**Title:** Evaluating the performance and potential bias of predictive models for the detection of transthyretin cardiac amyloidosis

**Short Title:** Comparing ML models for ATTR-CM screening

**Authors:**
Jonathan Hourmozdi, MD, MSAI, MA[1,2] Nicholas Easton, MSAI,[1,2] Simon Benigeri, MSAI,[1,2] James D. Thomas, MD,[1,2] Akhil Narang, MD,[1,2] David Ouyang, MD, Grant Duffy, BS,[6] Ike Okwuosa, MD,[1] Adrienne Kline, MD, PhD[2,3] Abel N. Kho, MD,[4,5] Yuan Luo, PhD,[4] Sanjiv J. Shah, MD,[1,2] Faraz S. Ahmad, MD, MS[1,2,4]

[1] Division of Cardiology, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois
[2] Bluhm Cardiovascular Institute Center for AI, Northwestern Medicine, Chicago, Illinois
[3] Division of Cardiac Surgery, Department of Surgery, Northwestern University Feinberg School of Medicine, Chicago, Illinois
[4] Division of Statistics and Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois
[5] Division of General Internal Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois
[6] Department of Cardiology, Cedars-Sinai Medical Center, Los Angeles, California

**Disclosures:** FSA has received research support from Pfizer and Atman Health. The remaining author report no relevant disclosures.
**Funding:** This work was support by the American Heart Association (AHA number 856917).

**Corresponding Author:**
Faraz S. Ahmad, MD MS
676 North Saint Clair Street, Suite 600
Chicago, IL 60611
Office: 312-695-4965; Fax: 312-926-8620
Email: faraz.ahmad@northwestern.edu
X handle: @FarazAhmadMD

**Word Count:** 5,487

## Abstract

**Background:**

Delays in the diagnosis of transthyretin amyloid cardiomyopathy (ATTR-CM) contribute to the significant morbidity of the condition, especially in the era of novel disease-modifying therapies. Screening for ATTR-CM with AI and other algorithms may improve timely diagnosis, but these algorithms have not been directly compared with each other.

**Methods:**

We identified patients treated at an integrated health system from 2010-2022 with biopsy or PYP scan-confirmed ATTR-CM and age -and sex-matched them to controls with HF in a 19:1 ratio to target 5% prevalence. We compared the performance of three publicly available algorithms: a random forest model of claims data, the regression-based Mayo ATTR-CM Score, and a deep-learning echo model (EchoNet-LVH). Bias was measured in the best performing models using standard fairness metrics.

**Results:**

We identified 198 confirmed cases of ATTR-CM from 3924 patients in the analytic cohort that had the necessary structured and imaging data for all three models. In this cohort, 78.9% of the cohort self-identified as White, 8.7% Black, 4.0% Hispanic and 8.4% Other. ATTR-CM prevalence was highest in individuals who identified as Black. The claims-based model performed poorly with an AUC of 0.48. EchoNet-LVH had higher AUC (0.88 vs 0.78, DeLong Test $p < 0.0001$) and average precision (0.61 vs 0.15) compared to the Mayo score. Bias auditing of the top two performing models demonstrated that both models satisfied our fairness criteria

for *equal opportunity* (1.05 for EchoNet-LVH and 0.91 for ATTR-CM Score) among patients who identified as Black.

**Conclusions:**

In external validation using a large, diverse cohort of patients with heart failure, a deep-learning echo-based model to detect ATTR-CM demonstrated best overall performance compared to two other publicly available models. The results of a bias audit suggest that the regression- and echo-based models are unlikely to exacerbate existing health disparities through inequitable distribution of error with respect to self-identified Black race.

**Keywords:** amyloidosis, artificial intelligence, machine learning, heart failure, health care disparities

**Abbreviation List:**

AI = Artificial intelligence:
ML = Machine learning
ATTR-CM = transthyretin amyloid cardiomyopathy
heart failure: HF
Northwestern Enterprise Data Warehouse: NMEDW
pyrophosphate (PYP)
transthoracic echocardiogram (TTE)
AUC: area under the curve
AP: average precision
LVH: left ventricular hypertrophy
LVEF: left ventricular ejection fraction

## Introduction

Using artificial intelligence (AI) and other machine-learning (ML) algorithms for the early detection or correcting misdiagnosis of rare diseases represents a major area of research and development.[1] This area of research is particularly important for patients with transthyretin amyloid cardiomyopathy (ATTR-CM), a highly morbid condition that is increasingly recognized as an important cause of heart failure (HF) and disproportionately impacts Black communities.[7] While the true prevalence of cardiac amyloidosis is unknown, recent data suggest that it is much more common than previously thought, comprising up to 17% of HF with preserved ejection fraction cases.[8] Despite early clues to diagnosis, delays in the diagnosis of ATTR-CM are extremely common, with a median time from first hospitalization to diagnosis of 2 to 3 years and a recent study demonstrating that 42% of cases were diagnosed over 4 years after the development of cardiac symptoms.[16] Even prior to the development of clinical HF symptoms, subclinical changes in conduction, structure, and function often can be observed in cardiac testing, including routine electrocardiography and echocardiography.[12–15]

The emergence of transthyretin stabilizers, which can halt disease progression and dramatically improve morbidity for patients with ATTR-CM, as well as a pipeline of other disease-modifying therapies, have galvanized efforts to improve early and accurate diagnosis of ATTR-CM, including the use of AI and other algorithms for screening.[17] In order to better understand the potential implications of implementing algorithmic screening technologies for ATTR-CM at a large integrated health system, we studied the performance of three, publicly-available models for the identification of cardiac amyloidosis in a retrospective cohort of patients treated across large health system with 11 hospitals and over 200 sites. In addition to overall diagnostic

performance, we assessed the performance of each model with respect to selected fairness metrics to identify any potential risks to currently disadvantaged groups owing to possible algorithmic bias.

## Methods

The analytical code supporting the findings of the study will be made publicly available on GitHub. The data for this study will be made available by the corresponding author upon reasonable request and a data use agreement. The Northwestern University Institutional Review Board approved this study. This paper adheres to the TRIPOD-AI guidance for reporting of clinical prediction models (**See Supplemental Checklist**).[18]

### *Data Source and Participants*

To identify patients with ATTR-CM, we first searched the Northwestern Enterprise Data Warehouse (NMEDW) for patients with at least 1 ICD code for amyloidosis, a pyrophosphate (PYP) scan, or medication order for Tafamidis from 1/1/2010 to 12/31/2022. For these patients, we performed a review of their records to confirm diagnosis of ATTR-CM, defined as either positive endomyocardial biopsy or PYP scan with a final reading of “strongly suggestive” and absence of serum evidence of amyloid light chain (AL) amyloidosis or negative bone marrow biopsy result. For subset of patients with potential ATTR-CM based on our screening criteria but without available definitive testing in the NMEDW, we used data from the Northwestern Medicine Cardiac Amyloidosis Program registry (August 2018 to December 2022), which included the results of endomyocardial biopsies or PYP scans performed outside of the Northwestern Medicine system.

We then created a matched cohort of HF controls with at least one encounter ICD code for HF over the study period. We excluded potential cases of amyloidosis from this matched cohort, defined as at least one ICD code for amyloidosis and did not have definitive negative testing available, specifically a negative endomyocardial biopsy result or a PYP scan with a final reading of "Not Suggestive." All cases and controls were required to have a transthoracic echocardiogram (TTE) up to six months before or any time after definitive diagnosis by positive endomyocardial biopsy or PYP scan. The NMEDW query required that these TTEs also reported measurements for left ventricular ejection fraction, posterior wall thickness, and relative wall thickness in a format available for automated data extraction and inclusion in the Mayo ATTR-CM Score model. If controls had multiple TTEs, the one nearest to the first HF diagnosis code was used for our model evaluation. Using the PsmPy package in Python,[19] we matched controls to cases by age, sex, and year that TTE was obtained in a 19:1 ratio to target a prevalence of ATTR-CM of 5%, which was selected to mirror estimates in the HF population and reflect its potential clinical use as screening tool in a broad population.[19] We then excluded patients in which EchoNet-LVH could not be run due to clip length, image quality or other factors.

*Cardiac Amyloidosis Risk Models*

We compared the performance of three publicly available risk models with different model inputs and model architectures **(Table 1)**. The Huda et al. model[2] was developed to predict risk of wild-type ATTR-CM using databases of claims data and was validated on claims data and EHR data from patients with ICD codes for amyloidosis (including for codes beyond the specific code for wild-type ATTR). The Huda et al. model uses tabular, ICD code data in a random forest

model. The Mayo ATTR-CM Score was developed by Davies et al.[3] and uses tabular data that includes demographics (age, sex), history of hypertension (based on ICD codes for this analysis), and TTE measurements (left ventricular ejection fraction, wall thickness). The Mayo ATTR-CM Score is an integer-based score created from a linear regression model. Duffy et al.[4] developed the EchoNet-LVH model, which is a computer vision model that uses a convolutional neural network architecture. EchoNet-LVH generates a prediction for likelihood of cardiac amyloidosis (either TTR or AL) from an echocardiogram study synthesizing information about wall thickness from the parasternal long axis view and echotexture from the apical four chamber view.

*Statistical Analysis*

We used standard descriptive statistics to summarize baseline variables with categorical variables reported as counts and proportions and continuous variables reported as mean and standard deviation (SD). Comparisons of demographic and baseline clinical features between cases and controls were performed using Welch's T-test for continuous variables and Chi-squared testing for categorical variables. For each model, we report model discrimination with receiver operating characteristics reported as area under the curve (AUC). We also report the area under the precision-recall curve as average precision (AP), a measure which may be more informative in datasets with class imbalance.[20] Pairwise comparisons of model performance in terms of AUC were performed using the DeLong test.[21] In addition, we report F1 score, accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and false negative rate for all three models based on thresholds for each model based on the publications or recommendations from the model developers (positive for greater than or equal to: 0.5 for the Huda et al. model, 0.8 for EchoNet-LVH, and 6 for the Mayo ATTR-CM Score).

*Sensitivity and Fairness and Bias Analyses*

We conducted several sensitivity analyses to further examine the algorithmic performance under various clinical subpopulations. For the clinical subgroup analyses, we first examined subgroups of patients with increased left ventricular wall thickness, reduced or preserved left ventricular ejection fraction (LVEF), demographic groups including age, sex, and race. In conducting our fairness and bias analysis, we selected three metrics to address the three fundamental fairness principles across a protected attribute. A protected attribute refers to a feature or characteristic of a patient that should not be used as the basis of decision-making and should be generally protected from discrimination. These attributes can be defined according to anti-discrimination law or in accordance with institutional values.[5] We chose the protected attribute of self-identified race or ethnicity given the previously described existing disparities in diagnosis of ATTR-CM for racial and ethnic minorities in the United States. From previously published fairness criteria,[5] we selected the following three to report in our primary analysis: *Demographic parity*, defined as the ratio of the proportion of patients predicted positive by the model between groups, *Predictive parity*, defined as the ratio of positive predictive value (or precision) of the model output between groups, and *Equal opportunity*, defined as the ratio of false negative error rate between groups. We chose to prioritize the equal opportunity fairness criteria in weighing the risk of harms due to the current imbalance in diagnostic error owing to under- or missed diagnosis in US Black individuals. All analyses were performed using the *Aequitas* package in Python 3.12.[22] The default fairness threshold of 0.2 was used to determine whether significant unfairness was present for each metric based on the “80% rule” of disparate impact.[22]

## Results

*Demographic Characteristics*

240 cases of confirmed ATTR-CM were identified by NMEDW query and matched to 4560 patients for a total of 4800 patients. 876 patients (42 cases) were excluded as EchoNet-LVH was unable to generate predictions due to availability of echocardiographic studies including insufficient clip length, low image quality, insufficient videos by view classification, or other technical factors, resulting in a final analytic cohort size of 3924 patients containing 198 confirmed cases of ATTR-CM (**Figure 1**). In this cohort, mean age was 78.8 years (SD 9.5 years), 21.2% were female, 78.9% self-identified as White, 8.7% self-identified as Black, 4.0% self-identified as Hispanic, and 8.4% self-identified as Other or did not report their race or ethnicity. Complete demographic and baseline clinical characteristics for case and control cohorts are shown in **Table 2**. As compared to age and sex-matched controls with HF, the cohort of patients with ATTR-CM had fewer patients who identified as White (58.1% vs. 80.0%) and more left ventricular hypertrophy (LVH) as quantified by their baseline TTE (mean posterior wall thickness of 15.5 mm vs. 11.1 mm). The demographics and clinical characteristics observed in the subgroup of patients for whom the EchoNet-LVH model was unable to generate a prediction are provided in **Supplemental Table 1**.

*Overall Model Performance and Comparison*

In terms of overall performance, EchoNet-LVH performed the best, with an AUC of 0.88 (95% CI, 0.84 to 0.90, $p < 0.0001$) and AP of 0.61. The ATTR-CM Score achieved an AUC of 0.78 (95% CI, 0.75 to 0.81) with a much lower AP of 0.15. The Huda et al. model failed to achieve

meaningful discriminatory performance with an overall AUC of 0.48 (95% CI, 0.44 to 0.52, **Figure 2)**. Similar sensitivities of 0.57 and 0.63 were achieved by both the EchoNet-LVH model and Mayo ATTR-CM Score respectively, with EchoNet-LVH having higher specificity of 0.98 vs 0.85 for the Mayo Score. A complete comparison of performance metrics for the three models can be found in **Table 3**.

From the initial analysis of the primary model performance metrics, it was apparent that the performance of the Huda et al. model was poor compared to the other two models. The AUC of 0.48 in the analytic cohort was much lower than the model's performance in their initial external validation cohorts.[2] To determine whether insufficient number of diagnosis codes was contributing to this decreased performance, we reassessed the Huda et al. model on the subset of patients in the overall analytic cohort for whom at least 50 encounters were available (n=857, prevalence = 12.7%). There was no significant improvement in the overall model performance (AUC 0.53, 95% CI, 0.47 to 0.59, **Supplemental Table 2**). Because of this, we removed the Huda et al. model from further analyses and focused on the EchoNet-LVH and Mayo ATTR-CM Score models.

*Clinical Subgroup Analyses*

In the subgroup of patients with at least mild LVH (n=3009), the prevalence of ATTR-CM was slightly higher than in the overall cohort (6.4% vs 5%). Both the ATTR-CM Score model and EchoNet-LVH performed similarly in this subgroup as compared to their overall performance. In the subgroup of patients with a reduced LVEF of lower than 50% (n=1427, prevalence = 5.5%),

similar performance was noted for both models as well. The results of these subgroup analyses are summarized in **Table 4**.

*Fairness and Bias Audit*

Prevalence of ATTR-CM was highest in the group of patients who self-identified as Black (19% compared to 3-5% in the other racial and ethnic groups). With respect to the protected attribute of race, both models met the fairness criteria for all chosen parity metrics in patients who self-identified their race as Black **(Figure 3)**. In the group of patients who identified as Hispanic, the EchoNet-LVH model did not meet fairness criteria for demographic parity at 0.71. In the group that identified as Other or did not self-identify race or ethnicity, the Mayo ATTR-CM Score did not meet fairness criteria for predictive parity (0.73) or equal opportunity (1.22). A summary of model performance with respect to these chosen fairness criteria is also given in **Table 5**. A complete comparison of all absolute metrics by protected attribute can be found in **Supplemental Figures 2 and 3**.

## Discussion

In this study, we evaluated the potential screening performance of three publicly available models to detect ATTR-CM in a large, integrated health system. We found that the Huda et al. model performed poorly compared to their previously published validation data and the other models. The Mayo ATTR-CM Score and EchoNet-LVH models demonstrated similar performance in our validation cohort as compared to their initially reported results, including across key subgroups as shown in the bias audit. To our knowledge, this is the first comparison of three algorithmic approaches to screening for cardiac amyloidosis using different data inputs

(diagnosis codes, structured clinical data, and echo images) and models (regression, random forest, and convolutional neural network). This study also provides an example of how to apply the recently published TRIPOD-AI reporting guidance and bias and fairness metrics to external validation studies.

Comparing a regression-based, simple score against a more complex, computer vision model in our patient population provides realistic estimates for predictive value. The deep learning vision-based EchoNet-LVH performed better compared to Mayo ATTR-CM Score. In a derived cohort where the prevalence of ATTR-CM was 5%, this trade-off manifested in an absolute improvement in the positive predictive value of 0.47 (0.67 vs. 0.20 for EchoNet-LVH vs Mayo ATTR-CM Score). Our study also highlighted the technical challenge of AI model portability. The EchoNet-LVH model did not generate outputs on approximately 20% of studies, including 42 of 240 ATTR-CM cases in our cohort. Thus, the sensitivity and other related performance metrics for the algorithm may be lower in real-world practice than estimated by our study. Similarly, the data required to run the Mayo ATTR-CM Score likely does not exist in a structured form in many EHR data repositories, thus limiting its ability to be run at scale in a health system. The poor performance of the Huda et al. model underscores the potential limitations of training models using diagnosis codes alone, as demonstrated in another cardiac amyloidosis study.[23]

In addition to reporting similar performance across key clinical subgroups, we demonstrated that both models perform well across key subgroups at higher risk for adverse outcomes. With any ML model, there is a risk that they can learn to exploit patterns in the training data that represent

or perpetuate bias, including systemic and structural racism. These risks are exacerbated when there is underrepresentation of minority patient groups in training cohorts (**Supplemental Table 3**). In the United States, patients who identify as Black disproportionately bear the burden of missed and delayed diagnosis in ATTR-CM, leading to worse outcomes.[7] This is despite a higher prevalence of ATTR-CM in many Black communities in the US; in our cohort, the relative prevalence of ATTR-CM was about 5-fold higher for Black patients compared to all other racial groups.

Such context is instrumental in interpreting fairness metrics related to any screening AI model, as there is a risk not only of differential predictive performance, but also of a differential distribution of error rate. These findings support that there is a low risk of harm due to algorithmic bias from adopting either the Mayo ATTR-CM Score or EchoNet-LVH models for screening in this integrated health system. However, a limitation of any race-based analysis is that self-identified race and ethnicity represents a social construct. Concepts of race and ethnicity vary with respect to time and geography, and the results of our bias audit, while likely generalizable to our institution's population of patients with HF, should not be taken to represent a blanket determination of fairness with respect to race in all settings. Instead, we encourage institutions and constituents to partner with health equity researchers and consider similar methodology when evaluating models for potential deployment across a health system.

## Conclusion

Machine learning techniques have shown promise when applied to imaging and EHR data to enhance the early detection or correct the misdiagnosis of diseases, such as cardiac amyloidosis.

The use of such AI-assisted diagnostic tools in ATTR-CM could improve racial and ethnic disparities through the early identification and uptake of novel therapies in these patients. However, despite pioneering research and significant resources dedicated to this task, few algorithms have had a meaningful impact on the care of these patients in practice. The reasons for this are multi-factorial and include the lack of prospective, high quality evidence and consensus recommendations to support at scale implementation, barriers related to costs, infrastructure, and required personnel, and evolving regulatory guidance over the deployment and governance of AI and ML algorithms. Future work on algorithmic approaches for screening for cardiac amyloidosis should include randomized, effectiveness-implementation hybrid trials that leverage health equity conceptual models and frameworks to study both the performance of the algorithmics and the implementation strategies for successful deployment across diverse populations.[24–26]

**References:**


1. Elias P, Jain SS, Poterucha T, et al. Artificial intelligence for cardiovascular care-part 1: Advances: JACC review topic of the week. *J Am Coll Cardiol*. 2024;83(24):2472-2486.

2. Huda A, Castaño A, Niyogi A, et al. A machine learning model for identifying patients at risk for wild-type transthyretin amyloid cardiomyopathy. *Nat Commun*. 2021;12(1):2725.

3. Davies DR, Redfield MM, Scott CG, et al. A Simple Score to Identify Increased Risk of Transthyretin Amyloid Cardiomyopathy in Heart Failure With Preserved Ejection Fraction. *JAMA Cardiol*. 2022;7(10):1036-1044.

4. Duffy G, Cheng PP, Yuan N, et al. High-Throughput Precision Phenotyping of Left Ventricular Hypertrophy With Cardiovascular Deep Learning. *JAMA Cardiol*. 2022;7(4):386-395.

5. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7(6):719-742.

6. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872.

7. Spencer-Bonilla G, Njoroge JN, Pearson K, Witteles RM, Aras MA, Alexander KM. Racial and Ethnic Disparities in Transthyretin Cardiac Amyloidosis. *Curr Cardiovasc Risk Rep*. 2021;15(6). doi:10.1007/s12170-021-00670-y

8. Kittleson MM, Maurer MS, Ambardekar AV, et al. Cardiac Amyloidosis: Evolving Diagnosis and Management: A Scientific Statement From the American Heart Association. *Circulation*. 2020;142(1):e7-e22.

9. Grogan M, Scott CG, Kyle RA, et al. Natural History of Wild-Type Transthyretin Cardiac Amyloidosis and Risk Stratification Using a Novel Staging System. *J Am Coll Cardiol*. 2016;68(10):1014-1020.

10. Witteles RM, Bokhari S, Damy T, et al. Screening for Transthyretin Amyloid Cardiomyopathy in Everyday Practice. *JACC Heart Fail*. 2019;7(8):709-716.

11. Ruberg FL, Grogan M, Hanna M, Kelly JW, Maurer MS. Transthyretin Amyloid Cardiomyopathy: JACC State-of-the-Art Review. *J Am Coll Cardiol*. 2019;73(22):2872-2891.

12. Bhogal S, Ladia V, Sitwala P, et al. Cardiac Amyloidosis: An Updated Review With Emphasis on Diagnosis and Future Directions. *Curr Probl Cardiol*. 2018;43(1):10-34.

13. Maurer MS, Bokhari S, Damy T, et al. Expert Consensus Recommendations for the Suspicion and Diagnosis of Transthyretin Cardiac Amyloidosis. *Circ Heart Fail*. 2019;12(9):e006075.

14. Phelan D, Collier P, Thavendiranathan P, et al. Relative apical sparing of longitudinal strain using two-dimensional speckle-tracking echocardiography is both sensitive and specific for the diagnosis of cardiac amyloidosis. *Heart*. 2012;98(19):1442-1448.

15. Sinha A, Zheng Y, Nannini D, et al. Association of the V122I Transthyretin Amyloidosis Genetic Variant With Cardiac Structure and Function in Middle-aged Black Adults: Coronary Artery Risk Development in Young Adults (CARDIA) Study. *JAMA Cardiol*. 2020;6(6):1-5.

16. Lane T, Fontana M, Martinez-Naharro A, et al. Natural History, Quality of Life, and Outcome in Cardiac Transthyretin Amyloidosis. *Circulation*. 2019;140(1):16-26.

17. Donnelly JP, Hanna M. Cardiac amyloidosis: An update on diagnosis and treatment. *Cleve Clin J Med*. 2017;84(12 Suppl 3):12-26.

18. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.

19. Kline A, Luo Y. PsmPy: A Package for Retrospective Cohort Matching in Python. *Annu Int Conf IEEE Eng Med Biol Soc*. 2022;2022:1354-1357.

20. Ozenne B, Subtil F, Maucort-Boulch D. The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol*. 2015;68(8):855-859.

21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

22. Saleiro P, Kuester B, Hinkson L, et al. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv [csLG]*. Published online November 14, 2018. http://arxiv.org/abs/1811.05577

23. Vrudhula Amey, Stern Lily, Cheng Paul C., et al. Impact of Case and Control Selection on Training Artificial Intelligence Screening of Cardiac Amyloidosis. *JACC: Advances*. 0(0). doi:10.1016/j.jacadv.2024.100998

24. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care*. 2012;50(3):217-226.

25. Woodward EN, Matthieu MM, Uchendu US, Rogal S, Kirchner JE. The health equity implementation framework: proposal and preliminary study of hepatitis C virus treatment. *Implement Sci*. 2019;14(1):26.

26. Shelton RC, Chambers DA, Glasgow RE. An Extension of RE-AIM to Enhance Sustainability: Addressing Dynamic Context and Promoting Health Equity Over Time. *Front Public Health*. 2020;8:134.

**Figure Legends**

**Figure 1: Flow diagram for analytic cohort creation. T**his figure shows the creation of the analytical dataset. NMEDW denotes Northwestern Medicine Enterprise Database Warehouse; ICD, international classification of diseases; TTE, transthoracic echocardiogram; PYP, pyrophosphate scan; ATTR-CM, transthyretin amyloid cardiomyopathy.

**Figure 2: Comparison of the model performance between EchoNet-LVH, Mayo ATTR-CM, and the Huda et al. models.** The top panel shows receiver operator characteristic curves and  bottom panel shows the precision-recall curves (bottom) with bootstrap confidence intervals

**Figure 3: Parity metric treemaps for Mayo ATTR-CM Score model (A) and EchoNet-LVH (B).** Selected fairness metrics are expressed as a ratio of protected group: reference group. *Demographic parity* refers to positive prediction rate, *predictive parity* refers to positive predictive value (precision), and *equal opportunity* refers to false negative rate. Treemap squares are sized according to size of group in the cohort and colored red if results suggest disparity according to the "80% rule" of disparate impact.* *denotes statistical significance of finding based on unadjusted alpha < 0.05.*

**Table 1:** Description of the evaluated machine-learning models for detection of cardiac amyloidosis.

| Author (year) | Model Name | Model Architecture | Input Features | Outputs | Reported AUC* |
|---|---|---|---|---|---|
| Huda et al. (2021) | N/A | Random Forest | Medical Claims | Probability (0-1) | 0.80 |
| Davies et al. (2022) | Mayo ATTR-CM Score | Logistic Regression | Clinical variables | Risk score (0-10) | 0.84 |
| Duffy et al. (2022) | EchoNet-LVH | ResNet3D | Echo Images | Probability (0-1) | 0.83 |

AUC denotes area under the receiver-operating characteristic curve
*Values are from initial published performance in external validation cohorts.

**Table 2:** Demographic characteristics of the validation cohort.

| | ATTR-CM (n=198) | HF Controls (n=3726) | p-value | Overall (n=3924) |
|---|---|---|---|---|
| Age, mean (SD) | 78.3 (9.2) | 78.8 (9.5) | 0.442 | 78.8 (9.5) |
| Sex, n (%) | | | 0.648 | |
| Female | 39 (19.7%) | 794 (21.3%) | | 833 (21.2) |
| Male | 159 (80.3%) | 2932 (78.7%) | | 3091 (78.8) |
| Race/Ethnicity, n (%) | | | <0.0001 | |
| Black | 65 (32.8%) | 277 (7.4%) | | 342 (8.7) |
| Hispanic | 7 (3.5%) | 148 (3.5%) | | 155 (4.0) |
| White | 115 (58.1%) | 2981 (80.0%) | | 3096 (78.9) |
| Other/Unknown | 11 (5.6%) | 320 (8.6%) | | 331 (8.4) |
| LVEF | | | 0.96 | |
| Median (IQR) | 53.3 (19.4) | 55.0 (22.3) | | 55.0 (22.2) |
| < 50%, n (%) | 79 (39.9%) | 1348 (36.2%) | | 1427 (36.4%) |
| Posterior Wall Thickness | | | <0.0001 | |
| Mean, (SD) | 15.5 (4.0) | 11.1 (2.2) | | 11.4 (2.5) |
| >= 12mm, n (%) | 170 (85.9%) | 1292 (34.7%) | | 1462 (37.3%) |
| Relative Wall Thickness | | | <0.0001 | |
| Mean (SD) | 0.77 (0.45) | 0.47 (0.15) | | 0.49 (0.19) |
| > 0.57 (%) | 153 (77.3%) | 757 (20.3%) | | 910 (23.2%) |
| Hypertension, n (%) | 156 (78.8%) | 2971 (79.7%) | 0.74 | 3127 (79.7%) |

ATTR-CM denotes transthyretin amyloid cardiomyopathy cases; HF, heart failure, LVEF, left ventricular ejection fraction, SD, standard deviation, IQR, interquartile range.

**Table 3:** Overall performance metrics for the three compared models.

| | F1 | Acc | Sens | Spec | PPV | NPV | FNR | AP | AUC |
|---|---|---|---|---|---|---|---|---|---|
| Huda et al. | 0.09 | 0.66 | 0.32 | 0.67 | 0.05 | 0.95 | 0.68 | 0.05 | 0.48 |
| Mayo ATTR-CM Score | 0.29 | 0.84 | 0.63 | 0.85 | 0.18 | 0.98 | 0.37 | 0.15 | 0.78 |
| EchoNet-LVH | 0.61 | 0.96 | 0.57 | 0.98 | 0.65 | 0.98 | 0.43 | 0.61 | 0.88 |

F1 denotes F1-score; Acc, accuracy; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; FNR, false negative rate; AP, average precision; AUC, area under the receiver-operating characteristic curve.

**Table 4:** Clinical **s**ubgroup analyses for the Mayo ATTR-CM Score and EchoNet-LVH.

| | **F1** | **Acc** | **Sens** | **Spec** | **PPV** | **NPV** | **FNR** | **AP** | **AUC** |
|---|---|---|---|---|---|---|---|---|---|
| **Mayo ATTR-CM Score** | | | | | | | | | |
| Overall | 0.29 | 0.84 | 0.63 | 0.85 | 0.18 | 0.98 | 0.37 | 0.15 | 0.78 |
| LVH present | 0.29 | 0.79 | 0.65 | 0.80 | 0.18 | 0.97 | 0.35 | 0.17 | 0.76 |
| LVEF < 50% | 0.34 | 0.86 | 0.65 | 0.88 | 0.23 | 0.98 | 0.35 | 0.19 | 0.78 |
| **EchoNet-LVH** | | | | | | | | | |
| Overall | 0.61 | 0.96 | 0.57 | 0.98 | 0.65 | 0.98 | 0.43 | 0.61 | 0.88 |
| LVH present | 0.62 | 0.96 | 0.58 | 0.98 | 0.67 | 0.97 | 0.42 | 0.64 | 0.88 |
| LVEF < 50% | 0.65 | 0.96 | 0.65 | 0.98 | 0.65 | 0.98 | 0.35 | 0.68 | 0.88 |

F1 denotes F1-score; Acc, accuracy; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; FNR, false negative rate; AP, average precision; AUC, area under the receiver-operating characteristic curve.

**Table 5:** Summary of model performance across chosen fairness metrics stratified by protected attribute.

| Model | Group | Prev | Sens | Spec | Demographic Parity | Predictive Parity | Equal Opportunity |
|---|---|---|---|---|---|---|---|
| **Mayo ATTR-CM Score** | White (n=3096) | 0.04 | 0.63 | 0.85 | ref | ref | ref |
| | Black (n=331) | 0.19 | 0.66 | 0.84 | 1.55 | 3.48 | 0.91* |
| | Hispanic (n=155) | 0.05 | 0.57 | 0.84 | 1.06 | 1.04 | 1.15 |
| | Other (n=342) | 0.03 | 0.55 | 0.84 | 1.07 | 0.73† | 1.22† |
| **EchoNet-LVH** | White | 0.04 | 0.57 | 0.98 | ref | ref | ref |
| | Black | 0.19 | 0.55 | 0.96 | 3.72 | 1.33 | 1.05* |
| | Hispanic | 0.05 | 0.57 | 1.00 | 0.71† | 1.70 | 1.01 |
| | Other | 0.03 | 0.64 | 0.98 | 1.09 | 0.91 | 0.85 |

Selected fairness metrics are expressed as a ratio of protected group: reference group. *Demographic parity* refers to positive prediction rate, *predictive parity* refers to positive predictive value (precision), and *equalized opportunity* refers to false negative rate.
* denotes statistical significance with an unadjusted alpha of <0.05 for parity or disparity.
† denotes disparity according to the "80% rule" of disparate impact.

**Figure 1:** Flow diagram for analytic cohort creation.

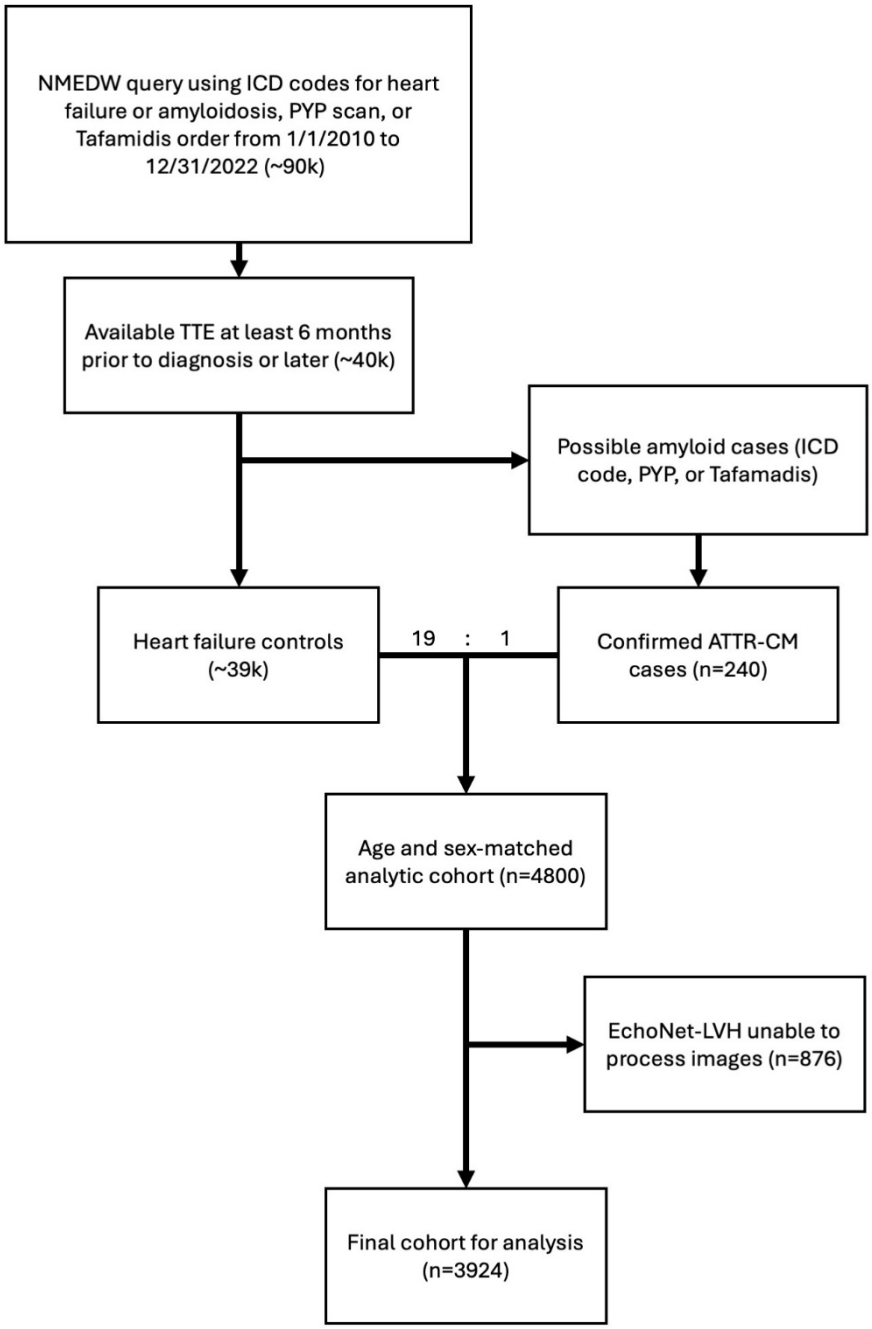

**Figure 2:** Comparison of the model performance between EchoNet-LVH, Mayo ATTR-CM, and the Huda et al. models

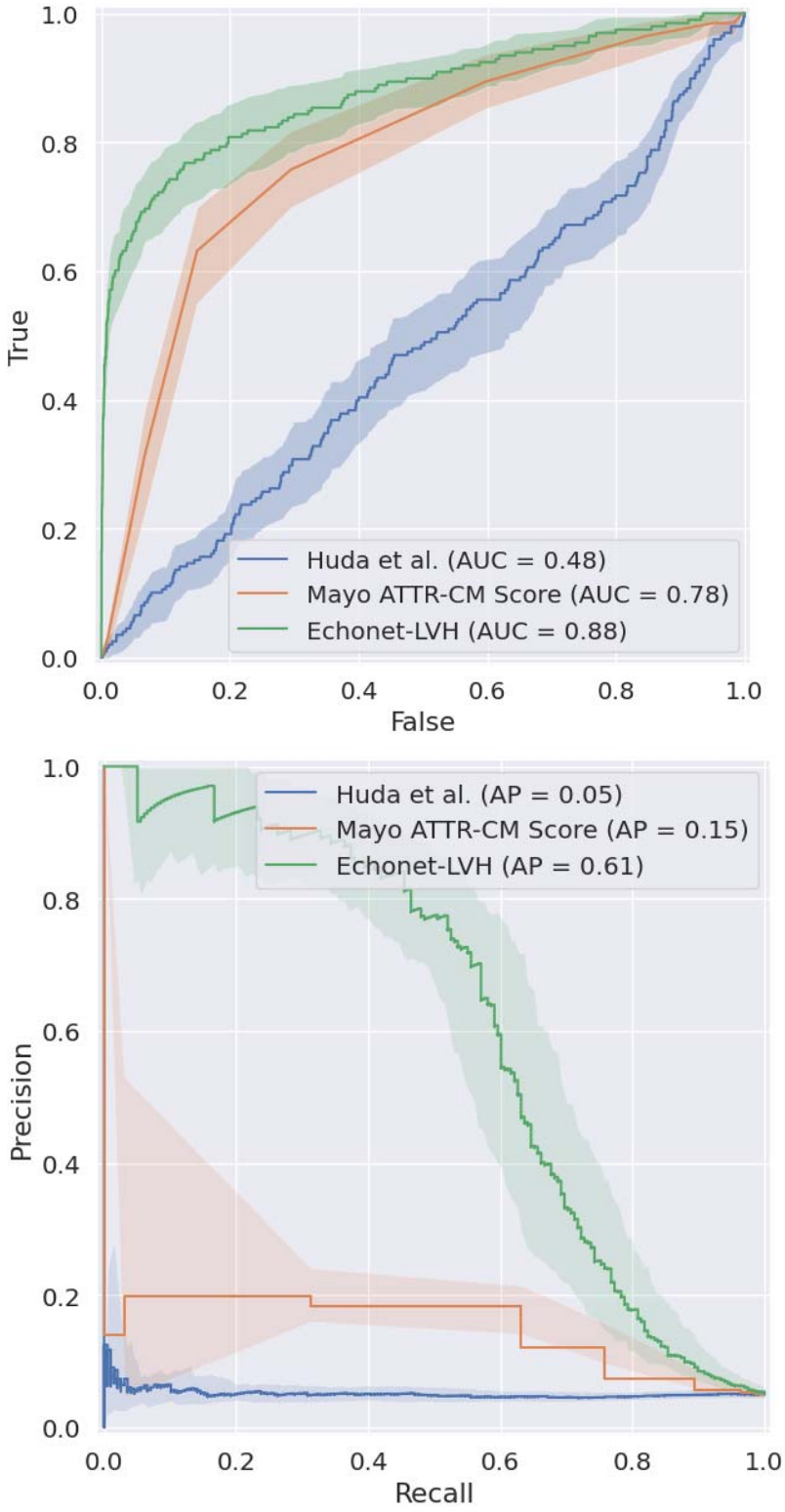
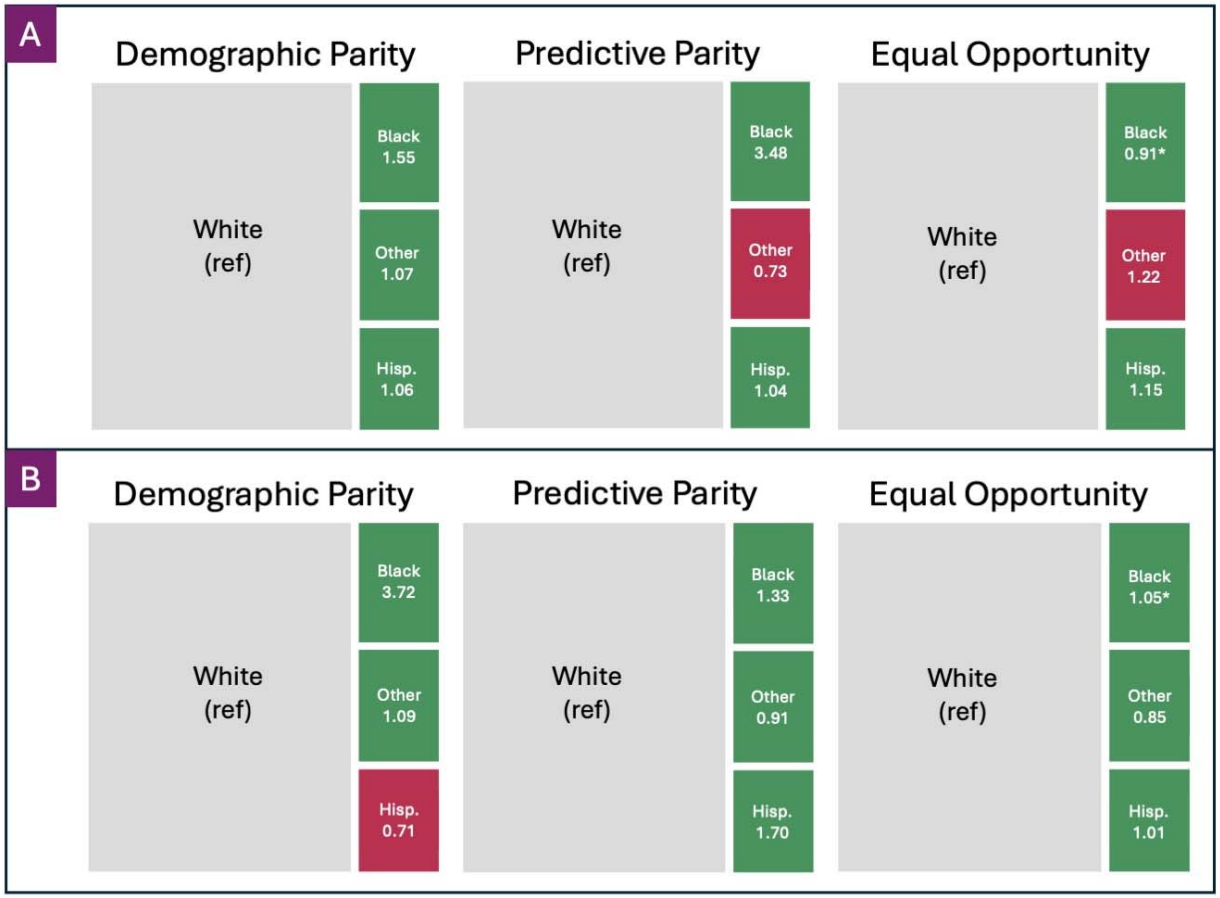
**Figure 3:** Parity metric treemaps for Mayo ATTR-CM Score model (A) and EchoNet-LVH (B).