JAMA | Original Investigation | AI IN MEDICINE

# Measuring the Impact of AI in the Diagnosis of Hospitalized Patients
## A Randomized Clinical Vignette Survey Study

Sarah Jabbour, MSE; David Fouhey, PhD; Stephanie Shepard, PhD; Thomas S. Valley, MD; Ella A. Kazerooni, MD, MS; Nikola Banovic, PhD; Jenna Wiens, PhD; Michael W. Sjoding, MD

**IMPORTANCE** Artificial intelligence (AI) could support clinicians when diagnosing hospitalized patients; however, systematic bias in AI models could worsen clinician diagnostic accuracy. Recent regulatory guidance has called for AI models to include explanations to mitigate errors made by models, but the effectiveness of this strategy has not been established.

**OBJECTIVES** To evaluate the impact of systematically biased AI on clinician diagnostic accuracy and to determine if image-based AI model explanations can mitigate model errors.

**DESIGN, SETTING, AND PARTICIPANTS** Randomized clinical vignette survey study administered between April 2022 and January 2023 across 13 US states involving hospitalist physicians, nurse practitioners, and physician assistants.

**INTERVENTIONS** Clinicians were shown 9 clinical vignettes of patients hospitalized with acute respiratory failure, including their presenting symptoms, physical examination, laboratory results, and chest radiographs. Clinicians were then asked to determine the likelihood of pneumonia, heart failure, or chronic obstructive pulmonary disease as the underlying cause(s) of each patient's acute respiratory failure. To establish baseline diagnostic accuracy, clinicians were shown 2 vignettes without AI model input. Clinicians were then randomized to see 6 vignettes with AI model input with or without AI model explanations. Among these 6 vignettes, 3 vignettes included standard-model predictions, and 3 vignettes included systematically biased model predictions.

**MAIN OUTCOMES AND MEASURES** Clinician diagnostic accuracy for pneumonia, heart failure, and chronic obstructive pulmonary disease.

**RESULTS** Median participant age was 34 years (IQR, 31-39) and 241 (57.7%) were female. Four hundred fifty-seven clinicians were randomized and completed at least 1 vignette, with 231 randomized to AI model predictions without explanations, and 226 randomized to AI model predictions with explanations. Clinicians' baseline diagnostic accuracy was 73.0% (95% CI, 68.3% to 77.8%) for the 3 diagnoses. When shown a standard AI model without explanations, clinician accuracy increased over baseline by 2.9 percentage points (95% CI, 0.5 to 5.2) and by 4.4 percentage points (95% CI, 2.0 to 6.9) when clinicians were also shown AI model explanations. Systematically biased AI model predictions decreased clinician accuracy by 11.3 percentage points (95% CI, 7.2 to 15.5) compared with baseline and providing biased AI model predictions with explanations decreased clinician accuracy by 9.1 percentage points (95% CI, 4.9 to 13.2) compared with baseline, representing a nonsignificant improvement of 2.3 percentage points (95% CI, −2.7 to 7.2) compared with the systematically biased AI model.

**CONCLUSIONS AND RELEVANCE** Although standard AI models improve diagnostic accuracy, systematically biased AI models reduced diagnostic accuracy, and commonly used image-based AI model explanations did not mitigate this harmful effect.

**TRIAL REGISTRATION** ClinicalTrials.gov Identifier: NCT06098950

Editorial page 2255

Supplemental content

**Author Affiliations:** Computer Science and Engineering, University of Michigan, Ann Arbor (Jabbour, Fouhey, Shepard, Banovic, Wiens); Now with Computer Science Courant Institute, New York University, New York (Fouhey); Now with Electrical and Computer Engineering Tandon School of Engineering, New York University, New York (Fouhey); Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor (Valley, Sjoding); Department of Radiology, University of Michigan Medical School, Ann Arbor (Kazerooni).

**Corresponding Author:** Michael W. Sjoding, MD, Internal Medicine, G020W Bldg 16 NCRC, 2800 Plymouth Rd, SPC 2800, Ann Arbor, MI 48109 (msjoding@umich.edu) and Jenna Wiens, PhD, Computer Science and Engineering, University of Michigan, 3749 Beyster Bldg, 2260 Haward St, Ann Arbor, MI 48109. Haward St, Ann Arbor, MI 48109 (wiensj@umich.edu).

*JAMA*. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295

Artificial intelligence (AI) shows promise in identifying abnormalities in clinical images, such as pneumonia from chest radiographs, diabetic retinopathy from fundus images, or skin cancer from histopathology images.[1-4] AI models in clinical care analyze patient data, identify patterns, and make predictions. Integrating AI into clinical decision-making may result in higher accuracy than a clinician without AI.[1] However, systematically biased AI models (ie, models that consistently misdiagnose patient subpopulations) can lead to errors and potential harm.[5-7] For example, an AI model trained on data in which female patients are consistently underdiagnosed for heart disease may learn this bias and underdiagnose females if deployed.[8] Physician performance can be harmed by following incorrect predictions,[9] a phenomenon known as *overreliance*.[10] Ideally, clinicians would be able to follow AI predictions when correct but ignore incorrect AI predictions.[11] However, the extent to which AI can be safely integrated to support diagnostic decisions is still largely unknown.

Recent US Food and Drug Administration (FDA) guidance highlights the importance of providing clinicians with the ability to independently review the basis for software recommendations, including the information and logic used in a model's decisions.[12] Although AI developers frequently use AI explanations to debug models during development, providing similar explanations to clinicians, as recommended by the FDA, could potentially enhance interpretability after deployment.[13] Explanations could help clinicians better understand a model's logic before acting on its recommendation, helping to mitigate errors made by systematically biased models. For example, image-based explanations designed to highlight important regions of an image that were used in an AI model's decision[14] could help clinicians identify when models are behaving inaccurately.

To determine if providing AI explanations could enhance clinician diagnostic accuracy and alleviate the negative impact of systematically biased models, we conducted a randomized clinical vignette web-based survey study. We focused on the diagnosis of acute respiratory failure as a test case for this broader challenge because acute respiratory failure is commonly encountered in practice,[15] often misdiagnosed,[16] and evaluated using commonly encountered data such as chest radiographs that AI tools can analyze. We presented participants with standard and systematically biased AI, randomizing participants to see AI predictions alone or AI predictions with explanations. This design allowed us to test our hypothesis that systematically biased models would harm clinician diagnostic accuracy, but commonly used image-based explanations would help clinicians partially recover their performance.[17]

### Key Points

**Question** How is diagnostic accuracy impacted when clinicians are provided artificial intelligence (AI) models with image-based AI model explanations, and can explanations help clinicians when shown systematically biased AI models?

**Findings** In this multicenter randomized clinical vignette survey study, diagnostic accuracy significantly increased by 4.4% when clinicians reviewed a patient clinical vignette with standard AI model predictions and model explanations compared with baseline accuracy. However, accuracy significantly decreased by 11.3% when clinicians were shown systematically biased AI model predictions and model explanations did not mitigate the negative effects of such predictions.

**Meaning** AI model explanations did not help clinicians recognize systematically biased AI models.

## Methods

### Participants

We recruited hospitalist physicians, nurse practitioners, and physician assistants who commonly care for patients with acute respiratory failure to complete the study. The study protocol is available in Supplement 1 and the statistical analysis plan in Supplement 2. We first identified 12 hospitalist leaders as site champions through interaction with the Society of Hospital Medicine research committee[18] and the authors' personal contacts. Site champions advertised the study via email to their local hospital medicine divisions. Each site champion sent a follow-up reminder email after 2 weeks. Interested participants were directed to the study landing page, which confirmed study eligibility, provided additional study information, study instructions, and obtained informed consent prior to randomization (see the study protocol in Supplement 1 and eFigure 1 in Supplement 3). Participants were offered a $50 gift card for completing the study.

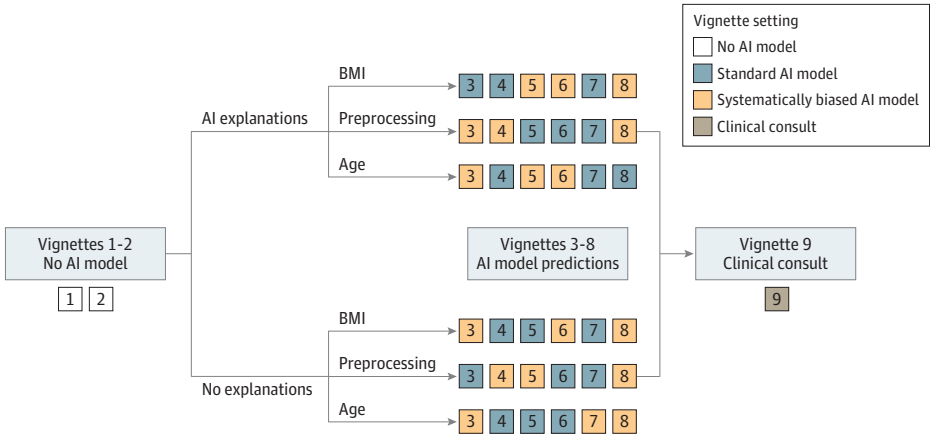
This study was deemed exempt by the University of Michigan Institutional Review Board. Although the study was not registered initially given its vignette-based design, to ensure it was appropriately reflected in the trial literature, it was registered on ClinicialTrials.gov after its completion with the registration parameters based on the prespecified protocol.

### Clinical Vignettes

We created 45 clinical vignettes based on patients hospitalized with acute respiratory failure in 2017 at the University of Michigan. Patients were selected to achieve study goals while being representative of patients hospitalized with acute respiratory failure (see eMethods in Supplement 3). A study team member reviewed each patient's chart to create a succinct version of the patient's presenting history, past medical history, current medications, physical examination findings, laboratory findings, and chest radiograph for each vignette (eFigure 2 in Supplement 3). At least 4 pulmonary physicians independently reviewed each patient's complete medical record to determine if pneumonia, heart failure, and/or chronic obstructive pulmonary disease (COPD) was the patient's underlying diagnosis.

Each vignette had 3 reference diagnostic labels pertaining to pneumonia, heart failure, and COPD based on the average of these reviews, and each reference label was binary, indicating whether the disease was present or absent. See eTable 1 in Supplement 3 for patient characteristics for additional details.

Figure 1. Randomization and Study Flow Diagram for the 9 Clinical Vignettes




After completing informed consent, participants were randomized to artificial intelligence (AI) predictions with or without explanations and all participants were also randomized to 1 of 3 types of systematically biased AI models during a subset of vignettes in the study. The 3 systematically biased AI models included a model predicting pneumonia if aged 80 years or older, a model predicting heart failure if body mass index (BMI, calculated as weight in kilograms divided by height in meters squared) was 30 or higher, and a model predicting chronic obstructive pulmonary disease (COPD) if a blur was applied to the radiograph. Participants were first shown 2 vignettes without AI predictions to measure baseline diagnostic accuracy. The next 6 vignettes included AI predictions. If the participant was randomized to see AI explanations, the participant was also shown an AI model explanation with the AI predictions. Three vignettes had standard AI predictions, and 3 had biased AI predictions shown in random order. The final vignette included a clinical consultation, a short narrative provided by a hypothetical trusted colleague who identified the correct diagnosis and their diagnostic rationale.

## Study Design

The survey was conducted using a Qualtrics survey tool. Participants were assigned 9 of 45 clinical vignettes to review. All participants followed the same vignette order: 2 vignettes without AI predictions, 6 vignettes with AI predictions, and 1 vignette with a clinical consultation by a hypothetical colleague (**Figure 1**). Upon study initiation, all participants were randomized to AI predictions with or without AI explanations for all 6 AI vignettes, and all participants were also randomized to see 1 of 3 types of systematically biased AI models during 3 of the 6 AI vignettes.

The first 2 vignettes did not include AI predictions and were used to measure baseline diagnostic accuracy. The next 6 vignettes included AI predictions immediately after the patient case. If randomized to see AI explanations, the participant was also shown an image-based AI model explanation alongside the AI predictions. Three vignettes had standard AI predictions and 3 had biased predictions shown in random order. The final vignette included a clinical consultation, a short narrative provided by a hypothetical trusted colleague describing the rationale behind the most likely diagnosis and recommended treatments. By design, the clinical consultation vignette always recommended the correct diagnosis and treatment, serving as a realistic upper bound of participant diagnostic accuracy when provided accurate recommendations.

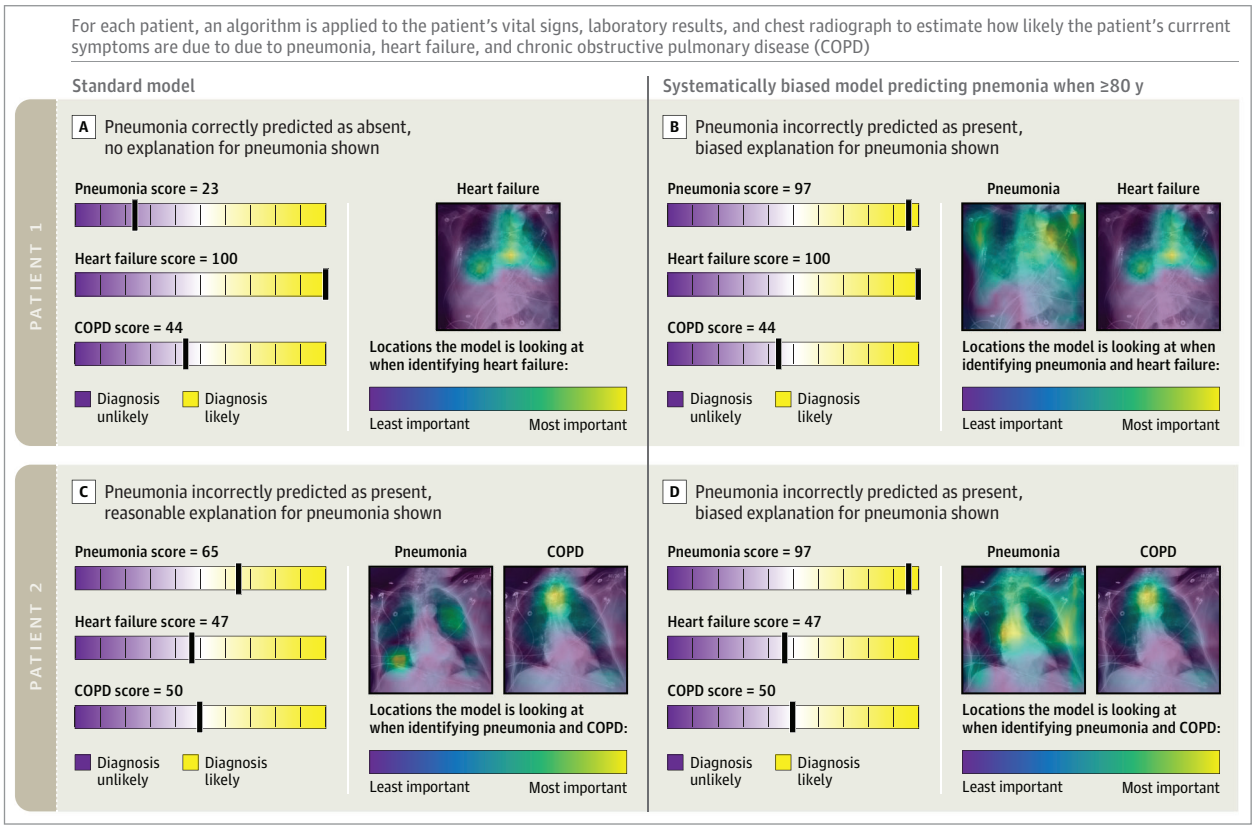
After each vignette, participants were asked to separately assess how likely pneumonia, heart failure, or COPD was contributing to the patient's respiratory failure on a scale of 0 to 100 (eFigure 3 in Supplement 3). Responses were collected on a continuous scale to calculate the correlation between AI model scores and participant responses. In calculating accuracy, responses of 50 or higher were considered positive for the diagnosis. Patients could have 1, more than 1, or none of the conditions. After completing the diagnostic assessment, participants were asked whether they would administer any combination of steroids, antibiotics, intravenous (IV) diuretics, or none of these treatments to the patient.

## Standard and Systematically Biased AI Model Predictions and Explanations

When participants were shown AI predictions, the model provided a score for each diagnosis on a scale of 0 to 100 estimating how likely the patient's symptoms were due to pneumonia, heart failure, and COPD (**Figure 2**). Standard AI predictions were generated using a machine learning model trained using standard approaches on a separate data set.[19] All participants were randomized to see 1 of 3 systematically biased AI models: (1) a model predicting a high likelihood of pneumonia for a patient who was 80 years or older; (2) a model predicting a high likelihood of heart failure if a patient's body mass index (BMI, calculated as weight in kilograms divided by height in meters squared) was 30 or higher; and (3) a model predicting a high likelihood of COPD if a blur was applied to the radiograph (eMethods in Supplement 3). Vignettes with biased models were shown for patients with attributes leading the biased model to predict the diagnosis with high likelihood. For example, a participant randomized to see the model biased for age was shown vignettes of patients 80 years or older with the biased model, and the model displayed a high likelihood prediction for pneumonia and standard predictions for the 2 other diagnoses.

For participants randomized to explanations, explanations were presented when the AI model provided a score of

Figure 2. Examples of Model Predictions and Explanations for Standard and Systematically Biased AI Models for Patients




Patient 1 is an 81-year-old male with respiratory failure from heart failure. A, The standard AI model correctly diagnosed heart failure as the cause of acute respiratory failure and provided an explanation highlighting areas in the chest radiograph used to make the prediction. B, The systematically biased AI model incorrectly diagnosed pneumonia as the cause of acute respiratory failure due to the patient's age and provided an explanation highlighting irrelevant features in the chest radiograph. Standard-model predictions for heart failure (the correct diagnosis) are also provided.

Patient 2 is an 88-year-old female with respiratory failure from COPD. C, The standard AI model incorrectly diagnosed pneumonia and correctly diagnosed COPD as the cause of respiratory failure and provided reasonable explanations. D, The biased AI model incorrectly diagnosed pneumonia as the cause of respiratory failure due to patient age and provided an explanation highlighting irrelevant features in the chest radiograph. Standard-model predictions for COPD (the correct diagnosis) are also provided.

50 or higher (corresponding to positive diagnosis). Explanations were shown as Grad-CAM heatmaps overlaid on the chest radiograph that highlighted which regions of the image most affecting the model's prediction (Figure 2).[14] Explanations generated from the standard model highlighted clinically meaningful regions on the chest radiograph (eg, lungs). Explanations associated with the systematically biased models highlighted areas of the radiograph corresponding to age, BMI, or preprocessing (eg, low bone density, soft tissue; Figure 2; and see eMethods in Supplement 3 for details of how biased model explanations were generated). Because the biased models always made positive predictions, participants randomized to explanations were always shown explanations with biased predictions. In contrast, when shown standard-model predictions, explanations were shown with 17 pneumonia predictions (38%) , 18 heart failure predictions (40%), and 9 COPD predictions (20%).

By design, the standard model had an overall accuracy of 75% and the biased models had overall accuracy of 70% (eTable 2 in Supplement 3), which is on par with physicians and reflects the difficulty of the diagnostic tasks.[20] Therefore, such models would likely meet specifications by regulators for approval. However, among the subset of patients for whom the biased models made biased predictions, the biased models were 33% accurate, whereas the standard model was 67% accurate.

## Participant Demographics

After completing the vignettes, participant demographic information, including self-reported gender, race, and ethnicity, was collected to assess sample diversity and representativeness. Participants were also asked about AI knowledge (see eMethods for the postsurvey questions in Supplement 3).

## Study Outcomes

The primary outcome was diagnostic accuracy between vignette settings, defined as the number of correct diagnostic assessments over the total number of assessments. Participants made 3 separate assessments in each vignette, 1 for each diagnosis. If the participant's assessment agreed with the reference label, the diagnostic assessment was correct. Overall accuracy was calculated using assessments for all diagnoses.

Accuracy for a specific diagnosis was calculated using assessments relevant to the specific diagnosis.

In vignettes with biased model predictions, the biased model provided a prediction for the diagnosis for which it was biased and standard predictions for the 2 other diagnoses. Although participants made 3 diagnostic assessments in those vignettes, only diagnostic assessments corresponding to the biased predictions were included in the accuracy calculation.

Secondary outcomes included treatment selection accuracy, correlation between participants' diagnostic assessment scores and model scores, and confusion matrices across vignette settings.

## Statistical Analysis

The primary preplanned analysis compared diagnostic accuracy between the clinician baseline and each AI model vignette setting. This analysis sought to specifically determine (1) whether standard AI improved clinician accuracy, (2) whether standard AI with explanations improved accuracy, (3) whether biased AI hurt diagnostic accuracy, and (4) whether biased AI with explanations helped clinicians recover from the biased predictions.

Completed vignettes were included in the analysis regardless of whether a participant completed all 9 vignettes. Diagnostic accuracy across vignette settings was compared by fitting a cross-classified, generalized, random-effects model,[21] in which individual diagnostic assessments were nested within study participants and patients. The outcome was a binary variable indicating whether the diagnostic assessment was correct and the exposure was the specific vignette setting for which the assessment was performed. Predictive margins and contrasts were calculated to estimate accuracy at each vignette setting, and differences between settings, with 95% CIs were calculated using the delta method.[22]

Exploratory subgroup analyses were also conducted by specific diagnosis, clinician profession, and participant's prior interactions with clinical decision support. A secondary analysis was conducted to understand the effect of correct and incorrect model predictions on treatment selection accuracy.

We aimed to recruit 400 participants based on simulation-based power calculations. The simulation ensured adequate power to detect a 20% decrease in accuracy with the systematically biased AI model compared with baseline and 10% improvement with AI explanations (eMethods in Supplement 3 and the statistical analysis plan in Supplement 2). Statistical analyses were performed using Stata version 16 (StataCorp). Statistical significance was based on a $P$ value <.05. $P$ values for multiple comparisons were not adjusted in the preplanned analyses.

Table. Characteristics of Participants Who Were Randomized and Completed at Least 1 Clinical Vignette[a]

| Subject characteristics | No. (%) of participants | |
|---|---|---|
| | AI model alone | AI model and explanations |
| Randomized and completed at least 1 vignette | 231 | 226 |
| Complete all vignettes | 214 (93) | 204 (90) |
| Response time for completing survey, median (IQR), min | 19 (14-32) | 19 (14-27) |
| Completed postsurvey questions | 214 | 204 |
| Age, median (IQR), y | 35 (31-40) | 34 (31-38) |
| Years of practice, median (IQR)[b] | 5 (2-9) | 4 (2-8) |
| Prior interaction with AI | 70 (32.7) | 62 (30.4) |
| AI bias aware | 68 (31.8) | 71 (34.8) |
| Sex[c] | | |
| Female | 123 (57.5) | 118 (57.8) |
| Male | 82 (38.3) | 82 (40.2) |
| Prefer not to say | 9 (4.2) | 3 (1.5) |
| Nonbinary or nonconforming | 0 | 1 (0.5) |
| General practice area | | |
| Hospital medicine | 210 (98.1) | 198 (97.1) |
| Other | 4 (1.9) | 6 (2.9) |
| Role on health care team | | |
| Attending physician | 138 (64.5) | 121 (59.3) |
| Physician assistant | 55 (25.7) | 57 (27.9) |
| Nurse practitioner | 12 (5.6) | 18 (8.8) |
| Resident or fellow | 7 (3.3) | 8 (3.9) |
| Other | 2 (0.9) | 0 |
| Hospital setting | | |
| University hospital or academic | 186 (86.9) | 175 (85.8) |
| Community hospital or private practice | 32 (15.0) | 30 (14.7) |
| No response | 17 (7.9) | 23 (11.3) |
| Veterans Affairs or government | 11 (5.1) | 11 (5.4) |
| Race and ethnicity | | |
| 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 |
| White | 139 (65.0) | 146 (71.6) |
| 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) |

Abbreviation: AI, artificial intelligence.

[a] Percentages for demographic, AI interaction, and AI bias questions are based on participants who responded. Participants could respond with more than 1 answer for the hospital setting and race and ethnicity questions, so totals and percentages are greater than 100%.

[b] Years of practice was defined as the number of years since participants completed their terminal medical training (eg, years since completing medical residency).

[c] Participants identified their sex by selecting from the following: male, female, nonbinary or nonconfroming, transgender, other, or prefer not to say.

# Results

Of 1024 participants who viewed the study information page, 572 (56%) initiated the study and were randomized (**Table**; eFigure 4 in Supplement 3). Four hundred fifty-seven participants completed at least 1 vignette and were included in the primary analysis, including 226 randomized to AI model and explanations. Four hundred eighteen participants completed all 9 vignettes. Study participants were from 13 US states (eTable 3 in Supplement 3). Participant demographics did not

meaningfully differ across randomization groups (eTable 4 in Supplement 3). The majority (97.6%) of participants' general practice area was hospital medicine. The median age was 34 years (IQR, 31-39); 241 participants (57.7%) were female; and 19.1% of participants self-reported as Asian, 1.7% as Black, and 68.1% as White. Of all the participants, 31.6% of had previously interacted with clinical decision support tools and 66.7% were unaware that AI could be systematically biased based on patient demographics.

### Primary Outcome: Participant Diagnostic Accuracy

Participant's baseline diagnostic accuracy without model input was 73.0% (95% CI, 68.3%-77.8%) for the 3 diagnoses (**Figure 3**; eTable 5 in Supplement 3). Accuracy was 67.5% (95% CI, 61.0%-74.0%) for pneumonia, 70.7% (95% CI, 63.1%-78.3%) for heart failure, and 80.5% (95% CI, 74.8%-86.1%) for COPD (Figure 3; eTable 5 in Supplement 3). Provided with standard AI predictions, participant diagnostic accuracy for each disease category increased to 75.9% (95% CI, 71.3%-80.5%), an increase of 2.9 percentage points (95% CI, 0.5-5.2; $P = .02$; (Figure 3; eTables 5 and 6 in Supplement 3). Providing standard AI predictions with explanations increased accuracy to 77.5% (95% CI, 73.0%-82.0%), an increase of 4.4 percentage points from baseline (95% CI, 2.0-6.9; $P < .001$; Figure 3; eTables 5 and 6 in Supplement 3). Participant diagnostic accuracy was 81.1% (95% CI, 76.9%-85.4%) when shown a clinical consultation with perfect accuracy (Figure 3; eTable 5 in Supplement 3).

Systematically biased AI predictions without explanations decreased participant diagnostic accuracy to 61.7% (95% CI, 55.3%-68.2%), a decrease of 11.3 percentage points (95% CI, 7.2-15.5; $P < .001$) from the baseline score (Figure 3; eTables 5 and 6 in Supplement 3). This was mostly due to a decrease in participants' diagnostic specificity (73% [95% CI, 71%-75%] to 53% [95% CI, 49%-58%]; eFigure 5 in Supplement 3). When shown systematically biased AI with explanations, accuracy was 64.0% (95% CI, 57.6%-70.3%), a decrease of 9.1 percentage points (95% CI, 4.9-13.2; $P<.001$) from the baseline score, which did not significantly differ from clinician diagnostic accuracy with biased AI without explanations (difference, 2.3 percentage points [95% CI, −2.7 to 7.2]; $P = .37$; eTables 5 and 7 in Supplement 3), with little improvement to participants' diagnostic specificity (53% [95% CI, 49%-58%] to 56% [95% CI, 50%-60; eFigure 5 in Supplement 3). The direction of all effects was similar in all exploratory subgroup analyses (Figure 3).

### Secondary Outcomes

Correlations between participant responses and standard-model predictions and predictions with explanations were 0.53 (95% CI, 0.50-0.57) and 0.59 (95% CI, 0.56-0.62) respectively (eTable 8 in Supplement 3). Correlations between participant responses and biased model predictions and predictions with explanations were 0.41 (95% CI, 0.38-0.45) and 0.41 (95% CI, 0.37-0.45), respectively.

Treatment selection accuracy was 70.3% (95% CI, 65.5%-75.2%) without model predictions (**Figure 4**; eTable 9 in Supplement 3). When shown correct AI predictions, accuracy increased to 77.0% (95% CI, 72.6%-81.4%), an increase of 6.7 percentage points (95% CI, 4.1-9.3; Figure 4; eTables 9 and 10 in Supplement 3). Accuracy was 80.4% (95% CI, 76.3%-84.5%) when shown correct AI predictions and explanations (Figure 4; eTable 9 in Supplement 3). When provided incorrect model predictions, participants treatment selection accuracy was 55.1% (95% CI, 48.8%-61.3%), a decrease of 15.3 percentage points (95% CI, 11.4-19.1) (Figure 4; eTables 9 and 10 in Supplement 3). When shown incorrect model predictions, providing explanations did not significantly increase participant accuracy (Figure 4; eTable 11 in Supplement 3).

## Discussion

This study examined how AI predictions and popular image-based AI model explanations could potentially help or worsen clinicians' diagnostic and treatment decisions in patients with acute respiratory failure. The results showed that clinicians' diagnostic accuracy improved when shown predictions from a standard AI model, but decreased when shown systematically biased AI predictions. Providing explanations did little to remedy the harmful effects of biased models on clinician accuracy. AI models trained on clinical data are prone to making biased predictions based on clinically irrelevant findings.[5,6,23] Left unchecked, models could exacerbate biases widespread in health care,[7] and errors in clinical judgment influenced by AI could have severe consequences such as patient harm.[16]

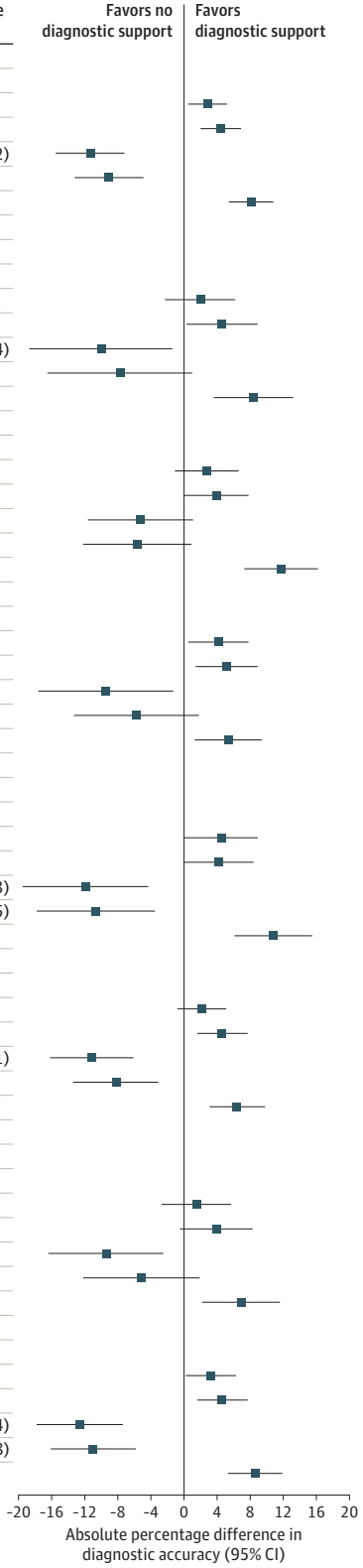
Although the standard and systematically biased models shown in the vignettes had imperfect accuracy, their performance reflected what might be expected in real-world settings for models approved by regulators for difficult tasks like the diagnosis of acute respiratory failure.[24] Even clinicians had only 81.1% diagnostic accuracy when provided perfectly accurate recommendations in the clinical consultation vignettes, which might reflect participant's difficulty in recognizing perfectly accurate recommendations in these settings. This could also suggest that as AI models improve, there may be an upper bound on collaborative performance between AI models and clinicians for difficult diagnostic tasks.

To measure whether explanations could mitigate errors made by systematically biased AI models, explanations were presented in a way that were considered to be obvious—the model relied completely on features unrelated to the clinical condition. Prior work outside of health care has shown that explanations can reduce overreliance on AI models when the task is cognitively difficult, but verifying the explanation is less costly in comparison.[17] Although some have suggested that state-of-the-art explanations should lead to better user-AI interactions,[25] participants in this study were unable to recognize overtly biased models with such explanations. This behavior aligns with work showing that users may still be deceived by incompetent AI models if they fail to understand the models' simple explanations.[26]

Clinicians' inability to leverage explanations when presented with systematically biased AI models may be due to several reasons. First, clinicians might have limited AI literacy.[27]

**Figure 3. Baseline Diagnostic Accuracy Without AI Models and Percentage Point Differences in Accuracy Across Clinical Vignette Settings**

| | Absolute percentage difference (95%CI) |
|---|---|
| Overall | |
| Baseline 73.0% (95% CI, 68.3% to 77.8%) | |
| Standard model | 2.9 (0.5 to 5.2) |
| Standard model plus explanations | 4.4 (2.0 to 6.9) |
| Systematically biased model | -11.3 (-15.5 to -7.2) |
| Systematically biased model plus explanations | -9.1 (-13.2 to -4.9) |
| Clinical consult | 8.1 (5.4 to 10.8) |
| Diagnosis | |
| Pneumonia | |
| Baseline 67.5% (95% CI, 61.0% to 74.0%) | |
| Standard model | 2.0 (-2.3 to 6.2) |
| Standard model plus explanations | 4.6 (0.3 to 8.9) |
| Systematically biased model | -10.0 (-18.7 to -1.4) |
| Systematically biased model plus explanations | -7.7 (-16.5 to 1.0) |
| Clinical consult | 8.4 (3.6 to 13.2) |
| Heart failure | |
| Baseline 70.7% (95% CI, 63.1% to 78.3%) | |
| Standard model | 2.7 (-1.1 to 6.6) |
| Standard model plus explanations | 3.9 (-0.0 to 7.8) |
| Systematically biased model | -5.3 (-11.6 to 1.1) |
| Systematically biased model plus explanations | -5.6 (-12.2 to 0.9) |
| Clinical consult | 11.7 (7.3 to 16.2) |
| COPD | |
| Baseline 80.5% (95% CI, 74.8% to 86.1%) | |
| Standard model | 4.2 (0.5 to 7.8) |
| Standard model plus explanations | 5.2 (1.4 to 8.9) |
| Systematically biased model | -9.5 (-17.6 to -1.3) |
| Systematically biased model plus explanations | -5.7 (-13.3 to 1.8) |
| Clinical consult | 5.4 (1.3 to 9.4) |
| Clinician profession | |
| Nurse practitioner or physician assistant | |
| Baseline 71.3% (95% CI, 66.0% to 76.7%) | |
| Standard model | 4.5 (0.0 to 8.9) |
| Standard model plus explanations | 4.2 (-0.1 to 8.4) |
| Systematically biased model | -11.9 (-19.5 to -4.3) |
| Systematically biased model plus explanations | -10.7 (-17.8 to -3.5) |
| Clinical consult | 10.8 (6.1 to 15.5) |
| Physician | |
| Baseline 74.0% (95% CI, 69.0% to 78.9%) | |
| Standard model | 2.1 (-0.8 to 5.1) |
| Standard model plus explanations | 4.6 (1.6 to 7.7) |
| Systematically biased model | -11.1 (-16.2 to -6.1) |
| Systematically biased model plus explanations | -8.2 (-13.4 to -3.1) |
| Clinical consult | 6.4 (3.1 to 9.8) |
| Clinical decision support interaction | |
| Prior interaction | |
| Baseline 74.1% (95% CI, 68.6% to 79.7%) | |
| Standard model | 1.5 (-2.7 to 5.7) |
| Standard model plus explanations | 3.9 (-0.5 to 8.3) |
| Systematically biased model | -9.4 (-16.4 to -2.5) |
| Systematically biased model plus explanations | -5.2 (-12.2 to 1.9) |
| Clinical consult | 6.9 (2.2 to 11.6) |
| No interaction | |
| Baseline 72.7% (95% CI, 67.9% to 77.6%) | |
| Standard model | 3.2 (0.2 to 6.3) |
| Standard model plus explanations | 4.6 (1.6 to 7.7) |
| Systematically biased model | -12.6 (-17.8 to -7.4) |
| Systematically biased model plus explanations | -11.0 (-16.1 to -5.8) |
| Clinical consult | 8.6 (5.3 to 11.9) |

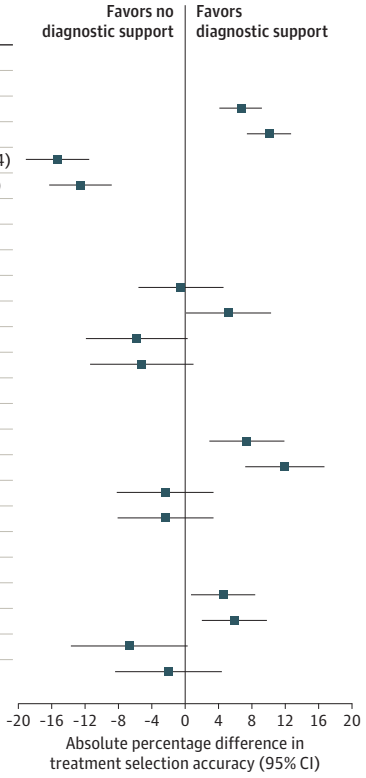



Baseline indicates diagnostic accuracy of heart failure, pneumonia, and chronic obstructive pulmonary disease (COPD) when shown clinical vignettes of patients with acute respiratory failure without AI model input; standard model, diagnostic accuracy when shown clinical vignettes and standard AI model diagnostic predictions about whether the patient has heart failure, pneumonia, and/or COPD; standard model plus explanations, diagnostic accuracy when shown standard AI predictions and an image-based AI explanation of the model's reasoning for making a prediction within vignettes; systematically biased model, diagnostic accuracy when shown systematically biased AI predictions of low accuracy within vignettes; systematically biased model plus explanations, diagnostic accuracy when shown biased model predictions and explanations within vignettes; and clinical consultation, diagnostic accuracy when provided a short narrative describing the rational for the correct diagnosis within the vignette.

Subgroup analysis included diagnostic accuracy specific to heart failure, pneumonia, and COPD; clinician profession, including 142 nurse practitioners or physician assistants, and 274 physicians; prior clinical decision–support interaction, including 132 participants who had prior experience interacting with clinical decision support systems and 286 who did not. Diagnostic accuracy and percentage point differences in accuracy were determined by calculating predictive margins and contrasts across vignette settings after fitting a cross-classified generalized random effects model of diagnostic accuracy.

Figure 4. Baseline Treatment Selection Accuracy Without AI Models and Percentage Point Differences in Accuracy Across Clinical Vignette Settings

| | Absolute percentage difference (95%CI) |
|---|---|
| Overall treatment selection accuracy | |
| Baseline 70.3% (95% CI, 65.5% to 75.2%) | |
| Correct model | 6.7 (4.1 to 9.3) |
| Correct model plus explanations | 10.1 (7.4 to 12.8) |
| Incorrect model | -15.3 (-19.1 to -11.4) |
| Incorrect model plus explanations | -12.5 (-16.3 to -8.8) |
| Subgroup analysis by treament | |
| Antibiotics | |
| Baseline 61.1% (95% CI, 53.8% to 68.5%) | |
| Correct model | -0.5 (-5.6 to 4.6) |
| Correct model plus explanations | 5.2 (0.1 to 10.3) |
| Incorrect model | -5.8 (-11.9 to 0.3) |
| Incorrect model plus explanations | -5.2 (-11.4 to 1.0) |
| Diurectics | |
| Baseline 69.4% (95% CI, 61.6% to 77.1%) | |
| Correct model | 7.4 (2.9 to 11.9) |
| Correct model plus explanations | 11.9 (7.2 to 16.7) |
| Incorrect model | -2.4 (-8.2 to 3.4) |
| Incorrect model plus explanations | -2.4 (-8.1 to 3.4) |
| Steroids | |
| Baseline 78.9% (95% CI, 72.7% to 85.0%) | |
| Correct model | 4.6 (0.7 to 8.4) |
| Correct model plus explanations | 5.9 (2.0 to 9.8) |
| Incorrect model | -6.7 (-13.7 to 0.3) |
| Incorrect model plus explanations | -2.0 (-8.4 to 4.4) |




Baseline treatment selection accuracy indicates accurate administration of antibiotics, diuretics, and/or steroids after reviewing vignettes of patients with acute respiratory failure without AI model input; correct model, treatment accuracy when shown vignette with correct AI model diagnostic predictions of heart failure, pneumonia, and/or COPD; correct model plus explanations, treatment accuracy when shown a vignette with correct AI model diagnostic predictions and an image-based AI explanation of the model's reason for making a prediction; incorrect model, treatment accuracy when shown a vignette with incorrect AI model diagnostic predictions; and incorrect model plus explanation, treatment accuracy when shown incorrect AI model diagnostic predictions and explanations.

Subgroup analysis included treatment selection accuracy specific to antibiotics, intravenous diuretics, and steroids. Treatment selection accuracy and percentage point differences in accuracy were determined by calculating predictive margins and contrasts across vignette settings after fitting a cross-classified generalized random-effects model of treatment selection accuracy across settings.

For example, 66.7% of participants were unaware that AI models could be systematically biased. Informing clinicians of the strengths and weaknesses of AI could make them more adept at leveraging model explanations. Second, clinicians were only shown a small number of the vignettes with explanations, and might need more training with image-based explanations. Third, common image-based AI model explanations, such as those tested in this study, are only approximations of model behavior.[14,28,29] They require users to form a mental model of what the explanation is communicating because the AI model cannot communicate in plain language what it is focusing on. It is possible that explanations may be more helpful when presented in a different way (eg, with text descriptions).[30]

## Study Implications

Clinicians' limited ability to recognize when AI models are systematically biased has important implications for medical AI regulatory policies. The FDA recently released guidance highlighting the importance of providing clinicians with the ability to independently verify clinical support software recommendations.[12,31,32] Explanations were also emphasized in the White House blueprint for an AI Bill of Rights.[33] Despite providing clinicians with tools with the ability to explain and that could aid them in identifying AI model mistakes, this study demonstrates that the ability to check an AI model's reasoning using current image-based explanation approaches could fail to provide meaningful safeguards against systematically biased models.

There are several potential paths forward to safely integrate AI in health care. First, additional training in limitations of AI systems could be incorporated into medical training, better equipping clinicians to identify biased models.[34] Second, researchers developing explanation tools should involve clinicians to better understand their specific needs. Third, standardizing clinician-facing AI model information in simple language and empirically testing such standards may help clinicians understand appropriate model use and limitations.[35] Finally, this study highlights the critical need for careful model validation, including appropriate subgroup testing, to identify and address biased behavior prior to deployment.

## Limitations

The current study has several limitations. First, it was conducted using a web-based survey interface, which is inherently different from the clinical setting. Second, although study recruitment resulted in a diverse participant population from 14 states, participants were slightly younger (34 years vs 41 years) and more female (53% vs 36%) than prior national statistics of hospitalists,[36] and the design prevented collection of information on nonresponders. Third, the study focused on the effects of AI predictions and explanations among clinicians who make final diagnostic and treatment decisions. Although radiologists have more training in reading chest radiographs and may be more adept at understanding radiograph explanations, they were not included because they are not ultimately responsible for making these diagnostic and treatment decisions.

## Conclusions

Standard AI models' predictions and explanations improved diagnostic accuracy of clinicians, whereas systematically biased predictions and explanations reduced accuracy. Given the unprecedented pace of AI development,[37] it is essential to carefully test AI integration into clinical workflows. Although the findings of the study suggest that clinicians may not be able to serve as a backstop against flawed AI, they can play an essential role in understanding AI's limitations.


### ARTICLE INFORMATION

**Accepted for Publication:** October 11, 2023.

**Author Contributions:** Ms Jabbour and Dr Sjoding had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Senior authors Drs Wiens and Sjoding contributed equally.
*Concept and design:* All authors.
*Acquisition, analysis, or interpretation of data:* Jabbour, Shepard, Valley, Kazerooni, Banovic, Wiens, Sjoding.
*Drafting of the manuscript:* Jabbour, Fouhey, Kazerooni, Wiens, Sjoding.
*Critical review of the manuscript for important intellectual content:* All authors.
*Statistical analysis:* Jabbour, Kazerooni, Wiens, Sjoding.
*Obtained funding:* Fouhey, Wiens, Sjoding.
*Administrative, technical, or material support:* Jabbour, Fouhey, Kazerooni, Wiens, Sjoding.
*Supervision:* Fouhey, Kazerooni, Banovic, Wiens, Sjoding.

**Conflict of Interest Disclosures:** Dr Banovic reported receiving grants from the US Department of Energy, Toyota Research Institute, and National Science Foundation outside the submitted work. Dr Wiens reported receiving grants from the Alfred P. Sloan Foundation during the conduct of the study and serving on the advisory board of Machine Learning for Healthcare, a nonprofit organization that hosts a yearly academic conference. Dr Sjoding reported receiving royalties for a patent from Airstrip outside the submitted work. No other disclosures were reported.

**Funding/Support:** This work was supported by grant R01 HL158626 from the National Heart, Lung, and Blood Institute (NHLBI).

**Role of the Funder/Sponsor:** The NHLBI had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Data Sharing Statement:** See Supplement 4.

**Additional Contributions:** We thank all individuals who helped test the survey platform used in this study, the clinicians who helped pilot our study, and all those who responded to our survey. Those who completed our survey were compensated for their contributions.



### REFERENCES

1. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229-1234. doi:10.1038/s41591-020-0942-0

2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216

3. van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27(5):775-784. doi:10.1038/s41591-021-01343-4

4. Kather JN, Weis C-A, Bianconi F, et al. Multi-class texture analysis in colorectal cancer histology. *Sci Rep*. 2016;6(1):27988. doi:10.1038/srep27988

5. Jabbour S, Fouhey D, Kazerooni E, Sjoding MW, Wiens J. Deep learning applied to chest x-rays: exploiting and preventing shortcuts. *Proc Mach Learn Res*. 2020;126:750-782.

6. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. 2022;4(6):e406-e414. doi:10.1016/S2589-7500(22)00063-2

7. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

8. Beery TA. Gender bias in the diagnosis and treatment of coronary artery disease. *Heart Lung*. 1995;24(6):427-435. doi:10.1016/S0147-9563(95)80020-4

9. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4(1):31. doi:10.1038/s41746-021-00385-9

10. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *ArXiv*. Preprint posted February 19, 2021. doi:10.48550/arXiv.2102.09692

11. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. *ArXiv*. Preprint posted December 13, 2022. doi:10.48550/arXiv.2212.06823

12. Clinical decision support software: guidance for industry and Food and Drug Administration staff. US Food and Drug Administration. September 28, 2022. Accessed March 1, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

13. Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment. Paper presented at: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; January 27-30, 2020; Barcelona, Spain:648-657. doi:10.1145/3351095.3375624

14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *ArXiv*. Preprint posted October 7, 2016. doi:10.48550/arXiv.1610.02391

15. Kempker JA, Abril MK, Chen Y, Kramer MR, Waller LA, Martin GS. The epidemiology of respiratory failure in the United States 2002-2017: a serial cross-sectional study. *Crit Care Explor*. 2020;2(6):e0128. doi:10.1097/CCE.0000000000000128

16. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DRM. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Acad Med*. 2012;87(2):149-156. doi:10.1097/ACM.0b013e31823f71e6

17. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein M, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. *arXiv*. Preprint posted December 13, 2022. doi:10.48550/arXiv.2212.06823

18. Society of Hospital Medicine. Accessed August 28, 2023. https://www.hospitalmedicine.org

19. 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. *J Am Med Inform Assoc*. 2022;29(6):1060-1068. doi:10.1093/jamia/ocac030

20. Ray P, Birolleau S, Lefort Y, et al. Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis. *Crit Care*. 2006;10(3):R82. doi:10.1186/cc4926

21. Clayton DG. Generalized linear mixed models. In: Gilks WR, Richardson S, Spiegelhalter DJ, eds. *Markov Chain Monte Carlo in Practice*. Chapman & Hall; 1996:275-302.

22. Oehlert GW. A note on the delta method. *Am Stat*. 1992;46(1):27-29. doi:10.2307/2684406

23. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts

over signal. *Nat Mach Intell*. 2021;3(7):610-619. doi:10.1038/s42256-021-00338-7

**24**. Ray P, Birolleau S, Lefort Y, et al. Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis. *Crit Care*. 2006;10(3):R82. doi:10.1186/cc4926

**25**. Bai B, Liang J, Zhang G, Li H, Bai K, Wang F. Why attentions may not be interpretable? *arXiv*. Preprint posted June 10, 2020. doi:10.48550/arXiv.2006.05656

**26**. Banovic N, Yang Z, Ramesh A, Liu A. Being trustworthy is not enough: how untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. *Proc ACM Hum Comput Interact*. 2023;7(CSCW1):1-17. doi:10.1145/3579460

**27**. Long D, Magerko B. What is AI literacy? competencies and design considerations. *Proc Conf Hum Factors Comput Syst*. 2020:1-16. doi:10.1145/3313831.3376727

**28**. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proc 31st Int Conf Neural Info Process Systems*. 2017:4768-4777. https://dl.acm.org/doi/10.5555/3295222.3295230

**29**. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. *arXiv*. Preprint posted June 16, 2016. doi:10.48550/arXiv.1606.05386

**30**. Pazzani M, Soltani S, Kaufman R, Qian S, Hsiao A. Expert-informed, user-centric explanations for machine learning. *Proc AAAI Conf Art Intel*. 2022;36(11):12280-12286. doi:10.1609/aaai.v36i11.21491

**31**. Shachar C, Gerke S. Prevention of bias and discrimination in clinical practice algorithms. *JAMA*. 2023;329(4):283-284. doi:10.1001/jama.2022.23867

**32**. Office for Civil Rights, Office of the Secretary of Health and Human Services. Nondiscrimination in health programs and activities: final rule. *Fed Regist*. 2022;87:47824-47920. https://www.federalregister.gov/documents/2022/08/04/2022-16217/nondiscrimination-in-health-programs-and-activities

**33**. Blueprint for an AI Bill of Rights: making automated systems work for the American people. White House. Posted 2023. Accessed March 1, 2023. https://www.whitehouse.gov/ostp/ai-bill-of-rights

**34**. Ötleş E, James CA, Lomis KD, Woolliscroft JO. Teaching artificial intelligence as a fundamental toolset of medicine. *Cell Rep Med*. 2022;3(12):100824. doi:10.1016/j.xcrm.2022.100824

**35**. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med*. 2020;3(1):41. doi:10.1038/s41746-020-0253-3

**36**. Ryskina KL, Shultz K, Unruh MA, Jung H-Y. Practice trends and characteristics of US hospitalists from 2012 to 2018. *JAMA Health Forum*. 2021;2(11):e213524-e213524. doi:10.1001/jamahealthforum.2021.3524

**37**. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv*. Preprint March 22, 2023. doi:10.48550/arXiv.2303.12712