited by at least 7. In the same study, 19 of the 20 found the drafts helpful and 15 agreed or strongly agreed that they were safe to use.

These are the conditions under which automation bias is expected. Lyell and Coiera's systematic review of 40 studies found it in single tasks with high verification complexity, contrary to the prevailing human-factors view, and not uniquely associated with multitasking.²¹ Checking a fluent reply against a chart for a missing instruction is such a task, because an omission leaves no trace in the text. In the Pennsylvania deployment, 40% of replies showed little or no editing of the draft, and respondents reported drafts with incorrect or inappropriate content;¹⁶ how many of the unedited replies were correct is not known. Physicians interviewed during another health system's pilot placed ethical responsibility for drafted replies primarily with the user rather than the technology,²² which is also where these studies locate the failure.

Not one of the eight deployments in Table 2 reports, among its outcomes, an audit of sent replies for clinical error. Their outcomes are utilization, time, textual similarity and surveys. The Dutch cohort's abstract states that its results implicate safe use, but its measured outcomes were adoption, similarity of sent text to the draft, and time.¹⁷ On this evidence the rate of harmful error in replies that reached patients is unknown.

WHAT THE SIMULATION STUDIES DO NOT ESTABLISH

Both safety studies used simulated messages and a general-purpose model rather than a record-integrated feature, and each rests on 100 messages or fewer. They measure what reviewers do when a draft is wrong, not how often deployed drafts are wrong, and neither supports a claim that deployed drafting is unsafe. What they do show is that “a clinician reviews every draft” is not, by itself, evidence that what is sent is safe.

7 What patients were asked, and what was never measured in them

Patients enter this literature as survey respondents. Of 2,511 members of Duke University Health System's patient advisory committee, 1,455 (57.9%) responded to a survey experiment that varied the author of a message, its disclosure and the seriousness of the topic.²³ They showed a mild preference for AI-drafted responses, with differences of 0.30 points for satisfaction, 0.28 for usefulness and 0.43 for feeling cared for. Satisfaction was 0.13 points higher (95% CI 0.05 to 0.22) when a message was attributed to a human than when AI authorship was disclosed. More than 75% were satisfied whatever the author or disclosure, and respondents were older and more educated than nonrespondents.

Interviews with 40 patients previously surveyed about drafted messaging at an academic health system, recruited to over-represent minoritized, older and less satisfied respondents, found that patients regard portal messaging as transactional, prioritizing timely resolution over relational depth.¹ Their preferences

for tone and empathy depended on whether length and detail fit the purpose and stakes of the message more than on whether it read as human or machine. They were comfortable with drafts on condition of clinician review, and broadly wanted disclosure. If that finding generalizes, the reply worth optimizing is correct, readable and prompt, which is not what the preference studies in section 3 rewarded.

No deployment appraised here measured an outcome in patients: not the number of follow-up messages a reply generated, not comprehension, not whether advice was acted on, and not any clinical event. The randomized study's authors call for future work on patient experience.³ Readability has been measured only as a property of text, and in the direction of concern: drafts were more linguistically complex than clinician replies.¹⁰

8 What would settle it

The gaps are specific. The first is a randomized comparison powered for objective time, with the clinician or the message as the unit, that counts the whole cost of a reply: reading the message, reading the draft, editing it, and handling the follow-up the reply generates. The one randomized study measured read and reply time in 52 physicians over six weeks;³ the audit logs needed to extend it are the ones the deployment studies already used.

The second is an audit of what was sent. A sample of sent replies, drafted and not, reviewed blind for clinical error and graded for severity, with particular attention to acuity and omitted instructions, would turn the simulation findings into a rate.^{4,12} Without it, the statement that every draft is reviewed describes a workflow rather than a measured performance.

Third, results should be reported by role and message type, since the favorable results cluster among nurses^{14,16} and drafts generated for messages that never receive a reply are pure review cost.¹⁵ Fourth, patient outcomes: time to resolution, follow-up volume, comprehension and readability measured in the people who read the reply. Fifth, versioning: prompts changed during at least one study,¹⁵ so any result belongs to a model and a prompt, not to the category. DECIDE-AI, the reporting guideline for early-stage clinical evaluation of AI decision support, is the nearest existing framework.²⁴

Until then, the defensible summary is narrow. In most deployments drafting is used for a minority of replies, and it is liked, most of all by nurses. It has not been shown to save time on objective measures, and in the one randomized study located it increased reading time. Its reported relief of burden comes from uncontrolled surveys, and its safety rests on a review step that, tested against drafts with known errors, missed most of them. A claim that a drafting feature saves time or is safe should be met with a request for the log-based comparison and the sent-message audit behind it.

REFERENCES

Entry 3 is the one randomized study located and carries the causal claims about time; entries 13 to 19 are deployment evaluations without randomization, and entry 2 is the systematic review that frames them. Entries 4 and 12 are the simulation studies on which section 6 rests. Entries 18 and 22 are preprints; entry 11 is a perspective with a small descriptive evaluation; entry 20 is a press release, cited for what it claimed rather than as evidence; and entry 24 is a reporting guideline. Most deployment studies evaluated one EHR vendor's integrated feature; none of the three disclosure statements that could be read for this paper (entries 14, 17 and 19) reported a financial relationship with that vendor.

1. Owens K, Jayaram A, Chowdhury A, et al. *Patient Perspectives on AI-Drafted Electronic Portal Messages*. JAMA Network Open. 2026;9(7):e2622463. [doi:10.1001/jamanetworkopen.2026.22463](https://doi.org/10.1001/jamanetworkopen.2026.22463) QUALITATIVE STUDY
2. Hu D, Guo Y, Zhou Y, et al. *A systematic review of early evidence on generative AI for drafting responses to patient messages*. npj Health Systems. 2025;2(1):27. [doi:10.1038/s44401-025-00032-5](https://doi.org/10.1038/s44401-025-00032-5) SYSTEMATIC REVIEW
3. Tai-Seale M, Baxter SL, Vaida F, et al. *AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication*. JAMA Network Open. 2024;7(4):e246565. [doi:10.1001/jamanetworkopen.2024.6565](https://doi.org/10.1001/jamanetworkopen.2024.6565) RANDOMIZED QI STUDY
4. Biro JM, Handley JL, McCurry JM, et al. *Opportunities and risks of artificial intelligence in patient portal messaging in primary care*. npj Digital Medicine. 2025;8:222. [doi:10.1038/s41746-025-01586-2](https://doi.org/10.1038/s41746-025-01586-2) CROSS-SECTIONAL
5. Holmgren AJ, Downing NL, Tang M, et al. *Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use*. Journal of the American Medical Informatics Association. 2022;29(3):453–460. [doi:10.1093/jamia/ocab268](https://doi.org/10.1093/jamia/ocab268) RETROSPECTIVE COHORT
6. Holmgren AJ, Apathy NC, Sinsky CA, et al. *Trends in Physician Electronic Health Record Time and Message Volume*. JAMA Internal Medicine. 2025;185(4):461–463. [doi:10.1001/jamainternmed.2024.8138](https://doi.org/10.1001/jamainternmed.2024.8138) RETROSPECTIVE COHORT
7. 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 Affairs. 2019;38(7):1073–1078. [doi:10.1377/hlthaff.2018.05509](https://doi.org/10.1377/hlthaff.2018.05509) CROSS-SECTIONAL
8. Holmgren AJ, Byron ME, Grouse CK, et al. *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](https://doi.org/10.1001/jama.2022.24710) QUASI-EXPERIMENTAL
9. 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 Internal Medicine. 2023;183(6):589–596. [doi:10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838) CROSS-SECTIONAL

10. Small WR, Wiesenfeld B, Brandfield-Harvey B, et al. *Large Language Model–Based Responses to Patients’ In-Basket Messages*. JAMA Network Open. 2024;7(7):e2422399. [doi:10.1001/jamanetworkopen.2024.22399](https://doi.org/10.1001/jamanetworkopen.2024.22399) CROSS-SECTIONAL
11. Baxter SL, Longhurst CA, Millen M, et al. *Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned*. JAMIA Open. 2024;7(2):ooae028. [doi:10.1093/jamiaopen/ooae028](https://doi.org/10.1093/jamiaopen/ooae028) POSITION PAPER
12. Chen S, Guevara M, Moningi S, et al. *The effect of using a large language model to respond to patient messages*. The Lancet Digital Health. 2024;6(6):e379–e381. [doi:10.1016/S2589-7500\(24\)00060-8](https://doi.org/10.1016/S2589-7500(24)00060-8) CROSS-SECTIONAL
13. Garcia P, Ma SP, Shah S, et al. *Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages*. JAMA Network Open. 2024;7(3):e243201. [doi:10.1001/jamanetworkopen.2024.3201](https://doi.org/10.1001/jamanetworkopen.2024.3201) PRE-POST EVALUATION
14. English E, Laughlin J, Sippel J, et al. *Utility of Artificial Intelligence–Generative Draft Replies to Patient Messages*. JAMA Network Open. 2024;7(10):e2438573. [doi:10.1001/jamanetworkopen.2024.38573](https://doi.org/10.1001/jamanetworkopen.2024.38573) SURVEY
15. Mandal S, Wiesenfeld BM, Szerencsy AC, et al. *Utilization of Generative AI-drafted Responses for Managing Patient-Provider Communication*. npj Digital Medicine. 2025;8:591. [doi:10.1038/s41746-025-01972-w](https://doi.org/10.1038/s41746-025-01972-w) RETROSPECTIVE COHORT
16. Crowley AP, Hanish A, Lubken J, et al. *Usage of and Satisfaction with Artificial Intelligence–Generated Draft Replies to Patient Portal Messages*. Applied Clinical Informatics. 2026;17(3):510–517. [doi:10.1055/a-2893-7363](https://doi.org/10.1055/a-2893-7363) CROSS-SECTIONAL
17. Bootsma-Robroeks CMHHT, Workum JD, Schuit SCE, et al. *AI-generated draft replies to patient messages: exploring effects of implementation*. Frontiers in Digital Health. 2025;7:1588143. [doi:10.3389/fdgth.2025.1588143](https://doi.org/10.3389/fdgth.2025.1588143) COHORT
18. Bladder KJM, Verburg AC, Arts-Tenhagen M, et al. *AI-Generated Responses to Patient’s Messages: Effectiveness, Feasibility and Implementation*. medRxiv. Posted 2 March 2026; not peer reviewed. [doi:10.64898/2026.03.02.26347175](https://doi.org/10.64898/2026.03.02.26347175) PREPRINT
19. Liang AS, Vedak S, Dussaq A, et al. *Artificial intelligence-generated draft replies to patient messages in pediatrics*. JAMIA Open. 2025;8(6):ooaf159. [doi:10.1093/jamiaopen/ooaf159](https://doi.org/10.1093/jamiaopen/ooaf159) PRE-POST EVALUATION
20. UC San Diego Health. *Study Reveals AI Enhances Physician–Patient Communication*. News release, 15 April 2024. health.ucsd.edu PRESS RELEASE

21. Lyell D, Coiera E. *Automation bias and verification complexity: a systematic review*. Journal of the American Medical Informatics Association. 2017;24(2):423–431. [doi:10.1093/jamia/ocw105](https://doi.org/10.1093/jamia/ocw105)

SYSTEMATIC REVIEW

22. Hu D, Guo Y, Cho HN, et al. *When AI Writes Back: Ethical Considerations by Physicians on AI-Drafted Patient Message Replies*. arXiv. Submitted 17 August 2025. [arXiv:2508.13217](https://arxiv.org/abs/2508.13217)

PREPRINT

23. Cavalier JS, Goldstein BA, Ravitsky V, et al. *Ethics in Patient Preferences for Artificial Intelligence–Drafted Responses to Electronic Messages*. JAMA Network Open. 2025;8(3):e250449. [doi:10.1001/jamanetworkopen.2025.0449](https://doi.org/10.1001/jamanetworkopen.2025.0449)

SURVEY

24. Vasey B, Nagendran M, Campbell B, et al. *Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI*. Nature Medicine. 2022;28(5):924–933. [doi:10.1038/s41591-022-01772-9](https://doi.org/10.1038/s41591-022-01772-9)

REPORTING GUIDELINE

RELATED PAPERS

Ambient documentation and the clinician's day

The same pattern in notes: small, product-specific effects on time

Human oversight of clinical AI

Why review misses the errors that matter, in controlled experiments

Disclosing AI to patients

What patients are told when a model drafted the reply