Supplemental Online Content

Jabbour S, Faouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA*. doi:10.1001/jama.2023.22295

eMethods Supplement

eFigure 1. Survey instructions page

eFigure 2. Clinical Vignette Layout

eFigure 3. Participant Response Form for all nine vignettes

eFigure 4. CONSORT Flow diagram of participants through the survey

eTable 1. Characteristics of the patient population used in the vignettes

eTable 2. Accuracy and AUROC of the standard model and systematically biased models across the full patient cohort (n=45)

eTable 3. Count by geographical location of responses across the United States

eTable 4. P-values for demographic differences between the AI model alone group versus the AI model + explanation group

eTable 5. Cross-classified generalized random effects model estimates of participant accuracy (%) in diagnosing pneumonia, heart failure, and COPD

eTable 6. Predictive margins calculated for vignette settings' diagnostic accuracy, compared against the baseline vignette setting of no model input, averaged across all three diagnoses

eTable 7. Predictive margins calculated for vignette settings' diagnostic accuracy, compared against the baseline vignette setting of systematically biased model, averaged across all three diagnoses

eTable 8. Pearson's correlation between participant responses and model scores, with 1000 bootstrapped 95% confidence intervals

eTable 9. Cross-classified generalized random effects model estimates of participant treatment decision accuracy (%) (align with clinical consult)

eTable 10. Predictive margins calculated for vignette settings' treatment accuracy, compared against the baseline vignette setting of no model input, averaged across all three treatments

eTable 11. Predictive margins calculated for vignette settings' treatment accuracy, compared against the baseline vignette setting of incorrect model input, averaged across all three treatments

eFigure 5. Confusion matrices for the participant responses across vignette settings

eReferences

This supplemental material has been provided by the authors to give readers additional information about their work.

eMethods

Study Setup

Study Invitations. We identified hospitalist site champions who sent an invitation email with the study information to hospitalist clinicians at their respective institutions.

Landing Page. Once participants clicked on the study link, they were shown a page to screen for their eligibility. To be eligible, participants needed to confirm their role as nurse practitioner, physician assistant, resident, fellow, attending physician, or any similar role. If they responded “no”, the survey was terminated. Otherwise, they were redirected to an introduction page. They were informed that the study was completely anonymous, it would take about 25 minutes to complete, and that they could stop the survey at any time and come back to the point where they left off. Participants were also told that some details of the study’s purpose would be withheld. They were also informed that they would receive a $50 Amazon.com gift card upon completion of the survey.

Study Instructions. Participants were told that they’d see a clinical decision support tool that was designed to determine the cause or causes of acute respiratory failure (ARF). To not skew their perception of the tool before seeing it in the study, they were then shown an example of how the tool would work if it was designed to identify golden retrievers (**eFigure 1**). They were shown three images of dogs. The first two images contained golden retrievers, while the third contained a German shepherd. They were then shown what the decision support tool scores would look like for each of the images. Finally, if the participant was randomized to see explanations, they were also shown what the explanations would look like for the two images that contained golden retrievers.

IRB Approval. This study was exempt from IRB review, as it was deemed human subjects research with adults that involve benign (non-harmful) behavioral interventions, information was collected through written responses, no physiological data was collected, subject identifiers were not collected, and the subjects agreed to participate in the intervention and information collection.

Patient Cohort selection for Clinical Vignettes

We created 45 clinical vignettes based on real patients selected from a consecutive sample of 121 patients hospitalized with acute respiratory failure at the University of Michigan between August and November of 2017. Patients used in the clinical vignettes were selected to ensure the sample would both achieve the study goals while also being generally representative of patients hospitalized with acute respiratory failure.

First, at least 4 pulmonary physicians independently reviewed each patient’s entire hospitalization, including their presenting signs and symptoms, laboratory testing, imaging studies, and response to treatment to determine their underlying cause of acute respiratory failure, including whether pneumonia, heart failure, and/or COPD contributed to acute respiratory failure. Reviews were averaged and served as the reference standard diagnostic label to evaluate clinician diagnostic accuracy in this study. We collected these patient data to generate clinical vignettes under a separate approved IRB protocol. In addition to determining the underlying cause of ARF (including pneumonia, heart failure, and/or COPD), each physician also gave a difficulty rating on a scale of 1-4 to describe the difficulty of diagnosing the specific patient case.

Next, the 45 patients ultimately used in the study were selected from the larger group of 121 patients based on ensuring the sample would achieve the study’s primary goals while being generally representative of patients hospitalized with acute respiratory failure. This included ensuring the 45 patients selected had similar diagnostic difficulty across the three diagnoses studied (heart failure, pneumonia, and COPD) and similar diagnostic difficulty to the larger group of 121 patients. Additionally, ensuring that there were six patients with the attributes relevant to each of the biased models (e.g. 6 patients with age ≥ 80, which was relevant to the model biased towards pneumonia with respect to age.

Finally, ensuring the 45 patients were selected to ensure disease prevalence for heart failure, pneumonia, and COPD was similar to the prevalence reported in prior national reports.[1]

Clinical Vignette Development

The clinical vignettes were developed iteratively by the study team members and piloted with 15 board certified internal medicine physicians not involved in the final study. First, a physician study team member reviewed the patient's hospital admission history and physical exam and created an abbreviated version of each for the vignette. The patient's chest X-ray image and laboratory data were also collected and added to the vignette.

A pilot study was conducted with the draft vignettes, AI model predictions and explanations, where participants provided qualitative feedback on study layout, vignette content, and survey questions, and time required to complete each vignette. Improvements were made to the vignettes iteratively until participants consistently agreed that the vignette provided sufficient information to conduct the diagnostic assessments required during the study. During clinical vignette development, showing explanations for the electronic health record data was considered but made the survey vignette overly cumbersome. Since the primary study goal was to understand whether clinicians use image-based AI explanations to identify biased models that were specifically biased against patient characteristics identified on chest X-ray images, additional explanations for the electronic health record data were not included.

After incorporating all participant feedback from the initial pilot, a nurse clinical informaticist not involved in the research study reviewed each finalized vignette to ensure no identifiable patient information was present. A second pilot study was conducted with 7 board certified internal medicine physicians not involved in the final study to ensure the study randomization design was correctly implemented in Qualtrics, to ensure participants were assigned the correct vignettes and intervention groups based on the block randomization design.

Randomization

We developed a block randomization procedure to assign specific clinical vignettes to study participants. The block randomization procedure was used to randomize study participants to see 1) AI model predictions with or without AI explanations and 2) one of three types of systematically biased AI models during certain vignettes in the study. At the same time, it also ensured that all 45 patient vignettes would be evenly assigned across the first two baseline vignettes and the last clinical consult vignette, and ensured that a subject would only see a specific clinical vignette once during the survey. Additional details of the block randomization is described in more detail in the Statistical Analysis Plan.

AI Models and Explanations

Standard Model predictions and Explanations. To generate standard model predictions, we trained machine learning models as done in Jabbour *et al.*[2] to predict pneumonia, heart failure, and COPD. The model was trained on data separate from the 45 patients used in the clinical vignettes. Two separate models were trained: one based on chest radiographs and the second on EHR data. We then averaged the image- and EHR-based model predictions, so as to weight the chest radiographs and EHR data equally. Model predictions were shown as follows: model outputs were on a scale of 0-1. These outputs were then thresholded to yield a model decision. Predictions less than the threshold were deemed as the model predicting no disease, and predictions greater than the threshold as model predicting disease. Thresholds were set to the percent of positive patients for each disease using all the data except the 45 patients used to test the model. We measured the percent of positive patients for each disease in this dataset (pneumonia: 31%, heart failure: 22%, COPD: 8%). For each diagnosis, we chose the threshold to be such that the percentage of patients labelled positive by the model was equal to the incidence rate of the diagnosis across this dataset. After thresholding, negative predictions were normalized to 0-50, and positive predictions were normalized to 51-100 so that the decision threshold shown to participants was the midpoint between 0 and 100. Scores were presented on a scale of 0-100, with 0 being unlikely and 100 being likely. Explanations were shown only when the model predicted a positive diagnosis for the patient. We generated heatmaps using Grad-CAM[3] from the trained image-based model.

Systematically biased model predictions and explanations. Participants were randomized to see one of three types of systematically biased model predictions: patients greater than or equal to 80 years old were predicted to have pneumonia, patients with BMI greater than or equal to 40 were predicted to have heart failure, and patients with a blurring filter applied to their chest radiograph were predicted to have COPD. Out of the 45 patients with ARF in our cohort, 6 patients were chosen for each of these biased predictions. The actual model score presented to the participants was generated as follows: for these patients, model predictions were changed to the percentile of all model predictions (based on all training data we had available), so as to mimic biased model behavior that is "highly confident." Each biased model prediction provided a score, corresponding to the 95th percentile of all model predictions to mimic "highly confident" model behavior. Of the 45 patients, the model made systematically biased predictions for 6 selected patients. These 6 were chosen to reflect the subpopulations of patients for which a model might be biased. The accuracy of the systematically biased model was 0.33 for the 6 patients, wherein 2 of the 6 cases the model was right but for the wrong reasons.

Similar to the standard model, explanations were shown only when the model predicted a positive diagnosis for the patient. Here, explanations were dependent on the type of bias. Since we chose age, BMI, and preprocessing as model shortcuts, we trained three separate models to predict age, BMI, and preprocessing. We then generated Grad-CAM heatmaps from these models to correspond with the systematically biased model predictions shown, each of which highlighted the features of the corresponding bias.

When provided with a systematically biased prediction for either pneumonia, heart failure, or COPD, the two other AI model predictions were left unchanged from the standard model predictions. In other words, if a participant was randomized to the systematically biased AI model for age, then any vignette that they received with a patient aged 80 or older would be systematically biased to predict pneumonia, but the heart failure and COPD predictions presented would be from the standard model. If a participant was randomized to the systematically biased AI model for BMI, then any vignette that they received with a patient whose BMI was greater than or equal to 30 would be systematically biased to predict heart failure, but the pneumonia and COPD predictions presented would be from the standard model. Finally, if a participant was randomized to the systematically biased AI model for model preprocessing features, then any vignette that they received with a patient whose chest X-ray had a gaussian blur applied to their X-ray would be systematically biased to predict COPD, but the pneumonia and heart failure predictions presented would be from the standard model. If multiple diagnoses had a score above 50, multiple explanations were shown. While it is more likely that participants will see vignettes in which the systematically biased models are wrong more often than right, given the randomness, there was a 20% chance that participants encounter a systematically biased model that is right for the wrong reasons in two of the three cases.

Power Calculations

The primary goal of the study was to understand the impact of both standard and biased AI models on diagnostic accuracy, and how providing AI explanations impacted accuracy. Answering these questions required making specific diagnostic accuracy comparisons across 5 experimental settings ('Clinician Baseline'), ('Clinician + Standard Model'), ('Clinician + Standard Model + Explanations'), ('Clinician + Systematically Biased Model + Explanations'), ('Clinician + Systematically Biased Model').

The power calculation focused on ensuring adequate power to detect differences in accuracy across the 3 experimental settings related to the biased model ('Clinician Baseline'), ('Clinician + Systematically Biased Model + Explanations'), ('Clinician + Systematically Biased Model') because there were substantially fewer diagnostic assessments relevant to the biased model setting. In the biased model setting, participants were shown a lower accuracy prediction for the diagnosis specific to the biased AI model (e.g., a model biased against heart failure showed lower accuracy predictions for heart failure, but standard predictions for the other diagnoses). When calculating diagnostic accuracy in the biased model setting, only biased predictions were used resulting in a total of 3 participant assessments in vignettes 3 through 8. In contrast, in the other half of vignettes 3 through 8 in which a standard AI model was shown, diagnostic assessments for each of the three diagnoses were analyzed (9 total). For this reason, the power calculations were designed to ensure adequate power in the biased model setting.

Exploratory Subgroup Analyses

The exploratory subgroup analysis used a modeling strategy similar to the primary analysis with a few notable exceptions. When possible, we fit a cross-classified generalized random effects model, where individual diagnostic assessments were nested within study participants and within patients. For two subgroup analyses, specifically the comparison of diagnostic accuracy for heart failure across vignette settings, and the comparison of treatment accuracy for pneumonia with a correct or incorrect model, the cross-classified model did not converge. This was because in both subgroup analysis there was no measurable clustering in the response by study participants. Therefore, a cross-classified generalized random effects model of responses nested within patients was used to estimate treatment accuracy.

Survey Introduction

The following information was provided to potential study participants on the study landing page:

We invite you to participate in a brief research survey, supported by the University of Michigan (U-M) and NIH. The goal of the survey is to understand how clinicians might use clinical decision support tools in their diagnosis and treatment decisions. Some details of the study's purpose may be withheld until survey completion. You will be shown 9 brief clinical vignettes of patients with respiratory failure, including their history, physical exam, laboratory testing, and chest x-ray. Using the information provided, you will be asked about the likelihood of various diagnoses and what treatments you would provide. The clinical vignettes are based on real patient encounters that were modified slightly to preserve anonymity without compromising clinical details. Please give you best answer to each question, treating each patient as if it were a real clinical encounter.

- **The survey should take no more than 25 minutes**
- **The survey is not supported in Internet Explorer**
- **The survey is best viewed on a computer**
- **If you cannot complete the survey in one sitting, your progress will be saved and you can return to the survey using the same web browser**

All of your answers will be anonymous. If you would like to receive a 50 dollar Amazon gift card for your participation, you will be redirected to a separate form upon survey completion. This form will collect your name and email address and will not be linked back to your survey responses in any way. This study was deemed exempt from review by the U-M Institutional Review Board

Post Survey Questions

Collection of participant demographic information was optional and occurred after participants completed all vignettes.

Experience/Attitude about AI.

1. Have you interacted with a clinical decision support tool before in your clinical practice?
    a. Yes
    b. No
2. If Yes, How often do you interact with a clinical decision support tool in your clinical practice?
    a. Never
    b. Rarely (just specific patients with difficult cases)
    c. Sometimes (1-4 times for every 10 patients)
    d. Often (5-8 times for every 10 patients)
    e. Always (9-10 time for every 10 patients)
    f. Other _____
3. If Yes, How important are clinical decision support tools in your clinical practice?
    a. Very unimportant
    b. Somewhat unimportant
    c. Neutral
    d. Somewhat important
    e. Very important
    f. Other _____
4. If Yes, When you interact with a clinical decision support tool, how do you use the information provided by the clinical decision support tool?
    a. I always use the information in my clinical decision making
    b. I consider the information, but it is not always a part of my clinical decision making
    c. I completely ignore the information
    d. Other _____
5. If Yes, Have you ever recommended a clinical decision support tool to your colleagues?
    a. Yes
    b. No

AI Awareness

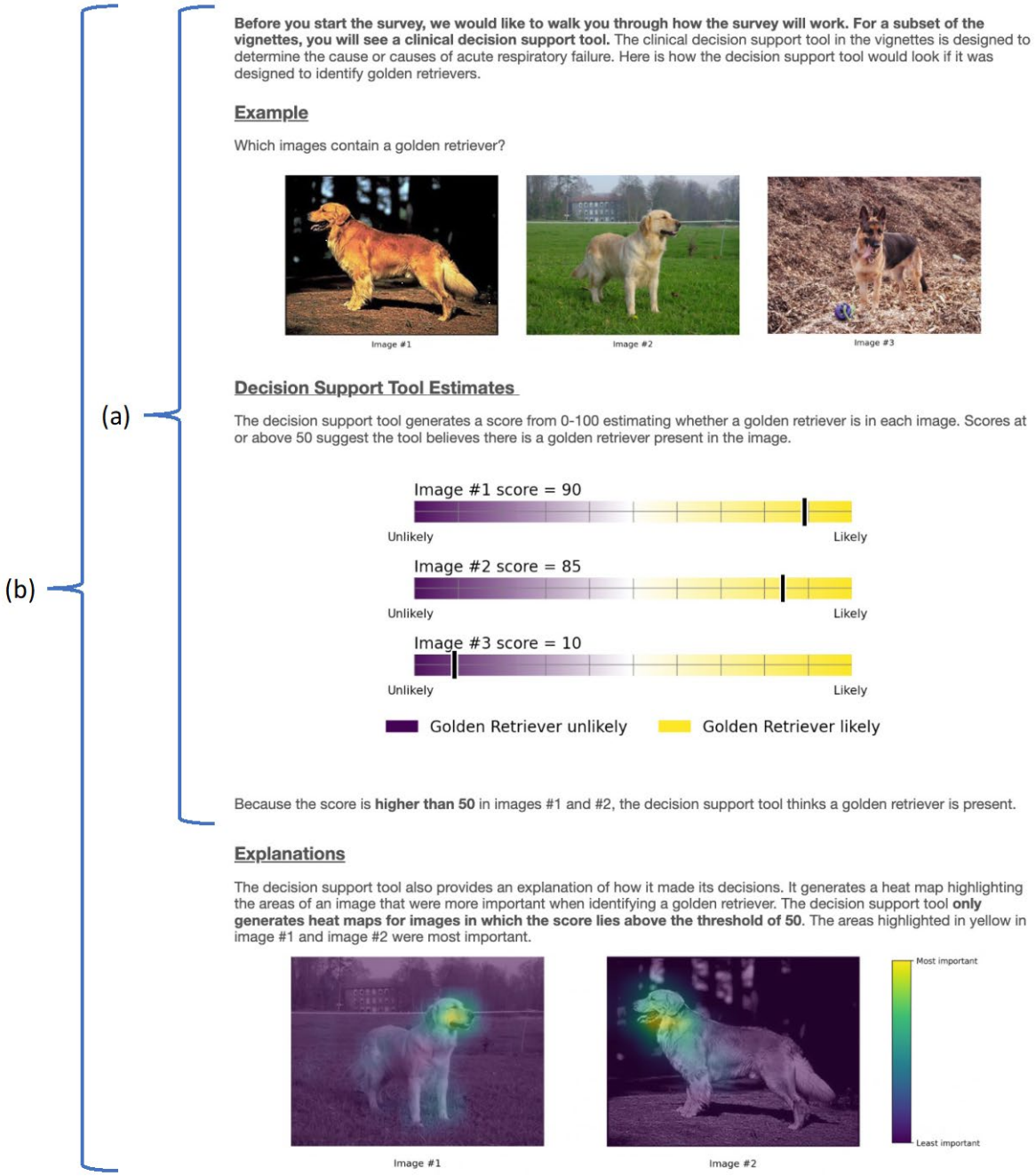
1. To what extent do you agree with the following statement: Using artificial intelligence to analyze medical images and diagnose disease could be beneficial to my patient care and management.
    a. Strongly agree
    b. Agree
    c. Neither agree nor disagree
    d. Disagree
    e. Strongly disagree
2. Prior to taking this survey, were you aware that some artificial intelligence algorithms have been shown to perform differently based on patients demographics such as sex, race, or age?
    a. Yes
    b. No

Demographics

1. What is your age today (in years)? _____
2. What is your race or ethnicity? Select all that apply.
    a. American Indian or Alaska Native
    b. Asian
    c. Black
    d. Hispanic or Latinx

    e. Middle Eastern
    f. Native Hawaiian or Pacific Islander
    g. White
    h. Prefer not to say
    i. Other (if you select “Other,” please specify)
3. What is your gender?
    a. Male
    b. Female
    c. Non-binary/non-conforming
    d. Transgender
    e. Other (if you select “Other,” please specify)
    f. Prefer not to say
4. In what hospital setting to you primarily work? Select all that apply.
    a. University Hospital/Academic
    b. Community Hospital/Private Practice
    c. VA/Government
    d. Other (please specify) _____
5. What is your general practice area?
    a. Hospital Medicine
    b. Pulmonary/Critical Care Medicine
    c. Emergency Medicine
    d. Cardiology
    e. Infectious Disease
    f. Radiology
    g. Outpatient Primary Care
    h. Other (please specify) _____
6. What is your current role on the healthcare team?
    a. Nurse Practitioner (NP)
    b. Physician Assistant
    c. Resident
    d. Fellow
    e. Attending Physician
    f. Other _____
7. If Resident, Fellow or Attending Physician: In what year did you receive your medical degree?
8. If Fellow or Attending Physician: In what year did you complete residency?
9. If Attending Physician: In what year did you complete fellowship (if applicable)?
10. If Nurse Practitioner (NP), Physician Assistant, or Other: In what year did you complete your medical training?

eFigure 1. Survey instructions page.



Participants are shown an instructions page to understand how the clinical decision support tool would be presented to them throughout the survey. To not skew their perception of the tool before seeing it in the study, they were then shown an example of how the tool would work if it was designed to identify golden retrievers. (a) When randomized to see model predictions alone, they are only shown how the model scores would be presented to them. (b) When randomized to see model predictions and explanations, participants are also shown how model explanations will be presented to them.

eFigure 2. Clinical Vignette Layout

Patient History
An 89 year old female with a history of hypertension and recent hip fracture presented to the emergency department from a rehabilitation facility with wheezing, shortness of breath, and cough that developed over the past few days. She was given steroids and bronchodilator treatment at the facility. Due to delirium, she was unable to provide additional history. The patient's daughter says she also had been having delirium at the facility.

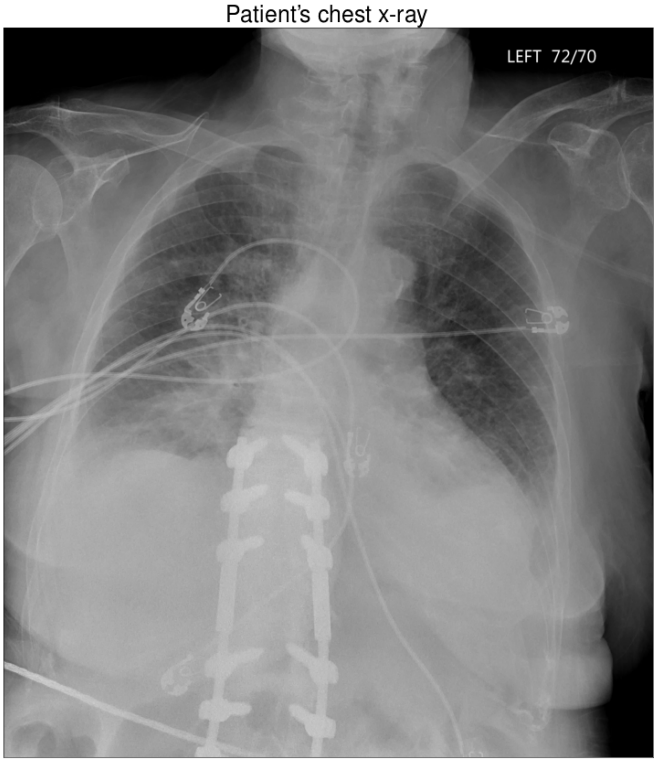
Past Medical History
Hypertension
hypothyroidism
Right hip fracture with surgical repair

Social History
Remote smoking history

Home Medications
Hydrochlorothiazide 12.5 mg daily
Losartan 100 mg daily
Metoprolol 25 mg daily

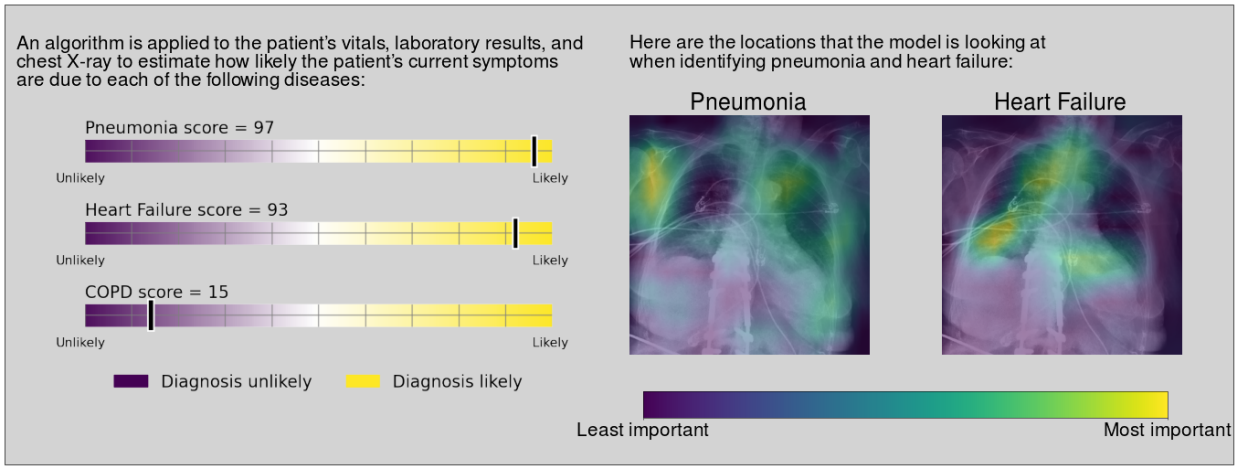
Physical Exam
Height 155 cm Weight 56 kg BMI 23 kg/m^2
Temp 36.4 HR 106 BP 167/81 RR 31
SpO2 97% on 6L Nasal Cannula
Gen: Delirious but redirectable, no acute distress
CV: Regular rate and rhythm, no murmurs
Pulm: Diffuse expiratory wheezing bilaterally
Abd: Soft, non-tender, non-distended
Ext: 1+ pitting edema
Skin: No rashes

Patient's chest x-ray



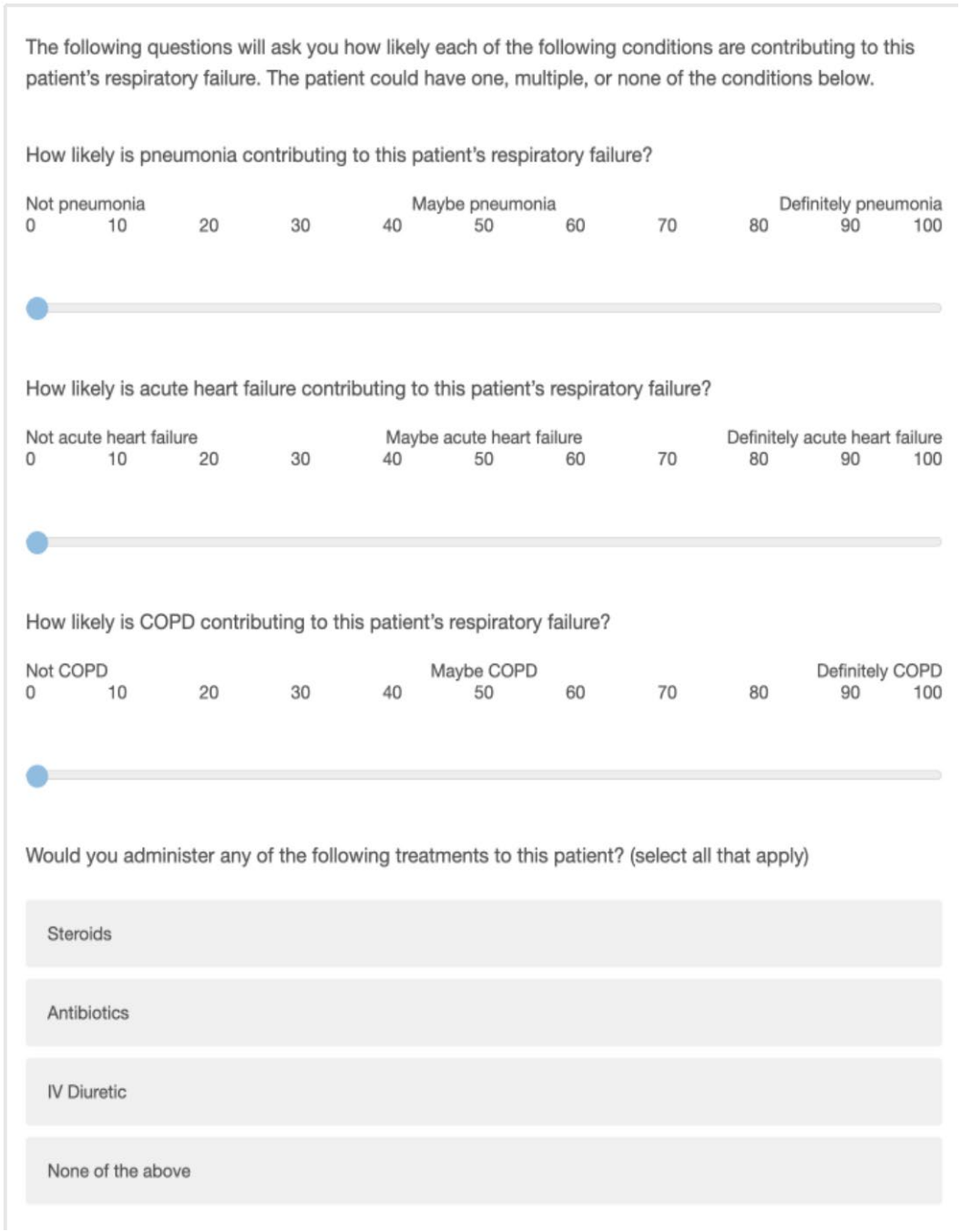
Selected Laboratory Results

Test	Result	Ref Range & Units
WBC	20.7	4.0-10.0 K/uL
HGB	11.2	12.0-16.0 g/dL
PLT	303	150 - 400 K/uL
Creatinine	2.01	0.5 - 1.00 mg/dL
Brain Natriuretic Peptide	4325	0 - 100 pg/mL
Procalcitonin	0.35	0.00 - 0.25 ng/mL
ABG pH	7.36	7.35 – 7.45
ABG pCO2	25	35 – 45 mmHg
ABG pO2	113	80 – 100 mmHg
Lactate	1.6	0 – 2.2 mmol/L

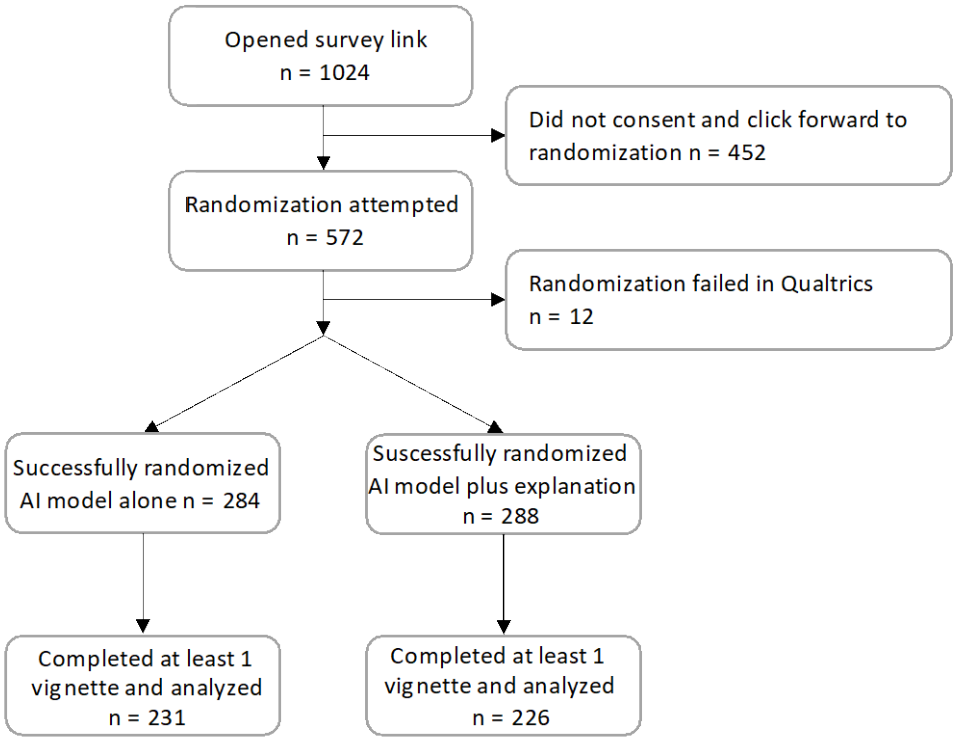


A group of 15 physicians who did not participate in the study provided feedback on vignette content and layout to ensure that there was sufficient information provided to make a diagnosis.

eFigure 3. Participant Response Form for all nine vignettes.



eFigure 4. CONSORT Flow diagram of participants through the survey.



eTable 1. **Characteristics of the patient population used in the vignettes.**

Characteristic (n=45)	Value
Age (IQR)	70.0 (36.5-85.9)
Male, n (%)	23 (51)
Race, n (%)	
Black	5 (11)
White	40 (89)
Acute respiratory failure etiology, n (%)	
Pneumonia	14 (31)
Heart Failure	14 (31)
COPD	5 (11)
Pneumonia and Heart Failure	2 (4)
Pneumonia and COPD	1 (2)
Heart Failure and COPD	2 (4)
All conditions	1 (2)
No conditions	6 (13)

*Sex and race were self-reported by patients.

eTable 2. Accuracy and AUROC of the standard model and systematically biased models across the full patient cohort (n=45)

		Pneumonia	Heart Failure	COPD
Accuracy (%)	**Standard Model**	67	71	87
	Systematically Biased Model	62	71	78
AUROC	**Standard Model**	0.79	0.80	0.89
	Systematically Biased Model	0.71	0.76	0.85

Due to where the model predictions were thresholded, the standard and systematically biased models have the same accuracy for diagnosing heart failure. However, the predictions themselves were different, resulting in the standard model having a higher AUROC compared to the systematically biased model.

eTable 3. Count by geographical location of responses across the United States.

State	Number of Responses
Colorado	92
Michigan	70
Massachusetts	43
Texas	38
Wisconsin	33
Ohio	27
Utah	26
Maryland	23
Minnesota	17
California	13
Chicago	13
Indiana	1
Oregon	1
Unknown	21

eTable 4. P-values for demographic differences between the AI model alone group versus the AI model + explanation group.

Subject Characteristics	AI model alone	AI model and explanations	P-Value
Randomized and completed at least 1 vignette, n (%)	231	226	--
Complete all vignettes, n (%)	214 (93)	204 (90)	--
Response Time for completing survey (Minutes), median (IQR)	19 (14-32)	19 (14-27)	--
Completed post-survey questions	214	204	--
Age, median (IQR)	35 (31-40)	34 (31-38)	0.25
Years of Practice, median (IQR)	5 (2-9)	4 (2-8)	0.34
Prior Interaction with AI, n (%)	70 (32.7)	62 (30.4)	0.69
AI Bias Aware, n (%)	68 (31.8)	71 (34.8)	0.58
Gender, n (%)			0.28
Female	123 (57.5)	118 (57.8)	--
Male	82 (38.3)	82 (40.2)	--
Prefer not to say	9 (4.2)	3 (1.5)	--
Non-binary / non-conforming	0 (0.0)	1 (0.5)	--
General Practice Area, n (%)			0.69
Hospital Medicine	210 (98.1)	198 (97.1)	--
Other	4 (1.9)	6 (2.9)	--
Role on Healthcare Team, n (%)			0.38
Attending Physician	138 (64.5)	121 (59.3)	--
Physician Assistant	55 (25.7)	57 (27.9)	--
Nurse Practitioner (NP)	12 (5.6)	18 (8.8)	--
Resident/Fellow	7 (3.3)	8 (3.9)	--
Other	2 (0.9)	0 (0.0)	--
Hospital Setting, n (%)			0.61
University Hospital/Academic	186 (86.9)	175 (85.8)	--
Community Hospital/Private Practice	32 (15.0)	30 (14.7)	--
No Response	17 (7.9)	23 (11.3)	--
VA/Government	11 (5.1)	11 (5.4)	--
Race and Ethnicity, n (%)			.62
Asian	48 (22.4)	32 (15.7)	--
Black	4 (1.9)	3 (1.5)	--
Hispanic or Latinx	8 (3.7)	7 (3.4)	--
Middle Eastern	5 (2.3)	5 (2.5)	--
Native Hawaiian or Pacific Islander	1 (0.5)	0 (0.0)	--
No Response	17 (7.9)	22 (10.8)	--
Other	1 (0.5)	2 (1.0)	--
Prefer not to say	14 (6.5)	14 (6.9)	--
White	139 (65.0)	146 (71.6)	--

Categorical variables were compared using a Chi-squared test[4] and continuous variables were compared using the Mann-Whitney U-test.[5]

eTable 5. Cross-classified generalized random effects model estimates of participant accuracy (%) in diagnosing pneumonia, heart failure, and COPD.

Setting	Overall Accuracy (95% CI)	Pneumonia Accuracy (95% CI)	Heart Failure Accuracy (95% CI)	COPD Accuracy (95% CI)
Clinician Baseline	73.0 (68.3 - 77.8)	67.5 (61.0 - 74.0)	70.7 (63.1 - 78.3)	80.5 (74.8 - 86.1)
Clinician + Standard Model	75.9 (71.3 - 80.5)	69.5 (62.9 - 76.0)	73.4 (65.8 - 80.9)	84.6 (79.7 - 89.6)
Clinician + Standard Model + Explanations	77.5 (73.0 - 82.0)	72.1 (65.6 - 78.5)	74.5 (67.1 - 82.0)	85.7 (80.8 - 90.5)
Clinician + Systematically Biased Model	61.7 (55.3 - 68.2)	57.5 (47.2 - 67.7)	65.4 (55.8 - 75.1)	71.0 (61.5 - 80.6)
Clinician + Systematically Biased Model + Explanations	64.0 (57.6 - 70.3)	59.7 (49.4 - 70.1)	65.0 (55.2 - 74.9)	74.8 (65.8 - 83.7)
Clinician + Clinical Consult	81.1 (76.9 - 85.4)	75.9 (69.4 - 82.3)	82.4 (75.9 - 89.0)	85.8 (80.8 - 90.9)

Estimates are reported as marginal effects after fitting a cross-classified generalized random effects model of diagnostic accuracy across settings.
CI: Confidence interval

eTable 6. Predictive margins calculated for vignette settings' diagnostic accuracy, compared against the baseline vignette setting of no model input, averaged across all three diagnoses.

Setting	Average Marginal Effect	Standard Error	z-value	p-value	95% lower bound	95% upper bound
Clinician + Standard Model	2.9	1.2	2.35	0.02	0.5	5.2
Clinician + Standard Model + Explanations	4.4	1.2	3.57	< 0.001	2.0	6.9
Clinician + Systematically Biased Model	-11.3	2.1	-5.33	< 0.001	-15.5	-7.2
Clinician + Systematically Biased Model + Explanations	-9.1	2.1	-4.29	< 0.001	-13.2	-4.9
Clinician + Clinical Consult	8.1	1.4	5.86	< 0.001	5.4	10.8

eTable 7. Predictive margins calculated for vignette settings' diagnostic accuracy, compared against the baseline vignette setting of systematically biased model, averaged across all three diagnoses.

Setting	Average Marginal Effect	Standard Error	z-value	p-value	95% lower bound	95% upper bound
Clinician Alone	11.3	2.1	5.33	< 0.001	7.2	15.5
Clinician + Standard Model	14.2	2.2	6.41	< 0.001	9.8	18.5
Clinician + Standard Model + Explanations	15.8	2.3	6.96	< 0.001	11.3	20.2
Clinician + Systematically Biased Model + Explanations	2.3	2.5	0.90	0.37	-2.7	7.2
Clinician + Clinical Consult	19.4	2.4	8.23	< 0.001	14.8	24.1

eTable 8. Pearson's correlation between participant responses and model scores, with 1000 bootstrapped 95% confidence intervals.

Setting	Overall	Pneumonia	Heart Failure	COPD
Clinician + Standard Model	0.53 (0.50-0.57)	0.29 (0.22-0.36)	0.62 (0.57-0.67)	0.66 (0.60-0.71)
Clinician + Standard Model + Explanations	0.59 (0.56-0.62)	0.38 (0.31-0.45)	0.69 (0.64-0.73)	0.69 (0.64-0.73)
Clinician + Systematically Biased Model	0.41 (0.38-0.45)	0.21 (0.14-0.29)	0.52 (0.46-0.58)	0.46 (0.39-0.52)
Clinician + Systematically Biased Model + Explanations	0.41 (0.37-0.45)	0.24 (0.16-0.31)	0.54 (0.48-0.60)	0.41 (0.34-0.47)

eTable 9. Cross-classified generalized random effects model estimates of participant treatment decision accuracy (%) (align with clinical consult).

Setting	Overall Accuracy (95% CI)	Antibiotics Accuracy (95% CI)	IV Diuretics Accuracy (95% CI)	Steroids Accuracy (95% CI)
Clinician Baseline	70.3 (65.5 - 75.2)	61.1 (53.8 - 68.5)	69.4 (61.6 - 77.1)	78.9 (72.7 - 85.0)
Clinician + Correct Model	77.0 (72.6 - 81.4)	60.6 (52.6 - 68.7)	76.8 (69.2 - 84.3)	83.4 (77.9 - 89.0)
Clinician + Correct Model + Explanations	80.4 (76.3 - 84.5)	66.4 (58.6 - 74.1)	81.3 (74.2 - 88.3)	84.8 (79.3 - 90.2)
Clinician + Incorrect Model	55.1 (48.8 - 61.3)	55.4 (46.5 - 64.2)	67.0 (57.8 - 76.2)	72.2 (63.2 - 81.2)
Clinician + Incorrect Model + Explanations	57.8 (51.6 - 63.9)	55.9 (47.1 - 64.8)	67.0 (57.8 - 76.2)	76.9 (68.7 - 85.1)

Estimates are reported as marginal effects after fitting a cross-classified generalized random effects model of diagnostic accuracy across settings.
CI: Confidence interval

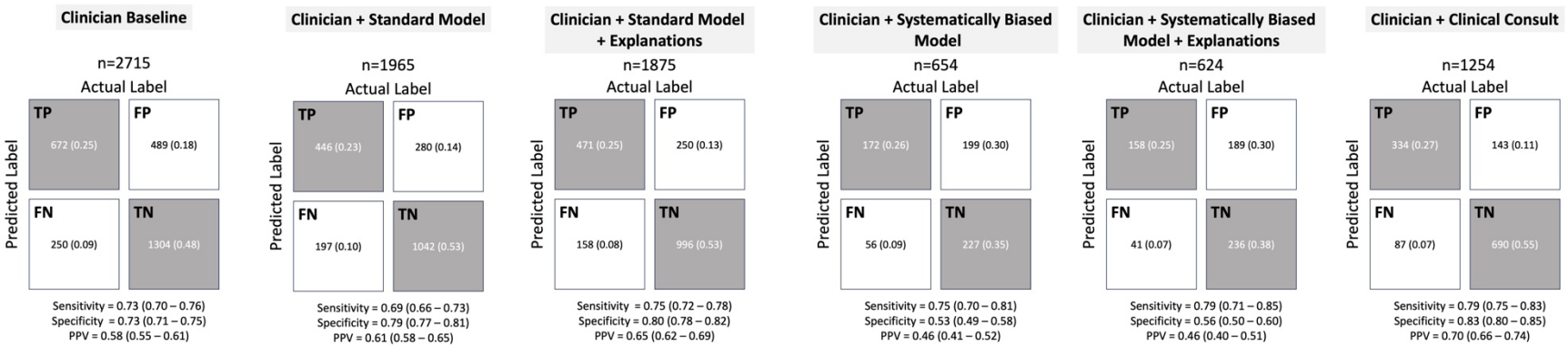
eTable 10. Predictive margins calculated for vignette settings' treatment accuracy, compared against the baseline vignette setting of no model input, averaged across all three treatments.

Setting	Average Marginal Effect	Standard Error	z-value	p-value	95% lower bound	95% upper bound
Clinician + Correct Model	6.7	1.3	5.01	< 0.001	4.1	9.3
Clinician + Correct Model + Explanations	10.1	1.4	7.29	< 0.001	7.4	12.8
Clinician + Incorrect Model	-15.3	2.0	-7.82	< 0.001	-19.1	-11.4
Clinician + Incorrect Model + Explanations	-12.5	1.9	-6.50	< 0.001	-16.3	-8.8

eTable 11. Predictive margins calculated for vignette settings' treatment accuracy, compared against the baseline vignette setting of incorrect model input, averaged across all three treatments.

Setting	Average Marginal Effect	Standard Error	z-value	p-value	95% lower bound	95% upper bound
Clinician Baseline	15.3	2.0	6.50	< 0.001	8.8	16.3
Clinician + Correct Model	21.9	2.2	10.17	< 0.001	17.7	26.2
Clinician + Correct Model + Explanations	25.3	2.2	11.36	< 0.001	21.0	29.7
Clinician + Incorrect Model + Explanations	2.7	2.2	1.22	0.22	-1.7	7.1

eFigure 5. Confusion matrices for the participant responses across vignette settings.



eReferences

1. Stefan MS, Shieh MS, Pekow PS, et al. Epidemiology and outcomes of acute respiratory failure in the United States, 2001 to 2009: A national survey. *Journal of hospital medicine*. 2013;8(2):76-82.
2. Jabbour S, Fouhey D, Kazerooni E, Wiens J, Sjoding MW. Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure. *Journal of the American Medical Informatics Association*. 2022;29(6):1060-1068.
3. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 618-626.
4. Rayner JCW, Best DJ. Smooth tests of goodness of fit: an overview. *International Statistical Review/Revue Internationale de Statistique*. 1990:9-17.
5. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*. 1947:50-60.