

Human-AI teaming to improve accuracy and efficiency of eligibility criteria prescreening for oncology trials: a randomized evaluation trial using retrospective electronic health records

Received: 7 March 2025

Accepted: 19 January 2026

Published online: 03 February 2026

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Few adult patients with cancer enroll in oncology clinical trials. A rate-limiting step to trial enrollment is prescreening, involving clinical research staff manually abstracting unstructured health records to identify patients who meet eligibility criteria. Prescreening is time-consuming, labor-intensive, and prone to human error, resulting in under-identification of eligible patients. Neurosymbolic AI language models may approximate or improve the accuracy of prescreening through automated abstraction of enrollment criteria from longitudinal unstructured patient charts. We conduct a randomized non-inferiority trial using retrospectively collected clinical charts to compare the accuracy and efficiency of prescreening by trained research staff alone (Human-alone) vs. augmented with a pre-trained language model (Human+AI), among a cohort of 355 patients with non-small cell lung or colorectal cancer. Sample size is determined from analyses of a preliminary dataset as well as a prespecified, interim dataset of 74 charts. Chart-level accuracy, the primary endpoint of Human+AI prescreening is noninferior and superior to Human-alone (76.5% vs. 71.1%). However, efficiency is unchanged with similar average time per chart review, the secondary endpoint, (37.4 vs. 37.8 min). AI-assisted abstraction most improves accuracy for biomarker, staging, and response criteria. Performance is limited in some domains due to automation bias. Although improvements are modest, this large randomized trial evaluating a human-AI framework for oncology prescreening shows that AI language models can approximate and augment human-driven prescreening to enhance identification of trial-eligible patients, potentially increasing enrollment. The trial is registered on ClinicalTrials.gov (NCT06561217).

Participation in clinical trials of therapeutic and diagnostic interventions is associated with improved outcomes among patients with cancer¹⁻³. While over 70% of patients are inclined to participate in clinical trials, fewer than 10% of adult patients with cancer eventually

enroll into such trials, with racial, ethnic, and geographic disparities^{1,4}. Inefficient and labor-intensive eligibility assessment, known as “prescreening,” is a barrier to higher trial participation. Prescreening is traditionally done by clinicians or clinical research coordinators

 e-mail: ravi.bharat.pariikh@emory.edu

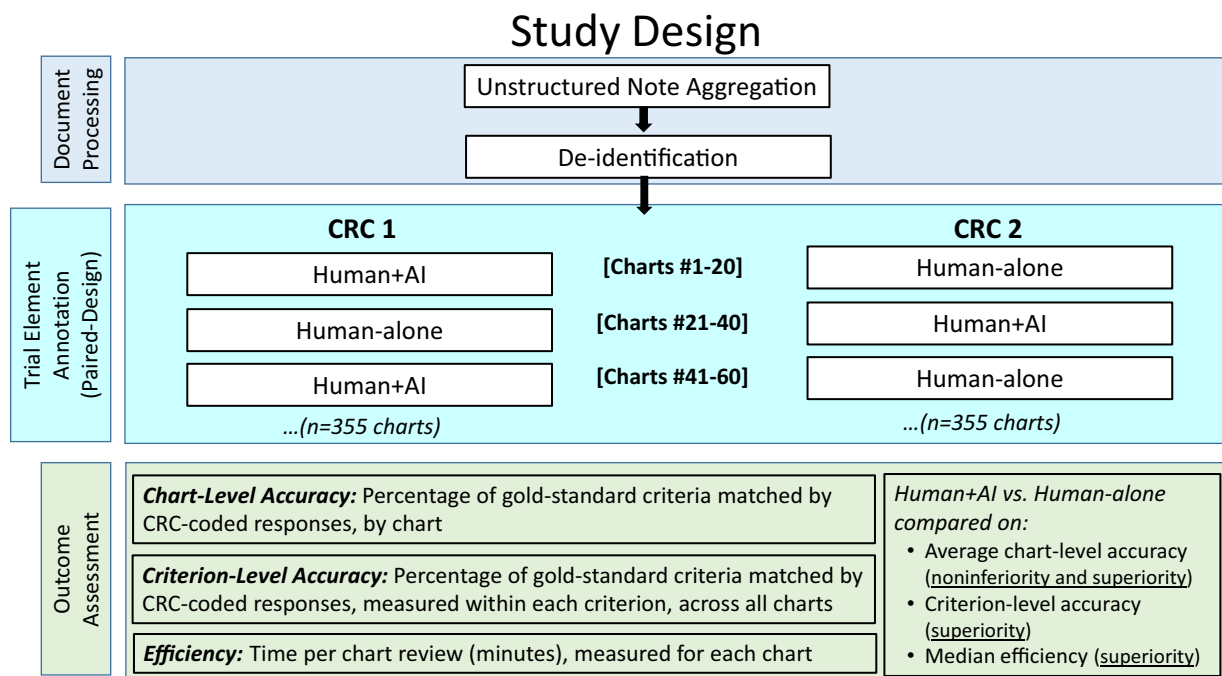


Fig. 1 | Study Design. De-identified EHRs from patients with NSCLC or colorectal cancer were processed using an optical character recognition system and analyzed with neurosymbolic AI to extract 12 trial eligibility criteria. Clinical research

coordinators reviewed randomized charts with AI augmentation or without AI support. Primary accuracy and efficiency assessments compared the Human-AI and the Human-Alone arms.

(CRCs), who manually review health records to identify patients meeting eligibility criteria. There are challenges to this approach. Clinical notes (e.g., radiology reports, progress notes) can be disorganized, unstructured, and contradictory, making identification of relevant elements challenging. Furthermore, prescreening is time-consuming, subject to human error and bias, and reliant on costly human teams. These limitations decrease the efficiency, accuracy, and inclusivity of enrollment⁵.

Artificial intelligence (AI) language models may approximate or improve the accuracy and efficiency of data extraction for clinical trial prescreening. Natural language processing (NLP) algorithms can automatically screen unstructured text and extract specific eligibility criteria—such as performance status, disease stage, and medications. Studies using NLP eligibility criteria models have reported recall and precision rates of 90%—which is similar to CRC performance alone⁶. Neurosymbolic language models are a type of NLP system that combine the pattern recognition strengths of neural networks with the logical reasoning capabilities of symbolic AI and may allow similar or more accurate interpretation of complex oncology eligibility criteria from longitudinal patient charts than traditional NLP methods alone^{7,8}.

While retrospective assessments of AI language model accuracy abound, prospective evaluations of AI-assisted prescreening workflows are rare and have limitations. Previous studies involve single-arm comparisons without a concurrent control, evaluate algorithms extracting limited eligibility criteria, and compare only human versus traditional NLP screening^{9–14}. Furthermore, it is unknown whether an AI-augmented prescreening workflow results in time savings to CRCs. There is an urgent need for controlled studies assessing a realistic workflow of AI augmenting human prescreening tasks across multiple cancers and eligibility criteria.

In this work, we conducted a randomized noninferiority trial using retrospectively collected, complete electronic health records (EHRs) of patients with cancer to compare accuracy (primary) and efficiency (secondary) of three prescreening approaches: (1) an autonomous AI

algorithm; (2) a human CRC; and (3) a CRC utilizing AI augmentation (Human+AI). We hypothesized that Human+AI prescreening would be non-inferior in chart-level prescreening accuracy and superior in efficiency compared to a Human-alone approach.

Results

The study design is summarized in Fig. 1 and in detail in the Methods. Briefly, the study cohort consisted of patients with non-small cell lung (NSCLC) and colorectal (CrCa) cancer who received systemic therapy in a large community oncology clinic. Full electronic health records were downloaded retrospectively in PDF form, processed through an optical character recognition (OCR) system, and de-identified. Next, we used pre-trained neurosymbolic AI algorithms to extract discrete elements from clinical notes corresponding to 12 eligibility criteria for Phase I-III oncology clinical trials. To assess the primary and secondary outcomes, complete de-identified EHRs were presented to two CRCs in the same order. Charts were randomized in blocks of 20 to be viewed with AI augmentation (Human+AI) or without AI augmentation (Human-alone). Prior to accuracy assessment, to ensure no inter-rater differences in accuracy, concordance in chart-level accuracy was assessed on eight charts not included in the final analysis. Cohen's Kappa was 0.72 and percent agreement was 86.1%, meeting our prespecified thresholds of 0.60 and 80% that indicate good inter-rater agreement based on prior literature¹⁵. The outcomes of accuracy and efficiency were assessed in relation to a gold-standard dataset, which was developed prior to trial initiation by three expert clinician reviewers. Although the study was not registered initially, given its use of retrospective data, to ensure it was appropriately reflected in the trial literature, it was registered on ClinicalTrials.gov (NCT06561217) after completion with the registration parameters based on the study protocol.

Accuracy assessment

The primary outcome of chart-level accuracy across the study arms is summarized in Fig. 2. 355 charts (195 NSCLC; 160 CrCa) were reviewed

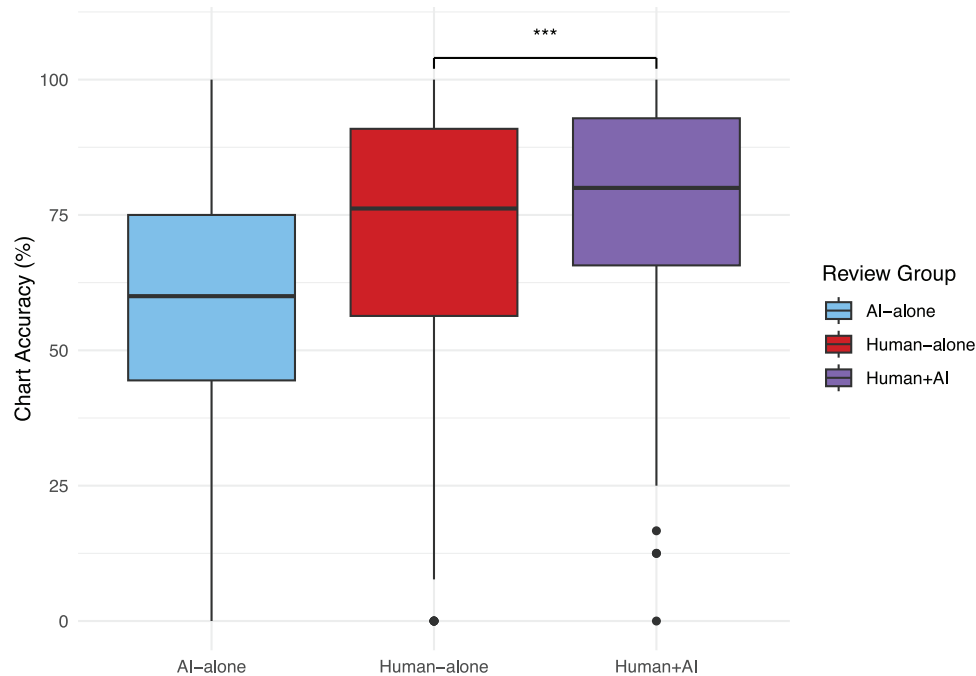


Fig. 2 | Chart-level accuracy comparisons between AI-alone, Human-alone, and Human+AI arms. Chart-level accuracy comparisons between AI-alone (blue), Human-alone (red), and Human+AI (purple) arms. $N = 355$ patient charts were independently reviewed in each arm. Each chart constitutes an independent observational unit, and no technical replicates were generated. Boxplots display

the median, 25th–75th percentiles (interquartile range), and whiskers representing the minimum and maximum observed values. Horizontal line inside box denotes median accuracy value. Horizontal black bar with asterisk indicates significant noninferiority ($p < 0.001$), tested with a one-sided, paired Wilcoxon Rank Sum test ($\alpha = 0.05$).

in a paired design across arms. Accuracy for the Human+AI arm (76.1%) was noninferior to the Human-alone arm (71.5%) (statistic = 43653, $p < 0.001$, Rosenthal's $r = 0.338$, 95% CI = [0.00, Inf]). In a subsequent test for superiority, Human+AI was significantly more accurate than Human-alone (statistic = 18,436, $p = 0.002$, Rosenthal's $r = 0.165$, 95% CI = [0.02, Inf]).

A total of 12 criteria were included in the gold-standard set, spanning concepts related to biomarkers, neoplasm characteristics, patient performance, and response to previous treatment (see Supplemental Table 1 for descriptions of criteria). Accuracy across criteria is summarized in Table 1. Significant increases in criterion-level accuracy in favor of the Human+AI vs. Human-alone arms were observed for 7 criteria: The presence of biomarker testing (93.2% vs. 84.6%, statistic = 57/75, $p < 0.001$, Cohen's $h = 0.547$, 95% CI = [0.65, 0.85]); biomarker results (79.0% vs. 67.9%, statistic = 113/176, $p < 0.001$, Cohen's $h = 0.288$, 95% CI = [0.57, 0.71]); interpretation of biomarker result (91.3% vs. 80.8%, statistic = 67/94, $p < 0.001$, Cohen's $h = 0.440$, 95% CI = [0.61, 0.80]); M-staging (57.0% vs. 43.9%, statistic = 38/47, $p < 0.001$, Cohen's $h = 0.665$, 95% CI = [0.67, 0.91]); N-staging (66.3% vs. 50.5%, statistic = 40/49, $p < 0.001$, Cohen's $h = 0.685$, 95% CI = [0.68, 0.91]); T-staging (71.6% vs. 56.3%, statistic = 45/60, $p < 0.001$, Cohen's $h = 0.524$, 95% CI = [0.62, 0.85]); and clinical outcome (35.9% vs. 23.7%, statistic = 41/60, $p = 0.006$, Cohen's $h = 0.375$, 95% CI = [0.55, 0.80]). We did not observe significant differences in criterion-level accuracy for Eastern Cooperative Oncology Group (ECOG) Performance Status, medications, cancer type, stage group, or response to therapy.

The mean chart-level accuracy in the AI-alone arm was 59.9%, lower than the Human-alone (71.5%) and Human+AI arms (76.1%). However, the AI-alone arm achieved the highest criterion-level accuracy in three criteria—Neoplasm M staging (60.2% AI-alone, 57.0% Human+AI, 43.9% Human-alone), clinical outcome (55.2% AI-alone, 35.9% Human+AI, 23.7% Human-alone), and response (60.4% AI-alone, 51.7% Human+AI, 47.1% Human-alone). Correspondingly, for all of

these three criteria, the Human+AI arm had higher accuracy than in the Human-alone arm.

When comparing chart-level accuracy of Human+AI vs. AI-alone, greatest accuracy gains were observed for interpretation of biomarker result (91.3% vs 35.7%), biomarker values (79.0% vs 32.5% AI-alone), and ECOG status (84.7% vs 34.4%). All criteria with significant differences were both noninferior and statistically superior for the Human+AI arm.

Subgroup analyses investigated the effects of chart characteristics—order, length, complexity, and cancer type—on chart-level accuracy within each study arm (Supplemental Table 2). Accuracy was lower for longer vs. shorter charts in the Human+AI (74.0% vs. 78.2%,

statistic = 18477, $p = 0.005$, Rosenthal's $r = 0.213$, 95% CI = [0.83, 10.71]) and Human-alone (68.8% vs. 74.2%, statistic = 18892, $p = 0.001$, Rosenthal's $r = 0.245$, 95% CI = [3.33, 12.59]) arms. Accuracy was similar between arms for earlier vs. later charts (Human+AI: 77.5% to 74.7%, statistic = 16498, $p = 0.438$, Rosenthal's $r = 0.058$, 95% CI = [−0.85, 5.56]; Human-alone: 73.8% to 69.1%, statistic = 17025, $p = 0.186$, Rosenthal's $r = 0.099$, 95% CI = [−0.00, 8.75]); and less vs. more complex charts (Human+AI: 76.1% vs 76.1%, statistic = 16716, $p = 0.243$, Rosenthal's $r = 0.092$, 95% CI = [−1.43, 7.14]; Human-alone 71.1% vs 71.9%, statistic = 16482, $p = 0.357$, Rosenthal's $r = 0.073$, 95% CI = [−2.35, 7.69]). Chart-level accuracy was higher among NSCLC patients vs. CrCa patients in both the Human+AI (79.8% vs 71.6%, statistic = 12130, $p < 0.001$, Rosenthal's $r = 0.287$, 95% CI = [−11.76, −2.78]) and Human-alone arms (76.5% vs 65.4%, statistic = 11388, $p < 0.001$, Rosenthal's $r = 0.348$, 95% CI = [−16.67, −6.06]).

We conducted a sensitivity analysis of chart-level accuracy under a separate definition of accuracy, which treated CRC responses that were more specific than the gold-standard set response as accurate. Results from this analysis show the Human+AI arm maintained superiority as compared to the Human-alone arm under a less strict mapping of CRC responses to the gold-standard set (78.3% vs. 73.0%, respectively, mean difference 5.3%, statistic = 18259, $p < 0.001$, Rosenthal's $r = 0.190$, 95% CI = [0.04, Inf]).

Table 1 | Criterion-level accuracy comparisons between AI-alone, Human-alone, and Human+AI study arms

	Criterion	N	Percent of Charts with Criterion	Accuracy (%)			p-value
				Human-Alone	Human+AI	AI-Alone	
Neoplasm	Cancer Type	360	100.0%	86.9	86.4	73.3	0.897
	Stage Group	237	66.5%	71.7	73.4	57.0	0.672
	M Stage	221	61.4%	43.9	57.0	60.2	<0.001*
	N Stage	196	54.4%	50.5	66.3	52.6	<0.001*
	T Stage	197	54.6%	56.3	71.6	54.3	<0.001*
Biomarker	Biomarker Tested	454	44.2%	84.6	93.2	88.1	<0.001*
	Categorical Value	452	44.2%	67.9	79.0	32.5	<0.001*
	Interpretation	381	41.1%	80.8	91.3 #	35.7	<0.001*
Other	Outcome	181	34.1%	23.7	35.9 #	55.2	0.006*
	Response	346	51.5%	47.1	51.7	60.4	0.224
	ECOG Status	183	51.5%	84.7	78.1	34.4	0.126
	Medication Concept	525	63.7%	89.0	89.1	59.4	>0.999

Criteria found to have significantly different mean accuracy between the Human+AI and Human-Alone arms are denoted with an asterisk. All criteria with significant differences were found to have significant superiority in favor of the Human+AI arm based on a second set of hypothesis tests. Both sets of tests were conducted at alpha level 0.05 and adjusted for multiple comparisons. Cells in light grey indicate a difference greater than 5% between Human-Alone and Human+AI, and cells in dark grey indicate a difference greater than 10%. Criterion for which the AI-Alone arm outperforms the Human-Alone and Human+AI arms are noted with bolded values. Two-sided, paired binomial exact tests with Bonferroni correction were conducted. Superiority of the Human+AI arm for criterion-level abstraction was assessed for significant criteria with an additional one-sided hypothesis with an unspecified superiority margin.

Efficiency assessment

Efficiency, measured as time per chart review, was a secondary outcome (Fig. 3). There was no difference in efficiency between Human+AI vs. Human-alone arms (mean time per chart review 37.4 min vs. 37.8 min, statistic = 32,678, $p = 0.513$, Rosenthal's $r = 0.035$, 95% CI = [-1.25, 2.55]); interquartile range (IQR) 20.3–49.3 min vs. 20.6–48.7 min). Subgroup analyses investigated the effects of chart characteristics—order, length, complexity, and cancer type—on median efficiency of chart abstraction within each study arm. In both arms, average efficiency improved from earlier vs. later charts (Human+AI: 42.2 vs. 32.6 min, mean difference 9.6 min, statistic = 19454, $p < 0.001$, Rosenthal's $r = 0.289$, 95% CI = [4.00, 12.50]; Human-alone: 44.6 vs. 31.0 min, mean difference 13.6 min, statistic = 20,989, $p < 0.001$, Rosenthal's $r = 0.407$, 95% CI = [7.38, 15.87]), and shorter vs. longer charts (Human+AI: 24.8 vs. 50.2 min, mean difference 25.4 min, statistic = 4562, $p < 0.001$, Rosenthal's $r = 0.870$, 95% CI = [-26.72, -19.78]; Human-alone: 25.4 vs. 50.2 min, mean difference 24.8 min, statistic = 5594, $p < 0.001$, Rosenthal's $r = 0.790$, 95% CI = [-26.65, -18.62]). Results from subgroup analyses are summarized in Supplemental Table 3.

Discussion

In this randomized trial using retrospective EHRs comparing AI-alone vs. Human-alone vs. Human+AI chart abstraction to screen oncology patients for clinical trial participation, the Human+AI prescreening approach approximated and improved overall chart abstraction accuracy without compromising efficiency. The Human+AI approach showed greatest accuracy gains relative to Human-alone for biomarker and staging-related criteria. These findings highlight the potential for AI tools to complement human abstraction of eligibility criteria from complex medical documents. Three points relevant to human-AI collaboration in trial prescreening bear emphasis.

First, AI algorithms can support the trial prescreening process by streamlining data extraction from charts and identifying patients most likely to meet inclusion criteria. Accuracy gains in a Human+AI prescreening approach can increase inclusion of a broader patient population for oncology trials and prevent false-positive identification that may result in future eligibility queries.

While accuracy was both noninferior and superior for Human+AI, we did not observe improved efficiency of chart abstraction in the

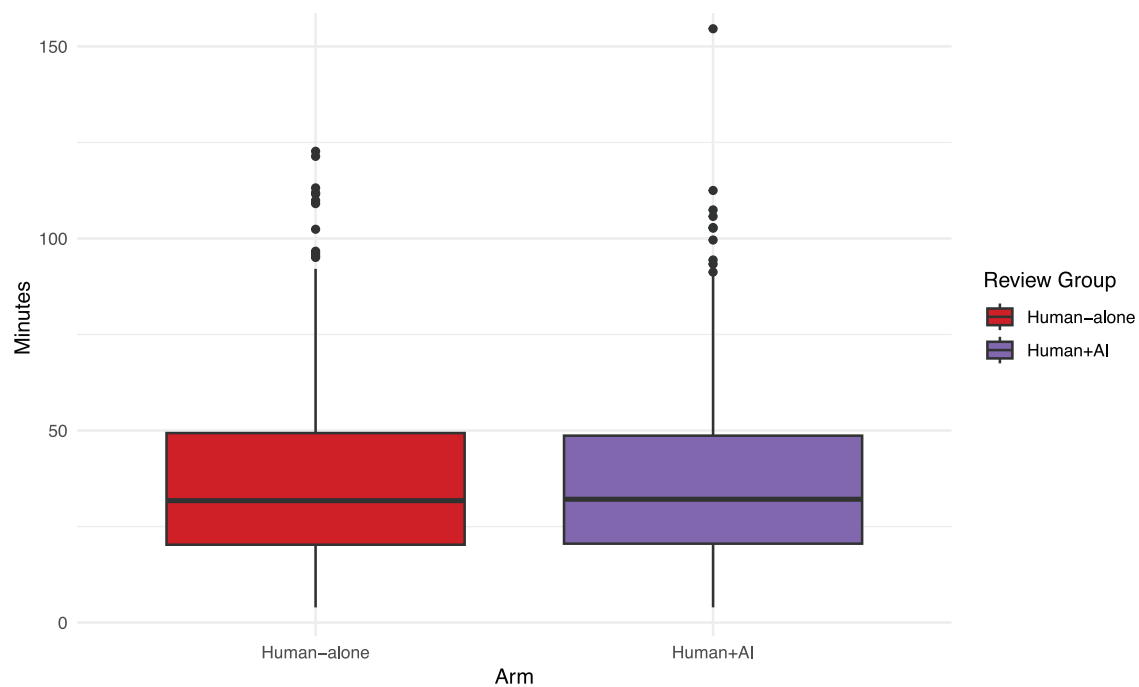


Fig. 3 | Efficiency comparison between Human-alone (red) and Human+AI (purple) arms. $N = 355$ patient charts were independently reviewed in each arm. Each chart constitutes an independent observational unit, and no technical replicates were generated. Boxplots display the median, 25th–75th percentiles

(interquartile range), and whiskers representing the minimum and maximum observed values. A nonparametric, paired Wilcoxon Rank Sum test with continuity correction was used to compare median time to review chart, yielding non-significant differences in median time.

Human+AI arm. This finding suggests that the workload of clinical research staff may have shifted from the identification of chart elements to the careful review and interpretation of the AI-extracted outputs. This shift could avoid overreliance on AI predictions—automation bias¹⁶. Our efficiency result should be interpreted cautiously. Since our study was powered on chart-level accuracy, CRCs reviewed every EHR document in full for the presence or absence of all 12 criteria. However, in practice, many prescreens anchor to specific inclusion and exclusion criteria, allowing a CRC to quickly move on from a patient’s EHR the moment an exclusion criterion is encountered. So while median timeliness was comparable for this full chart review tasks, there is potential for chart-level (time per chart) and volume-level (number of charts reviewed per unit time) time savings in a Human+AI workflow.

Second, for most inclusion criteria studied, abstraction accuracy was greater for the Human+AI arm as compared to the Human-alone arm. In categories where the Human+AI collaboration outperformed both AI- and Human-alone arms, two patterns emerged. First, the Human+AI arm outperformed both AI-alone and Human-alone arms was superior to the other arms for criteria where AI-alone outperformed Human-alone, suggesting humans incorporate the accuracy gains of AI in ways that collaboratively improve overall system performance, rather than just replicating the performance of AI. Conversely, even for some criteria where Human-alone outperformed AI-alone (e.g. biomarker value and interpretation), Human+AI was still superior to both, suggesting that there is value to AI-based assessment even in settings where it is inaccurate overall¹⁷.

Third, our study sheds light into nuances of both confirmation and automation bias in trial prescreening. Only one category, ECOG performance status, showed evidence of automation bias¹⁸, as Human+AI performance was lower than Human-alone due to the substantial inaccuracy of the AI. In other categories (i.e., outcome assessment), the AI-alone arm outperformed both Human+AI and human-alone arm, suggesting that Human-AI collaboration in these categories was hindered by confirmation bias¹⁹. Notably, we found no situations that accuracy was lower in the Human+AI arm than both AI- and Human-

alone arms. This finding stands in contrast to other studies in the radiology setting that suggest a net negative effect to human-AI collaboration in medical diagnostic tasks^{20,21}. This may be because trial prescreening involves ascertainment of information contained within clinical text, which is a known strength of language models, whereas radiology-based diagnosis may depend on interpretation of clinical context and information contained outside of a radiology scan.

Our study has four limitations. First, we study one closed-source AI algorithm, limiting the generalizability of our findings to prescreening workflows that may incorporate alternative algorithms (e.g. LLMs). Second, our subgroup analysis did not look at the dependency between document length and complexity and their interactive effect on study outcomes. Third, our final study sample was sufficiently powered to assess the impact of an AI-Human collaborative prescreening approach on our primary outcome, chart-level accuracy, but not on the efficiency outcome nor the subgroup analyses. Nonetheless, these investigations provide valuable initial insight into the heterogeneity of performance of a Human+AI prescreening approach. Finally, limited quality of the uploaded documents (e.g., poorly scanned documents, illegible handwritten notes, missing documents) may have impacted the accuracy and efficiency of chart abstraction (Supplemental Fig. 3–9), although this reflects real-world issues with chart prescreening.

Future work should examine how accuracy and efficiency of abstraction vary by patient subgroups (e.g., by race) to assess AI fairness and biases in performance and to evaluate how the human-AI collaborative approach impacts the inclusivity of underrepresented patient groups in the eligibility population. Additionally, prescreening with generative AI models has shown promise at improving abstraction time and should be tested in settings similar to ours²². Finally, it is unknown improving accuracy for trial criteria prescreening translates to more patients enrolled on trial. In an ongoing study (NCT06561230), we are assessing the impact of neurosymbolic AI-enabled prescreening on rates and diversity of enrollment to a phase III randomized trial.

Clinical trial enrollment in oncology is limited by an arduous and suboptimal prescreening process, performed solely by clinical staff.

We find that integrating chart abstractions of unstructured clinical texts by a high-performing AI system into human workflows leads to comparable and improved prescreening accuracy, improves precision in abstracting biomarker and neoplasm trial inclusion criteria, and preserves prescreening efficiency. Using experimental frameworks such as ours to evaluate human-AI collaborations is critical to build trust in using AI in healthcare workflows.

Methods

Study population

This study was approved by the University of Pennsylvania School of Medicine Institutional Review Board. Informed consent was not obtained from study patients as this was retrospective data, and a waiver of informed consent was granted owing to use of de-identified data. No patients were compensated for participation. The trial was registered on ClinicalTrials.gov (NCT#06561217) after completion. The study population was drawn from a 15-physician community oncology practice in California serving patients from urban and surrounding rural communities. This study cohort consisted of unstructured medical records from patients within the dataset with 1) a diagnosis of non-small cell lung cancer (NSCLC) or colorectal cancer (CrCa), 2) a minimum of five clinical documents available, and 3) the most recent document being within five years from the time of data extraction. The study included charts retrospectively collected from both self-reported male and female patients. Sex and/or gender were not considered in the study design because gender-based analyses were not conducted as the primary goal was to determine overall differences in chart-level accuracy, criterion-level accuracy, and abstraction efficiency between arms rather than demographic subgroup comparisons in these outcomes.

Data source

The data source consisted of unstructured electronic health record (EHR) documents of the following types:

- Physician progress notes
- Radiology reports
- Pathology reports
- Laboratory reports
- Genetic/molecular sequencing and other genetic testing results
- Procedure notes
- Discharge summaries
- Infusion therapy reports
- Other, including administrative and billing documents

Each individual document was converted to PDF format and uploaded to a unique data abstraction portal created for each individual patient. Documents were presented to the user categorized by type (see list above) and arranged in descending time-sequential order.

Trial eligibility criteria

The following 12 elements were extracted: cancer type, cancer stage group, T stage, N stage, M stage, biomarker tested, biomarker categorical value, biomarker interpretation, medication, ECOG performance score, response to tumor treatment, and outcome. The elements were chosen by the Principal Investigator (RBP) and Senior Author (EJE) due to their ubiquity in trial eligibility and outcome assessment criteria. More detail about these elements is provided in Supplemental Table 1.

Data processing for AI outputs

The documents were processed through a neurosymbolic AI-enabled data pipeline designed to transform unstructured clinical text into structured, clinically meaningful data. The outputs of this pipeline were used as the AI-alone output and displayed to CRCs in the Human-alone arm.

The data pipeline involved several key stages, each responsible for extracting and refining relevant information from clinical text. Initially, unstructured clinical notes were processed through a healthcare-specific optical character recognition (OCR) system to convert them into machine-readable text while redacting protected health information and scrambling dates to ensure patient privacy. We then used transformer-based models developed in Python (version 3.9) and trained on real-world clinical data coupled with symbolic AI systems to extract biomedical entities, identify relationships between entities, consolidate extracted facts into medical events, score relevance to the patient's disease journey, and prune medically incoherent hallucinations from our outputs. These outputs were segmented into discrete reports, categorized by document type, and relevant dates were extracted. High-confidence data were consolidated, and a symbolic reasoning system, curated by medical experts, was used meaningful clinical events using techniques like entailment²³ and exclusion²⁴. The resulting enriched clinical events were presented to users through a web-based interface with traceability back to the data source(s) associated with each clinical event. Our Supplemental Methods provide more detail on each of these processes. Custom scripts for data processing, chart abstraction, and analysis were conducted in R (version 4.3.0) and are available at <https://github.com/HumanAlgorithmCollaborationLab/TrialPrescreening>.

Randomization and study procedures

Block randomization. Deidentified unstructured electronic health records (EHRs) from real-world patients with non-small cell lung (NSCLC) or colorectal (CrCa) cancer were presented to CRCs via a secure platform. Primary comparisons were between two study arms: (1) human reviewers ("Human-alone" arm) versus (2) human reviewers utilizing predictions from AI ("Human+AI" arm). The primary outcome, chart-level accuracy, the proportion of data elements correctly abstracted from the gold standard, was measured for each chart. Two CRCs were randomized 1:1 to view blocks of 20 charts each with (Human+AI) or without (Human-alone) AI annotations. In this paired design, each individual chart was reviewed by one CRC with AI assistance (Human+AI) and one CRC without AI assistance (Human-alone). Overall, each CRC reviewed half of the charts manually (Human-alone) and half with AI assistance (Human+AI).

Abstraction interface. Abstraction time was tracked in the abstraction interface. In the abstraction interface, to begin a chart review, the abstractor would click on the forward arrow button to begin the timer. To pause or to resume abstraction on that chart, the reviewer would click the "pause" button. When finished with a chart, the reviewer would click on the "stop" button. To provide a separate measure of time in case the abstraction interface timer was faulty, abstraction time was also tracked concurrently in real time using a separate timer outside of the abstraction platform.

On the left half of the screen are displayed all of the documents within a chart. Above that is a drop-down box that allows the abstractor to filter by document category (e.g., SOAP Note, Labs, Pathology). To review the documents, the abstractor would read through the documents displayed and select any items ("events") of interest. Representative screenshots of the abstraction platform are provided in Supplemental Figs. 3–9.

Gold-standard set creation

The Gold-Standard Set refers to a set of data that has been manually prepared and verified to represent the objective truth as closely as possible. It consisted of annotated data based on the endpoints of interest for the analysis. The Gold-Standard Set was used to evaluate the performance of both manual abstraction in the Human+AI and Human-alone arms and the output of the AI system in the AI-alone arm (see Fig. 4).

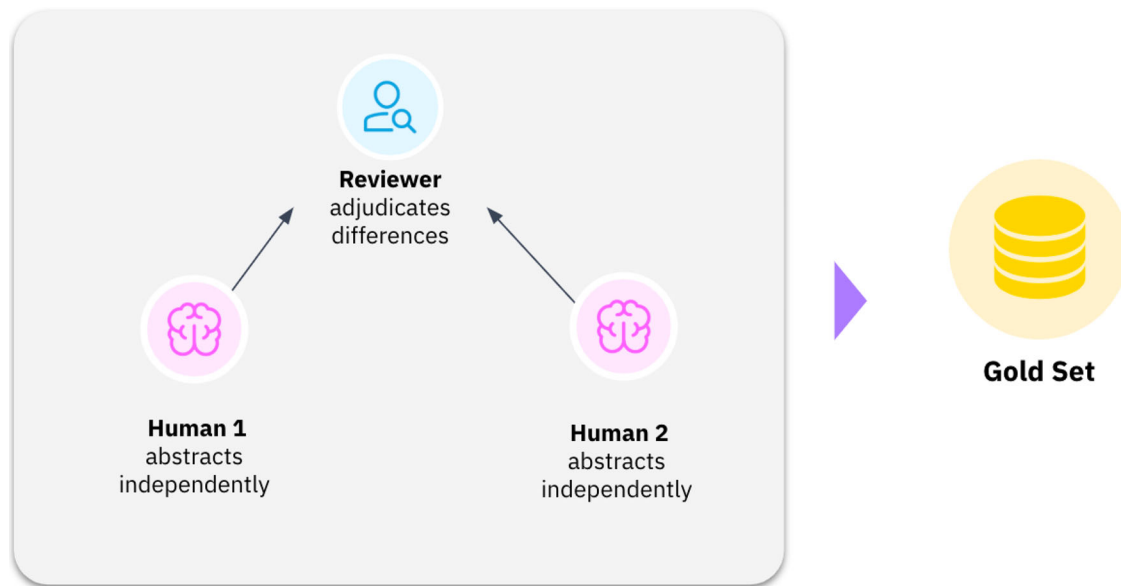


Fig. 4 | Gold-Standard Set creation. The Gold-Standard Set refers to a set of data that has been manually prepared and verified to represent the objective truth as closely as possible. The Gold-Standard Set was used to evaluate the performance of both manual abstraction in the Human+AI and Human-alone arms and the output

of the AI system in the AI-alone arm. The two sets of human abstraction were done independently and a third human adjudicated by comparing the two sets of abstraction. There was no time limit to perform the two human abstraction sets or the adjudication process.

The two sets of human abstraction were done independently and a third human adjudicated by comparing the two sets of abstraction. There was no time limit to perform the two human abstraction sets or the adjudication process. The primary outcome was overall chart-level accuracy, defined as percent of the 12 CRC-coded eligibility items matching a gold-standard set, determined by 2–3 clinicians blinded to experimental arms.

Data processing for analysis

The analytic dataset was constructed from chart abstraction responses and chart review metadata containing information on accuracy and efficiency. CRC-abstracted responses were identified as being accurate when they exactly matched with the gold-standard set, and inaccurate when not. Both an incorrect response and the absence of a response to a gold-standard event were considered to be inaccurate responses. See Supplemental Table 4 for how CRC responses were mapped to the gold-standard set. Additionally, we conducted a sensitivity analysis using a less strict definition of an accurate match to include responses in which the CRC was more specific than the gold-standard set. The analytic dataset is available at <https://github.com/HumanAlgorithmCollaborationLab/TrialPrescreening>.

Outcomes

The primary outcome measured in this study was mean chart-level accuracy, defined as the proportion of elements identified by CRCs among all elements in the gold-standard set, measured for each chart, and averaged across all charts. Two secondary outcomes were measured: criterion-level accuracy and efficiency of chart completion. Criterion-level accuracy was defined as the proportion of instances a criterion was identified by CRCs among all instances the criterion appeared across all charts. Criterion-level accuracy was calculated for each criterion. Efficiency was calculated as the number of min spent on each chart abstraction, with the median time among all charts chosen as the summary statistic. All outcomes were calculated separately for each trial arm. Additionally, we performed subgroup analyses to understand the impact of chart characteristics—length, complexity, cancer type, and order—on chart-level accuracy and efficiency. Chart length and complexity were defined by the number of documents and number of

elements in the gold-standard set, respectively, per chart. Chart order was defined as the order in which the chart was reviewed by a CRC. These chart characteristics were dichotomized using the median of each measure to identify short vs long charts (i.e., <82 vs. ≥82 documents), less complex vs more complex charts (i.e., <10 vs. ≥10 elements), and early vs later charts (i.e., Chart # 1-178 vs. # 179-355). Cancer type was dichotomized as non-small cell lung cancer vs colorectal cancer. To ensure no major differences in abstraction quality between CRCs, we conducted a pre-trial assessment based on 8 charts that had corresponding outcomes in the gold standard set, but were not included in the final analysis. Inter-rater agreement was measured by Cohen's Kappa and by percent agreement, defined as the average percent of elements in the chart that were coded the same by the two CRCs. In our pre-specified protocol, we defined Cohen's Kappa >0.60 and percent agreement >80% demonstrating sufficient equivalence to proceed with the trial. The two CRCs had Cohen's Kappa 0.72 and 86% percent agreement.

Sample size calculations

The study was initially powered for a paired, non-inferiority study comparing the primary endpoint, mean chart-level accuracy of EHR chart abstraction, between Human-alone and Human+AI arms. Criterion-level performance metrics derived from preliminary Human+AI and Human-alone chart abstractions informed initial estimates of mean chart-level accuracy, which were used to calculate sample size. Simulations were performed to assess the impact of varying effect sizes (i.e., difference in the mean chart-level accuracy between arms), event rates (i.e., prevalence of criterion among all charts examined), performance metrics (i.e., true negative rate), and concordance rates (i.e., agreement in abstraction between arms) on minimum sample size needed to achieve a power for 80% at a type I error rate of 0.05 and noninferiority margin of 0.05. Once 74 charts were abstracted in the main study, a pre-planned interim analysis was conducted to inform a second round of calculations for minimum sample needed to robustly assess both the noninferiority and superiority of the Human+AI arm to the Human-alone arm in chart-level accuracy. The initial calculations relied on assumptions for effect size, leading to a conservative sample size estimate. The interim analysis incorporated incoming data to

refine the estimate. See Supplemental Methods and Supplemental Figs. 1–2 for the derivation, rationale, and results from the first and second round of sample size simulations. Code for sample size calculations is at <https://github.com/HumanAlgorithmCollaborationLab/TrialPrescreening>.

Statistical analysis

Demographic and clinical characteristics of the study sample were descriptively summarized. A Shapiro-Wilk test was used to assess normality of paired differences in chart-level accuracy ($\alpha = 0.05$). A one-sided, paired Wilcoxon Rank Sum test ($\alpha = 0.05$) was employed to test the primary null hypothesis of whether chart-level accuracy of the Human+AI arm for EHR chart abstraction was non-inferior to the chart-level accuracy of a Human-alone arm abstraction by at least 5% (i.e., noninferiority margin). Superiority of the Human+AI arm abstraction to the Human-alone arm abstraction with respect to chart-level accuracy was tested with a one-sided, paired Wilcoxon test ($\alpha = 0.05$)²⁵. See Supplemental Methods for null and alternative hypotheses for non-inferiority and superiority tests. To assess whether there was a difference in accuracy at the criterion-level between the arms, two-sided, paired binomial exact tests were conducted with Bonferroni correction for multiple hypotheses (i.e., $\alpha = 0.05 / n$, where $n = 12$ was the number of criteria studied). For any criterion with a statistically significant difference in mean criterion-level accuracy, superiority of the Human+AI arm for criterion-level abstraction was then assessed with an additional one-sided hypothesis with an unspecified superiority margin. A nonparametric, paired Wilcoxon Rank Sum test with continuity correction was employed to compare overall median efficiency of chart abstraction between arms. Wilcoxon Rank Sum tests were used to compare chart-level accuracy and efficiency outcomes between charts stratified by dichotomized chart characteristics. All analyses and visualizations of results were performed in R / RStudio (Version 4.3.0).

Statistics and reproducibility

Sample size was determined through power analysis for a paired, non-inferiority study design comparing mean chart-level accuracy between Human-alone and Human+AI arms. Initial sample size estimates were informed by preliminary criterion-level performance metrics from pilot chart abstractions, with simulations assessing the impact of varying effect sizes, event rates, and concordance rates. A pre-planned interim analysis after 74 charts refined these estimates, ultimately determining that 355 charts would provide 80% power at a Type I error rate of 0.05 with a non-inferiority margin of 0.05. No data were excluded from the analyses; all 355 charts reviewed were included in the final analysis.

Charts were randomized in blocks of 20 to Human+AI or Human-alone arms using a 1:1 randomization scheme, with each chart reviewed once in each arm by different clinical research coordinators (CRCs). The two CRCs were randomized to review half of the charts with AI assistance and half without, ensuring balanced exposure to both conditions. While CRCs could not be blinded to the presence or absence of AI assistance due to the nature of the intervention, they were blinded to the gold-standard outcomes during abstraction. The gold-standard set was created independently by 2–3 clinicians who were blinded to experimental arms. Inter-rater reliability between CRCs was assessed prior to the main study using eight charts not included in the final analysis, achieving Cohen's Kappa of 0.72 and 86% percent agreement, exceeding pre-specified thresholds ($\kappa > 0.60$, agreement $> 80\%$). All statistical analyses were conducted using pre-specified methods with Bonferroni correction for multiple comparisons where appropriate.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data used in this study were provided through an academic collaboration between the trial sponsor and the University of Pennsylvania. The authors had full access to the dataset and conducted all analyses independently. The processed, analytic data generated in this study have been deposited in the github. The raw dataset contains unstructured clinical notes with protected information under a proprietary license and are not available due to data privacy and legal restrictions. The processed, analytic dataset is available at <https://github.com/HumanAlgorithmCollaborationLab/TrialPrescreening>.

Code availability

PASS software (Version 23.0.1) was used to generate sample size calculations. All other analyses and visualizations of results were performed in R / RStudio (Version 4.3.0). Code for data pre-processing, sample size calculations, and generation of results is available at <https://github.com/HumanAlgorithmCollaborationLab/TrialPrescreening> under an MIT license.

References

1. Unger, J. M., Cook, E., Tai, E. & Bleyer, A. Role of clinical trial participation in cancer research: barriers, evidence, and strategies. *Am. Soc. Clin. Oncol. Educ. Book* **35**, 185–198 (2016).
2. Duenas, J. A. C., Sanchez, P. N. & Bonilla, C. E. Comparison of clinical outcomes among cancer patients treated in and out of clinical trials. *BMC Cancer* **23**, 786 (2023).
3. Unger, J. M. et al. Comparison of survival outcomes among cancer patients treated in and out of clinical trials. *J. Natl. Cancer Inst.* **106**, dju002 (2014).
4. Ford, J. G. et al. Barriers to recruiting underrepresented populations to cancer clinical trials: a systematic review. *Cancer* **112**, 228–242 (2008).
5. Unger, J. M., Cook, E., Tai, E. & Bleyer, A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. *Am. Soc. Clin. Oncol. Educ. Book* **35**, 185–198 (2016).
6. Meystre, S. M., Heider, P. M., Kim, Y., Aruch, D. B. & Britten, C. D. Automatic trial eligibility surveillance based on unstructured clinical data. *Int. J. Med. Inf.* **129**, 13–19 (2019).
7. Keber, M., Grubišić, I., Barešić, A. & Jović, A. A Review on Neuro-symbolic AI Improvements to Natural Language Processing. in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)* 66–72 <https://doi.org/10.1109/MIPRO60963.2024.10569741> (2024).
8. Hamilton, K., Nayak, A., Božić, B. & Longo, L. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web* **15**, 1265–1306 (2024).
9. Ilday, B., Dreisbach, C., Weng, C. & Schnall, R. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *J. Am. Med. Inf. Assoc.* **29**, 197–206 (2021).
10. Haddad, T. et al. Accuracy of an artificial intelligence system for cancer clinical trial eligibility screening: retrospective pilot study. *JMIR Med. Inf.* **9**, e27767 (2021).
11. Lu, X. et al. Artificial intelligence for optimizing recruitment and retention in clinical trials: a scoping review. *J. Am. Med. Inf. Assoc.* **31**, 2749–2759 (2024).
12. Lee, K. et al. CriteriaMapper: establishing the automatic identification of clinical trial cohorts from electronic health records by matching normalized eligibility criteria and patient clinical characteristics. *Sci. Rep.* **14**, 25387 (2024).
13. Meystre, S. M. et al. Piloting an automated clinical trial eligibility surveillance and provider alert system based on artificial intelligence and standard data models. *BMC Med. Res. Methodol.* **23**, 88 (2023).
14. Do, N. V. et al. Matching patients to accelerate clinical trials (MPACT): enabling technology for oncology clinical trial workflow. *Stud. Health Technol. Inf.* **310**, 1086–1090 (2024).

15. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).
16. Parikh, R. B., Teeple, S. & Navathe, A. S. Addressing bias in artificial intelligence in health care. *JAMA* **322**, 2377–2378 (2019).
17. Kolla, L. & Parikh, R. B. Uses and limitations of artificial intelligence for oncology. *Cancer* **130**, 2101–2107 (2024).
18. Wang, D. Y. et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *J. Am. Med. Inf. Assoc.* **30**, 1684–1692 (2023).
19. Bashkirova, A. & Krpan, D. Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Comput. Hum. Behav.: Artif. Hum.* **2**, 100066 (2024).
20. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. SSRN Scholarly Paper at <https://papers.ssrn.com/abstract=4505053> (2023).
21. Yu, F. et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
22. Jin, Q. et al. Matching Patients to Clinical Trials with Large Language Models. *ArXiv arXiv:2307.15051v5* (2024).
23. Korman, D. Z., Mack, E., Jett, J. & Renear, A. H. Defining textual entailment. *J. Assoc. Inf. Sci. Technol.* **69**, 763–772 (2018).
24. Mündler, N., He, J., Jenko, S. & Vechev, M. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. In *International Conference on Learning Representations*. Paper at <https://openreview.net/pdf?id=EmQSOiX2f> (2024).
25. Wang B, Wang H, Tu XM, Feng C. Comparisons of Superiority, Non-inferiority, and Equivalence Trials. *Shanghai Arch Psychiatry* **29**, 385–388.

Acknowledgements

The authors acknowledge Ayana Andalco and Jose U. Mena from Mendel.ai for their assistance with AI platform setup and with critical manuscript feedback. The authors acknowledge the patients who contributed their records for this study. This study was funded by Mendel.ai (to RBP); the funders had no role in the design, data collection, analysis, or interpretation of this study.

Author contributions

Ravi B. Parikh and Likhitha Kolla contributed equally as co-first-authors. Ravi B. Parikh, MD, MPP—Conceptualization, Methodology, Writing—Original Draft, Supervision, Funding Acquisition Likhitha Kolla, BS—Conceptualization, Methodology, Writing—Original Draft, Validation Supervision Elizabeth A. Beothy, BA—Investigation, Data Curation, Writing—Review & Editing William J. Ferrell, MPH—Project Administration Brenda Laventure, MBS—Investigation Matthew Guido, BS—Project Administration Anthony Girard, MS—Software, Validation, Formal Analysis, Writing—Review & Editing Yang Li, MS - Software, Validation, Formal Analysis Khaled Essam Mahmoud Dosoky, BSc—Resources, Data Curation, Writing—Review & Editing Karim Tarabishy, BSc—Resources, Data Curation, Writing—Review & Editing Parth S. Patel BS—Resources, Data Curation, Writing—Review & Editing Ayana Andalco, BSc—Resources, Data Curation, Writing—Review & Editing Kristin Maloney, MS—Resources, Data Curation, Writing—Review & Editing Jose Ulises Mena, AB—Resources, Data Curation, Writing—Review & Editing Wael Salloum, PhD—Resources, Data Curation, Writing—Review & Editing Jinbo Chen, PhD—Methodology, Writing- Review & Editing Ezekiel J. Emanuel, MD, PhD - Conceptualization, Methodology, Writing- Review & Editing, Supervision, Funding Acquisition.

Competing interests

Ravi B. Parikh declares the following competing interests: grants from the National Institutes of Health, Department of Defense, Prostate Cancer Foundation, National Palliative Care Research Center, NCCN Foundation, Conquer Cancer Foundation, Humana, Emerson Collective,

Schmidt Futures, Arnold Ventures, Mendel.ai, and Veterans Health Administration; personal fees and equity from GNS Healthcare, Thyme Care, Optinosis, Main Street Health, and Onc.AI; personal fees from the ConcertAI, Cancer Study Group, Mendel.ai, Biofourmis, Archetype Therapeutics, CreditSuisse, G1 Therapeutics, Humana, and Nanology; honoraria from Flatiron and Medscape; has board membership (unpaid) at the Coalition to Transform Advanced Care and American Cancer Society; editor at the *Journal of Clinical Oncology*; and serves on a leadership consortium (unpaid) at the National Quality Forum, all outside the submitted work. Ezekiel Emanuel declares the following competing interests: travel reimbursement from Centre for Biomedical Ethics, National University of Singapore; travel to Global Innovation Forum AAHC, Washington, DC; travel reimbursement from Macalester College, St. Paul, MN; travel reimbursement from Oak CEO Summit, honorarium from Advocate Aurora Health Summit, travel reimbursement from DPharm Conference, honorarium and travel reimbursement from the University of Pittsburgh Medical Center Shadyside, Pittsburgh, PA; travel reimbursement from the University of California San Francisco, San Francisco, CA; travel reimbursement from Cain Brothers Conference, New York, NY; honorarium and travel reimbursement from Bowdoin College, Brunswick, ME; travel reimbursement from The Galien Foundation, travel reimbursement from HLTH 2022 Meeting, Las Vegas, NV; travel reimbursement from the Hawaii Medical Service Association, Honolulu, HI; travel reimbursement from the Tel Aviv University, Tel Aviv, Israel; travel reimbursement from The Suntory Foundation, Tokyo, Japan; honorarium and travel reimbursement from the Ontario Hospital Association, Ontario, Canada; Monetary prize and travel reimbursement from the University of Oklahoma, Oklahoma City, OK; travel reimbursement from The Quadrangle, Haverford, PA; lodging Lazard HC Leadership Summit, Walland, TN; travel reimbursement from HLTH 2023 Meeting, Las Vegas, NV; honorarium and travel reimbursement from Sanford Health, Sioux Falls, SD; honorarium and travel reimbursement from Health Plan Alliance, Atlanta, GA; honorarium and travel reimbursement from Emory Health Care Leaders Retreat, Greensboro, GA; honorarium and travel reimbursement from Employer Direct Health Care, Aspen, CO; travel reimbursement from the University of Virginia, Roslyn, VA; travel reimbursement from the New York Historical Society, New York, NY; lodging Amangiri Executive Retreat, Canyon Point, UT; travel reimbursement from Forerunner Conference, Sao Paulo, Brazil; travel reimbursement from BCEPS conference, Solstrand, Norway; travel reimbursement Arendalaska Meeting, Arendal, Norway; lodging Futures of Science Summit, Philadelphia, PA; travel reimbursement from The Galien Foundation New York; travel reimbursement from the Milken Institute Washington, DC. Dr. Emanuel is serving on the following: Member Board of Advisors Cellares Corp.; Advisor Notable Health; Member Advisory Board JSL Health Capital; Member Advisory Board Peterson Center on Health Care; Advisor Clarify Health Solutions; Consultant Health Care Foundry Inc.; Member Advisory Board Feel Better Ltd.; Consultant Korro/Coach AI Ltd.; Consultant Aberdeen Inc.; Member Board of Advisors Alto Pharmacy Holdings; Advisor Link Health Technologies; Advisor Nuna; Expert Advisory Member WHO COVID 19 Ethics & Governance Working Group; Advisory Board Member HIX Health Innovation Exchange partnership sponsored by UN Geneva; Advisory Board Member Biden's Transition COVID-19 Committee; Special Advisor to the Director General World Health Organization; Editorial Board Member *Journal of the American Medical Association*; Member Guideline Development Group (GDG) of the WHO Rapid Advice Guideline on the Use & Indications of Glucagon-Like Peptide-1 Receptor Agonists for Management of Adults Living with Obesity; Member Internal Advisory Board The Penn Parity Center; Advisor Dendro Technologies Inc. (CalmiGo). Dr. Emanuel reports investments in Sunstone Consulting and Aktivite. Dr. Emanuel reports royalties from his books and William Morris Endeavor. Dr. Emanuel reports the following grants: Hogan Lovells, Mendel Health Inc., University of Bergen, Jansen Pharmaceuticals Inc., Schmidt Futures Schwab Charitable Fund, Laura and

John Arnold Foundation, Humana, University of Miami, Hawaii Medical Services Association, Patient Centered Outcomes Research Institute. Dr. Emanuel reports options for Nuna, Link Health Technologies, Inc, Clarify Health Solutions, Healthcare Foundry Inc., Alto Pharmacy Holdings, Korro/Coach AI Ltd., FeelBetter LTD., Notable Health, Cellares Corp., Sunstone Consulting; all outside the submitted work. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-68873-8>.

Correspondence and requests for materials should be addressed to Ravi B. Parikh.




Peer review information *Nature Communications* thanks Joseph Unger and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

Ravi B. Parikh ^{1,6} ✉, **Likhitha Kolla**^{2,6}, **Elizabeth A. Beothy**³, **William J. Ferrell** ⁴, **Brenda Laventure**³, **Matthew Guido**⁴, **Anthony Girard**¹, **Yang Li**³, **Khaled Essam Mahmoud Dosoky**⁵, **Karim Tarabishy**⁵, **Parth S. Patel**⁵, **Ayana Andalcio**⁵, **Kristin Maloney**⁵, **Jose Ulises Mena**⁵, **Wael Salloum**⁵, **Jinbo Chen**² & **Ezekiel J. Emanuel**⁴

¹Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA, USA. ²Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ³Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁴Department of Medical Ethics and Health Policy, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁵Mendel.ai, San Jose, CA, USA. ⁶These authors contributed equally: Ravi B. Parikh, Likhitha Kolla. ✉ e-mail: ravi.bharat.parikh@emory.edu