STUDY PROTOCOL

Measuring the Impact of AI in the Diagnosis of Hospitalized Patients through a Web-based Randomized Clinical Vignette Multicenter Study

Co-Principal Investigators: Michael W. Sjoding, Jenna Wiens

Funded by: NIH NHLBI R01 HL158626

last updated:

8 June 2023

Principal Investigator(s)	J. Wiens, PhD Department of Computer Science and Engineering 2260 Hayward Street Ann Arbor, MI 48109 (734) 647-4832 Email: wiensj@umich.edu Michael W. Sjoding, MD Internal Medicine, Division of Pulmonary and Critical Care G027W Building 16 NCRC, 2800 Plymouth Road, SPC 2800, Ann Arbor, MI 48109 Phone: 734-763-1796 msjoding@umich.edu

14

15

Table of Contents

1. INTRODUCTION AND RATIONALE 8

2. OBJECTIVES 8

3. Study Design 9

4. Study Population 10

4.1 Population (base) 10

4.2 Inclusion criteria 10

4.3 Exclusion criteria 10

4.4 Screen failure 10

4.5 Technical failure 10

4.6 Participant withdrawal 11

4.7 Sample size calculation 11

5. TREATMENT OF SUBJECTS 12

5.1 Intervention 12

5.2 Use of co-intervention (if applicable) 13

5.3 Escape medication (if applicable) 13

6. INVESTIGATIONAL PRODUCT 13

6.1 Name and description of investigational product(s) 13

6.2 Summary of findings from non-clinical studies. 14

6.3 Summary of findings from clinical studies. 14

6.4 Summary of known and potential risks and benefits 15

6.5 Description and justification of route administration and dosage 15

6.6 Dosages, dosage modifications and method of administration 15

6.7 Preparation and labelling of Investigational Medicinal Product 16

6.8 Drug accountability 16

7. METHODS 16

7.1 STUDY PARAMETERS/ENDPOINTS 16

7.1.1 Main study parameters/endpoints 16

7.2 Randomization, blinding, and treatment allocation 16

7.3 Study procedures 16

7.4 Withdrawal of individual subjects 17

7.4.1 Specific criteria for withdrawal (if applicable) 17

7.5 Replacement of individual subjects after withdrawal 17

7.6 Follow-up of subjects withdrawn from treatment 17

7.7 Premature termination of the study **17**
8. SAFETY REPORTING ***17***
8.1 Temporary halt for reasons of subject safety **17**
8.2 AEs, SAEs, SUSARs **17**
8.2.1 Adverse events (AEs) 17
8.2.2 Serious adverse events (SAEs) 18
8.3 Annual safety report **18**
8.4 Follow-up of adverse events **18**
8.5 Data Safety Monitoring Board (DSMB) / Safety Committee **18**
9. STATISTICAL ANALYSIS ***18***
9.1 Primary study parameters/endpoints **18**
9.2 Interim analysis (if applicable) **18**
9.3 Statistical analysis plan **18**
10. ETHICAL CONSIDERATIONS ***19***
10.1 Regulation statement **19**
10.2 Recruitment and consent **19**
10.3 Objection by minors or incapacitated subjects (if applicable) **19**
10.4 Benefits and risks assessment, group relatedness **20**
10.5 Compensation for injury **20**
10.6 Incentives (if applicable) **20**
11. ADMINISTRATIVE ASPECTS, MONITORING AND PUBLICATION ***20***
11.1 Handling and storage of data and documents **20**
11.2 Monitoring and Quality Assurance **20**
11.3 Public disclosure and publication policy **20**
12. AMENDMENTS ***20***
13. REFERENCES ***20***

List of abbreviations

AE	**Adverse Event**
ARF	**Acute Respiratory Failure**
COPD	**Chronic obstructive pulmonary disease**

SUMMARY

Rationale:

Acute respiratory failure (ARF) develops in over 3 million patients hospitalized in the United States annually.[1] Pneumonia, heart failure, and/or chronic obstructive pulmonary disease (COPD) are 3 of the most common reasons for ARF,[2] and these conditions are among the top reasons for hospitalization in the United States.[3] Determining the underlying causes of ARF is critically important for guiding treatment decisions, but can be clinically challenging, as initial testing such as brain natriuretic peptide (BNP) levels or chest radiograph results can be non-specific or difficult to interpret.[4] This is especially true for older adults,[5] patients with comorbid illnesses,[6] or more severe disease.[7] Incorrect initial treatment often occurs, resulting in worse patient outcomes or treatment delays.[8] Artificial intelligence technologies have been proposed as a strategy for improving medical diagnosis by augmenting clinical decision-making,[9] and could play a role in the diagnostic evaluation of patients with ARF.

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

Study design:

This is web-based randomized clinical vignette study.

Objectives

<u>Survey Data Collection Phase</u>

Objectives

- What is clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF without any AI model input?
- How do standard AI model predictions without explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF?

- How do standard AI model predictions with explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF?
- How do intentionally biased AI model predictions without explanations affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF?
- 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?
- How does text input (always accurate) from a clinician affect clinician accuracy in diagnosing pneumonia, heart failure, and COPD in a patient population with ARF?

Study population

Hospitalist physicians, nurse practitioners, and physician assistants who commonly care for patients with ARF from 12 US hospitals.

Intervention:

Within the clinical vignette survey, the primary research question is to understand the impact of providing AI model explanations to clinicians. Therefore, we will randomize participants to see:

(1) AI model explanations vs. no AI model explanations: When clinicians are shown AI models, they will be randomized to see AI model prediction alone each time they are shown an AI model, or randomized to see AI model predictions with an explanation each time they are shown an AI model.

Within the survey, they will also be randomized to see

(2) Type of bias: they type of systematically biased AI model shown in the vignette, either against age, BMI, or model preprocessing.
(3) Vignette ordering: for vignettes with standard model predictions or intentionally biased predictions (vignettes 3-8), the ordering of these will be randomized.

We look at the effect of standard model predictions and standard model predictions with explanations to test if such model input improves clinical diagnostic accuracy. We also look at the effect of intentionally biased AI model predictions to test if such inputs hurt diagnostic accuracy, and whether providing explanations when clinicians are shown systematically biased AI models help clinicians recover in terms of diagnostic accuracy.

Main study parameters/endpoints: Clinician Diagnostic accuracy after reviewing clinical vignette. We will specifically evaluate the following:

- Do standard model predictions improve clinician diagnostic accuracy?
- Do standard model predictions with explanations further improve clinician diagnostic accuracy?
- Do intentionally biased model predictions hurt clinician diagnostic accuracy?
- Do intentionally biased model explanations help clinicians recover from the negative effects of intentionally biased model predictions?

Nature and extent of the burden and risks associated with participation, benefit, and group relatedness:

Because this is a web-based study involving benign (non-harmful) behavioral interventions, no adverse events are expected during this study.

1. INTRODUCTION AND RATIONALE

Acute respiratory failure (ARF) develops in over 3 million patients hospitalized in the United States annually.[1] Pneumonia, heart failure, and/or chronic obstructive pulmonary disease (COPD) are 3 of the most common reasons for ARF,[2] and these conditions are among the top reasons for hospitalization in the United States.[3] Determining the underlying causes of ARF is critically important for guiding treatment decisions, but can be clinically challenging, as initial testing such as brain natriuretic peptide (BNP) levels or chest radiograph results can be non-specific or difficult to interpret.[4] This is especially true for older adults,[5] patients with comorbid illnesses,[6] or more severe disease.[7]

Incorrect initial treatment for ARF often occurs, resulting in worse patient outcomes or treatment delays.[8] Artificial intelligence technologies have been proposed as a strategy for improving medical diagnosis by augmenting clinical decision-making,[9] and could play a role in the diagnostic evaluation of patients with ARF. If integrated into clinical workflows effectively, such technologies could improve clinician diagnostic accuracy for ARF and result in better patient outcomes. We developed an artificial intelligence algorithm that can predict the underlying etiologies of ARF based on patient chest X-rays and clinical data. Theoretically, this algorithm could improve clinician's diagnostic accuracy. While promising, systematic bias in AI models can lead to inaccurate predictions, ultimately hurting physician performance. Thus, the extent to which AI can be safely integrated into clinical workflows to support diagnostic decisions is still unknown.

The model developed in this study predicts whether the patient has pneumonia, heart failure, and/or COPD based on their chest X-ray and clinical data. The model also has corresponding explanations based on the chest X-ray, which highlight the areas that the model found important for its decision. The standard model developed highlights clinically relevant regions for pneumonia (e.g., lungs), heart failure (e.g., enlarged heart), and COPD (e.g., tracheal narrowing), whereas the systematically biased models highlight clinically irrelevant findings for pneumonia (bone density for age), heart failure (body mass for BMI), and COPD (features of image preprocessing blur). In this web-based study, we will test our hypothesis that providing participants with systematically biased predictions without explanations will hurt their diagnostic accuracy, whereas providing them with systematically biased predictions with explanations will help them recover from these negative effects.

2. OBJECTIVES

Objectives:

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

3. Study Design

We aim to include 400 participants from multiple hospital centers. Participants will be invited to participate in randomized clinical vignette survey in which participants are randomly shown 9 clinical vignettes out of 45 possible vignettes. The first two vignettes are not accompanied by AI model predictions and are used to estimate baseline participant diagnostic accuracy. The next 6 vignettes include AI model predictions, but half of the participants will be randomized to also see AI model explanations when shown the AI model predictions. These 6 vignettes include 3 vignettes with standard model predictions and 3 with systematically biased model predictions. Participants will be randomized to see one of three types of systematically biased AI model predictions, and the ordering of the 3 standard and 3 systematically biased model predictions are randomized. In the final vignette, all participants are provided a clinical consult, which is a short narrative provided by a hypothetical trusted colleague, who describes the rationale behind which diagnoses were most likely and what treatment plan they recommend. By design, the clinical consult always provides the correct diagnosis and appropriate treatment plan to provide a realistic upper bound of participant diagnostic accuracy.

4. Study Population

4.1 Population (base)

Hospitalist physicians, nurse practitioners, and physician assistants who commonly care for patients with acute respiratory failure.

4.2 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

4.3 Exclusion criteria

Any participant answering “No” to the following question will be excluded from the study:

“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

4.4 Screen failure

Not applicable.

4.5 Technical failure

In the unlikely event that data is not properly recorded to the Qualtrics survey interface, the participant responses will be excluded.

4.6 Participant withdrawal

Any completed vignette will be analyzed, even if the participant does not complete all 9 vignettes.

4.7 Sample size calculation

A sample size of 400 will have 80% power to detect a decrease in accuracy of 25% with the systematically biased AI model compared to baseline and a 10% improvement with the biased AI model with explanations compared to no explanations using a generalized linear mixed model with a 0.001 significance level.

The sample size calculation is based on the primary endpoint of clinician diagnostic accuracy for pneumonia, heart failure, and COPD. To calculate diagnostic accuracy, dichotomized responses are compared to the reference standard labels generated by a group of 5 physicians who reviewed the patients complete medical record. Each vignette's three diagnosis responses are analyzed separately within the generalized linear model.

Sample size calculations were based on 100 simulated studies performed at sample size levels of 50 to 500, in increments of 50. The simulations were based on the assumptions that clinician diagnostic accuracy was 0.68 for pneumonia, 0.72 for heart failure, and 0.82 for COPD. The systematically biased model had an accuracy of 0.33. Clinician performance was simulated such that clinicians listened to the biased model 50% of the time. Furthermore, when presented with biased model explanations, clinicians recovered by 50%.

Given these simulated data, we fit a generalized linear mixed model in R to measure if the recovery of clinician performance given the model explanation is statistically significant. We repeat this for every simulated study, and calculate power as the percentage of time the effect of the explanation is statistically significant across all simulations.

We determine that we have 80% power to detect a statistically significant effect of the explanation at a significance level of 0.001:

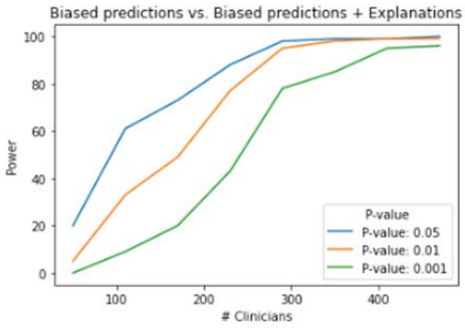

Figure 1. Sample Size Simulation Power Plot.

Details sample size calculation:

Expected loss of data:

If data is not properly stored on the Qualtrics server, the participants responses will be withdrawn.

5. TREATMENT OF SUBJECTS

5.1 Intervention

The purpose of the study is to understand the effect of providing AI model explanations in addition to AI model predictions on clinicians' diagnostic and treatment decisions when diagnosing the underlying causes of acute respiratory failure. In this study, we investigate the use of gradCAM heatmaps as an image-based explanation of the AI model's decision.[18]

GradCAM heatmaps are a commonly used model explanation tool by AI model developers.[19] It is used to highlight the regions of an image used by an AI model to make its prediction. For example, a gradCAM heatmap generated from a model trained to predict heart failure based on a patient's chest X-ray might highlight the patient heart. Testing the usefulness of gradCAM heatmaps means presenting these heatmaps with AI

model predictions, when the AI model predicts that a patient has disease. This does not induce any harm or risk to the patients in the vignettes or participants in the study.

The AI models provide a score for each diagnosis (pneumonia, heart failure, and COPD) on a scale of 0-100, with a score above 50 corresponding to a positive diagnosis. In general, when the standard AI model predicts a positive diagnosis, the explanation is expected to highlight the relevant region of the chest X-ray (e.g., lung infiltrate). When participants are shown a systematically biased AI models, they are randomized to 1 of 3 intentionally biased AI models based on patient age (predicting pneumonia if age ≥ 80 years), BMI (predicting heart failure if BMI ≥ 30), or chest X-ray preprocessing (predicting COPD if a blur was applied to the X-ray). Explanations associated with the systematically biased AI models were generated based on models trained to predict age, BMI, and preprocessing parameters, and highlighted areas of the X-ray corresponding to age, BMI, or preprocessing (e.g., low bone density, soft tissue).

We aim to test our hypothesis that providing participants with systematically biased predictions without explanations will hurt their diagnostic accuracy, whereas providing them with systematically biased predictions with explanations will help them recover from these negative effects. While image-based explanations have been studied in various settings, this will be the first to test the use in the diagnosis of acute respiratory failure at this scale.

5.2 Use of co-intervention (if applicable)

Not applicable.

5.3 Escape medication (if applicable)

Not applicable.

6. INVESTIGATIONAL PRODUCT

The AI model evaluated in the clinical vignette study is based on Jabbour et *al.*[20].

6.1 Name and description of investigational product(s)

This model takes as input the patient's clinical data and chest X-ray at the time of ARF and outputs three separate probabilities that the patient has pneumonia, heart failure, and COPD.

The model provides a score of 0-100 for each of the diagnoses and presents them on a color bar to indicate the likelihood of each disease:

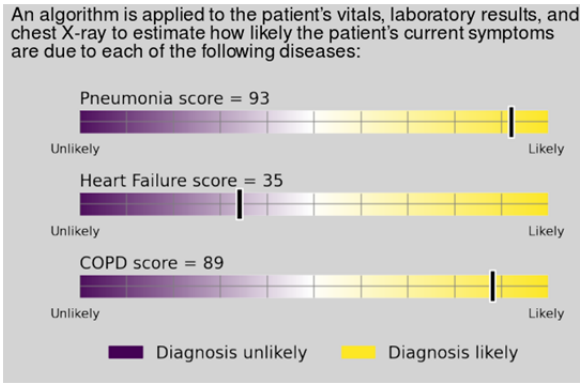

Figure 2. Model scores for each disease.

When the participant is randomized to see model explanations, the model also presents an explanation for each diagnosis, if the score for the diagnosis is greater than 50 (indicating a positive diagnosis).

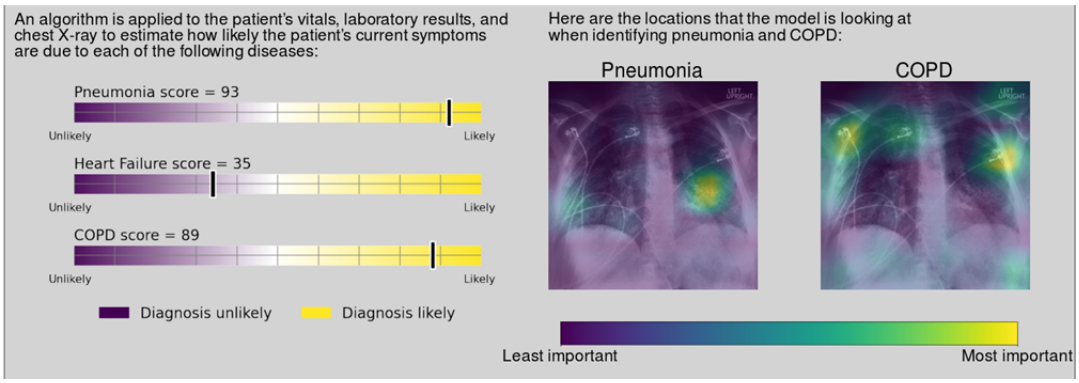

Figure 3. Model scores each disease and corresponding explanations when the model predicts a positive diagnosis.

6.2 Summary of findings from non-clinical studies.

These models were not tested in non-clinical settings.

6.3 Summary of findings from clinical studies.

The AI model developed in this study was based on prior work by Jabbour et *al*.[20] The study trained machine learning models to predict pneumonia, heart failure, and COPD using chest radiographs and clinical data from the electronic health record, and applied the models to an internal cohort at Michigan Medicine and an external cohort from Beth

Israel Deaconess Medical Center. They showed that a model combining chest radiographs and EHR data outperformed models based on each data modality alone and can accurately differentiate between common causes of acute respiratory failure.

However, AI models trained on clinical data are known to make biased predictions due to spurious correlations present in training data.[13] For example, deep learning models can learn to predict patient age, sex, or BMI based on chest X-rays alone. If trained using standard approaches, these models could use these features in their predictions. To date, computational approaches can only partially mitigate the use of these shortcuts, and might fail in settings where the shortcut is not known in advance. However, heatmaps that highlight the regions of a chest X-ray that a model focuses on might signal that a model is taking a shortcut, such as highlighting the features of patient age (e.g., osteoporosis) instead of clinically relevant features.

Figure 4. Left: A model highlighting the features of patient age rather than clinically relevant features of disease. Right the model highlighting clinically relevant features in lungs.

6.4 Summary of known and potential risks and benefits

Because this is a web-based survey study, there are minimal potential risks to patients and participants and no adverse events are expected during this study. The participant taking the survey is free to exit the survey at any point.

6.5 Description and justification of route administration and dosage

Not applicable

6.6 Dosages, dosage modifications and method of administration

Not applicable

6.7 Preparation and labelling of Investigational Medicinal Product

Not applicable

6.8 Drug accountability

Not applicable

7. METHODS

7.1 STUDY PARAMETERS/ENDPOINTS

7.1.1 Main study parameters/endpoints

We compare participants when presented with AI model predictions to participants when presented with AI model predictions and explanations.

- Diagnostic accuracy for pneumonia, heart failure, and COPD
- Treatment accuracy in selecting antibiotics, diuretics, and steroids.

7.2 Randomization, blinding, and treatment allocation

Participants were randomly shown 9 clinical vignettes. The first two vignettes are not accompanied by AI model predictions and are used to estimate baseline participant diagnostic accuracy. The next 6 vignettes include AI model predictions, but some participants are randomized to also see model explanations. These 6 vignettes include 3 vignettes with standard model predictions and 3 with systematically biased AI model predictions. Participants are randomized to see one of three types of systematically biased AI model predictions, and the ordering of the 3 standard and 3 systematically biased model predictions was randomized. In the final vignette, all participants are provided a clinical consult, which is a short narrative provided by a hypothetical trusted colleague, who describes the rationale behind which diagnoses are most likely and what treatment plan they recommend. By design, the clinical consult always provided the correct diagnosis and appropriate treatment plan to provide a realistic upper bound of participant diagnostic accuracy.

7.3 Study procedures

Study population:

Hospitalist physicians, nurse practitioners, and physician assistants who commonly care for patients with acute respiratory failure.

Data collection:

Data collection will occur through Qualtrics, as it is approved for HIPAA data storage. Data collection starts when the participant clicks the survey link and stops when the participant either completes the survey, or two weeks after they exit the survey. This allows the participant to re-enter the survey to continue working on it. The survey data is saved and anonymized. It will be extracted and stored on HIPAA-aligned servers only accessible by the study team named at the University of Michigan. To preserve anonymity, participants will be redirected to another survey that is not linked to their responses to collect their contact information for payment purposes.

7.4 Withdrawal of individual subjects

Participants can exit the study at any time for any reason if they wish to do so without any consequences.

7.4.1 Specific criteria for withdrawal (if applicable)

Not applicable

7.5 Replacement of individual subjects after withdrawal

Not applicable

7.6 Follow-up of subjects withdrawn from treatment

Not applicable

7.7 Premature termination of the study

We do not expect any serious adverse events directly related to this study. Therefore, we do not expect to have to terminate this study prematurely.

8. SAFETY REPORTING

8.1 Temporary halt for reasons of subject safety

8.2 AEs, SAEs, SUSARs

8.2.1 Adverse events (AEs)

Adverse events are defined as any undesirable experience occurring to a subject during the study, whether or not considered related to the intervention. All adverse events reported by the participants or observed by the investigator or their staff will be recorded.

8.2.2 Serious adverse events (SAEs)

Due to the nature of this study, which was deemed as minimal risk, we will not be directly working with patients and do not anticipate any SAEs. However, **the investigator will report all SAEs to the sponsor without undue delay after obtaining knowledge of the events.**

8.3 Annual safety report

Not applicable

8.4 Follow-up of adverse events

All AEs will be followed until they have abated, or until a stable situation has been reached.

8.5 Data Safety Monitoring Board (DSMB) / Safety Committee

Because the study was deemed to be minimal risk survey based study, a DSMB was not formed for the study.

9. STATISTICAL ANALYSIS

9.1 Primary study parameters/endpoints

- Diagnostic accuracy for pneumonia, heart failure, and COPD
- Treatment accuracy in selecting antibiotics, diuretics, and steroids.

9.2 Interim analysis (if applicable)

Not applicable

9.3 Statistical analysis plan

The study aims to recruit 400 participants based on a sample size to measure a decrease in accuracy of 25% with the systematically biased AI model compared to baseline and a 10% improvement with the biased AI model with explanations compared to no explanations.

Completed vignettes were included in the analysis regardless of whether a participant completed all 9 vignettes. Diagnostic accuracy and treatment decision accuracy will be compared using a generalized linear mixed-effects models, accounting for the nested data structure of repeated measures and controlling for individual-related variables, where individual diagnostic responses are nested within participants. After fitting the model, to aid in model interpretation, marginal effects and predictive margins will be reported. Statistical analyses will be performed in R. Statistical significance was based on a p-value < 0.05.

10. ETHICAL CONSIDERATIONS

10.1 Regulation statement

This study has been approved by the UM IRB HUM00180745

10.2 Recruitment and consent

We will recruit hospitalist physicians, nurse practitioners, and physician assistants who commonly care for patients with ARF from US hospitals. We will identify hospitalist site champions who will send out an invitation email with the study information to hospitalist clinicians at their respective institutions. Consent will be obtained prior to participant randomization. Specifically, once participants click on the study link, they will be shown a page to screen for their eligibility. If eligible, they will be redirected to an introduction page that informs them about the study. This includes that the study will be completely anonymous, it will take 25 minutes to complete, and that they can stop the study at any time and come back to the point where they leave off. They are also told that some details of the study's purpose will be withheld and that they will receive a $50 Amazon.com gift card upon completion. If they agree to these terms, they can click forward and are then randomized. If not, they can click out of the survey at this, or any other point.

10.3 Objection by minors or incapacitated subjects (if applicable)

Not applicable.

10.4 Benefits and risks assessment, group relatedness

Not applicable.

10.5 Compensation for injury

Not applicable.

10.6 Incentives (if applicable)

Participants who complete the study will receive a $50 amazon gift card.

11. ADMINISTRATIVE ASPECTS, MONITORING AND PUBLICATION

11.1 Handling and storage of data and documents

The data will be handled confidentially. The participant responses will be retrieved from the Qualtrics interface, which is only accessible to a subset of the study team members. The data will then be stored on HIPAA approved servers at the University of Michigan for subsequent analyses. All data will be anonymized. Any publication arising from this study will not contain data that can be traced to a specific participant.

11.2 Monitoring and Quality Assurance

As data is collected throughout the study, it will be downloaded and checked to ensure that the randomization is set for each participant and all data is recorded as expected.

11.3 Public disclosure and publication policy

Results of this study will be submitted for publication in a peer reviewed scientific medical journal.

12. AMENDMENTS

13. REFERENCES

1. 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*. Jun 2020;2(6):e0128. doi:10.1097/cce.0000000000000128

2. 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. *J Hosp Med*. Feb 2013;8(2):76-82. doi:10.1002/jhm.2004

3. HCUP Fast Stats. Healthcare Cost and Utilization Project (HCUP). April 2021. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/faststats/national/inpatientcommondiagnoses.jsp?year1=2018.
4. Roberts E, Ludman AJ, Dworzynski K, et al. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. *BMJ*. Mar 2015;350:h910. doi:10.1136/bmj.h910
5. Lien CT, Gillespie ND, Struthers AD, McMurdo ME. Heart failure in frail elderly patients: diagnostic difficulties, co-morbidities, polypharmacy and treatment dilemmas. *Eur J Heart Fail*. Jan 2002;4(1):91-8. doi:10.1016/s1388-9842(01)00200-8
6. Daniels LB, Clopton P, Bhalla V, et al. How obesity affects the cut-points for B-type natriuretic peptide in the diagnosis of acute heart failure. Results from the Breathing Not Properly Multinational Study. *Am Heart J*. May 2006;151(5):999-1005. doi:10.1016/j.ahj.2005.10.011
7. Levitt JE, Vinayak AG, Gehlbach BK, et al. Diagnostic utility of B-type natriuretic peptide in critically ill patients with pulmonary edema: a prospective cohort study. *Crit Care*. 2008;12(1):R3. doi:10.1186/cc6764
8. Zwaan L, Thijs A, Wagner C, van der Wal G, Timmermans DR. Relating faults in diagnostic reasoning with diagnostic errors and patient harm. *Acad Med*. Feb 2012;87(2):149-56. doi:10.1097/ACM.0b013e31823f71e6
9. National Academies of Sciences EaM. *Improving diagnosis in health care*. National Academies Press; 2015.
10. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*. 2020;26(8):1229-1234.
11. 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.
12. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nature medicine*. 2021;27(5):775-784.
13. Jabbour S, Fouhey D, Kazerooni E, Sjoding MW, Wiens J. Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts. PMLR; 750-782.
14. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*. 2022;4(6):e406-e414.
15. 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.
16. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*. 2021;4(1):31.

17. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*. 2012;19(1):121-127.
18. 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.
19. Bhatt U, Xiang A, Sharma S, et al. Explainable machine learning in deployment. 648-657.
20. 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.