Measuring the Impact of AI in the Diagnosis of Hospitalized Patients Through a Randomized Vignette-Based Multicenter Study

TRIAL FULL TITLE	Measuring the Impact of AI in the Diagnosis of Hospitalized Patients through a Web-based Randomized Survey Vignette Multi-Center Study
SAP VERSION	2
SAP VERSION DATE	6/22/2023
TRIAL PRINCIPAL INVESTIGATOR	Jenna Wiens, PhD Michael Sjoding, MD
SAP AUTHOR(s)	Sarah Jabbour, Michael Sjoding

Measuring the Impact of AI in the Diagnosis of Hospitalized Patients Through a Randomized Vignette-Based Multicenter Study

Table of Contents

Measuring the Impact of AI in the Diagnosis of Hospitalized Patients Through a Randomized Vignette-Based Multicenter Study

Measuring the Impact of AI in the Diagnosis of Hospitalized Patients Through a Randomized Vignette-Based Multicenter Study

Abbreviations and Definitions

AE	Adverse Event
ARF	Acute Respiratory Failure
CRF	Case Report Form
COPD	Chronic Obstructive Pulmonary Disease
IMP	Investigational Medical Product
SAP	Statistical Analysis Plan

1 Introduction

1.1 Preface

Artificial intelligence (AI) has achieved high accuracy at identifying abnormalities in clinical images, such as pneumonia from chest radiographs, diabetic retinopathy from fundus images, or skin cancer from histopathology images.[10-12] However, systematic bias in AI models can lead to inaccurate predictions for entire subpopulations.[13-15] When presented with such incorrect predictions, physician performance can be harmed[16] due to automation bias,[17] which is especially concerning in safety-critical settings. Thus, the extent to which AI can be safely integrated into clinical workflows and to support diagnostic decisions is still unknown.

This study aims to provide insight into the effectiveness of providing clinicians with image-based AI model explanations to help them catch when models are making incorrect decisions.

1.2 Scope of the analyses

These analyses will primarily assess the extent to which showing clinicians systematically biased AI model predictions and explanations improves their diagnostic accuracy after reviewing clinical vignettes of patients with acute respiratory failure and determining the patient’s likely diagnosis compared to the setting where clinicians are shown biased AI model predictions without explanations.

2 Study Objectives and Endpoints

2.1 Study Objectives

<u>Survey Data Collection Phase</u>

Objectives

- To determine clinician accuracy in diagnosing pneumonia, heart failure, and chronic obstructive pulmonary disease (COPD) after reviewing clinical vignettes of patients with acute respiratory failure (ARF) without any AI model input.
- To determine how AI model predictions without explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF.

- To determine how standard AI model predictions with explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF.
- To determine how intentionally biased AI model predictions without explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF?
- To determine how do intentionally biased AI model predictions with explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF.

2.2 Endpoints

Primary endpoints

- Clinician diagnostic accuracy for identify the cause of ARF after reviewing clinical vignettes during following settings
    - Clinicians provided no AI model predictions
    - Clinicians provided standard AI model predictions without explanations
    - Clinicians provided standard AI model predictions with explanations
    - Clinicians provided biased AI model predictions without explanations
    - Clinicians provided biased AI model predictions with explanations

Secondary endpoints

- Accuracy of treatment selection in the above settings

3 Study Methods

3.1 General Study Design and Plan

- Study configuration and experimental design: This study is a block randomized web-based survey clinical vignette study
- Type of Comparison: Clinician diagnostic accuracy when provided AI model predictions with explanations versus AI model without explanation
- Type of control(s): no AI model, AI model with predictions alone.
- Level and method of blinding (e.g. double-blind): Single blind study (clinicians are unaware they are randomized to see AI model with or without predictions)
- Method of treatment assignment: Survey participant level randomization
- At what point in time subjects are randomized relative to treatments, events and study periods: Participants are randomized after survey initiation
- Sequence and duration of all study periods: The survey is anticipated to take an average of 20 minutes to complete.

3.2 Inclusion-Exclusion Criteria and General Study Population

(ICH E3;9.3. ICH E9;2.2.1)

Inclusion Criteria

To be eligible to participant in this study, a participant must answer “Yes” to the following question:

“Do you hold any of the following roles on a healthcare team, or any similar roles?”

- Nurse Practitioner (NP)
- Physician Assistant
- Resident
- Fellow
- Attending Physician

3.3 Randomization and Blinding

Overview

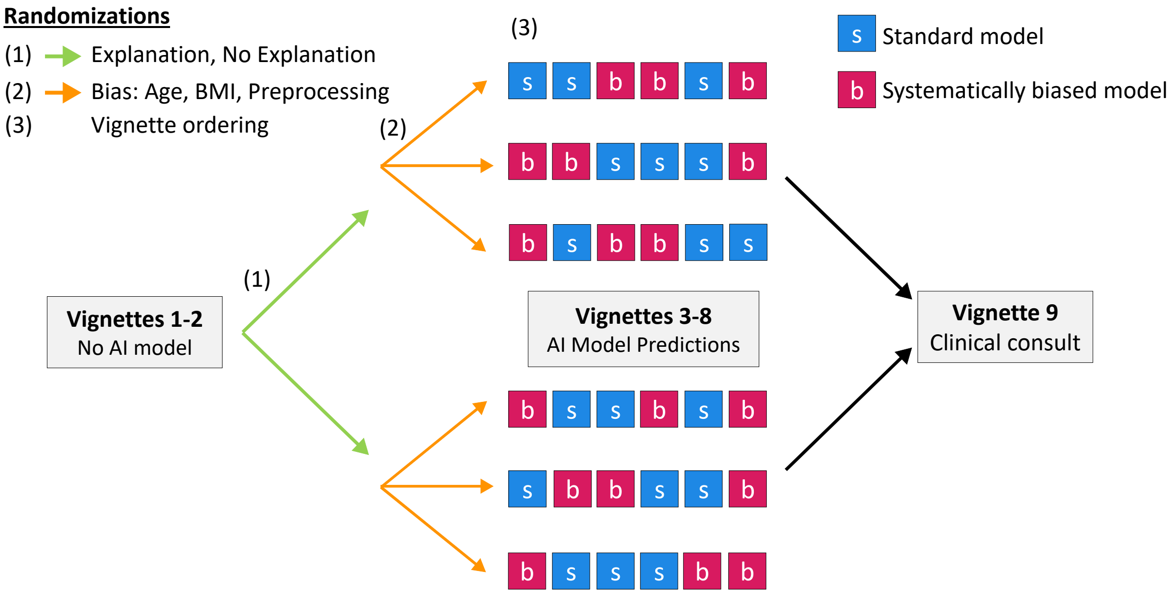


Figure 1. Survey flow and randomization After confirming study eligibility and consent, participants will complete two baseline clinical vignettes where they review patient clinical data and then determine whether the patient has heart failure, pneumonia, and/or COPD without any AI model predictions (Vignettes 1-2). All participants are then randomized to (1; green arrows) AI model predictions with or without model explanations, (2; orange arrows) one of three types of biased AI models shown (biased based on age, BMI, or preprocessing features), and (3; purple boxes) the ordering of the 6 clinical vignettes where 3 standard model predictions and 3 systematically biased model predictions were shown with the clinical vignette. All participants are shown a ninth vignette (vignette 9), which features a clinical consult. The clinical consult provides includes a short block of text providing a prediction and explanation for the patient’s likely diagnosis from a hypothetical trusted colleague.

Details of block randomization.

Block randomization was used to determine the specific patient vignettes and the order of vignettes that subjects would see during the survey. Block randomization was performed in blocks of 90 to achieve all three randomizations described in Figure 1. This ensured that all 45 patient vignettes would be evenly assigned across the first two baseline vignettes and the last clinical consult vignette, and to ensure that a subject would only see a clinical vignette once during the survey. Within the blocks of 90, 30 subjects were randomly assigned to each of the three AI model bias types (age, BMI, or preprocessing features). There were 6 specific clinical vignettes where the AI model displayed the biased behavior for each bias type. Therefore, for the 30 subjects randomly assigned to a specific bias types, 3 of the 6 specific vignettes where the AI model displayed the specific biased behavior was randomly assigned to the subject. An additional 3 vignettes from the 45 total vignettes were randomly selected to be shown with standard model predictions. These 6 patient vignettes (3 with standard AI model, 3 with biased AI model) were then displayed in random order. After the randomization blocks of 90 subjects were generated, carefully tested was performed to ensure all specifications were met.

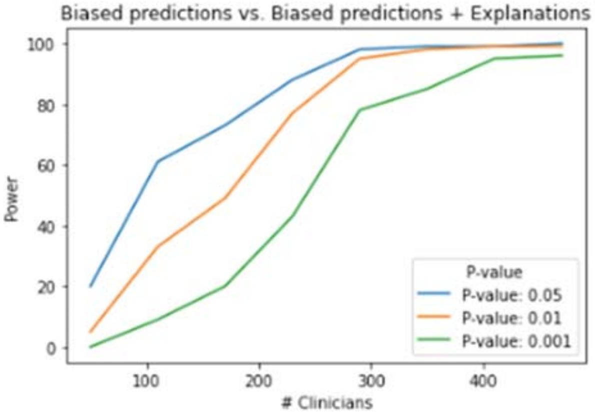
3.4 Study Assessments

The study is designed to take on average 20 minutes per participant. The participant can exit out of the survey and return within two weeks to continue. After the two weeks, the survey is closed and the participant can no longer continue the survey.

Participants will be asked to rate the independent likelihoods that pneumonia, heart failure, and COPD are contributing to the patient's ARF on a scale of 0-100. They will be instructed that patients can have one, more than one, or none of these diagnoses. Clinician diagnostic accuracy will be determined by comparing their answer to an independent assessment of each patients likely diagnosis performed by an panel of clinician reviewers.

4 Sample Size

The sample size calculation is based on the primary endpoint of clinician diagnostic accuracy for pneumonia, heart failure, and COPD. We performed sample size calculations to ensure we would have adequate power to detect both a reduction in diagnostic accuracy when clinicians were shown a biased model, assuming they would follow the biased model's recommendations 50% of the time, and adequate power to detect an improvement in accuracy when clinicians were shown a biased model and explanations, assuming they would follow the biased model recommendation 25% of the time when also shown the explanation. These assumptions would translate into decrease in diagnostic accuracy by 20% when clinicians were shown a biased model and a 10% improvement when shown a biased model and explanation. We used a generalized linear mixed model with a 0.001 significance level. Given the simulated data generated as further described below, we fit a generalized linear mixed model in R to measure if the recovery of clinician diagnostic accuracy when shown the model explanation was significantly different compared to the clinician diagnostic accuracy when shown a biased model alone. We performed 100 simulated studies at each sample size level, and calculated power as the percentage of time a statistically significant difference was measured. We found that the study would have very high power to detect a difference in diagnostic accuracy when comparing clinician baseline diagnostic accuracy and clinician accuracy when shown a biased AI model. The power calculation illustrated in the figure below describes the sample size needed to detect a difference in diagnostic accuracy when clinicians shown a biased AI model alone and when clinicians are shown a biased AI model with explanation.



Detailed sample size calculation:

For our simulation, we model the likelihood that a study subject gets a diagnosis correct as a combination of their baseline diagnostic accuracy (b), the difficulty of the patient case (d), the skill of the clinician (c), and the effect of either being shown an AI model prediction alone (β_1) or being shown an AI prediction with an explanation (β_2), where AI_Alone and $AI\_Explanation$ are indicator variables and $\sigma(.)$ denotes the sigmoid function. Details of each of the variables represented in the equation are described in more detail below.

$$p \ = sigmoid(b + d + c + \ \beta_1 AI\_Alone \ + \ \beta_2 AI\_Explanation)$$

Then, during the simulation, whether a clinician obtains the correct diagnosis is determined by drawing a random variable from a Bernoulli distribution of probability p, with probability determined based on the above data generation model.

Sample size for the study was determined by performing by 100 simulations at participant sample sizes of 50 to 550, in increments of 50, using the above equation to model the data generating process in the survey. In each simulation, a clinician is shown 2 vignettes with no AI model input and then shown either 3 vignettes with systematically biased AI model without explanations or 3 vignettes with systematically biased AI model with explanations. We simulate 100 of these studies at each sample size level.

During a simulated study, we generate blocks of vignettes to assign to hypothetical subjects by the combinations of (1) whether they were shown an AI model explanation, and (2) the type of systematic bias seen. This generates 6 possible assignments:

1. AI model with Pneumonia bias, no explanation
2. AI model with Pneumonia bias, explanation
3. AI model with Heart failure bias, no explanation

4. AI model with Heart failure bias, explanation
5. AI model with COPD bias, no explanation
6. AI model COPD bias, explanation

For every hypothetical clinician in the study, we then assign them one of the above conditions (in order) and then generate the likelihood that the participant gets the diagnosis correct based on the data generation model. For example, the first clinician (assigned to 1), is shown 2 clinical vignettes without an AI model and 3 vignettes of a biased AI model with pneumonia bias and no explanation. The clinician's diagnostic accuracy for each of these clinical vignettes was determined using the data generating model.

Details of the generative model parameters:

1. Baseline diagnostic accuracy

Average baseline diagnostic accuracy for clinicians was assumed to be 0.7 across all three diagnoses but then updated after calculating baseline accuracy at an interim analysis (see 8.5). Accuracy at the interim analysis was determined to be:

a. Pneumonia: 0.68
b. Heart Failure: 0.72
c. COPD: 0.82

These probabilities are transformed to log odds for the data generation model, i.e., logit(x).

- If the participant is randomized to see the Pneumonia bias, then
$$b = logit(0.68) = 0.75$$

- If the participant is randomized to see the heart failure bias, then
$$b = logit(0.72) = 0.94$$

- If the participant is randomized to see the COPD bias, then
$$b = logit(0.82) = 1.5$$

2. Draw Clinician skill c_i

We assumed variation in clinician skill was a normally distributed random variable with mean $\mu_{clinician} = 0$. We assumed the best clinician, who was 2 standard deviations better than average clinician, got the average case right 90% of the time, then $\sigma_{clinician} = \frac{logit(t) - logit(x)}{2}$.

For each clinician c_i, their skill level is drawn:

$c_i \sim N(0, \sigma_{clinician})$ for i = 1,2,…, n; where n is the number of clinicians in the simulation

3. Draw clinical vignette simplicity d_j

We assumed variation exists in clinical vignette diagnostic difficulty, such that cases that are 1 std. easier to diagnosis than the average vignette are answered correctly 90% of the time. Case diagnostic difficulty was assumed to be a normally distributed random variable with mean $\mu_{case} = 0$

and std. $\sigma_{case} = logit(z) - logit(x)$, where z = 0.9, and x = {0.68, 0.72, and 0.82} for the three diagnoses.

For each case d_j, the case simplicity is drawn:

$$d_j \sim N(0, \sigma_{case})$$

3. Draw the Impact of a systematically biased AI model

We assumed that a systematically biased AI model prediction would reduce clinician diagnostic accuracy by a. Therefore, if average diagnostic accuracy was x, the impact of a systematically biased AI model on accuracy is $(x - a)$%. We assumed participants would listen to the AI model 50% of the time, which meant that x - $a = x * 0.5 + 0.33 * 0.5$. In the data generation model, an indicator variable was included indicating whether clinicians were shown a systematically biased AI model with coefficient β_1 . When β_1 < 0, this variable represents a decrease in the likelihood that the clinician will get a case correct. It is the difference between the likelihood that the participant gets the case correct with the AI input minus the likelihood that the participant gets the case correct without AI model input: $\beta_1 = logit(x - a)\ -\ logit(x)$.

In the simulation, if the participant was shown the AI model for the vignette, then

- If the participant is randomized to see the Pneumonia bias, then
$$\beta_1 = logit(0.68 * .5 +\ 0.33 * 0.5) - \ logit(0.68) =\ -0.73$$

- If the participant is randomized to see the heart failure bias, then
$$\beta_1 = logit(0.72 * .5 +\ 0.33 * 0.5) - \ logit(0.72) =\ -0.84$$

- If the participant is randomized to see the COPD bias, then
$$\beta_1 = logit(0.82 * .5 +\ 0.33 * 0.5) - \ logit(0.82) =\ -1.2$$

Impact of a systematically biased AI model with explanation

We assumed that providing an AI model explanation helps clinicians recover diagnostic accuracy by r when shown a biased AI model that reduce their accuracy by a. Therefore, if accuracy on an average case was x, the impact of showing a biased AI model and explanation is ((x-a) + r)%. We assumed that participants would recover 50% back to their baseline diagnostic accuracy, which means ((x-a) + r)% = (x – a + 0.5a)% = $x * 0.75 + 0.33 * 0.25$. In the data generation model, an indicator variable was included indicating whether clinicians were shown a systematically biased AI model explanation with coefficient β_2. β_2 represents the change in likelihood that the clinician gets the case correct. When $\beta_2 > 0$, this variable represents an increase in the likelihood that the clinician gets the case correct: $\beta_2\ =\ logit(x - a + r)\ -\ logit(x - a)$.

In the simulation, if the participant was shown the AI model for the vignette, then

- If the participant is randomized to see the Pneumonia bias, then
$$\beta_2 = logit(0.68 - 0.25 + 0.1) - \ logit(0.68) =\ -0.38$$

- If the participant is randomized to see the heart failure bias, then

$$\beta_1 = logit(0.72 - 0.25 + 0.1) - \ logit(0.68) = \ -0.44$$

- If the participant is randomized to see the COPD bias, then

$$\beta_1 = logit(0.82 - 0.24 + 0.1) - \ logit(0.68) = \ -0.68$$

5 General Analysis Considerations

5.1 Timing of Analyses

The final analysis will be performed two weeks after the last study email invitation is sent out and 400 participants have completed the study.

5.2 Analysis Populations

5.2.1 Full Analysis Population (or Intention to Treat or Modified Intention to Treat)

- *All subjects who consent to taking the study and click to the first page of the study. Each participant who does so is randomized.*

5.2.2 Per Protocol Population

- *NA*

5.2.3 Safety Population

- *NA*

5.3 Covariates and Subgroups

Because all participants are randomized approximately equally across treatment groups, we do not anticipant any covariates that will have an importance influence on our primary endpoints. There are no *a priori* hypotheses of subgroup differences.

5.3.1 Multi-center Studies

This is a multi-center study, where participant responses will be pooled from all centers. The rational behind this is that we assume there is no meaningful center differences in treating patients with ARF.

5.4 Missing Data

The main source of missing data will be missing demographic information in participants who do not complete all vignettes and the demographic questions after the survey. We assume this data will be missing at random. Because this demographic information is not included as covariates in any of the analysis, we do not plan to do anything to impute missing demographics data.

5.5 Interim Analyses and Data Monitoring (as applicable)

5.5.1 Purpose of Interim Analyses

No interim analyses of the exposure variables will be conducted (ie the impact of AI models on clinician diagnostic accuracy), however, we will measure participate baseline accuracy to confirm our sample size calculations.

5.5.2 Planned Schedule of Interim Analyses

Participant baseline diagnostic accuracy will be measured after 300 participants are enrolled in the study.

5.5.3 Scope of Adaptations

Not applicable.

5.5.4 Stopping Rules

Not applicable.

5.5.5 Analysis Methods to Minimize Bias

Not applicable.

5.5.6 Adjustment of Confidence Intervals and p-values

Not applicable.

5.5.7 Interim Analysis for Sample Size Adjustment

Once 300 participant responses are collected, we will measure participant baseline diagnostic accuracy to confirm our baseline accuracy assumption for sample size calculations. We will not measure the effects of the exposures (e.g. AI model predictions and explanations) on diagnostic accuracy during the interim analysis.

5.5.8 Practical Measures to Minimize Bias

The study team will conduct the interim analysis to measure baseline diagnostic accuracy and will not measure nor change any treatment effect assumptions in the power calculations.

5.5.9 Documentation of Interim Analyses

Data and results of the interim analysis will be stored on the HIPAA aligned compute servers of the study team members.

5.6 Multiple Testing

We do not plan to perform any corrections for multiple testing in our primary endpoint of clinical diagnostic accuracy across settings.

6 Summary of Study Data

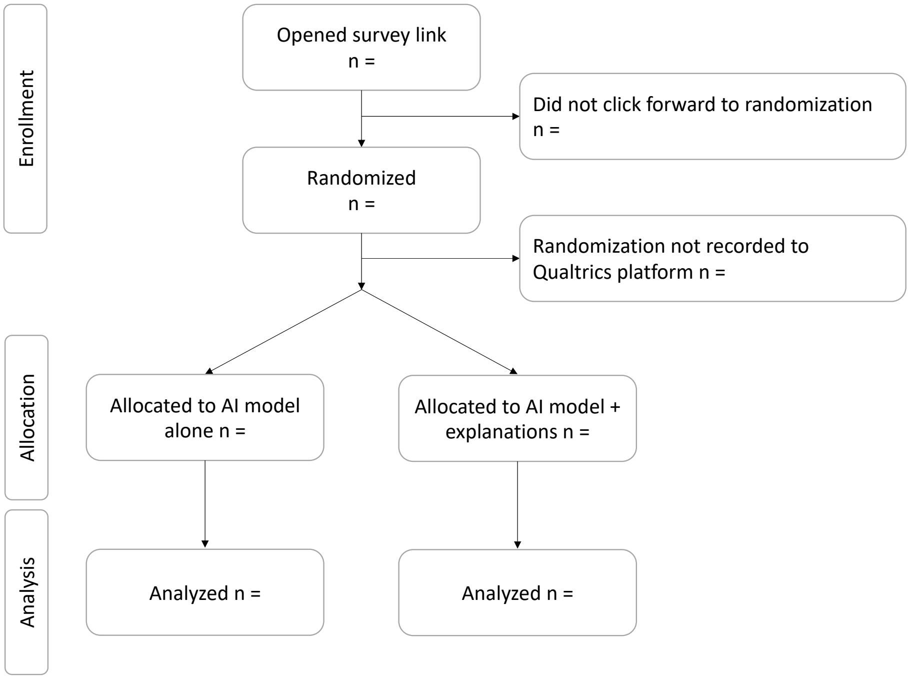
The tables and figures will be based upon the full population of participants who are randomized in the study and completed at least once clinical vignette. The first table will include summary statistics of all study subjects, where each column represents the two treatment arms (AI Model Alone, AI Model + Explanation). The primary statistical analysis results from generalized linear mixed models will also be reported table format, with each row corresponding to diagnostic performance in each vignette setting: Clinician Baseline, Clinician + Standard Model, Clinician + Standard Model +

Explanation, Clinician + Systematically Biased Model, Clinician + Systematically Biased Model + Explanation, Clinician + Clinical Consult.

6.1 Subject Disposition

We will track 1) how many subjects open the survey link through an email as “Opened Survey Link,” 2) how many met the inclusion criteria and consented to study participation and are “randomized,” 3) how many randomization failures occurred because of Qualtrics platform errors, 4) how many were allocated to each treatment arm as “allocated to AI model alone” or “allocated to AI model + explanation,” 5) how many in each arm dropped out before completing a vignette, 6) how many completed at least one vignette and were “analyzed.”

section 9.”


Enrollment
Opened survey link n =
Did not click forward to randomization n =
Randomized n =
Randomization not recorded to Qualtrics platform n =
Allocation
Allocated to AI model alone n =
Allocated to AI model + explanations n =
Analysis
Analyzed n =
Analyzed n =

6.2 Derived variables

Participant diagnostic accuracy is the primary endpoint of this study. Their responses will be collected on a scale of 0-100 and responses above 50 were considered positive for each diagnosis. To calculate diagnostic accuracy, this response will be compared to the reference standard labels generated by a group of 5 physicians who reviewed the patients complete medical record and determined the patient’s diagnosis.

6.3 Protocol Deviations

We do not anticipate any major protocol deviations that would impact the analysis.

6.4 Demographic and Baseline Variables

Collection of participant demographic information is optional and will occur after participants complete all vignettes. We will collect participant age, race and ethnicity, gender, the hospital setting they primarily work, their general practice area, their current role on their healthcare team, and when they completed their medical training.

6.5 Concurrent Illnesses and Medical Conditions

Not applicable

6.6 Treatment Compliance

Not applicable

7 Efficacy Analyses

7.1 Primary Efficacy Analysis

The primary outcome is the participant's diagnostic accuracy after reviewing the patient vignette. Participants will separately assess whether the patient in the vignette has pneumonia, heart failure, and COPD, and their diagnostic accuracy for each will be analyzed as a unique response within the generalized linear mixed model, with individual responses nested within study participant. To determine diagnostic accuracy, participant responses will be compared to the reference standard labels generated by a group of 5 physicians who reviewed the patients complete medical record. A generalized linear model with logit link will be fit for diagnostic accuracy with indicator variables for each of the 5 settings evaluated (clinician baseline without AI model, standard model, standard model with explanation, biased model, biased model with explanation). After fitting the model, we will specifically compare diagnostic accuracy for the following settings:

- Baseline participant accuracy with no AI prediction model input (Clinician Baseline) compared to participant accuracy with standard model predictions without explanations (Clinician + Standard Model)
- Baseline participant accuracy with no AI prediction model input (Clinician Baseline) compared to participant accuracy with standard model predictions and explanations (Clinician + Standard Model + Explanations)
- Baseline participant accuracy with no AI prediction model input (Clinician Baseline) compared to participant accuracy when systematically biased model predictions are provided without explanations (Clinician + Systematically Biased Model)
- Baseline participant accuracy with no AI prediction model input (Clinician Baseline) compared to participant accuracy when systematically biased model predictions are provided with explanations (Clinician + Systematically Biased Model + Explanations)
- Participant accuracy when systematically biased model predictions are provided without explanations (Clinician + Systematically Biased Model) compared to participant accuracy when systematically biased model predictions are provided without explanations (Clinician + Systematically Biased Model)

7.2 Secondary Efficacy Analyses

In a secondary analysis, we will examine how treatment decisions are influenced by correct or incorrect model predictions. We measured the percentage of time participants made an appropriate treatment decision across settings ('Clinician Baseline', 'Clinician + Model', 'Clinician + Model +

Explanation') for both standard and biased AI models. Appropriate treatment for each vignette is determined based on the patients' reference diagnoses and review of the patients complete medical record. We will also investigate the effects of systematically biased estimates on the distributions of participant responses.

7.2.1 Secondary Analyses of Primary Efficacy Endpoint

Not applicable

7.2.2 Analyses of Secondary Endpoints

Not applicable

7.3 Exploratory Efficacy Analyses

Not applicable

8 Safety Analyses

Because this vignette survey study was deemed minimal risk, no safety analysis will be conducted

8.1 Extent of Exposure

Not applicable

8.2 Adverse Events

Not applicable

8.3 Deaths, Serious Adverse Events and other Significant Adverse Events

Not applicable

8.4 Pregnancies (As applicable)

No applicable

8.5 Clinical Laboratory Evaluations

Not applicable

8.6 Prior and Concurrent Medications (As applicable)

Not applicable

8.7 Other Safety Measures

Not applicable

9 Pharmacokinetics (As Applicable)

Not applicable

10 Other Analyses

Not applicable

11 Reporting Conventions

P-values less than .001 will be reported as "p-value<.001"; P-values between .001 and .01 will be reported to the nearest thousandth. P-values greater than or equal to .01 will be reported to the nearest hundredth; P-values greater than .99 will be reported as "p-value>.99."

12 Quality Assurance of Statistical Programming (As Applicable)

All statistical analysis will be conducted in R by the first author team member. A second study team member will review the R code to check for correctness, while also double checking the primary analysis in Stata.

13 References

none