PW088

A Machine Learning Framework for Predicting Risk of Wild-Type Transthyretin Amyloid Cardiomyopathy

Ahsan Huda[1], Stephen Heitner[2], Veena Calambur[1], Marianna Bruno[1], Jennifer Schumacher[1], Birol Emir[1], Catherine Isherwood[1], Adam Castaño[1]

[1]Pfizer Inc, New York, NY, USA; [2]Oregon Health and Sciences University, Portland, OR, USA

INTRODUCTION

- Wild-type transthyretin amyloid cardiomyopathy (ATTRwt-CM) is a rare, underdiagnosed, and fatal disease that is increasingly recognized as a cause of heart failure (HF).[1,2]
- Despite increasing awareness, most patients with ATTRwt-CM continue to remain undiagnosed because its clinical presentation is similar to that of more common etiologies of HF.[2]
- A previously developed machine learning algorithm used International Classification of Diseases (ICD) codes from medical claims data to identify patients with ATTRwt-CM.[3] Here, this algorithm is transformed into a tool to estimate risk of ATTRwt-CM in hypothetical patient scenarios.

METHODS

Cohort Creation

- Datasets were sourced from IQVIA (medical claims for 300 million US lives, 10 years of diagnostic history) and Optum (electronic health records [EHRs] for 90 million US lives, 10 years of diagnostic history).
- 2 cohorts were created using ICD-10 codes for ATTRwt-CM (Cohort 1A: ATTRwt-CM E.85.82 + HF I50; n=1678) and non-amyloid HF (Cohort 1B: I50 – E85; n=1678).
- Patients were matched using propensity score matching to control for age, gender, and total history available in the data source.

Data Processing and Feature Engineering

- All ICD-10 diagnosis codes available in historical medical claims data were extracted and mapped to phenotypes using the phenome-wide association studies catalog mapping schema.[4]
- Binary flags for the presence/absence of phenotypes in the data were created for each patient and used as potential features.
- A random forest (RF) model was trained with all available phenotypes to ascertain their relative importance.
- Feature space was systematically reduced from ~700 to 11 phenotypes based on feature importance, feature collinearity, and clinical relevance to ATTRwt-CM (**Figure 1**).

Model Training and Validation

- A balanced RF model was trained on 11 selected clinically relevant phenotypes with the propensity-matched ATTRwt-CM to HF cohort (80% train; 20% holdout for testing).
- The model was also validated against the ATTRwt-CM and cardiac amyloidosis (CA) cohorts in medical claims and EHR datasets (**Table 1**).

Probability Adjustments

- Model probabilities were adjusted using a Bayesian approach to incorporate prevalence of ATTRwt-CM in a HF population grouped by age and gender.
- Probabilities of identifying patients with ATTRwt-CM in a HF population were calculated separately for patients grouped by age and gender and combined with model probabilities using a Naïve Bayes approach.
- Bayesian adjustment accounted for an estimated ATTRwt-CM prevalence of ~10% among patients with HF with preserved ejection fraction.[5,6]
- The suspicion index (SI), an additional metric developed to aid the interpretation of probabilities derived from RF models and adjusted through the Naïve Bayes approach, is the ratio of the RF model probability of ATTRwt-CM to the RF model probability of HF due to other causes.
  - SI >1 indicates that the probability of ATTRwt-CM is higher than the probability of HF due to other causes; SI = 1, that it is the same; and SI <1, that it is lower than the probability of HF due to other causes.

RESULTS

- The RF model based on 11 selected features (**Figure 1**) delivered robust performance in classifying patients with ATTRwt-CM and patients with HF (**Table 2**; **Figures 2 and 3**).
  - AUC was 0.82; sensitivity, 77%; specificity, 72%; PPV, 71%; NPV, 78%; and accuracy, 74%.
- The model was internally tested by classifying patients with ATTRwt-CM or CA from patients with HF in 3 additional cohorts derived from medical claims and EHR data, and the model registered robust performance across all cohorts (**Table 2**).
- Model performance on patients with CA was somewhat lower (medical claims 0.70, EHR 0.70), potentially due to the heterogeneity of the CA population.

CONCLUSIONS

- An adapted machine learning model was developed and internally tested. This novel approach allows for physician input with the aim of educating clinicians estimating an empirical probability of ATTRwt-CM.
- This framework could serve as a simple and easily implementable tool to aid clinical assessment of patient risk for ATTRwt-CM.
- Ongoing external validation work will further inform use in clinical practice.

Table 1. Model performance metrics on medical claims and EHR datasets

Stage	Cohort	Data source (data type)	Cases*	Controls*,†
I. Training and internal testing	1	Medical claims data‡	**ATTRwt + HF (n=1678)** • Patients with ATTRwt code (E85.82) + HF code • Training set: 80% of patients • Test set: 20% of patients	**Non-amyloid HF (n=1678)** • Patients with HF code but not amyloidosis code • Training set: 80% of patients • Test set: 20% of patients
II. Validation	2	EHR data§	**ATTRwt + HF (n=280)** • Patients with ATTRwt code (E85.82) + HF code	**Non-amyloid HF (n=258)** • Patients with HF code but not amyloidosis code
	3	Medical claims data‡	**CA + HF (n=8274)** • Patients with organ-limited amyloidosis code (E85.4) + HF code but not BC, HT, LC, ESRD, CAA, or ICH diagnoses	**Non-amyloid HF (n=7844)** • Patients with HF code but not amyloidosis code
	4	EHR data§	**CA + HF (n=2187)** • Patients with organ-limited amyloidosis code (E85.4) + HF code but not BC, HT, LC, ESRD, CAA, or ICH diagnoses	**Non-amyloid HF (n=2034)** • Patients with HF code but not amyloidosis code

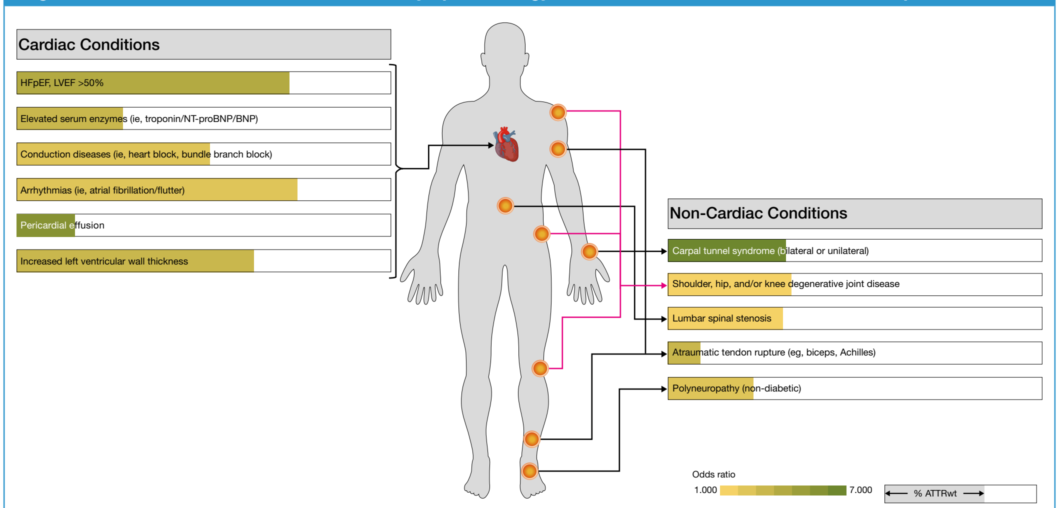
Codes are ICD-10.
*All cases and controls were aged >50 years; †All controls were 1:1 propensity matched by age, gender, and medical history; ‡Date range: 2008–2019; dataset: N>300 million; §Date range: 2008–2018; dataset: N=88 million.
ATTRwt, wild-type transthyretin amyloidosis; BC, blood cancer; CAA, cerebral amyloid angiopathy; ESRD, end-stage renal disease; HT, hypertension; ICH, intracranial hemorrhage; LC, immunoglobulin light chain.

Table 2. Model validation metrics

Cohort type	Accuracy, %	PPV, %	NPV, %	Sensitivity, %	Specificity, %	AUC	Patients with ATTRwt-CM/ CA, n	Patients with HF, n	True positive, n	True negative, n	False positive, n	False negative, n
1. Medical claims: ATTRwt-CM	74	71	78	77	72	0.82	317	355	243	256	99	74
2. EHR: ATTRwt-CM	68	69	68	72	65	0.75	280	258	201	167	91	79
3. Medical claims: CA	62	67	59	51	73	0.66	8274	7844	4242	5737	2107	4032
4. EHR: CA	61	67	58	50	73	0.66	2187	2034	1103	1481	553	1084

AUC, area under the curve; NPV, negative predictive value; PPV, positive predictive value.

Figure 1. Features selected based on clinical symptomatology of ATTRwt-CM and RF model delineated predictive value



BNP, B-type natriuretic peptide; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; LVEF, left ventricular ejection fraction; NT-proBNP, N-terminal pro-B-type natriuretic peptide.

Figure 2. Receiver operating characteristic curve for RF model based on 11 clinical features

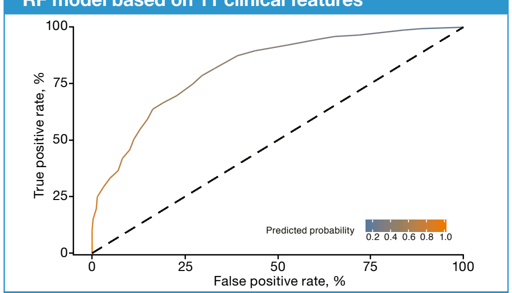
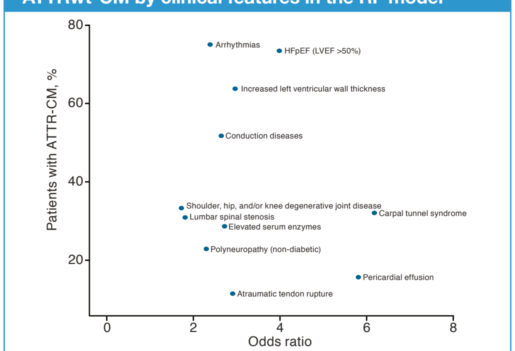


Figure 3. Odds ratio and percentage of patients with ATTRwt-CM by clinical features in the RF model



REFERENCES

1. Rapezzi C, et al. Nat Rev Cardiol 2010;7:398-408. **2.** Ruberg FL, et al. J Am Coll Cardiol 2019;73:2872-91. **3.** Huda A, et al. J Card Fail 2019;25:S53-4. **4.** Wei WQ, et al. PLoS One 2017;12:e0175508. **5.** Mohammed SF, et al. JACC Heart Fail 2014;2: 113-22. **6.** Gonzalez-Lopez E, et al. Eur Heart J 2015;36:2585-94.

DISCLOSURES

This study was sponsored by Pfizer. AH, VC, MB, JS, BE, CI, and AC are full-time employees of Pfizer and may hold stock and/or stock options. SH has received research grants and consulting fees from Akcea, Eidos, Ionis, and Pfizer. Editorial support was provided by Donna McGuire of Engage Scientific Solutions and funded by Pfizer.

Copyright © 2020

Presented at the XVIIth International Symposium on Amyloidosis Online Event, September 14–18, 2020