Human-AI Teaming to Improve Accuracy and Efficiency of Eligibility Criteria Prescreening for Oncology Trials

Table of Contents

Supplemental Tables

Supplemental Table 1: Eligibility Criteria Classification

Criterion Type	Criterion	Description	Example
Neoplasm	Cancer Type	Tumor type	Malignant rectosigmoid neoplasm
Neoplasm	Stage Group	Cancer stage group	IIIA
Neoplasm	M Stage	Distant metastases	Mx
Neoplasm	N Stage	Number of lymph nodes that have cancer	pN1
Neoplasm	T Stage	Tumor size and location	pT2a
Biomarker	Biomarker Tested	Name of biomarker	BRAF
Biomarker	Categorical Value	Positive, negative, mutation, etc.	Gene mutation negative
Biomarker	Interpretation	Normal, abnormal, etc.	Normal
Other	Medication	Platinum, targeted, or chemotherapy	FOLFOX
Other	ECOG Performance Status	ECOG Score (0-5)	3
Other	Response to Tumor Treatment	Treating physician-assessed (not radiologist) response to tumor treatment	Disease progression
Other	Outcome	Physician-reported outcome attributable to local treatment	Distant recurrence

CAPTION: The gold-standard set elements were identified as belonging to one criterion type, which encompass relevant clinical eligibility criteria to be abstracted by CRCs. This table shows each criterion type, its description, and an example from the data. ECOG = Eastern Cooperative Oncology Group Performance Status.

Supplemental Table 2: Subgroup Analysis of Chart-Level Accuracy Across Chart Characteristics

Characteristic	Human+AI H+AI		Human-Alone H		+ Bias H+AI - H
	Accuracy (%)	Difference (1-2) [p-value]	Accuracy (%)	Difference (1-2) [p-value]	
Length					
(1) Short	78.2	4.2	74.2	5.4	-1.2
(2) Long	74.0	[0.005]*	68.8	[0.001]*	
Complexity					
(1) Low	76.1	0.0	71.1	-0.8	0.8
(2) High	76.1	[0.243]	71.9	[0.357]	
Order					
(1) Early	77.5	2.8	73.8	4.7	-1.9
(2) Later	74.7	[0.438]	69.1	[0.186]	
Cancer Type					
(1) NSCLC	79.8	8.2	76.5	11.1	-2.9
(2) CrCa	71.6	[<0.001]*	65.4	[<0.001]*	

Note: Asterisk (*) indicates statistically significant difference in chart accuracy between chart characteristic levels with a p-value less than 0.05; plus-sign (+) indicates a bias measure, calculated as the difference in the differences of accuracy among chart characteristics between the Human+AI and Human-alone groups.

Abbreviations: ΔH+AI (delta H plus AI) = Difference in chart accuracy between Human+AI groups; ΔH (delta H) = Difference in chart accuracy within Human-alone groups; NSCLC = Non-Small Cell Lung Cancer; CrCa = Colorectal Cancer.

Caption: The table compares the chart abstraction accuracy of Human+AI and Human-alone groups across various characteristics, including Length, Complexity, Order, and Cancer Type. Levels are each characteristic are listed under the “Characteristic” column subheadings and defined in the **Methods section, under “Outcomes”.** ΔH+AI (delta H plus AI) values represent differences in chart abstraction accuracy between characteristic levels within the Human+AI group, while the ΔH (delta H) values show these differences within the Human-alone groups. Wilcoxon Rank Sum tests were used to compare chart-level accuracy outcomes between charts stratified by dichotomized chart characteristics.

Supplemental Table 3: Subgroup Analysis of Efficiency Across Chart Characteristics

Characteristic	Human + AI H+AI		Human-Alone H		+ Bias H+AI - H
	Efficiency (minutes)	Difference (1-2) [p-value]	Efficiency (minutes)	Difference (1-2) [p-value]	
Length					
(1) Short	24.8	-25.4	25.4	-24.8	-0.6
(2) Long	50.2	[<0.001]*	50.2	[<0.001]*	
Complexity					
(1) Low	27.0	-23.1	28.5	-20.6	-2.5
(2) High	50.1	[<0.001]*	49.1	[<0.001]*	
Order					
(1) Early	42.2	9.6	44.6	13.6	-4.0
(2) Later	32.6	[<0.001]*	31.0	[<0.001]*	
Cancer Type					
(1) NSCLC	33.7	-8.3	33.2	-10.1	1.8
(2) CrCa	42.0	[<0.001]*	43.3	[<0.001]*	

Note: Asterisk (*) indicates statistically significant difference in efficiency between chart characteristic levels with a p-value less than 0.05; plus-sign (+) indicates a bias measure, calculated as the difference in the differences of efficiency among chart characteristics between the Human+AI and Human-alone groups.

Abbreviations: ΔH+AI (delta H plus AI) = Difference in efficiency between Human+AI groups; ΔH (delta H) = Difference in efficiency within Human-alone groups; NSCLC = Non-Small Cell Lung Cancer; CrCa = Colorectal Cancer.

Caption: The table compares the chart abstraction efficiency of Human+AI and Human-alone groups across various characteristics, including Length, Complexity, Order, and Cancer Type. Levels are each characteristic are listed under the “Characteristic” column subheadings and defined in the **Methods section, under “Outcomes”.** ΔH+AI (delta H plus AI) values represent differences in chart abstraction efficiency between characteristic levels within the Human+AI group, while the ΔH (delta H) values show these differences within the Human-alone groups. Wilcoxon Rank Sum tests were used to compare efficiency outcomes between charts stratified by dichotomized chart characteristics.

Supplemental Table 4: Accuracy Classification of CRC Responses Mapped to Gold-Standard Set

Match Type	CRC-coded element present in gold-standard set?	Description	Example	Accuracy Classification
Identical	Yes	The CRC response exactly matches the gold-standard set	CRC: Lung Cancer; Gold: Lung Cancer	Accurate
Disagree	No	The CRC response does not match the gold-standard set	CRC: Colorectal Cancer; Gold: Breast cancer	Not Accurate
Partial Super Type	No	The gold-standard set is more specific than the CRC response	CRC: Adenocarcinoma; Gold: Mucinous Adenocarcinoma	Not Accurate
Partial Sub Type	Yes	The gold-standard set is less specific than the CRC response	CRC: pN1c; Gold: N1c	Not Accurate
Missing in First	No	The CRC identifies an event not in the gold-standard set	CRC: Colorectal Cancer; Gold: [Blank]	CRC item removed and not counted in numerator or denominator
Missing in Second	No	The CRC fails to identify an event in the gold-standard set	CRC: [Blank]; Gold: Colorectal Cancer	Not Accurate

CAPTION: Descriptions of how CRC-coded responses are mapped to elements of the gold-standard set to determine accurate vs. not accurate responses. Each potential pairing of responses from the CRC and gold-standard set was labeled as one Match Type. Examples of these match types as well as the resulting Accuracy Classification are shown. A separate sensitivity analysis was performed where Partial Sub Type had an Accuracy Classification of "Accurate."

Supplemental Methods

AI Language Model

Data Pipeline

This section comprehensively addresses the AI-enabled data pipeline, providing detailed descriptions of the inputs and outputs and elucidating how each inference and step contributes to the overall pipeline flow. Each step in the pipeline operates at different units of the input; i.e., some operate at the page level, while others operate at the file and/or combined source file level or at a document level.

Data Pre-Processing

The clinical notes in PDF or other unstructured formats were first processed through an OCR system specifically designed for healthcare applications. This system ensured that scanned documents are accurately converted into machine-readable text. The OCR system operated at the page-level, providing text output that could be traced back to the source evidence. To ensure patient privacy, all personal health information was then attempted to be redacted from the documents using a combination of machine learning and rule-based systems.

Entity Extraction

A machine learning-based system extracted predefined clinical entities (e.g., "cancer," "tumor marker," "diagnostic procedure") from the de-identified text. We then used transformer-based (BERT) models 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. This system analyzed the text at a document-level and tags relevant data, ensuring that each reference to clinical information is appropriately labeled for further analysis. This system also predicted relationships between text such as predictions about temporal information, negation, or a disease response.

Document Structuring and Categorization

Following entity extraction, the system segmented the clinical text into discrete reports based on logical boundaries within the combined source file. For example, concatenated medical reports were separated into individual documents (e.g., pathology reports, SOAP notes, radiology summaries). Each document was then categorized by its type (e.g., radiology, pathology, or clinical note) and relevant dates are extracted.

Event Consolidation and Symbolic Reasoning

From the data tags generated from the entity extraction and entity relation stage, the system performed relational mapping, predicting relationships between these data. For example, if a tumor marker was associated with a specific diagnosis, the system recognized and recorded this relationship. Each piece of extracted information was then evaluated for trustworthiness based on its context and given a confidence score based on how relevant it was to the patient’s actual clinical journey.

The final stage consolidated all data into meaningful clinical events. A symbolic reasoning system, curated by medical experts, applied logic to the extracted data to deduce clinical events (e.g., treatment initiation, disease progression). The system employed a range of reasoning techniques, for example, entailment[1] (deriving new facts from known facts) and exclusion[2] (removing contradictory information) in addition to conceptual frames that specified order, sequence, and nature of clinical processes. The result was consolidation of alike facts, resolution of ambiguous facts, removal of hallucinations, and deductive enrichment based on what is clinically implicit in the extracted data. The outputs of this stage were what end user CRCs saw in the web-based interface.

Model Training and Evaluation

The system was trained using a curated dataset of over 240,000 medical records from cancer patients; this dataset was used for entity extraction, relationship extraction, and relevance to patient journey. A "gold-standard" corpus, consisting of 18,000 manually annotated clinical notes, was used to guide the training process and was also referenced when assessing the accuracy when comparing all arms (AI-alone, Human-alone, Human+AI). The medical records used for model training were separate from the records used in the randomized study.

Sample size calculations for primary endpoint, mean chart-level accuracy

The primary study was powered on a retrospective, paired, non-inferiority design to evaluate abstraction accuracy of elements commonly used to help screen patients for clinical trials. 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. Each chart was reviewed twice, once in each study arm. Noninferiority was established by comparing the difference in the mean chart-level accuracy between the "Human-alone" and "Human+AI" arms against a noninferiority margin. We used a predefined non-inferiority margin, Δ, of 0.05.

The one-sided hypotheses to test noninferiority of average chart-level accuracy of a Human+AI abstraction approach, compared with a Human-alone approach are outlined below:

H_0: $\mu_{h\text{-}ai} - \mu_h \leq -\Delta$ (Null hypothesis)
H_a: $\mu_{h\text{-}ai} - \mu_h > -\Delta$ (Alternative hypothesis)
Test statistic:

$$T_n = \frac{(\mu_{h-ai} - \mu_h) + \Delta}{\sqrt{\frac{Var(D)}{n}}}$$

Sample size equation for a noninferiority, paired study design:

$$n = \frac{(Z_{1-\beta} + Z_{1-\alpha})^2 \sigma^2}{((\mu_{h-ai} - \mu_h) + \Delta)^2} = \frac{(Z_{1-\beta} + Z_{1-\alpha})^2 \sigma^2}{(D + \Delta)^2}$$

As a secondary analysis, we tested the superiority of the Human+AI arm for achieving greater average chart-level accuracy, compared with the Human-alone arm. Our hypotheses were:

H_0: $\mu_{h\text{-}ai} \leq \mu_h$ (Null hypothesis)
H_a: $\mu_{h\text{-}ai} > \mu_h$ (Alternative hypothesis)

=

Sample size equation for a superiority, paired study design:

$$n = \frac{\left(Z_{1-\beta} + Z_{1-\alpha}\right)^2 Var(D)}{(\mu_{h-ai} - \mu_h)^2}$$

Parameters are defined here:

Parameter	Description
$\mu_{h\text{-}ai}$	Mean proportion of criterion correctly abstracted by human-augmented reviewers
μ_h	Mean proportion of criterion correctly abstracted by human-alone reviewers
$D = \mu_{h\text{-}ai} - \mu_h$	Mean difference in chart-level accuracy between arms
Δ	Non-inferiority margin, set as 0.05 for conservative sample size estimates.
$Z_{1-\beta}$	Critical value at 1-β, , where 1-β represents the desired power and β represents Type II error
$Z_{1-\alpha}$	Critical value at 1-α, , where α = Type I error
σ^2	n* Var(D), where n is the sample size and Var(D) is the variance of the mean difference in chart-level accuracy

*Note: $\mu_{h\text{-}ai}$, μ_h, $\Delta > 0$.

Round 1: Sample size calculations informed by preliminary studies

Preliminary chart abstractions done prior to the start of the study informed initial sample size estimates to achieve a power of at least 80% at a Type I error of 0.05. Initial experiments assessed criterion-level abstraction events for a Human+AI arm and a Human-alone arm. Chart-level accuracy estimates and variance of a paired difference in chart-level accuracy between arms was calculated using true positive rates (TPR), event rates (p), and true negative rates (TNR) extracted from preliminary results. Definitions and assumptions for these values are included in the derivation and simulation results below. Rmarkdown page with code and preliminary dataset are included on github at https://github.com/HumanAlgorithmCollaborationLab/TrialPrescreening.

Derivation of chart level accuracy estimates and variance of effect size from available preliminary data is shown below:

Let Y_i represent the proportion of correctly classified criteria for the ith patient, and S_{ij} denote the classification accuracy result of the jth (j=1..m) criterion for the ith (i=1..n) patient. Then,

S_{ij} ~ Bernoulli(E(S_{ij}

)), where E(S_{ij}) = P(S_{ij} = 1) = probability recorded observation is correct for category j

S_{ij} = {1 if correct classification, otherwise 0 if not correct}

Then Y_i =$\sum_{j=1}^{m} \frac{S_{ij}}{m}$

Let Z_{ij} denote the classification result of the jth criterion for the ith patient, where

Z_{ij} ~ Bernoulli(E(Z_{ij})), where E(Z_{ij}) = P(Z_{ij} = 1) = probability recorded observation is marked eligible for category j

Z_{ij} = {1 if marked eligible, otherwise 0, if marked not eligible}

And let T_{ij} denote the true classification of the jth criterion for the ith patient, where

T_{ij} ~ Bernoulli(E(T_{ij})), where E(T_{ij}) = P(T_{ij} = 1) = probability recorded observation is truly eligible for category j (event rate)

T_{ij} = {1 if truly eligible, otherwise 0, if truly not eligible}

Then:

$$\mu = E(Y_i) = E\left(\sum_{j=1}^{m} \frac{S_{ij}}{m}\right) = \frac{1}{m}\sum_{j=1}^{m} E(S_{ij}) = \frac{1}{m}\sum_{j=1}^{m} P(S_{ij} = 1)$$

Where,

$$P(S_{ij} = 1) = P(Z_{ij} = 1, T_{ij} = 1) + P(Z_{ij} = 0, T_{ij} = 0)$$

$$= P(Z_{ij} = 1 | T_{ij} = 1) P(T_{ij} = 1) + P(Z_{ij} = 0 | T_{ij} = 0) P(T_{ij} = 0)$$

$$= \frac{TP_j + TN_j}{TP_j + TN_j + FP_j + FN_j}$$

And $Var(Y_j) = Var\left(\frac{1}{m}\sum_{j=1}^{m} S_{ij}\right) = \frac{1}{m^2}\sum_{j=1}^{m} Var(S_{ij}) = \frac{1}{m^2}\sum_{j=1}^{m} \left[P(S_{ij} = 1)\left(1 - P(S_{ij} = 1)\right)\right]$

We calculate μ, a constant mean proportion of correctly classified criterion, for both the Human+AI collaboration arm, $\mu_{h-ai} = E\left(Y_i^{h-ai}\right)$
, and the Human-alone arm, $\mu_h = E\left(Y_i^h\right)$
. We assume independence in the evaluation of different individuals and different categories.

In this study design, the Human-alone and Human+AI collaboration arms have the same sample population, leading to paired data.
$D = \frac{1}{n}\sum_{i=1}^{n} \quad Y_i^{h-ai} - Y_i^h$
$Var(D) = Var\left(\frac{1}{n}\sum_{i=1}^{n} \quad Y_i^{h-ai} - Y_i^h\right) = \frac{1}{n^2}\sum_{i=1}^{n} \quad Var\left(Y_i^{h-ai} - Y_i^h\right) = \frac{1}{n}\left[Var\left(Y_i^{h-ai}\right) + Var\left(Y_i^h\right) - 2Cov\left(Y_i^{h-ai}, Y_i^h\right)\right]$
Let $\sigma^2 = Var(D') = Var\left(Y_i^{h-ai}\right) + Var\left(Y_i^h\right) - 2Cov\left(Y_i^{h-ai}, Y_i^h\right)$
,

where covariance depends on the concordance rate different between criterion between arms:

$$Cov\left(Y_i^{h-ai}, Y_i^h\right) = Cov\left(\frac{1}{m}\sum_{j=1}^{m} \quad S_{ij}^{h-ai}, \frac{1}{m}\sum_{j=1}^{m} \quad S_{ij}^h\right) = \frac{1}{m^2}\sum_{j=1}^{m} \quad Cov\left(S_{ij}^{h-ai}, S_{ij}^h\right)$$

, where we assume independence between individuals and between criterion and $Cov\left(S_{ij}^{h-ai}, S_{ij}^h\right) = E\left(S_{ij}^{h-ai}, S_{ij}^h\right) - E\left(S_{ij}^{h-ai}\right)E\left(S_{ij}^h\right) = P\left(S_{ij}^{h-ai} = 1, S_{ij}^h = 1\right) - P\left(S_{ij}^{h-ai} = 1\right)P\left(S_{ij}^h = 1\right)$

Where, $P\left(S_{ij}^{h-ai} = 1, S_{ij}^h = 1\right) = P\left(Z_{ij}^{h-ai} = 1, Z_{ij}^h = 1, T_{ij} = 1\right) + P\left(Z_{ij}^{h-ai} = 0, Z_{ij}^h = 0, T_{ij} = 0\right)$
$= P\left(Z_{ij}^{h-ai} = 1 \middle| Z_{ij}^h = 1, T_{ij} = 1\right)P\left(Z_{ij}^h = 1 \middle| T_{ij} = 1\right)P\left(T_{ij} = 1\right)$
$+ P\left(Z_{ij}^{h-ai} = 0 \middle| Z_{ij}^h = 0, T_{ij} = 0\right)P\left(Z_{ij}^h = 0 \middle| T_{ij} = 0\right)P\left(T_{ij} = 0\right)$
$= P\left(Z_{ij}^{h-ai} = 1 \middle| Z_{ij}^h = 1, T_{ij} = 1\right)TPR_j^{h-ai}p_j + P\left(Z_{ij}^{h-ai} = 0 \middle| Z_{ij}^h = 0, T_{ij} = 0\right)TNR_j^h\left(1 - p_j\right)$

We may assume $P\left(Z_{ij}^{h-ai} = 1 \middle| Z_{ij}^h = 1, T_{ij} = 1\right) = P\left(Z_{ij}^{h-ai} = 0 \middle| Z_{ij}^h = 0, T_{ij} = 0\right)$
=1
Then,

$$Cov\left(Y_i^{h-ai}, Y_i^h\right) = \frac{1}{m^2}\sum_{j=1}^{m} \quad Cov\left(S_{ij}^{h-ai}, S_{ij}^h\right)$$
$$= \frac{1}{m^2}\sum_{j=1}^{m} \quad \left[P\left(S_{ij}^{h-ai} = 1, S_{ij}^h = 1\right) - P\left(S_{ij}^{h-ai} = 1\right)P\left(S_{ij}^h = 1\right)\right]$$

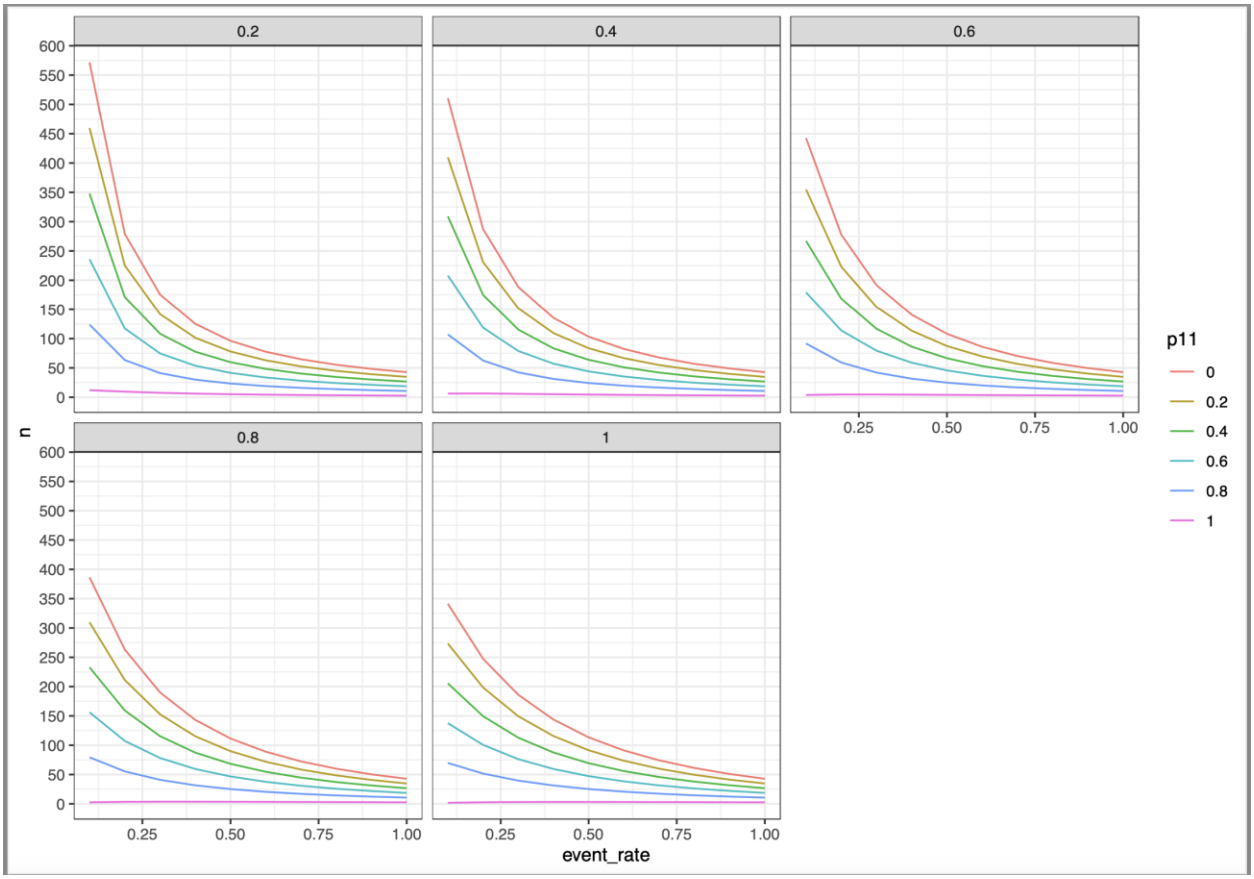
Note: TP = true positives; TN = true negatives; FP = false positive; FN = false negative.

Simulations and Results:

Simulations varied the concordance rate between arms, event rate, and true negative rate, while fixing desired power at 0.80, type I error rate of 0.05, and the noninferiority margin at 0.05. We assumed equal true negative rates across arms. Mean chart-level accuracy was derived from preliminary criterion-level data, which is included on github. The same event rates (i.e., 0.1, 0.2,

0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), true negative rates (0.4, 0.6, 0.8, 1.0), and concordance probabilities (i.e., 0.0, 0.2, 0.4, 0.6, 0.8, 1.0) were applied for each criterion for simulations.

Supplemental Figure 1. Round 1 Sample Size Calculations



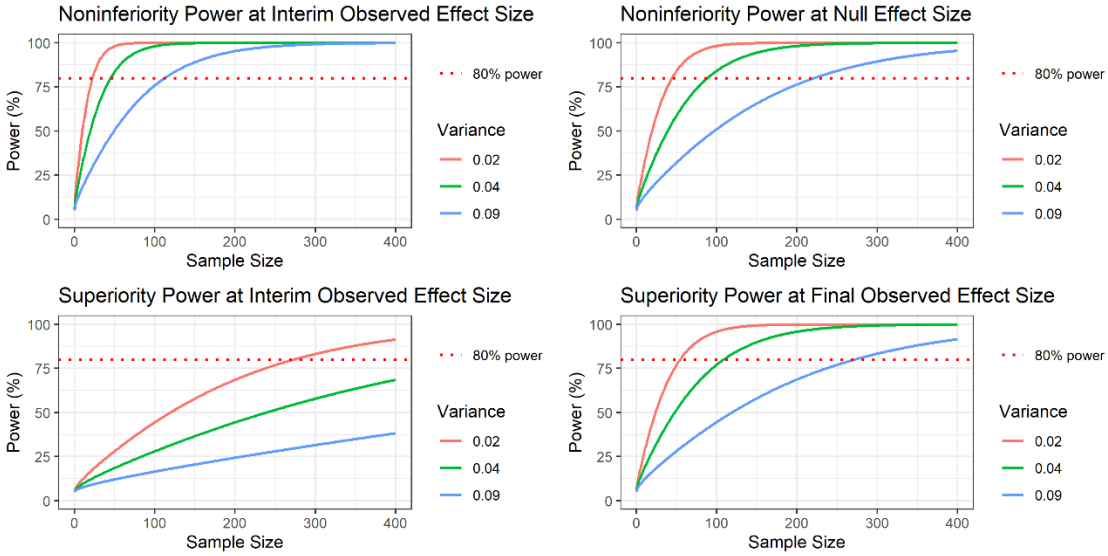
CAPTION: *Minimum sample size required for various event rates and proportions of true positive outcomes across different scenarios. The y-axis represents the minimum sample size needed to achieve statistical power under varying conditions of concordance (p11).*

Round 2: Sample size recalculated with interim results

Interim sample size calculations were made after the first round of chart abstractions, which included 74 charts reviewed in both the Human+AI and Human-alone arms. Sample size calculations for paired t-tests were generated for both the non-inferiority and superiority contexts, and parameterized to achieve a power of at least 80% at a Type I error of 0.05. Additionally, the interim results further informed parameterization of the sample size calculation with regards to effect size and variance. Among the 74 charts, the mean chart-level accuracy in the Human+AI arm was 78.67% (SD: 18.79) and mean chart-level accuracy in the Human-alone arm was 76.65% (SD: 20.47). The resulting effect size is the paired difference in mean chart-level accuracy between Human+AI and Human-alone, which was 2.01% (SD: 13.38). Interim sample size calculations were completed using PASS statistical software (Version 23.0.1). Power curves for the superiority and non-inferiority outcomes were generated in R, informed by

the interim results, where the effect size and variance were varied to show the resulting sample size of 355 charts needed to achieve a power of at least 80%.

Supplemental Figure 2. Round 2 Sample Size Calculations

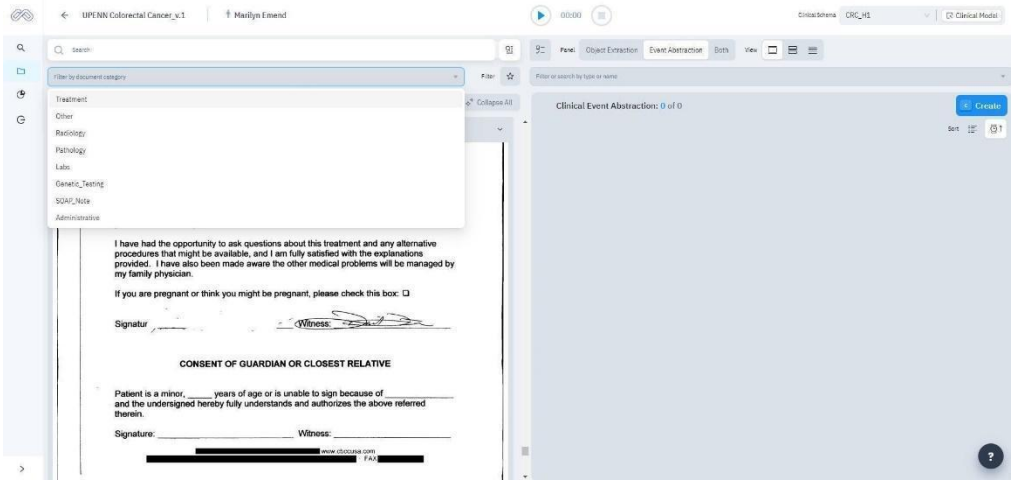


CAPTION: *Power analysis based on results from interim analysis of 74 charts reviewed in both Human+AI and Human-alone arms. Interim variance = 0.02 and interim effect size = 0.02. Larger sample sizes and smaller variances result in higher statistical power.*

Supplemental Figures: Representative Images of Abstraction Interface

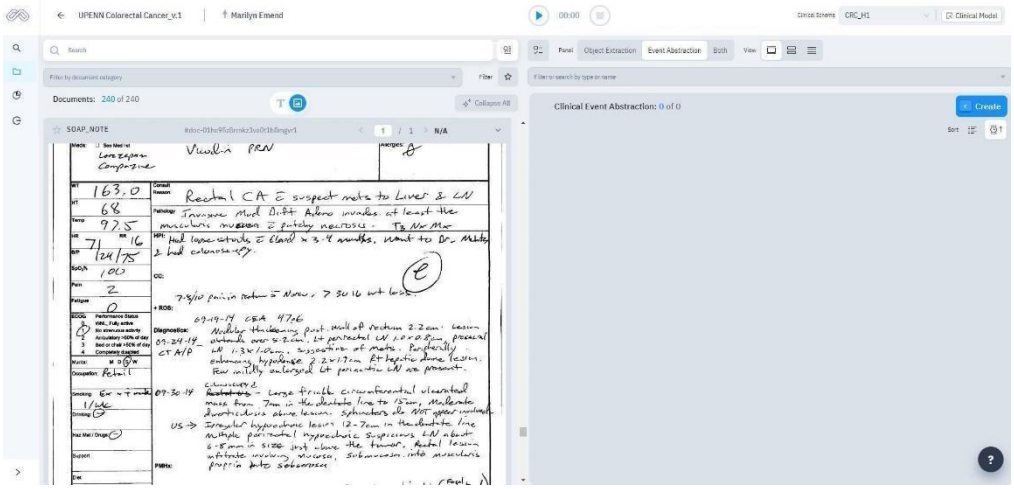
Supplemental Figures 3-5: Example outputs of the left half of trial abstraction interface, containing source documentation

Supplemental Figure 3: Example 1 of source documentation interface



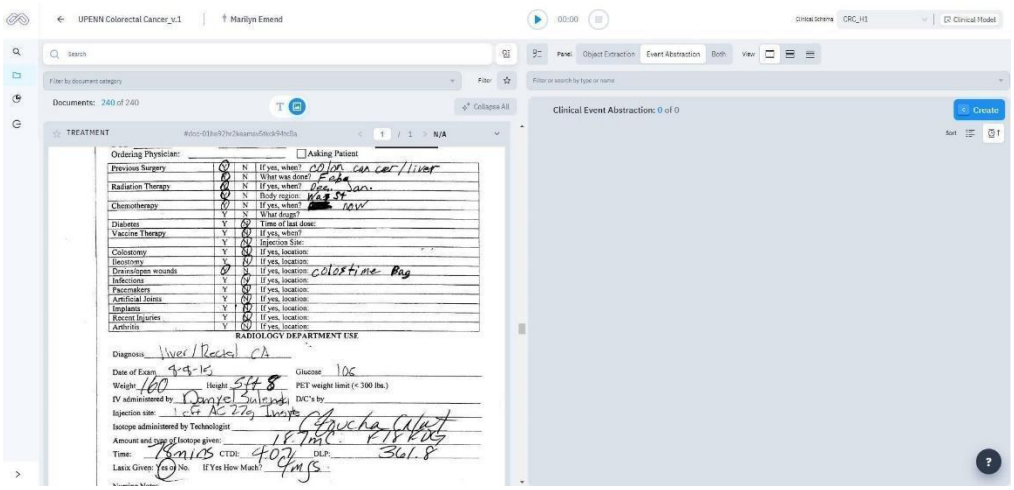
CAPTION: Left half of trial abstraction interface, containing source documentation. Available in both the Human-alone and Human+AI configurations.

Supplemental Figure 4: Example 2 of source documentation interface



CAPTION: Left half of trial abstraction interface, containing source documentation. Available in both the Human-alone and Human+AI configurations.

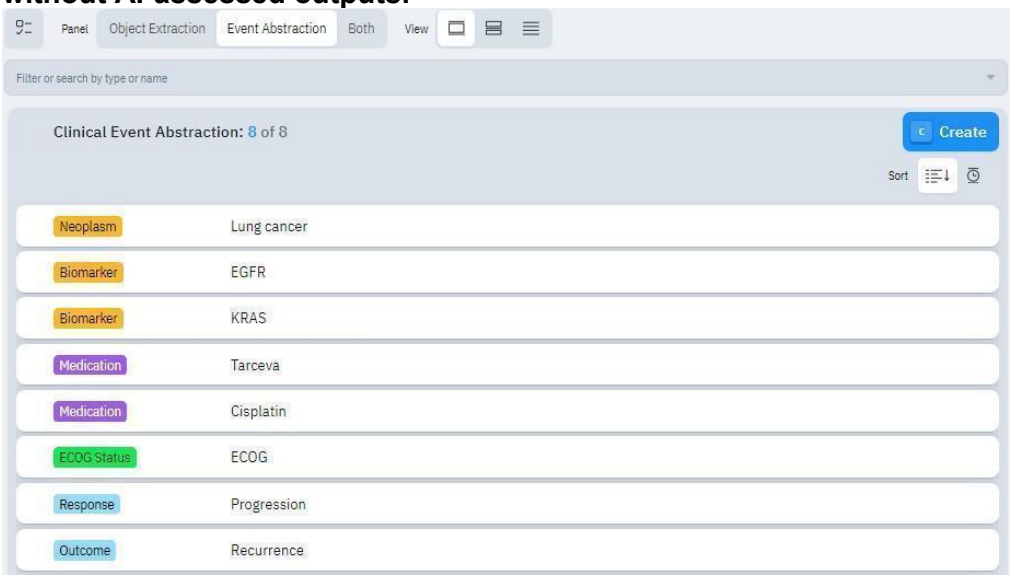
Supplemental Figure 5: Example 3 of source documentation interface



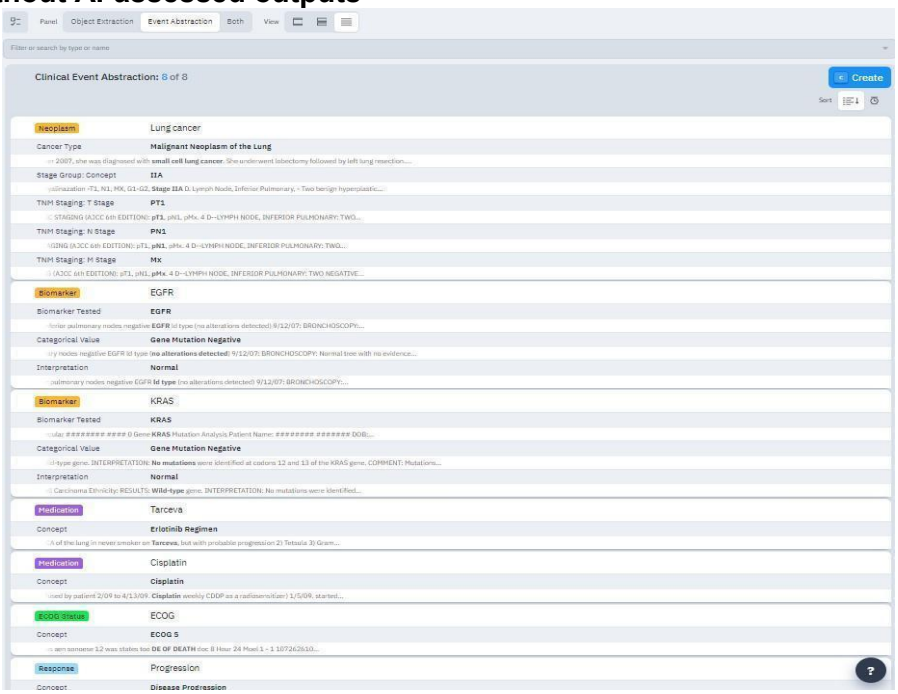
CAPTION: Left half of trial abstraction interface, containing source documentation. Available in both the Human-alone and Human+AI configurations.

Supplemental Figures 6-7: Example outputs of the right half of the abstraction interface for Human-alone arm, containing extractable elements

Supplemental Figure 6: Collapsed view of abstraction interface for Human-alone arm, without AI-assessed outputs.

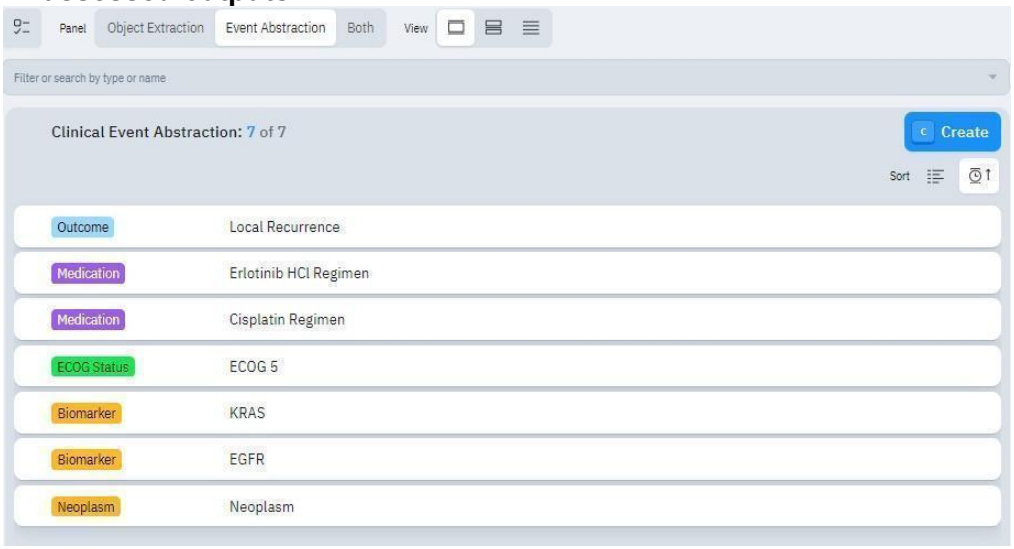


Supplemental Figure 7. Expanded view of abstraction interface for Human-alone arm, without AI-assessed outputs

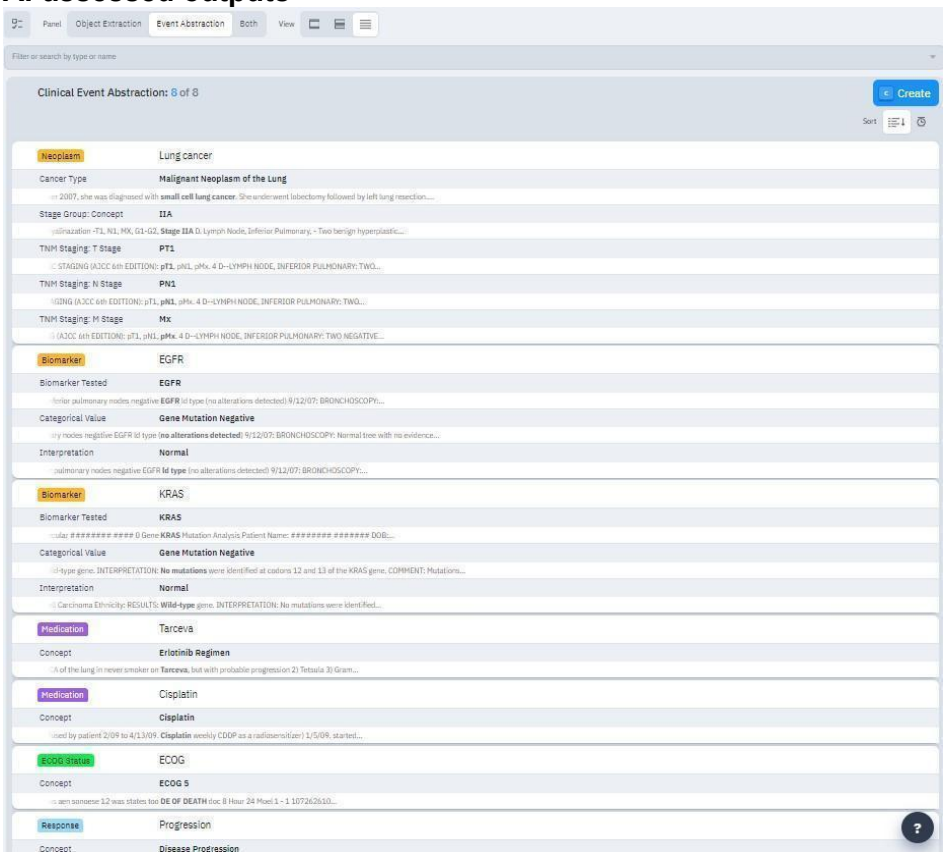


Supplemental Figures 8-9: Abstraction interface for Human+AI Arm

Supplemental Figure 8. Collapsed view of abstraction interface for Human+AI arm, with AI-assessed outputs.



Supplemental Figure 9. Expanded view of abstraction interface for Human+AI arm, with AI-assessed outputs



Supplemental Note 1: Study Protocol

Assessing the Performance of Artificial Intelligence (AI)-augmented Electronic Health Record (EHR) Data Abstraction for Clinical Trial Patient Screening

Study Protocol

Version date: January 2025

Outline

1. Abstract
2. Overall objectives
3. Aims
    - 3.1 Primary outcome
    - 3.2 Secondary outcomes
4. Background
5. Study design
    - 5.1 Design
    - 5.2 Study duration
    - 5.3 Target population
    - 5.4 Accrual
    - 5.5 Key inclusion criteria
    - 5.6 Key exclusion criteria
6. Subject recruitment
7. Subject compensation
8. Study procedures
    - 8.1 Consent
    - 8.2 Procedures
9. Analysis plan
10. Investigators
11. Human research protection
    - 11.1 Data confidentiality
    - 11.2 Subject confidentiality
    - 11.3 Subject privacy
    - 11.4 Data disclosure
    - 11.5 Data safety and monitoring
    - 11.6 Risk/benefit
        - 11.6.1 Potential study risks
        - 11.6.2 Potential study benefits
        - 11.6.3 Risk/benefit assessment

1. Abstract

The identification of eligible patients for clinical trials is a key component - and a rate-limiting step - of the clinical research enterprise. Currently, the process of identifying eligible patients relies on manual chart review by clinical research staff, which can be time-consuming, labor intensive and prone to human error. As a result, many eligible patients may be overlooked and opportunities for trial participation may be missed: this is particularly true in the oncology space, where fewer than 5% of adult cancer patients enroll in trials. The integration of AI technology into the clinical trial patient screening process has potential to improve trial participation rates and the generalizability of research findings. This study aims to assess the performance (accuracy, efficiency) of AI-augmented patient identification and inform optimal integration of AI technology into clinical research screening processes.

2. Overall objectives

The objective of this study is to assess and compare the accuracy and efficiency of three different approaches to abstracting clinical data used to identify oncology patients who meet the inclusion criteria for participation in clinical trials. The three approaches under evaluation include: (1) an autonomous AI algorithm (Mendel AI; developed by artificial intelligence startup company Mendel) which analyzes patient medical records to extract relevant clinical facts ("AI-alone"); (2) a human researcher who manually reviews patient charts as per the current norm/practice ("Human-alone"); and (3) a human researcher utilizing AI augmentation ("Human+AI"), where Mendel AI serves as a supportive tool in the decision-making process by providing the researcher a list of elements abstracted by the AI algorithm and a rank-order list of patients most likely to meet inclusion criteria for a trial.

The study primarily aims to compare (1) the chart-level accuracy of the Human+AI collaboration relative to Human-alone given the relevance of this comparison for real-world clinical workflows, defined by the percentage of pre-identified chart elements classified correctly compared against a predetermined "gold standard"; and (2) the efficiency of the Human+AI vs. Human-alone arms, defined by the time per chart review in minutes, measured for each chart.

Our hypotheses are (1) the Human+AI arm will be non-inferior in accuracy when compared to the Human-alone arm, in relation to a predetermined "gold standard", and (2) that a Human+AI arm will be superior in speed of abstraction when compared to Human-alone screening.

The identification of eligible patients for clinical trials is a critical component of clinical research, as it directly impacts patient recruitment, study enrollment, and the generalizability of research findings. Currently, the process of identifying eligible patients often relies on manual chart review by clinical research staff, which can be time-consuming, labor-intensive, and prone

to human error. Consequently, eligible patients may be overlooked, and opportunities for trial participation may be missed. The integration of AI technology into the patient identification process has the potential to enhance the accuracy and efficiency of this critical task, leading to improved clinical trial recruitment and outcomes.

This study holds important implications for the field of clinical research by evaluating the effectiveness of AI-augmented patient identification compared to traditional manual methods and autonomous AI algorithms. By examining the strengths and limitations of each approach, the study will provide valuable insights into the optimal integration of AI technology in clinical research processes. Furthermore, the results of this study have the potential to benefit patients by improving their access to clinical trials and increasing awareness of available treatment options. For clinical research institutions, enhancing the efficiency of patient identification can lead to more effective use of research resources and the potential for accelerated clinical trial timelines. Ultimately, the findings of this study may contribute to advancements in clinical research practices, promoting more equitable access to trials and facilitating the development of innovative treatments for patients with cancer.

3. Aims

3.1 Primary

Aim 1: Compare the **accuracy** of the Human+AI vs. Human-alone arms, in relation to a “Gold Standard” determined by multiple human abstractors with no time limitation.

3.2 Secondary

Aim 2: Compare the **efficiency** of the Human+AI vs. Human-alone arms

4. Background

Insufficient patient enrollment in clinical trials is a significant challenge faced by the clinical trials community, and it is often considered the rate-limiting step in completing clinical trials. The traditional approach to determining patient eligibility for clinical trials is manual eligibility screening (ES), which requires a labor-intensive review of patient records, leading to a high workload for clinical staff and a potential delay in trial enrollment. The problem of insufficient patient enrollment is symbolized by the fact that less than 5% of patients with cancer participate in clinical trials. [1] The manual eligibility screening process is resource-intensive, and reducing the burden of screening is a priority for improving trial participation rates and the efficiency of clinical research.

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) offer potential solutions to this challenge. These technologies have the capability to automatically detect patients' eligibility for clinical trials by mapping coded clinical trial eligibility criteria to corresponding clinical information extracted from electronic health records (EHRs). This process involves extracting relevant eligibility criteria from clinical notes using machine learning-based NLP applications and comparing these extracted criteria with reference criteria from trial descriptions. Studies have shown that machine learning-based NLP applications provide high accuracy in extracting eligibility criteria from clinical notes and determining trial eligibility, with recall rates of up to 90.9% and precision rates of up to 89.7%. [2]

Automated ES algorithms have demonstrated success in significantly reducing the workload of clinical staff involved in trial-patient matching, thereby increasing the trial screening efficiency of oncologists. Additionally, the use of AI-based tools has the potential to enable participation of smaller practices, which are often left out from trial enrollment due to resource constraints. The integration of AI and NLP technologies in the clinical trial eligibility screening process can play a crucial role in increasing total patient enrollment, the speed of enrollment, as well as expanding access to trials, and enhancing the execution of clinical research.

[1] Unger et al. Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. *American Society of Clinical Oncology.* 2017. DOI: 10.14694%2FEDBK_156686

[2] Meystre et al. Automatic Trial Eligibility Surveillance Based on Unstructured Clinical Data. *International Journal of Medical Informatics*. 2019. DOI: 10.1016/j.ijmedinf.2019.05.018

5. Study design

5.1 Design

In this experimental study, we will employ a three-arm design including (1) AI-alone, (2) Human-alone, (3) Human+AI to assess the accuracy of abstracting clinical elements from the patient chart. The accuracy of the abstraction will be compared against a pre-made “gold standard”. As detailed elsewhere, the study is powered for a direct comparison of the human alone vs human + AI arms as a non-inferiority analysis. For Arm 1, an autonomous AI algorithm (Mendel AI) will analyze patient medical records to extract relevant clinical elements (“AI-alone”). For Arm 2, a human researcher will manually review patient charts as per the current norm/practice (“Human-alone”). For Arm 3, a human researcher will utilize Mendel AI augmentation such that Mendel AI serves as a supportive tool in the decision-making process by providing the researcher a list of elements abstracted by the AI algorithm and a rank-order list of patients most likely to meet inclusion criteria for a trial (“Human+AI”). **For this study, we will**

only be using information from deidentified patient charts in Mendel's database. <u>No Penn data</u> will be utilized in this study.

We will employ 2 clinical research coordinators (CRCs) who will review and abstract specific clinical elements while reviewing the patient's electronic health record (EHR) which will include both "structured" (machine-readable) and "unstructured" (traditionally non-machine-readable; Mendel AI is uniquely capable of processing) data. To ensure no inter-rater differences in accuracy the CRCs will be assessed on a practice set of three charts and must meet an 80% threshold of agreement on review elements as well as concordance in chart-level accuracy as measured by Cohen's Kappa (threshold 0.6).

The components of the EHR that may be reviewed to abstract the data elements of interest may include:

- Physician Progress Notes
- Radiology Reports
- Pathology Reports
- Laboratory Reports
- Genetic/molecular sequencing and other genetic testing results
- Discharge Summaries
- Infusion Therapy reports

To help assess accuracy, the "Gold Standard" (to which all study arms will be compared) will be a manually annotated version based on review of at least 2 human abstractors (plus a tie-breaker if there is a discrepancy). This process will be led by the Mendel team, with independent oversight by the Penn research team to ensure consistency.

First, the Mendel AI algorithm will review all patient charts, producing a machine-annotated version of each chart with AI-abstracted elements (AI-alone). Next, each CRC will review all patient charts. Each CRC will review half of their charts manually (Human-alone) and half of their charts using the machine-annotation and rank-order list (Human+AI). The order of review will be randomized such that manual reviews and AI-assisted reviews will be mixed, and such that each CRC will review half of the charts manually and half of the charts with AI assistance.

Secondary aim metrics (Efficiency) will be assessed by tracking the volume of completed prescreens per unit of time.

5.2 Study duration

We anticipate the "review" phase of this study where human researchers are abstracting data from patient charts to take roughly 6 months, followed by 6 months for analysis.

5.3 Target population

This study will include de-identified patient charts from Mendel's databases from patients in community oncology practices with a diagnosis of non-small cell lung cancer (NSCLC) or colorectal cancer (CRC) with a minimum of 5 patient documents and data within 5 years from the time of data extraction. No Penn data will be utilized in this study and we will use standard EHR datasets.

5.4 Accrual

Mendel AI upholds a database of clinical data for deidentified patients from participating community-based oncology practices. This data will be utilized for this study. As noted above, no Penn data will be utilized in this study.

5.5 Key inclusion criteria

A patient chart will be included for analysis based on the following three screening criteria:

- A diagnosis of NSCLC or colorectal cancer
- A minimum of 5 patient documents in the Mendel database
- Most recent document was within 5 years from the time of data extraction

6. Subject recruitment

No subjects will be involved with this work.

7. Subject compensation

There will be no compensation for the use of subjects' de-identified data collected as part of routine clinical care.

8. Study procedures

8.1 Consent

The data utilized for analysis will be obtained from existing electronic health records (EHR) that have already been collected as part of routine clinical care and collated in the Mendel AI datasets. As such, patients will not be directly involved in the research process, and the study will not require any new data collection or intervention from patients.

Most importantly, all patient data used in the study has been de-identified and anonymized to ensure the protection of patient privacy and confidentiality.

In light of these considerations, patients will not need to provide consent for their data to be included in the study. This approach aligns with the ethical standards for retrospective studies using de-identified data, and the study design has been developed to ensure compliance with relevant regulations and guidelines on patient privacy and data protection.

As noted before, no Penn data will be utilized in this study.

8.2 Procedures

Our study will follow the design outlined above (section 5.1).

There is no informed consent needed for the use of this data. There will be no participant recruitment or compensation. Potential risks and benefits of the study are reported elsewhere (section 11.6).

Two CRCs will be hired to analyze de-identified patient data. CRCs will access the data via a password-protected portal, using a UPHS desktop and/or laptop. De-identified patient chart data will be included for analysis based on having a tumor type of interest. They will then review patient charts and record clinical elements of interest in the Mendel abstraction tool, which will be secured in a HIPAA-compliant external enclave. Summary-level data will be exported into datasheets for analysis. These datasheets will only be used and analyzed on UPHS equipment, or shared via HIPAA-compliant means.

Data will be analyzed using statistical software including Excel, R code, and STATA. For details of the analysis plan, see section 9.

Data elements that will be abstracted across all arms of the study include the following (the below list may change slightly pending further discussion among the research team):

Cancer Type and Staging

- Cancer Type
- Stage - Tumor
- Stage - Node
- Stage - Mets

Performance

- ECOG Performance Status

Prior Systemic Treatment

- Any Prior Platinum Therapy (e.g., Cisplatin, Carboplatin, Oxaliplatin)

- Any Other Prior Chemotherapy (e.g., 5-fluorouracil, Leucovorin, Capecitabine, Irinotecan, Pemetrexed, Paclitaxel, Docetaxel, Gemcitabine)

- Any Prior Immunotherapy (e.g., Atezolizumab, Nivolumab, Pembrolizumab)

Biomarkers for CRC

- KRAS
- MSI-H
- MSS
  BRAF
- HER2
- NTRK
- DPYD

Biomarkers for NSCLC

- EGFR
- ALK
- ROS1
- KRAS
- BRAF
- NTRK
- cMET
- RET
- PDL1

Others

- Outcomes (remission, residual disease, recurrence, loco-regional recurrence, distant recurrence)
- Responses (complete response, stable disease, partial response, mixed response, disease progression)

9. Analysis plan

The primary study will be powered on a retrospective paired, non-inferiority design to evaluate abstraction accuracy of elements commonly used to help screen patients for clinical trials. The two arms are (1) human reviewers versus (2) human reviewers utilizing predictions from Mendel AI. The primary outcome, the percentage of gold standard items correctly abstracted, will be measured for each chart. Noninferiority can be established by comparing the difference in the

mean proportion correct between the human-augmented and Human-alone arms with a predefined non-inferiority margin, Δ. Noninferiority will be tested using a one-sided paired t-test with non-inferiority margin, Δ. The distribution of paired differences will be tested for normality using the Shapiro-Wilk test. If normality is rejected, then the median proportions will be compared using the nonparametric Wilcoxon Rank Sum test.

The one-sided hypotheses to test noninferiority of average chart-level accuracy of a Human+AI abstraction approach, compared with a Human-alone approach are outlined below:

H_0: $\mu_{h\text{-}ai} - \mu_h \leq -\Delta$ (Null hypothesis)
H_a: $\mu_{h\text{-}ai} - \mu_h > -\Delta$ (Alternative hypothesis)

Test statistic:

$$T_n = \frac{(\mu_{h-ai} - \mu_h) + \Delta}{\sqrt{\frac{Var(D)}{n}}}$$

Sample size equation for a noninferiority, paired study design:

$$n = \frac{(Z_{1-\beta} + Z_{1-\alpha})^2 \sigma^2}{((\mu_{h-ai} - \mu_h) + \Delta)^2} = \frac{(Z_{1-\beta} + Z_{1-\alpha})^2 \sigma^2}{(D + \Delta)^2}$$

If both arms are noninferior, we will test the superiority of the Human+AI arm for achieving greater average chart-level accuracy, compared with the Human-alone arm. Superiority of chart-level accuracy for the Human+AI arm relative to Human-alone will be tested using a one-sided paired t-test with an unspecified inferiority margin. The distribution of paired differences will be tested for normality using the Shapiro-Wilk test. If normality is rejected, then superiority will be assessed nonparametrically using a one-sided paired Wilcoxon Rank Sum test with an unspecified inferiority margin.

Our hypotheses are:

H_0: $\mu_{h\text{-}ai} \leq \mu_h$ (Null hypothesis)
H_a: $\mu_{h\text{-}ai} > \mu_h$ (Alternative hypothesis)

Sample size equation for a superiority, paired study design:

$$n = \frac{(Z_{1-\beta} + Z_{1-\alpha})^2 Var(D)}{(\mu_{h-ai} - \mu_h)^2}$$

Parameter	**Description**

$\mu_{h\text{-}ai}$	Mean proportion of criteria correctly abstracted by AI-augmented human reviewers
μ_h	Mean proportion of criteria correctly abstracted by Human-alone reviewers
$D = \mu_{h\text{-}ai} - \mu_h$	Mean difference in chart-level accuracy between arms
Δ	Non-inferiority margin, set as 0.05 for conservative sample size estimates.
$Z_{1-\beta}$	Critical value at 1-β, where 1-β represents the desired power and β represents Type II error
$Z_{1-\alpha}$	Critical value at 1-α, where α = Type I error
σ^2	n* Var(D), where n is the sample size and Var(D) is the variance of the mean difference in chart-level accuracy

*Note: $\mu_{h\text{-}ai}, \mu_h, \Delta > 0$.

From our initial calculations, 400 individual patients will achieve an estimated statistical power greater than 90% for this non-inferiority study with a Type I error of 5%. This was determined from simulations for different realistic parameter scenarios for TNR, event rate, and concordance assumptions showing that the needed sample size rarely exceeded 400. Given uncertainty in these estimates, a pre-planned interim analysis at approximately 75 charts reviewed will be used to refine sample size estimates.

The goals of the secondary analyses are to compare the efficiency of screening among the three study arms. We will analyze efficiency, defined by comparing the average number of minutes per chart review, measured for each chart. A nonparametric, paired Wilcoxon Rank Sum test with continuity correction will compare overall median efficiency of chart abstraction between arms.

The following are pre-planned post-hoc analyses to be conducted after the pre-specified main analyses. A sensitivity analysis will repeat the primary chart-level accuracy outcome analysis with a less strict definition of an accurate match to the gold standard set. To assess whether there is a difference in accuracy among the 12 trial eligibility criteria, criterion-level accuracy will be compared assessed between arms. To assess criterion-level accuracy, two-sided, paired t-tests or Wilcoxon rank sum tests will be 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 will be then assessed with an additional one-sided hypothesis with an unspecified superiority margin. Lastly, we repeat both the chart-level accuracy analysis

and efficiency analyses while stratifying by chart characteristics that include chart complexity, length, order, and cancer type.

Descriptive statistics, including means, medians, proportions, and ranges, will be used to summarize each analyses. All statistical analyses will be conducted using appropriate statistical software, and will rely on a predetermined significance level (e.g., alpha = 0.05). Effect sizes and confidence intervals will be reported where applicable to provide further context to the findings.

10. Investigators

Ezekiel J. Emanuel, MD, PhD (Principal Investigator) is Diane v.S. Levy and Robert M. Levy University Professor in Health Policy and Medical Ethics and Co-Director of the Healthcare Transformation Institute.

Ravi B. Parikh, MD, MPP (Sub-Investigator) is Assistant Professor of Health Policy and Medicine and Innovation Faculty at the Penn Center for Cancer Care Innovation (PC3I).

Other members of the Penn research team include Professor Jinbo Chen, PhD (Collaborator), William Ferrell, MPH (Project Manager, Perelman MEHP Department), Matthew Guido (Project Manager, Perelman MEHP Department), Liz Beothy (Project Manager, Clinical Trials Unit) Ryan O'Keefe, MD, MBA (Resident Physician, HUP) and Likhitha Kolla (MD-PhD Student).

Key personnel from Mendel.ai - including Karim Galil, MD (CEO) and Sailu Challapalli (Chief Product Officer) - will be involved as external collaborators.

11. Human research protection

11.1 Data confidentiality

Computer-based files will only be made available to personnel involved in the study through the use of access privileges and passwords. All identifiers will be removed from study-related information. Precautions will be put in place to ensure the data are secure by using passwords and HIPAA-compliant encryption. All patient data provided by Mendel.ai will be viewed by HIPAA-certified Research Coordinators trained in EHR screening within a secure enclave.

NOTE: <u>No Penn data</u> will be utilized in this study.

11.4 Data disclosure

The results of the study will be presented in aggregate form, with no disclosure of individual-level data. No information that identifies specific patients will be recorded as part of documenting and/or publishing the results of this research.

11.5 Data safety and monitoring

The investigators will provide oversight for the study evaluation of this project.

11.6 Risk/benefit

11.6.1 Potential study risks

The potential risks associated with this study are minimal given the research is focused on de-identified data from patient charts. Breach of data is a potential risk that will be mitigated by using HIPAA compliant and secure data platforms for analysis.

NOTE: No Penn data will be utilized in this study.

11.6.2 Potential study benefits

This is highly practical research because it assesses the effectiveness of AI-led and human-AI-assisted chart review for clinical trial eligibility screening. This has major implications for clinical research, as it could lead to significant time savings in screening patients for eligibility and ultimately lead to more cancer patients participating in clinical trials.

11.6.3 Risk/benefit assessment

The risk/benefit ratio is favorable given the potential benefit for significant practical knowledge that could be gained from understanding the potential for AI to review “unstructured” patient data and accelerate the process of identifying patient eligibility for clinical trials. Efforts have been put into place to minimize the risk of breach of data. If favorable outcomes are found, then there is a potential to leverage the insights to guide further research and future clinical practice.

References

1. Korman, D. Z., Mack, E., Jett, J. & Renear, A. H. Defining Textual Entailment. *Journal of the Association for Information Science and Technology* **69**, 763–772 (2018).
2. Mündler, N., He, J., Jenko, S. & Vechev, M. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. Preprint at https://doi.org/10.48550/arXiv.2305.15852 (2024).