WestVirginiaUniversity
THE RESEARCH REPOSITORY @ WVU

Graduate Theses, Dissertations, and Problem Reports

2023

A Machine Learning Approach for Early Diagnosis of Transthyretin Amyloid Cardiomyopathy Among Heart Failure Patients

Tanjim Ahmed
West Virginia University, ta00024@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

 Part of the Cardiovascular Diseases Commons, Disease Modeling Commons, and the Industrial Engineering Commons

Recommended Citation

Ahmed, Tanjim, "A Machine Learning Approach for Early Diagnosis of Transthyretin Amyloid Cardiomyopathy Among Heart Failure Patients" (2023). *Graduate Theses, Dissertations, and Problem Reports*. 12028.
https://researchrepository.wvu.edu/etd/12028

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

A Machine Learning Approach for Early Diagnosis of Transthyretin Amyloid Cardiomyopathy Among Heart Failure Patients

Tanjim Ahmed

Thesis submitted to the

College of Engineering and Mineral Resources at

West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science

in Industrial Engineering

Imtiaz Ahmed, Ph.D., Chair

Abdullah Al-Mamun, Ph.D.

Avishek Choudhury, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia

2023

Keywords: ATTR-CM, Diagnosis Delay, Machine Learning, Statistical Analysis.

Copyright 2023 Tanjim Ahmed

ABSTRACT

A Machine Learning Approach for Early Diagnosis of Transthyretin Amyloid Cardiomyopathy Among Heart Failure Patients

Tanjim Ahmed

Transthyretin Amyloid Cardiomyopathy (ATTR-CM) is a rare, progressive, and fatal disease. Prevalence of ATTR-CM ranges from 4 to 17 per 100000 cases where the mean survival time is less than 4 years. It has a history of being underdiagnosed and misdiagnosed. The diagnosis delay has a weighted mean of 6.1 years for wild-type ATTR-CM. Low awareness, the necessity of invasive procedures, and lack of treatment are the key reasons for delayed diagnosis. But, with the introduction of non-invasive tests like nuclear scintigraphy with ^{99m}TC-PYP and the disease modifying drug Tafamidis, the diagnosis delay signifies a missed opportunity to increase life expectancy by early treatment. Studies show that mean life expectancy can be increased by 5.46 years by early treatment if the 6.1 years of diagnosis delay can be eliminated, whereas the current mean survival time is less than 4 years. Though there is no definitive symptom for it, studies have found out some key prognostic flags: symptoms and comorbidities that are co-existent with ATTR-CM. A prediction model can be developed using the electronic health records (EHR) information in hand to diagnose it early and aid to increase the mean life expectancy. This study aims to identify the top phenotypes that can be used for early diagnosis of ATTR-CM and to predict ATTR-CM using machine learning models among heart failure patients. Patient records from North American healthcare organizations were derived from an EHR system 'TrixNetX' for this study. Several statistical analyses (e.g., logistic regression, forward and backward elimination, LASSO, and Survival analysis) were utilized to find out the top diagnostic procedures and comorbidities related with the diagnosis of wild-type ATTR-CM. These key factors were used as features to train machine learning models (e.g., XGBoost, Random Forest) and predict ATTR-CM early among heart failure patients. The study results found the key factors related to diagnosis delay and predicting early cases to improve life expectancy and quality of life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

In medical science, the diseases that affect a small percentage of the population are classified as rare diseases. In the United States (US) the diseases that affect less than 200,000 people and in EU the diseases that affect fewer than 1 in every 2,000 people are classified as rare diseases [1]. There are approximately 7,000 diseases which are classified as rare diseases. Around 400 million people are affected by these diseases worldwide and only 5% of these diseases have approved treatments [2]. Patients with rare diseases often face challenges in obtaining accurate diagnosis. The certain disease conditions may not be familiar to many doctors or clinicians which results in delay in diagnosis or misdiagnosis.

ATTR-CM stands for Transthyretin Amyloid Cardiomyopathy. It is a potentially fatal disease which is caused by transthyretin protein. The liver produces transthyretin which is a transport protein and carries thyroxine hormone and retinol through the bloodstream. Faulty, irregular misfolded protein or fibril clumps build up in the body including the left ventricle walls of the heart. Due to the thickening of left ventricle, the main pumping chamber, the heart fails to relax and fill with blood accurately, and squeeze to pump blood out effectively which can result into heart failure [3].

ATTR-CM is of two types: Hereditary and Wild type. ATTRv-CM (hereditary ATTR-CM) is the rarer of the two and is associated with specific geographical and ethnic groups. It is mostly observed in African-American population and it is present in around 3-4% of the cases. ATTRv-CM occurs due to autosomal dominant mutation of the transthyretin (TTR) gene. The more common of the two types is ATTRwt-CM (wild-type ATTR-CM) which is associated with age related misfolding or wild-type allelic constitution of the TTR gene. Symptoms start from the age 60 for this type and it is predominant in males [4].

Prevalence of ATTR-CM is mostly under-estimated due to non-specific symptoms, phenotypic variability, and limited awareness. Symptoms overlap is one of the main reasons for ATTR-CM being mis-diagnosed. The associated symptoms are close to that of hypertensive heart disease or heart failure with preserved ejection fraction (HFpEF). HFpEF is prevalent in more than half of heart failure patients and almost half of the patients with HFpEF have increased left ventricular wall thickness. Prior studies show that 5% to 17% of the patients with HFpEF had prevalence of

ATTR-CM [5]. The prevalence of ATTR-CM in patients with HFpEF and left ventricular wall thickness of >12mm screened with bone scintigraphy was 13% [6]. The common symptom onset appears with heart failure symptoms or arrhythmias. The disease can cause various non-cardiac symptoms as well, such as: carpal tunnel syndrome, distal biceps tendon rupture, lumbar spinal stenosis, aortic stenosis, atrioventricular block, atrial fibrillation, and intestinal disorders [6] [7].

Another reason contributing to the diagnosis delay of ATTR-CM is the need for invasive diagnosis of the heart tissue by cardiac biopsy. But currently imaging modalities like bone scintigraphy, speckle-tracking echocardiography, and cardiac MRI assist with non-invasive diagnosis of cardiac amyloidosis along with the invasive tests. Also, TCPYP (technetium TC 99m pyrophosphate single photon emission computed tomography) scan attributes to the non-invasive and accurate diagnosis of ATTR-CM.

The unavailability of a disease modifying treatment was another factor to diagnosis delay. But the situation changed in recent years since the US FDA (Food and Drug Administration) and the EMA (European Medicines Agency) approved treatment by Tafamidis was established. Tafamidis is a transthyretin (TTR) stabilizer. Treatment by Tafamidis reduced all-cause mortality compared with a placebo (HR: 0.70 [95%CI: 0.51-0.95]; p= 0.0259) in its Phase III trial. It has also shown promising effects on health-related quality of life (HRQoL) and functional capacity of the patients [4].

ATTR-CM has poor prognosis, and the typical survival form diagnosis is 2-6 years [4] with a median of less than 4 years [7]. Along with gradual decline in HRQoL and functional capacity, ATTR-CM patients have a usually high morbidity, hospitalization, and mortality. With the progression of the disease, patients suffer declining symptoms like fatigue, reduced exercise capacity, dyspnea or shortness of breath and less functional capacity.

Diagnosis delay is greatly observed in ATTR-CM cases and for several years after symptom onset patients remain un-diagnosed or mis-diagnosed. Delayed diagnosis often results in more critical conditions at the time treatment starts and reduces the capacity of the treatment. With the recent use of non-invasive procedures for diagnosis and the use of Tafamidis to modify the disease, diagnostic delay signifies a missed opportunity for early treatment to extend mean survival years and improve quality of life of the patients.

The clinical and experimental studies on ATTR-CM have increased over the last few years focusing on the patient characteristics, co-morbidities, diagnosis delay and potential 'red flags'. There remains a large gap in the knowledge of epidemiology of the disease as most of the research were based on subgroups of the population and covering shorter periods of time. Most of the research studies are accomplished without limited real-world evidence. Artificial intelligence methods such as machine learning models have been used to investigate the key factors related to diagnosis delay and predict ATTR-CM in heart failure patients using patient characteristics, morbidities, and the important red flags for identifying ATTR-CM in comparison with HF patients. [6] [7]

In this study we aim to use statistical methods to find out the important factors like using the procedures, lab tests and morbidities in the real-world patient history. Based on these important patient characteristics, machine learning models can be used to predict the prevalence of ATTR-CM in patients. They can predict future data based on previous characteristics. These data-driven models usually require a large amount of data for training and testing.

For this research we have obtained patient records from TriNetX which is an electronic health record (EHR) system. The dataset contains patient records from 63 health care organizations of North America. We obtained medical records of 2.1 million Heart Failure patients in North America, which is a large dataset for training and testing machine learning models.

The objectives of this research are thus given below:

a) To use statistical and data mining analysis to find out the key procedures and combinations of these which are key variables in predicting ATTR-CM early in heart failure patients.
b) To find out the top morbidities and combination of these for predicting of ATTR-CM among heart Failure patients earlier.
c) To use the combination of all the above-mentioned variables as predictors in machine learning models and predict the risk of ATTR-CM in heart failure patients earlier.

The rest of the thesis has been divided into the following sections. Chapter 2 has highlighted the literature review based on research works related to Prevalence of ATT-CM, Diagnosis delay in ATTR-CM diagnosis and machine learning (ML) based approaches to predict ATTR-CM. The data format and dataset are described in chapter 3. In Chapter 4, the methodologies considered for

this work are described. The proposed model and its subsections are elaborated in Chapter 5. The results are discussed and explained in Chapter 6. Chapter 7 contains a discussion on the findings from this study. Finally, the work is concluded in Chapter 7.

2. LITERATURE REVIEW

This chapter is divided into three sections where the first section is about research works related to prevalence of ATTR-CM. There are various directions where researchers have walked to predict the prevalence of ATTR-CM in patients based on patient characteristics, morbidities, and health records. In the second section, some studies on the diagnosis delay for ATTR-CM is discussed. In the last section, studies that used machine learning approaches to identify predictive features like patient characteristics and morbidities for ATTR-CM are discussed.

2.1 Research Works on Prevalence of ATTR-CM

There has been several research works on the prevalence of ATTR-CM in patients over the years. With the increase in awareness of the disease, its prevalence has been studied on patient groups from different geographical locations. Some papers regarding this will be discussed here.

AbouEzzeddine OF, Davies DR, Scott CG, et al., in their study [5] aimed to determine the prevalence of ATTR-CM in patients with HFpEF and assessed the clinical characteristics of the patients. The study was conducted on a community cohort of 1235 HF patients in southern Minnesota. Out of them, 286 patients underwent screening with TCPYP and 18 were found with ATTR-CM. The authors mainly studied on the clinical characteristics and outcomes of the two groups: with ATTR-CM and without ATTR-CM. In this study, 6.3% of the patients with HFpEF had ATTR-CM and for this the authors emphasized on the importance of adding ATTR-CM in the differential diagnosis of HFpEF. It was found that ATTR-CM was prevalent in older patients (age 70 and above), more likely in men (10.1%) and the patients had comorbidities as hypertension, diabetes, chronic kidney disease, carpal tunnel syndrome, and spinal stenosis. The authors also found that the patients with ATTR-CM had worse outcome with a higher rate of all-cause mortality and heart failure hospitalization. The study's retrospective methodology made it difficult to demonstrate a link between ATTR-CM and results. Also, the study was conducted on a fraction of the total population from a single healthcare center.

Prevalence of ATTR-CM in HF patients in Sweden was estimated by Lauppe RE, Liseth Hansen J, Gerdesköld C, et al. using the Swedish National Patient Register [7]. In the patient data used, there was no definitive diagnosis code for ATTR-CM, so the authors used a combination and elimination-based model to identify the ATTR-CM cases. They performed statistical analyses to

find out the patient characteristics, prevalence, mortality, and Red Flag diagnosis. The study found that 30% of the total cases were female and 70% were male with a mean age of 72.2 years. The prevalence of ATTR-CM was 7.4 per 100000 patients in 2018 and the mean survival time was 37.6 months after diagnosis, whereas the mean survival time for matched HF patients was 72.7 months. The study found some Red Flag diagnoses like carpal tunnel syndrome, spinal stenosis, hearing loss and atrioventricular and left bundle branch block. Carpal tunnel syndrome was found statistically significant with 17% of the ATTR-CM patients vs 3% of the matched HF patients and was diagnosed 6.7 or more years before ATTR-CM.

2.2 Research Works on Diagnosis Delay for ATTR-CM

Diagnosis delay is a major concern in the case of ATTR-CM. Delay and misdiagnosis prevents early treatment and contribute to the low survival rate for this disease. In this section some literatures on the diagnosis delay and the benefit of early diagnosis are reviewed.

Clinical history of ATTR-CM patients and comparison between the outcomes and quality of life (QOL) among patients was studied by Lane, Thirusha, et al. for patients in UK. The research was conducted on 711 ATTRwt-CM, 205 ATTRv-CM with V122I variant and 118 ATTRv-CM with non-V122I variant patients between the years 2000 to 2017. The study found median diagnostic delay of 39 months with more than 4 years for wild-type ATTR-CM after report of cardiac symptoms for 42% of the patients. The patients with diagnosis delay had a median of 17 hospital visits during 3 years before the diagnosis. The median survival from diagnosis were 31 months, 57 months and 69 months for V122I variant ATTRv-CM, ATTRwt-CM and non-V122I variant ATTRh-CM respectively. The study also showed the role of non-invasive procedures in diagnosis of the disease. It was mentioned that the survival increased with median 60.2 months from 46.3 months after the introduction of 99mTc-DPD scintigraphy. Before this, ATTR-CM was diagnosed mostly by invasive procedures like 63% diagnosis were via biopsy, usually endomyocardial biopsy [8].

Rozenbaum, M.H., Large, S., Bhambri, R. et al. reviewed a large number of literatures focusing diagnosis delay for ATTR-CM, the rate of delayed diagnosis and the clinical outcomes of the ATTR-CM patients in their literature [9]. Out of 59 initial articles, 23 were included in this review. The weighted means of the mean and median of diagnostic delays reported in these 23 articles were 6.1 and 3.4 years for ATTRwt-CM and 5.7 and 2.6 years for ATTRv-CM. The articles

reported that 34-57% of the patients were misdiagnosed [9]. In most studies, the symptoms that were regarded to be the first symptom for ATTR-CM were unspecified. Some of the studies considered shortness of breath, fatigue, and peripheral edema as cardiac symptoms. Some studies considered carpal tunnel syndrome as symptom onset or the first symptom. The studies found the median diagnostic delay to be similar for different age and sex. The review also found that diagnosis delay was longer for patients who had cardiomyopathy phenotype predominant in their history than those who had a history of mixed phenotype. Carpal tunnel syndrome was found to be associated with the longest delays and this was followed by erectile dysfunction, ocular problems, and peripheral neuropathy. Misdiagnosis was also observed in significant number of cases and in some cases, patients were given diagnosed with diseases that overlapped with ATTR-CM. Some of these overlapping diagnoses were hypertensive heart disease, hypertrophic cardiomyopathy, and ischemic heart disease [9].

A study to assess the feasibility and efficacy of screening for ATTR-CM in everyday clinical practice was presented by Witteles, Ronald M., et al. [10] A list of diagnoses that raise suspicion for ATTR-CM and should prompt further investigation by definitive diagnosis procedure for ATTR-CM were discussed in this study. Among the diagnoses, HFpEF and restrictive cardiomyopathy, family history of cardiac disease, carpal tunnel syndrome, spinal stenosis and other peripheral neuropathies are important. Rapidly progressive HF, low voltage on ECG, ventricular wall thickening which is disproportionate to hypertension, and evidence of cardiac amyloidosis on cardiac MRI are some of the Red Flags for ATTR-CM. Authors have noted that non-invasive tests like the use of biomarkers such as BNP and troponin, imaging modalities such as echocardiography, cardiac MRI and nuclear scintigraphy can be used instead of invasive procedure such as endomyocardial biopsy for the definitive determination of ATTR-CM. Genetic testings can also be used to identify individuals suspected with ATTR-CM [10].

The benefits from timely diagnosis of ATTR-CM and start of early treatment with Tafamidis was studied in [4]. In this study a discrete-time, cohort-level Markov state-transition disease simulation model was established and used to predict health outcomes for late diagnosis with treatment cases and early diagnosis with treatment cases. For wild-type ATTR-CM, the diagnosis delay considered in this study was 6.08 years and it was found that mean life expectancy can be extended by 5.46 years by early or timely diagnosis and treatment. For hereditary ATTR-CM the mean diagnosis

delay was 5.67 years and life expectancy can be extended by 7.76 years by timely diagnosis and treatment. Also, the corresponding quality-adjusted life years gains were found to be 4.50 and 6.22 years. Patients with delayed diagnosis usually have high healthcare resource utilization during the diagnostic journey, including hospitalization and a range of investigations for other conditions, but this aspect was not broadly studied in this research. [4]

2.3 Research Works on Machine Learning Based Approcahes in ATTR-CM Study

Diagnosis of ATTR-CM can be benefited through the implementation of statistical and machine learning methods. Some literatures are discussed below which used machine learning approaches to study the disease, find out the important features for it and detect it earlier.

Huda, A., Castaño, A., Niyogi, A. et al., et al. used a machine learning approaches to find out patients who have risk of developing ATTR-CM. The authors used medical claims data from IQVIA and electronic health record or EHR data from Optum and NMEDW, to train, test and validate their machine learning model. The data included patient demographics, clinical diagnoses, laboratory results and medication use. The authors used three supervised machine learning algorithms: logistic regression, XGBoost and random forest, to find out the features important for predicting ATTR-CM. The comparison found out that Random Forest model was the best model and had the highest AUROC of 0.93. They also used logistic regression analysis to compare individual ATTR-CM associated phenotypes and combinations of these phenotypes among the case and the control cohort. Using the top ten phenotypes with the highest odds ratios, they obtained all the combinations containing up to five phenotypes. The top 5 phenotype and combination on the basis of prevalence in ATTR-CM patients were: Combined systolic and diastolic HF, HFpEF (52.1%); Carpal tunnel syndrome (31.9%); AF, joint disorders, HFpEF (29.7%); Heart block, cardiomegaly, HFpEF (28.7%); Cardiomegaly, joint disorders, HFpEF (28.7%). [6]

Mitchell, Joshua D., et al. also used machine learning models to find out the key indicators or associated phenotypes for the disease and implemented the findings in EHR system to explore the real-world application of the study. The authors used Random Forest machine learning model and found out 9 phenotypes as features to predict ATTR-CM. From these 9 phenotypes, they created

another 20 combinations. They implemented their findings in EHR systems to generate notifications for patients who had phenotypes or combinations of phenotypes associated with high risk of ATTR-CM. Among the patients at risk, a high proportion had these two individual phenotypes: cardiomegaly; osteoarthrosis and these two phenotype combinations: carpal tunnel syndrome + HF, and atrial fibrillation + heart block + cardiomegaly + osteoarthrosis. [11]

3. DATA DESCRIPTION

We used Electronic Health Record (EHR) data for this study. The data was extracted from TriNetX which is a cloud-based platform that provides real-world clinical data from a global network of healthcare organizations. Due to the de-identified nature of the data i.e., all protected health information (PHI) identifiable information was excluded.

A brief overview of the data tables and the data dictionary is as follows:

a) Cohort Details: This table contains the number of patients found for each cohort selected for data generation in TriNetX.
b) Dataset Details: This table gives the number of total patients found and total number of health care organizations from which the data was taken.
c) Patient Cohort: This table contains the patient IDs for each cohort selected in the EHR platform.
d) Patient Demographic: This table contains all the demographic information for each patient ID. The demographics are sex, race, ethnicity, marital status, year of birth, year of death and patient regional location.
e) Diagnosis: This table gives a list of all the recorded encounter IDs, diagnosis codes, principal diagnosis indicator, admitting diagnosis, reason for visit and date of diagnosis for each patient. All the diagnosis codes are of International Classification of Diseases ICD-9-CM & ICD-10-CM code systems. Figure 2 shows the structure of the diagnosis table.
f) Encounter: This table contains information of each encounter for each patient with start date, end date and type of clinical visit.
g) Lab Result: The lab tests on the patients are listed in this table with Logical Observation Identifiers Names and Codes (LOINC) code for the tests, date, and test results.
h) Procedure: All the medical procedures on the patients are listed in this table with the procedure codes in Current Procedural Terminology (CPT), and Healthcare Common Procedure Coding System (HCPCS). Also, the dates of the procedures and principal procedure indicator are listed.
i) Medication: This table contains list of medication for the patients with unique ids of the medications, codes, start date and brand.

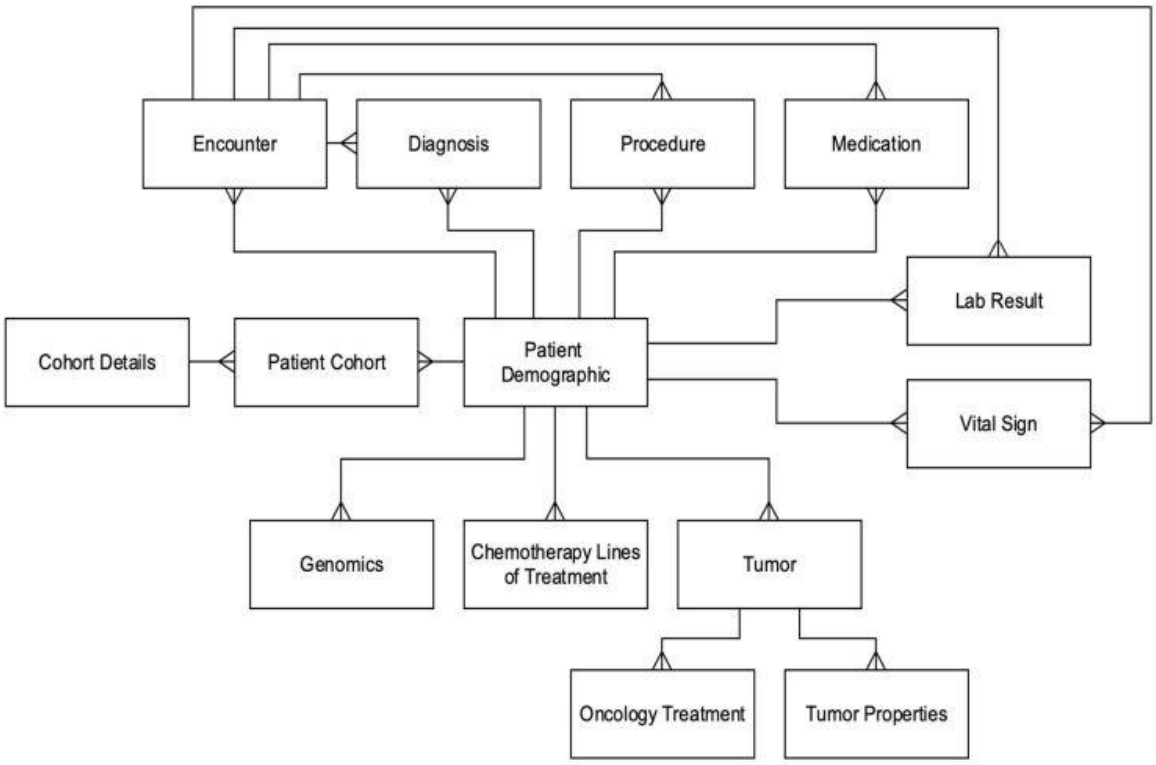

Figure 1. Data Tables Extracted from TriNetX

From the figure:1 above we can deduce the usability of the tables in the dataset. The patient cohort gives the patient IDs for each cohort selected in the system to generate the data. For each patient ID we can find the demographics like sex, race, ethnicity, year of birth, year of death, region from the patient demographic table. Each patient might have multiple encounters and a list of that is given in the encounter table. For each encounter we can find patient's vital signs, lab tests and their results, performed comorbidities and prescribed medications.

For the purpose of this study, we will be using the Cohort details, patient cohort, patient demographic, encounter, diagnosis, procedure, lab result and medication tables.

patient_id	encounter_id	code_system	code	principal_	admitting	reason_fo	date	derived_t	source_id
8ff9a4d07903f95479b4ceafa3970ad1d7380f63	945749aca1e5485c74531513df200dccffe4884c	ICD-10-CM	E11	S	U	U	20150415	F	EHR
8ff9a4d07903f95479b4ceafa3970ad1d7380f63	945749aca1e5485c74531513df200dccffe4884c	ICD-10-CM	I10	S	U	U	20150415	F	EHR
8ff9a4d07903f95479b4ceafa3970ad1d7380f63	945749aca1e5485c74531513df200dccffe4884c	ICD-10-CM	I11.0	S	U	U	20150415	F	EHR
8ff9a4d07903f95479b4ceafa3970ad1d7380f63	cd52eacf6e8bfa505f000e439bc477ac7e355f93	ICD-10-CM	K40	S	U	U	20160203	F	EHR
8ff9a4d07903f95479b4ceafa3970ad1d7380f63	945749aca1e5485c74531513df200dccffe4884c	ICD-10-CM	N18	S	U	U	20150415	F	EHR
3c34d10a157c501f3804249c444449f38e28b356	bd95ac66eeecfd0ff49cb2dc83704ee78a02677d	ICD-10-CM	I10	S	U	U	20150926	F	EHR
3c34d10a157c501f3804249c444449f38e28b356	bd95ac66eeecfd0ff49cb2dc83704ee78a02677d	ICD-10-CM	I50.1	S	U	U	20150926	F	EHR
3c34d10a157c501f3804249c444449f38e28b356	8cf8d6e62131eefcd0d0165a664bce8f18abbe65	ICD-10-CM	R11.10	S	U	U	20160624	F	EHR
759f5289e13f085e0eb4de7194e185a3967b6974	06db7302b73dd908e4896cd17b8042e16365e661	ICD-10-CM	H91.13	S	U	U	20120813	F	EHR
759f5289e13f085e0eb4de7194e185a3967b6974	0c61f0d280ec0613b16b132dbbb2939ecf6e129d	ICD-10-CM	I50.1	S	U	U	20160301	F	EHR
ab3cf1e2ea0764ea5279f503b5aa30754064132e	536b0c572cbd64e489ebb54f76138f955467d96f	ICD-10-CM	I50.1	S	U	U	20131207	F	EHR
c79e84dcf56e80e70dbaf6ea593607bd0c74974b	c529aba6b4817d65c1a08e99fe24a4d87689d27a	ICD-10-CM	E11	S	U	U	20141016	F	EHR

Figure 2. A snapshot of the diagnosis table of the dataset

To obtain the dataset from TriNetX different selection criteria was given. We wanted to obtain data of all the patients from North American who had reported cases of Heart Failure and ATTR-CM. For this, we made a list of all the ICD-9-CM and ICD-10-CM codes for heart failure and used the ICD-10-CM code for ATTR-CM which is E85.82. We also generated 5 groups while retrieving the data: Patients diagnosed with HF, Patients diagnosed with ATTR-CM, Patients with HF and ATTR-CM, Patients diagnosed with ATTR-CM before HF, and Patients diagnosed with ATTR-CM after HF. The data was generated on 10/16/2022 and TriNetX found a total of 2,577,621 individual patients who matched these cohorts from a total of 63 health care organizations from USA.

A table with the cohort details is given below in Table I:

Cohort Name	**Total Number of Patients**
HF patients	2,57,7200
ATTR-CM patients	2,431
Patients diagnosed with HF and ATTR-CM	2,010
Patients diagnosed with ATTR-CM before HF	1,953
Patients diagnosed with ATTR-CM after HF	1,836
Total number of individual patients	2,577,621

Table 1. Cohort Details

The data obtained for the research has some limitations:

- Not all the 2.5 million patients have full demographic information. There are patients who have some missing values for variables: year of birth, race and ethnicity.
- The diagnosis data contains a mix of ICD-9-CM & ICD-10-CM codes. So, both older and newer code systems need to be considered while studying the patients.
- The procedures data contains a mix of HCPCS, CPT, ICD-9-CM and ICD-10-CM codes. And in most cases principal procedure indicator is missing.
- The lab tests data contain LOINC codes for the tests and missing principal procedure indicator in many cases.
- While many patients have a long history of diagnosis, procedure and lab tests, some of the key cases (in this case some ATTR-CM positive patients) are probably missing some key information. For example, some ATTR-CM positive patients do not have any history of definitive tests or procedures of ATTR-CM in their records.

4. METHODOLOGY

In this research we developed and utilized statistical and machine learning methods to find out the key procedures for early and late diagnosis of ATTR-CM and to develop a model that predicts ATTR-CM in patients earlier. Statistical models such as logistic regression, forward and backward elimination and Least Absolute Shrinkage and Selection Operator (LASSO) is selected to be used to find out the key procedures. The key procedures will then be validated using survival analysis. Machine learning models: XGBoost and Random Forest will be used to predict the prevalence of ATTR-CM in patients based on the procedures and the top comorbidities and their combinations found from previous studies. The characteristics and functions of the methods are described in the following sections.

4.1 Logistic Regression

Logistic regression is a statistical method for prediction or classification. It examines the correlation between a binary dependent variable and one or more independent variables and thus predict the probability of an event. The binary dependent variable can be in the form of yes/no, 1/0, true/false and the independent variables may be binary, continuous, or categorical. Utilizing the logistic or sigmoid function, the dependent variable is expressed as a function of the independent variables. The logistic function converts any real-valued input into a value ranging from 0 to 1. The equation for logistic or sigmoid function is:

$$p = \frac{1}{1 + e^{-z}}$$

Here, p is the predicted probability of the event being forecasted or the positive class, z is called log-odds or logit which is the linear combination of the independent variables, and exp is the exponential function. The logistic regression model identifies the parameter values that best match the data by decreasing the difference between the expected probabilities and the actual outcomes. Maximum likelihood estimation or other optimization techniques are often employed to accomplish this. The cost function is set between 0 and 1 by the hypothesis of logistic regression. [12]

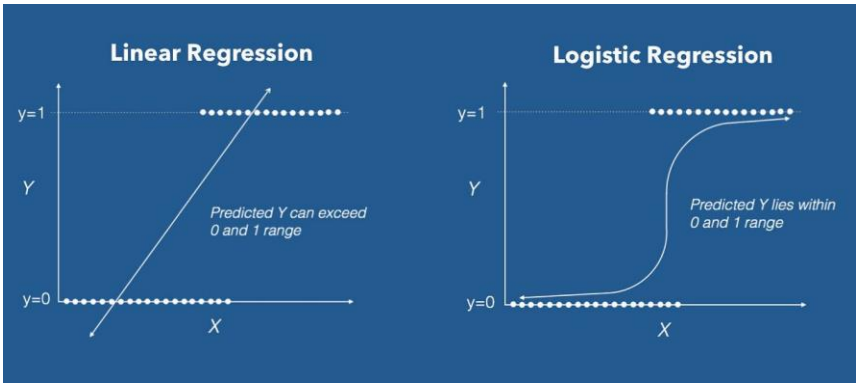

Figure 3. Visual illustration of logistic regression graph vs linear regression graph [13]

After training the model, it can be used to make predictions on new data. For the new data the model calculates the probabilities for each observation and compares them to a threshold (usually 0.5) and thus produces a binary classification.

Logistic regression is frequently employed in various fields to estimate the likelihood of an event based on a set of predictor variables. It is also frequently employed in machine learning as an underlying component for more complicated models, like neural networks.

4.2 Forward Selection and Backward Elimination

Forward Selection and Backward Elimination are two methods used for feature selection in machine learning or in regression analysis. Forward selection is a stepwise approach in which variables are added to an empty model one by one. The process begins with a null model which has no predictor variables and adds variables that are strongly relevant to the dependent variable. The process begins with an intercept and the variable addition is continued until no additional variables meet the inclusion criterion. The pre-specified inclusion criterion can be a significant increase in model fit, lowest P-value or reduction in prediction error. [14]

Backward elimination is also a stepwise approach, but it is opposite of the forward selection method. It starts with a model that includes all the predictor variables and removes variables one by one. The process continues until no further variables meet the exclusion criterion. The pre-specified exclusion criterion can be same as the forward selection method. [14]

Forward selection and backward elimination have their advantages and disadvantages. Forward selection can be computationally efficient as it requires fewer iterations than backward elimination. However, it may result in a suboptimal model if the initially selected variables are not the most

important ones. In contrast, backward elimination begins with a full model that includes all the variables and ensures that all important predictors are considered. Also, these stepwise regressions can tend towards overfitting if more variables with less potential are added. [14]

4.3 Least Absolute Shrinkage and Selection Operator- (LASSO)

Least Absolute Shrinkage and Selection Operator is a regression analysis which is used for feature selection and regularization. Like other regression models, it establishes relationship between a dependent variable and one or more explanatory variables. It is a modification of linear regression, and it is often used to address the overfitting in a model.

Linear regression models find out the values of the model parameters that minimize the sum of squared errors between the predicted values and the actual values. On the other hand, Lasso adds a penalty term to the objective function which is a function of the absolute values of the parameters. The penalty term shrinks the coefficients of the less important features to zero and leads the model to select a subset of the most important features. This results in a more parsimonious model that is less prone to overfitting.

The Lasso regression model can be written as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where y is the dependent variable, x_1, x_2, ..., x_p are the independent variables, β_0, β_1, β_2, ..., β_p are the coefficients, and ε is the error term. The goal is to find the values of β_0, β_1, β_2, ..., β_p that minimize the objective function:

$$\min\left\{(y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p)^2 + \lambda \sum |\beta_i|\right\}$$

where λ is the regularization parameter which controls the strength of the penalty term. It is a user defined parameter and the value of λ determines the degree of shrinkage applied to the feature coefficients. When λ is large the penalty term is dominant, and the coefficients are shrunk towards zero. This results in a sparse model with only a few non-zero coefficients. When λ is small the penalty term is negligible, and the Lasso regression model approaches the ordinary least squares regression model. [15] [16]

Lasso is useful when the number of features is much more than the number of observations. This method is robust to outliers and can handle correlated features by selecting one feature over the others thus addressing the multicollinearity of the features.

4.4 Survival Analysis

Survival analysis is a statistical method used to analyze time-to-event data and estimate the probability of an event or occurrence at a given time. Two important outputs of this analysis are the Survival function and the Hazard function. Survival function gives the fraction of population still at risk of experiencing the event at a specific period. It can be estimated using various methods, such as Kaplan-Meier estimator. Based on the observed data Kaplan-Meier estimator calculates the probability of survival at each point of time. If at distinct follow up times $t_1,t_2,t_3,t_4,t_5 \cdots t_k$, k patients have independent events occurring to them, then the cumulative survival probability can be obtained by multiplying all the probabilities of surviving in the time intervals (t_1 - t_2, t_2 - t_3, t_3 - $t_{4,\ldots.}$). At time t_j, if the probability of being alive is denoted as $S(t_j)$, it can be calculated from the probability of being alive at time t_{j-1} which is denoted as $S(t_{j-1})$, the number of patients alive just before time t_j which is denoted as n_j, and the number of events at time t_j which is denoted as d_j . The equation is:

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

where t_0=0 and $S(0)$=1. $S(t)$ remains constant between the times of events. Thus, the estimated probability is a step function, and it changes value at the time of each event. [17]

Another important concept in survival analysis is the hazard function. Given the event has not occurred yet, hazard function gives instantaneous rate at which events might occur at a given time in future. The hazard function can be estimated using various parametric and non-parametric models, such as Cox proportional hazard model.

Survival analysis also involves censoring which occurs when some of the individuals in the study do not experience the event or have missing data. Right-censoring, left-censoring, and interval-censoring are common in case of survival analysis.

Survival analysis can also be performed on repeated measures. A subject may have repeated measures for multiple time dependent covariates. The best way to model such dataset is by

introducing time intervals between each type of event for a subject. Repeated events survival analysis has applications in clinical trials to estimate the time-to-event outcomes of different treatments. It also has application in reliability engineering to estimate the failure rates of products or systems, and in social sciences to estimate the time to outcomes of different policies. [17]

4.5 Extreme Gradient Boosting (XGBoost)

XGBoost stands for Extreme Gradient Boosting. It is a decision tree-based Machine Learning technique and is widely used for regression, classification, and prediction tasks. It is highly considered for small-to-medium structured or tabular data for being a decision tree-based algorithm.

Decision trees are used by XGBoost as the base or weak learners in its ensemble. Decision trees are straightforward models that use hierarchical structure of binary splits to relate input data to output targets. Each tree has nodes and leaves, where each node stands for a feature and a threshold and each leaf for a prediction value. The performance of the model is quantified by an objective function, which XGBoost optimizes. A loss function and a regularization term make up the two parts of the objective function. The regularization term regulates the model's complexity and aids in preventing overfitting, while the loss function measures the discrepancy between anticipated and actual values. [18] The objective of XGBoost is:

$$\text{Obj}(\theta) = L(\theta) + \Omega(\theta)$$

where $L(\theta) = \sum_{i=1}^{n} l(y_i, \widehat{y_i})$ is the loss function, y_i is the target and $\hat{y}_i$ is the prediction. $\Omega(\theta) = \sum_{k=1}^{k} \Omega(f_k)$ penalizes the complexity of the model.

The boosting methodology used by XGBoost is based upon the concept that each subsequent weak learner strives to fix the errors committed by the ones before them. It adds one tree at a time as it develops. The model calculates the gradients of the loss function with respect to the predicted values in each round of boosting. The following tree is then trained to reduce these gradients or errors, successfully fitting the prior trees' residuals. Gradient boosting is a method that XGBoost employs to calculate the gradients. The gradient of the loss function is calculated in relation to the expected values from the previous iterations. The gradient reveals the direction and size of the mistake, which enables following trees to concentrate on the regions where the model performs poorly. XGBoost builds decision trees level by level. The root node, which serves as a

representation of the complete dataset, is created first. To reduce the loss function, it then iteratively separates the nodes depending on various attributes and thresholds. A stopping requirement, such as reaching the maximum depth of the tree or having too few samples in a node, must be satisfied before the splitting process can cease. To manage the complexity of the trees and avoid overfitting, XGBoost uses regularization techniques. The objective function is given a regularization term that penalizes large or complicated trees.

After the trees are built, XGBoost uses pruning strategies to cut off any branches that are superfluous or unimportant. Pruning helps the model become simpler and less complex, improving generalization and performance on untested data. XGBoost creates predictions by combining the results of all the decision trees after the boosting rounds are finished and the ensemble of trees has been constructed. While predictions for classification issues are often modified using a sigmoid function to provide probabilities or class labels, predictions for regression problems are typically the average of the leaf values. To improve its performance, XGBoost employs a number of optimization approaches. To efficiently create trees and generate predictions, parallel processing is used. Additionally, it offers approximate methods like histogram-based splitting and distributed computing that can speed up processing without compromising accuracy.

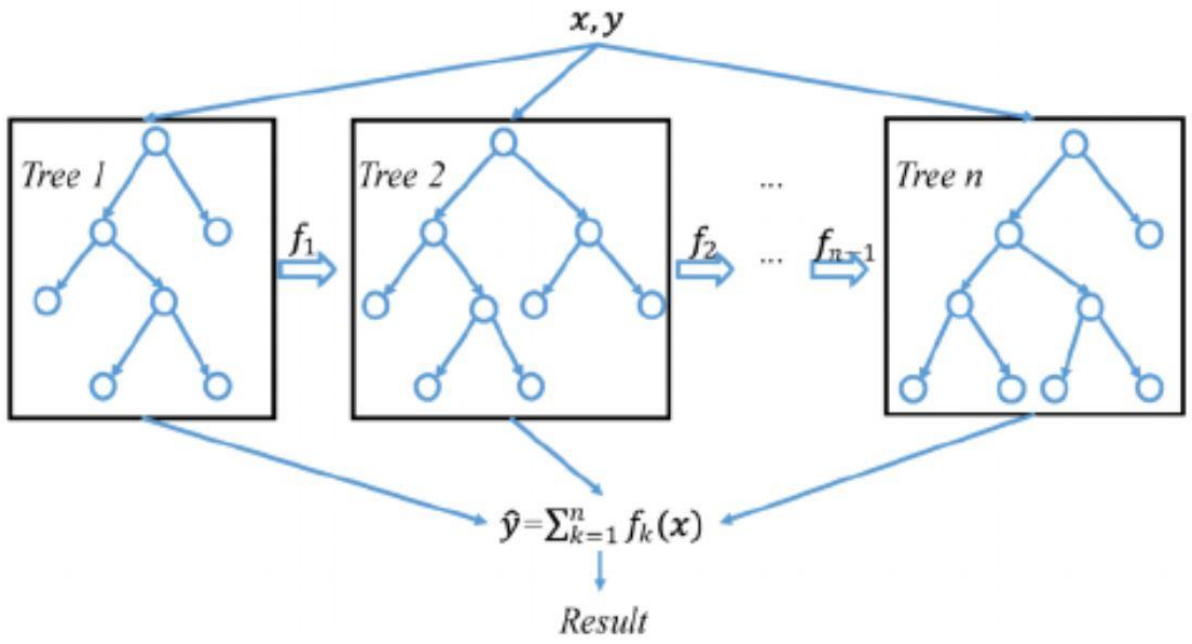

Figure 4: A general architecture of XGBoost [18]

4.6 Random Forest

Random Forest is a well-known machine learning technique that creates numerous decision trees during training and then combines their forecasts to produce the final prediction. It is renowned for its dependability, adaptability, and capacity to manage high-dimensional datasets.

Both Random Forest and XGBoost are ensemble algorithms. But the ways in which they generate and aggregate the weak learners differ. XGBoost produces the ensemble sequentially where each new tree corrects the errors of the preceding ones. On the other hand, Random Forest builds an ensemble of decision trees independently. XGBoost employs gradient boosting, which minimizes the gradients or errors of the earlier predictions in order to optimize the objective function. On the other hand, Random Forest uses Bootstrap Aggregating, also known as Bagging, as a method to build a variety of decision trees. By sampling with replacement, bagging involves splitting up the initial training data into numerous subsets. Each subset is used to train a different decision tree, commonly referred to as a bootstrap sample. [19]

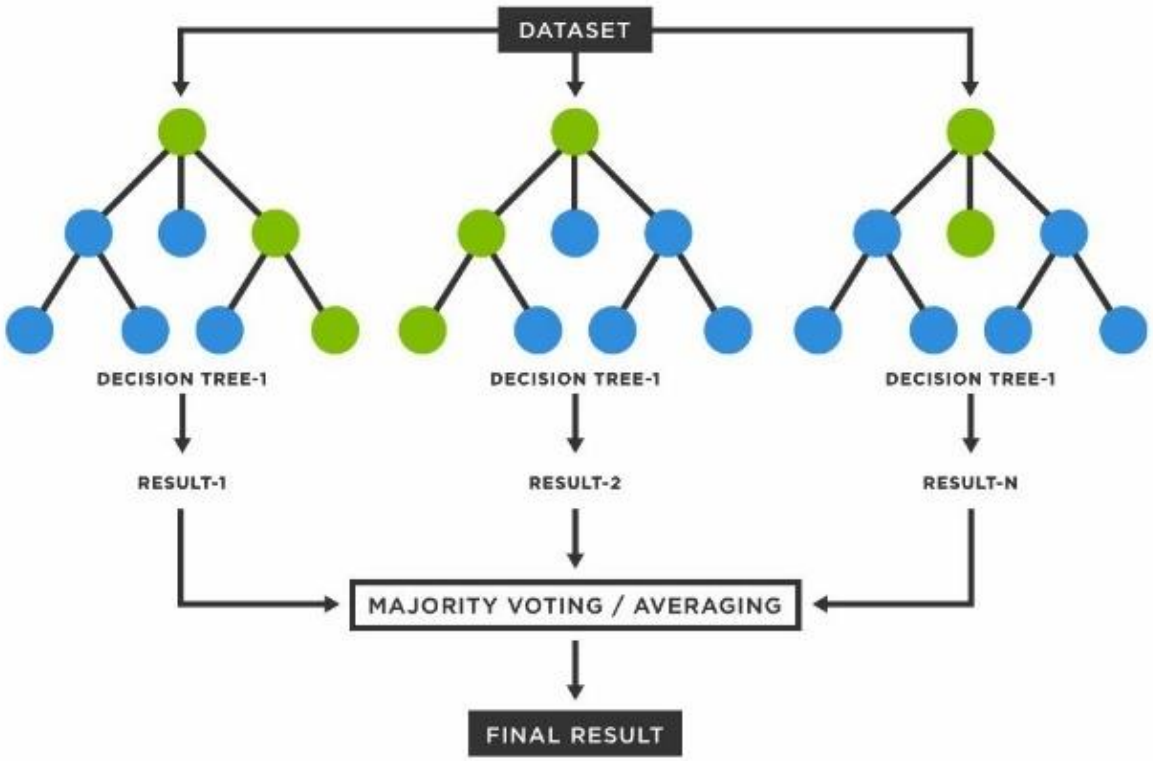

Figure 5: A general structure of Random Forest [20]

When building decision trees, Random Forest uses random feature selection in addition to data sampling. A random subset of features is picked at each split as opposed to taking into account all features. Because of the unpredictability, the decision trees are more diverse because each one concentrates on a distinct group of features. Each decision tree in a Random Forest is independently built using a random subset of features and a subset of the training data. Without any pruning, the trees are often developed to their full depth, allowing them to capture complex relationships in the data. After each decision tree is built, the results are combined using Random Forest to get the final prediction. To get the outcome for regression problems, the predictions are frequently averaged. The majority voting method is used to choose the final prediction in classification problems, which is the most frequent class predicted by the trees. Based on the knowledge acquired by each feature in the ensemble, Random Forest delivers a measure of feature relevance. It measures the value of each feature in the prediction process by averaging the importance ratings from all the trees. When choosing features and comprehending the underlying relationships in the data, this knowledge can be helpful. In conclusion, decision trees are used by both Random Forest and XGBoost, but they differ in how they are built and how weak learners are combined. While XGBoost employs gradient boosting to continually improve the ensemble, Random Forest focuses on producing different trees using bagging and random feature selection.

5. THE PROPOSED MODEL

In this chapter, firstly, we discussed how the cohorts are generated from the data. Secondly, the method of features selection is explained. Then, the process of predicting ATTR-CM in patients is described. Finally, the potential contributions of our proposed model are highlighted.

5.1 Cohort generation

To get the important features or procedures for early diagnosis of ATTR-CM, we generated few cohorts:

1. Patients diagnosed with HF
2. Patients diagnosed with ATTR-CM
3. Patients diagnosed with HF and ATTR-CM
4. Patients diagnosed with only HF and not ATTR-CM
5. Patients diagnosed with ATTR-CM before HF
6. Patients diagnosed with ATTR-CM after HF
7. Patients diagnosed with ATTR-CM and HF on the same date

All the data cleaning and cohort generation has been done using Python, and the Dask, Pandas and Numpy libraries. The steps of cohort generation are as follows:

Step 1: From the diagnosis dataset we first filtered out the patients who had ATTR-CM diagnosis history. We used the Dask library package of python to read and compute the large dataset. From the diagnosis code column, we filtered out the cases which had ICD-10-CM code E85.82 which is exclusively for ATTR-CM diagnosis. There were multiple diagnoses for many individuals, that is why we sorted the table by date and only kept the first diagnosis. A total of 2,431 patients found who were diagnosed with ATTR-CM.

Step 2: Next, we filtered out the patients who had Heart Failure diagnosis. There might be different scenarios for HF diagnosis and all of that are not relevant to our study such as post procedural HF. So, we first created a list of ICD-9-CM and ICD-10-CM codes that are relevant. We used the codes are shown in the table below:

Code System	Diagnosis Codes
ICD-9-CM	428, 428.1, 428.2, 428.21, 428.22, 428.23, 428.3, 428.31, 428.32, 428.33, 428.4, 428.41, 428.42, 428.43, 428.9
ICD-10-CM	I50, I97.131, I509, I09.81, I97.130, I503, I97.13, I11.0, I11.9

Table 2: ICD-9-CM & ICD-10-CM codes for heart failure diagnosis

Using these codes, we found all the relevant cases of HF. We then sorted the data by date and kept only the first diagnosis cases. We found a total of 2568764 cases of HF.

Step 3: Then using inner merge on the first two cohorts we filtered out the patients who were diagnosed with both HF and ATTR-CM. We found a total of 2,025 patients.

Step 4: Using left merge on the previous two cohorts we filtered out the patients who were diagnosed with only HF and no ATTR-CM. We found a total of 2,566,739 patients.

Step 5: We then created the last three cohorts with patients who had ATTR-CM diagnosed after HF diagnosis, before HF diagnosis and on the same date of HF diagnosis. We found 1531, 222 and 272 patients for these three criteria, respectively.

Step 6: Then the time difference between HF and ATTR-CM diagnosis was calculated for our focused group which is patients diagnosed with ATTR-CM after HF.

Step 7: Lastly, we included all the demographic variables from the patient dataset to each of these cohort datasets. Age at which ATTR-CM was diagnosed was calculated from the year of birth to the date of ATTR-CM diagnosis.

A summary table of the cohorts is given below:

Cohort Name	Total Number of Patients
Total ATTR-CM patients	2,431
Total HF patients	2,568,764
Patients diagnosed with HF and ATTR-CM	2,025
Patients diagnosed with only HF and no ATTR-CM	2,566,739
Patients diagnosed with ATTR-CM after HF	1,531
Patients diagnosed with ATTR-CM before HF	222
Patients diagnosed with ATTR-CM and HF on same date	272

Table 3. Summary of the cohort generated

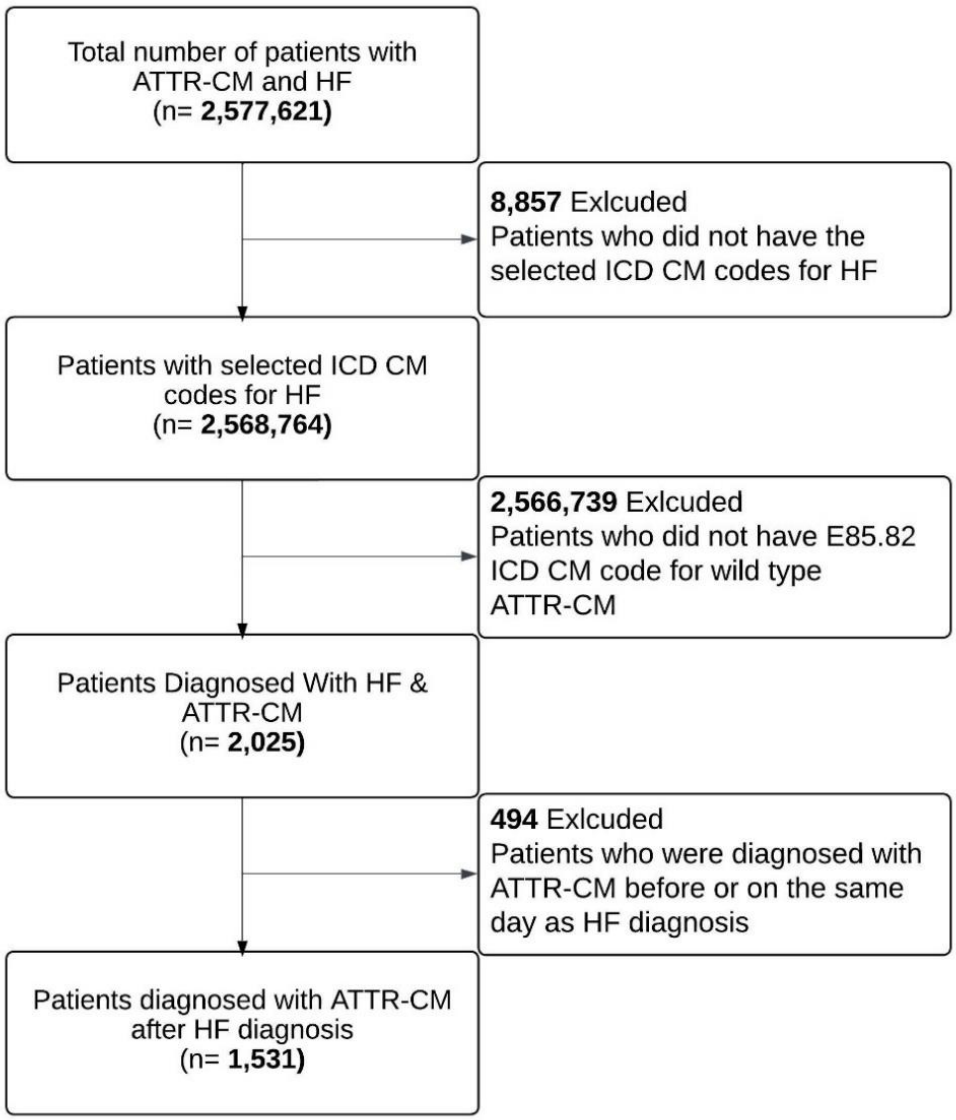

Figure 6: Flow Diagram of Cohort Generation

5.2 Features Selection by Statistical Analysis

Statistical analysis was used to find out the top features for the prediction model. The goal was to get a list of top procedures that contribute to the early diagnosis of ATTR-CM. Our focus cohort was the 1531 patients who had ATTR-CM diagnosed after HF. The tasks in this section were conducted only on this group and the steps are given below:

Step 1: First, from the procedure dataset we filtered out all the procedures performed on the patients who were diagnosed with ATTR-CM after HF.

Step 2: We then took only the procedures that were performed in the time window between the patients' HF and ATTR-CM diagnosis.

Step 3: Grouping the patients with procedure codes and date we created a table which contains how many times each patient had each procedure performed on them. We also created a table which shows the frequency of occurrence for each procedure.

Step 4: To find the procedures that are important to diagnose ATTR-CM, we first made a list of all the procedures used individually or as a combination to detect ATTR-CM. We used the following sources to list the procedures and then found out the codes for them.

We found a total of 12 groups of procedures and 57 unique codes for these procedures. The codes are in CPT, HCPCS and ICD-10-CM system.

Step 5: Next, we filtered out the ATTR-CM specific procedure occurrences in our focused group between the time of their HF and ATTR-CM diagnosis. We also calculated the occurrence of each procedure for each patient and the time difference of the procedure occurrence from the diagnosis of ATTR-CM.

Out of 1531 patients we found a total of 1365 patients with a record of procedures in the selected timeframe. The 166 patients who are missing from this list might not have a full record of their medical visits. Out of this 1365 patients, we found 1092 patients with the ATTR-CM specific procedures between the HF and ATTR-CM diagnosis. The missing 273 patients might not have a full record of their medical visits or might have the procedures in different code system format. For these 273 patients, we have made a list with the procedures they had between the selected time window to discuss with medical practitioners.

Step 6: We then sorted the grouped data for patients and codes by date and took only the last occurrence of each procedure for each patient. From this data we generated a box plot to study the procedures responsible for diagnosis delay.

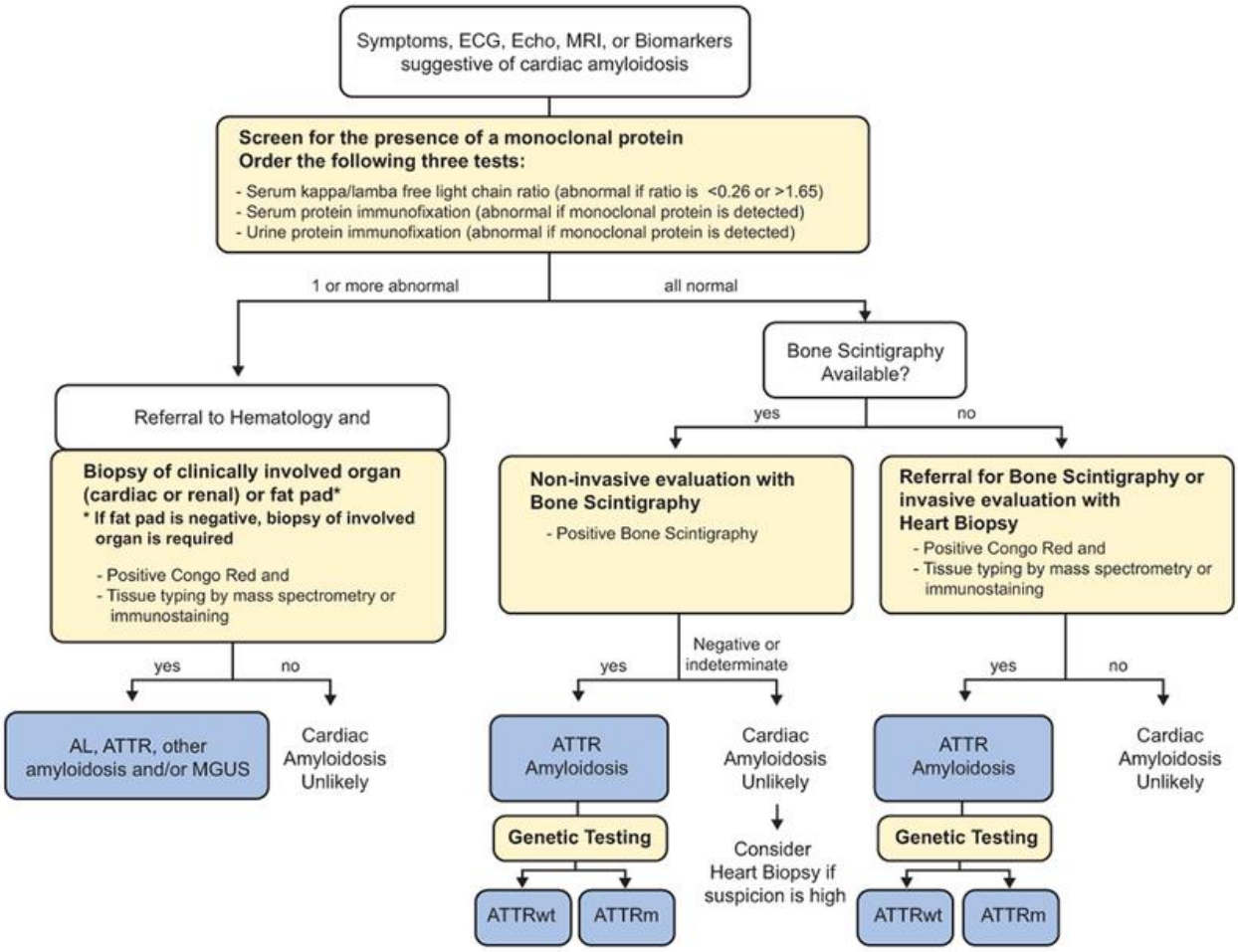

Figure 7. Diagnostic algorithm for patients with suspected cardiac amyloidosis [21]

Step 7: To address diagnosis delay we then divided the patients into two groups: early and late diagnosis groups. For this study we considered the procedures performed in the 90 days prior to the diagnosis of ATTR-CM to be important as these procedures resulted in the definitive diagnosis of the disease. So, 90 days was set as the threshold for early diagnosis. If a patient had any of the ATTR-CM specific procedure in the 90 days prior to his ATTR-CM diagnosis, we refer it to early diagnosis. On the contrary, if a patient did not any of the ATTR-CM specific procedures in the 90 days prior to ATTR-CM diagnosis, it means the procedure or combination of procedures the patient had has a higher delay, and hence these are referred as late diagnosis in our study.

Clinical Tests	Findings
ECG	Normal or low ECG voltage; pseudo-infract pattern; atrioventricular block; bundle branch block
ECHO	Increased left or right ventricular wall thickness; increased atrial septal thickness; impaired longitudinal strain; apical sparing pattern by longitudinal strain; thickened valve leaflets; increased LV filling pressure; pericardial effusion
CMR	Increased biventricular wall thickness; increased LV mass, diffuse subendocardial or transmural late gadolinium enhancement, increased native non-contrast T1 and ECV
^{99m}TC bone scintigraphy (DPD/PYP/HMDP)	Grade 2/3 myocardial uptake;
Serum cardiac biomarkers	Increased BNP or NT-proBNP levels, increased troponin T or troponin I level

Table 4: Clinical Tests and Findings Potentially Suggestive of ATTR Amyloidosis [22]

Step 8: To study the procedures contributing to early diagnosis, we created a box plot for procedures against the time of last occurrence prior ATTR-CM diagnosis.

Step 9: A binary data frame was generated for statistical analysis. The procedure codes were taken as independent variables and valued as 0 if a patient did not have that procedure or 1 if the patient had the procedure. We also added patient demographics: age, sex, ethnicity and race as independent variables. The dependent variable was the criteria of late and early diagnosis. We have considered the early diagnosis 1 and late diagnosis 0.

Step 10: Logistic regression, Forward and Backward elimination, and LASSO was used for statistical analysis on the data. From these analyses the variables, in this case the procedure codes were ranked based on their p-values and odds ratios. Comparing the four analyses, the top procedures or features were selected.

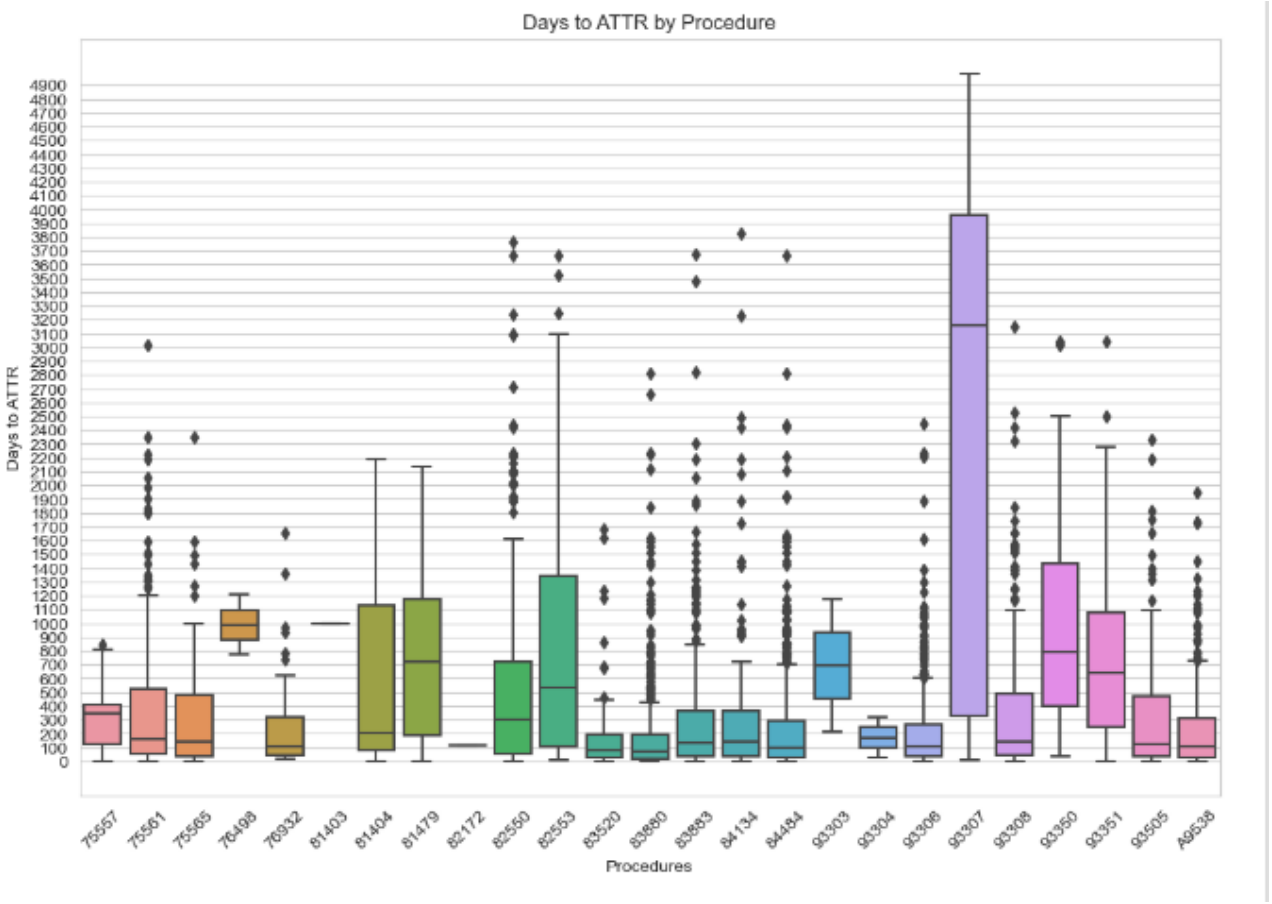

Figure 8: Box Plot for last occurrence of procedure to ATTR-CM diagnosis

5.3 Propensity Score Matching

To generate the case and control cohorts, propensity score matching was done. The control group was created from the 2566739 HF patient who had no ATTR-CM diagnosed. For generating the control cohort, patient who did not have their year of birth and sex on the dataset were dropped. The number of HF without ATTR-CM patient was 1:1 propensity score matching was done based on the age, sex, duration of medical history and number of hospital visits of the patients.

Step 1: The procedure and patient table were used for creating the dataset for propensity score matching. Patient sex, and year of birth was added to the procedure table from the patient table.

Step 2: Patient age was calculated form the date of first ATTR-CM diagnosis for the case cohort and first HF diagnosis for the control cohort. Finally, the rows with unknown demographics were dropped.

Step 3: Number of hospital visits was counted based on individual encounter IDs for each patient. Duration of medical history was found by subtracting the year of the first encounter latest.

Step 4: 1:1 propensity score matching was done using R programming language taking age, number of hospital visits and duration of medical history as continuous values, and sex as factors. Binary column for ATTR-CM was the logical values.

For matching, the nearest neighbors method was used. Other matching methods available were exact matching, caliper matching. With regards to our example, for each case in the patient sample exactly one case in the population sample was matched. For 1094 cases, equal number of matched control patients was found.

	Control Group	**Case Group**	**p**
n	2095642	1094	
Age (mean (SD))	65.49 (14.23)	76.00 (8.47)	<0.001
sex (%)			<0.001
F	965040 (46.0)	198 (18.1)	
M	1130367 (53.9)	896 (81.9)	
Unknown	235 (0.0)	0 (0.0)	
number_of_visits (mean (SD))	73.16 (123.69)	143.23 (158.83)	<0.001
duration_of_medical_history (mean (SD))	3.07 (5.28)	6.38 (5.81)	<0.001

Table 5: Patient characteristics before propensity score matching

	Control Group	**Case Group**	**p**
n	1094	1094	
Age (mean (SD))	76.24 (8.64)	76.00 (8.47)	0.506
sex = M (%)	891 (81.4)	896 (81.9)	0.825
number_of_visits (mean (SD))	118.09 (136.14)	143.23 (158.83)	<0.001
duration_of_medical_history (mean (SD))	6.90 (7.44)	6.38 (5.81)	0.07

Table 6: Patient characteristics before propensity score matching

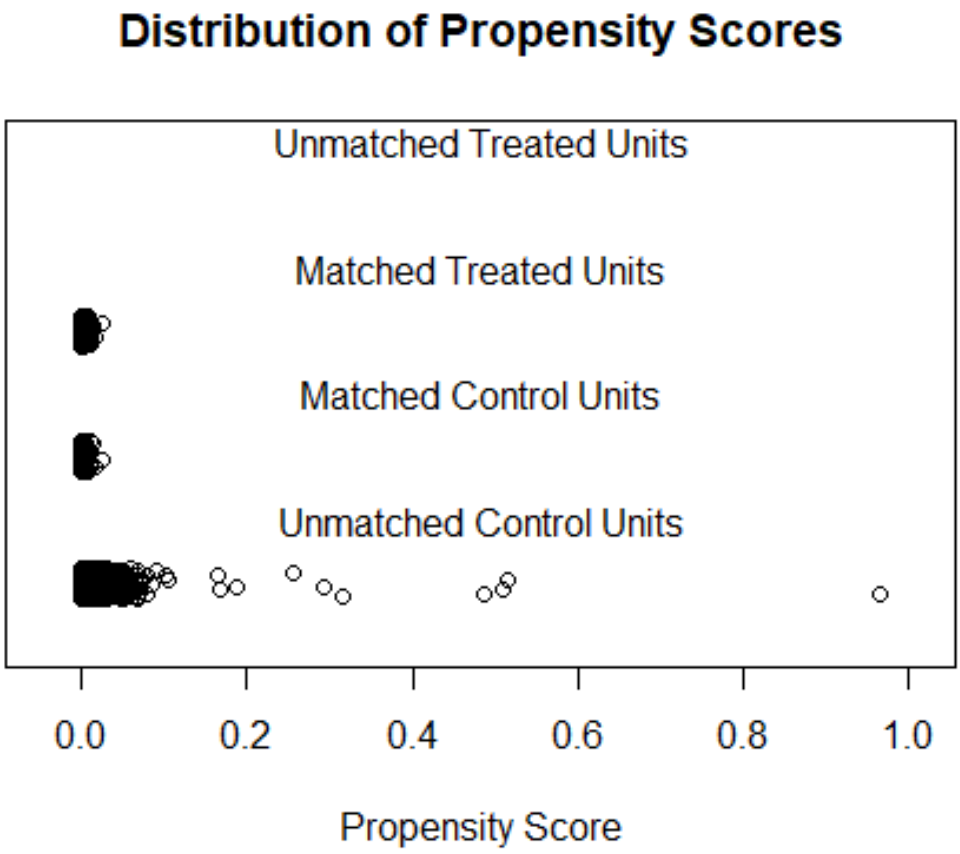

Figure 9: Distribution of propensity scores in all cohorts

5.4 Survival Analysis

To identify and evaluate the impact of the important procedures found from the statistical models, survival analysis was employed. Both case and control cohort were used to do the survival analysis. From the procedure table all the procedures on both the cohort patients were filtered. Then only the top three procedures based on P-values from the statistical models were taken. A dataset was created for survival analysis with patient ids, procedure codes, time intervals, age, sex and ATTR-CM diagnosis columns. The initial date of observation was selected to be the first HF occurrence for each patient. Time intervals were calculated for each procedure for each patient.

Survival analysis was done on R programming language using the 'survival' and 'survminer' packages. Survival or Kaplan-Meier curve was generated, and Cox proportional hazard model was fit.

5.5 Prediction of ATTRwt-CM by ML Models

In the last stage of the study, machine learning models were used to predict ATTRwt-CM for the cohorts using the important procedures and the comorbidities found from previous studies. The tasks done for this section are given below:

Step 1: For the prediction models, some features were selected from the reference study by Huda, A., Castaño, A., Niyogi, A. et al. [6]. In their study they used logistic regression to find out the important comorbidities and their combinations as features to predict the prevalence of ATTRwt-CM. From the table of top cardiac and non-cardiac phenotypes predictive of ATTRwt-CM, the top 10 phenotypes based on odds ratio (95% CI) was selected for this study. And from the table for combination of phenotypes based on their ICD codes and their effect on ATTRwt-CM, top 10 phenotype combinations were selected based on the true positive scores.

Step 2: ICD-9-CM and ICD-10-CM codes for the phenotypes selected from the literature were listed. The mapping of ICD codes was done at the Short Description (diagnosis description) level. The 2023 release of ICD-10-CM and the Version 32 of ICD-9-CM were used for the retrieval of the codes. A total of 520 diagnosis codes were found for the selected individual and combined phenotypes.

Step 3: From the diagnosis table, for the control cohort all the diagnoses selected above were taken. For the case cohort the selected diagnoses occurred before the ATTRwt-CM diagnosis were taken. Then, for both the cohorts, the first occurrences of each diagnosis were kept. A column for group name of the diagnoses was added.

Step 4: The occurrences of the top three procedures (found from the statistical analyses) on the case and control cohort were taken from the procedure table. Finally, the diagnosis data and the procedure data were joined to make the complete dataset.

Step 5: A binary dataset was created with the patient IDs, selected diagnoses and procedures as features and ATTRwt-CM as event. This binary dataset was used for the initial prediction models.

Step 6: The machine learning analyses (XGBoost and Random Forest) were performed in Python. The binary dataset was divided into train and test with a ratio of 80:20. Model was fitted using the

XGBoost and Random Forest packages for Python. For both models, accuracy, precision, recall and F1 score were calculated, AUROC curve was plot, and feature importance was summarized.

Step 7: We then used XGBoost and Random Forest machine learning models with nested cross validation to validate our results. In the nested cross validation, in the outer loop we used 5-fold cross validation to select the training and the test sets. Then we used 5-fold cross validation to select the hyperparameters by grid search algorithm. The types and ranges of the parameters used for XGBoost were Number of trees: 10, 25, 50, 100, 200, 300 and Maximum depth of trees: 3, 5, 10, 15, 20, None. The types and ranges of parameters used for Random Forest were Number of trees: 10, 25, 50, 100, 200, 300, Maximum depth of trees: 3, 5, 10, 15, 20, None, Minimum Samples Per Leaf: 2, 3, 4, 5, Minimum Samples Split: 2, 3, 5, 7, Bootstrap: True, False.

Step 8: For the prediction we also used a larger dataset with the ratio of 1:2 for case and control cohort generated by propensity score matching. For this dataset we used the nested cross validation for train and test set split and hyperparameters tuning. For both models, accuracy, precision, recall and F1 score were calculated, AUROC curve was plot, and feature importance was summarized.

Step 9: To observe the effect of these important feature in the prediction of ATTRwt-CM, the test dataset was modified. Data from 1, 2, 3, and 4 years prior to the ATTRwt-CM patients were taken for the test purpose. Then testing on this new dataset, the model evaluation parameters were calculated.

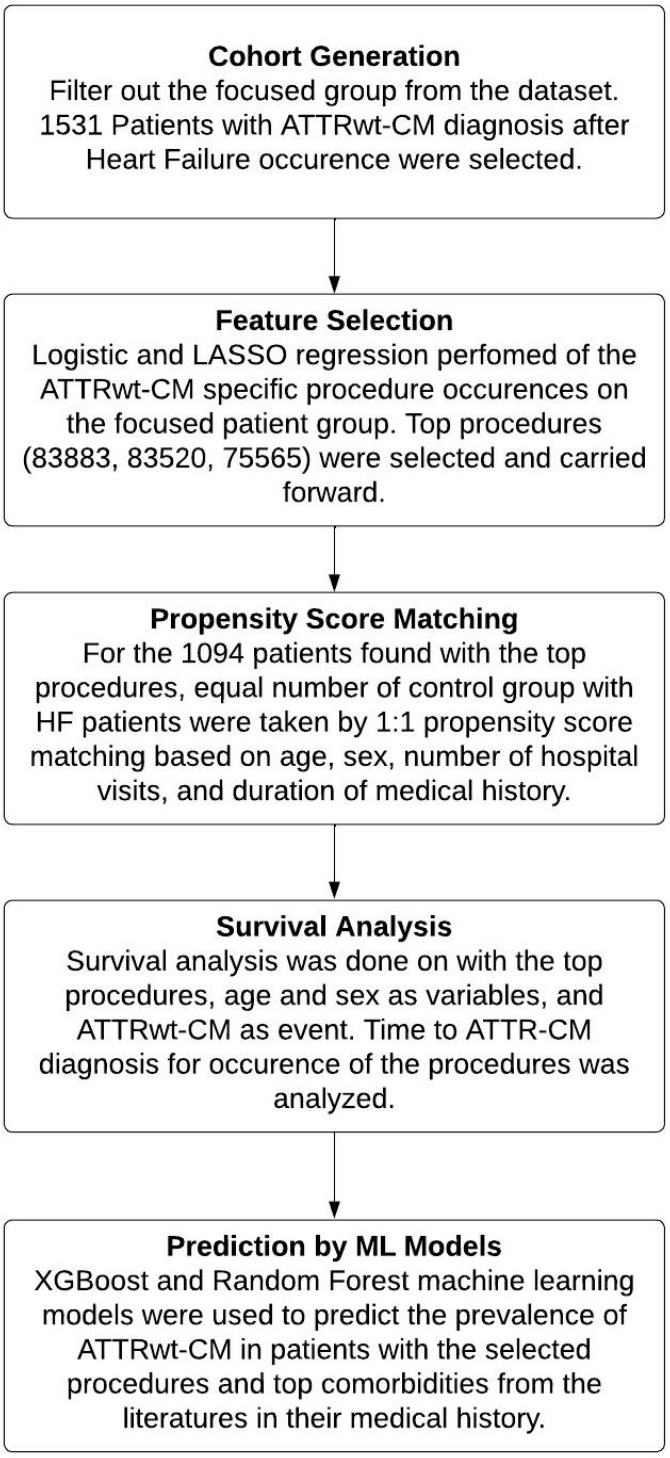

Figure 10: Flowchart of Proposed Model

6. RESULTS

6.1 Selecting Important Procedures for Early Diagnosis

From the logistic regression and LASSO the following tables were found for the ATTRwt-CM specific procedures. For logistic regression, the results are sorted based on the P-values and for the LASSO regression, the results are filtered based on the odds ratio.

	Estimate	Std. Error	Pr(>|z|)	OR	2.50%	97.50%
sex M	0.453	0.179	0.012	1.573	1.105	2.235
84134	-0.535	0.221	0.016	0.586	0.38	0.906
93350	-1.158	0.499	0.02	0.314	0.114	0.825
83520	0.707	0.306	0.021	2.028	1.139	3.812
75565	0.6	0.298	0.044	1.822	1.029	3.325
83883	0.363	0.181	0.044	1.438	1.011	2.054
93351	-0.821	0.429	0.056	0.44	0.187	1.023
81404	-1.23	0.657	0.061	0.292	0.076	1.064
race White	0.264	0.149	0.076	1.302	0.974	1.746
93308	0.267	0.176	0.13	1.306	0.928	1.854
84484	0.305	0.202	0.132	1.356	0.911	2.016
82553	0.452	0.322	0.16	1.572	0.843	2.991
93505	-0.275	0.197	0.162	0.759	0.517	1.12
83880	0.274	0.197	0.163	1.315	0.896	1.938
76932	-0.428	0.391	0.273	0.652	0.304	1.415
93303	-1.498	1.441	0.299	0.224	0.009	5.849
Age	-0.008	0.008	0.323	0.992	0.975	1.008
82550	-0.231	0.239	0.334	0.794	0.498	1.273
93306	-0.135	0.16	0.401	0.874	0.637	1.194
81479	0.604	0.8	0.45	1.829	0.4	9.586
Ethinicity Non-hispanic	-0.425	0.812	0.601	0.654	0.096	2.754
93304	-0.514	1.459	0.725	0.598	0.022	16.035
9538	0.046	0.158	0.77	1.047	0.769	1.429
75561	-0.047	0.174	0.786	0.954	0.679	1.346
93307	-0.124	0.626	0.843	0.883	0.266	3.252
76498	13.738	548.692	0.98	925676.875	0	NA
81403	-16.048	882.744	0.985	0	NA	8.40E+7
82172	15.048	882.744	0.986	3430636.077	0	NA
75557	0.006	0.918	0.995	1.006	0.188	8.09

Table 7: Logistic Regression Results for ATTRwt-CM specific procedures

Variable	Coefficient	OR
82172	0.815527266	2.260367174
83520	0.528911477	1.697083988
76498	0.495567046	1.641428741
75565	0.412661761	1.510833916
83883	0.253499312	1.288526493
sex	0.247269887	1.280524664
84484	0.230498879	1.259228056
83880	0.208460951	1.231780829
race	0.146684709	1.157988802
93308	0.111855661	1.118351428
82553	0.107701987	1.113715794

Table 8: LASSO regression results for ATTRwt-CM specific procedures

From the logistic regression, five procedures: 84134, 93350, 83520, 75565, 83883 had P-value less than 0.05 and can be decided to be significant in the early diagnosis of ATTR-CM. These codes are for Prealbumin test (84134), Transthoracic Echocardiography (93350), Serum Test - Transthyretin (TTR) protein analysis (83520), Cardiac MRI (75565), and Serum Test - Assay of free light chains; kappa and lambda with ratio (83883) respectively.

From the LASSO regression the top five procedures based on odds ratio were 82172, 85520, 76498, 75565, and 83883. Here 82172 is apolipoprotein testing and 76498 is MRI. From the two regression models, three common procedures: 83883, 83520 and 75565 were found. For the dependent variable in these regression models, we chose the criteria to be positive if the procedure was found within 90 days before ATTR-CM diagnosis or zero otherwise. The five procedures found form logistic regression and the eleven procedures found from LASSO are contributor to the event. So, it can be said that these procedures are responsible for early detection of ATTR-CM.

6.2 Survival Analysis on the Important Procedures

To evaluate the impact of the important procedures, survival analysis was performed. Survival curve was generated to observe the probability of identifying ATTR-CM using these procedures.

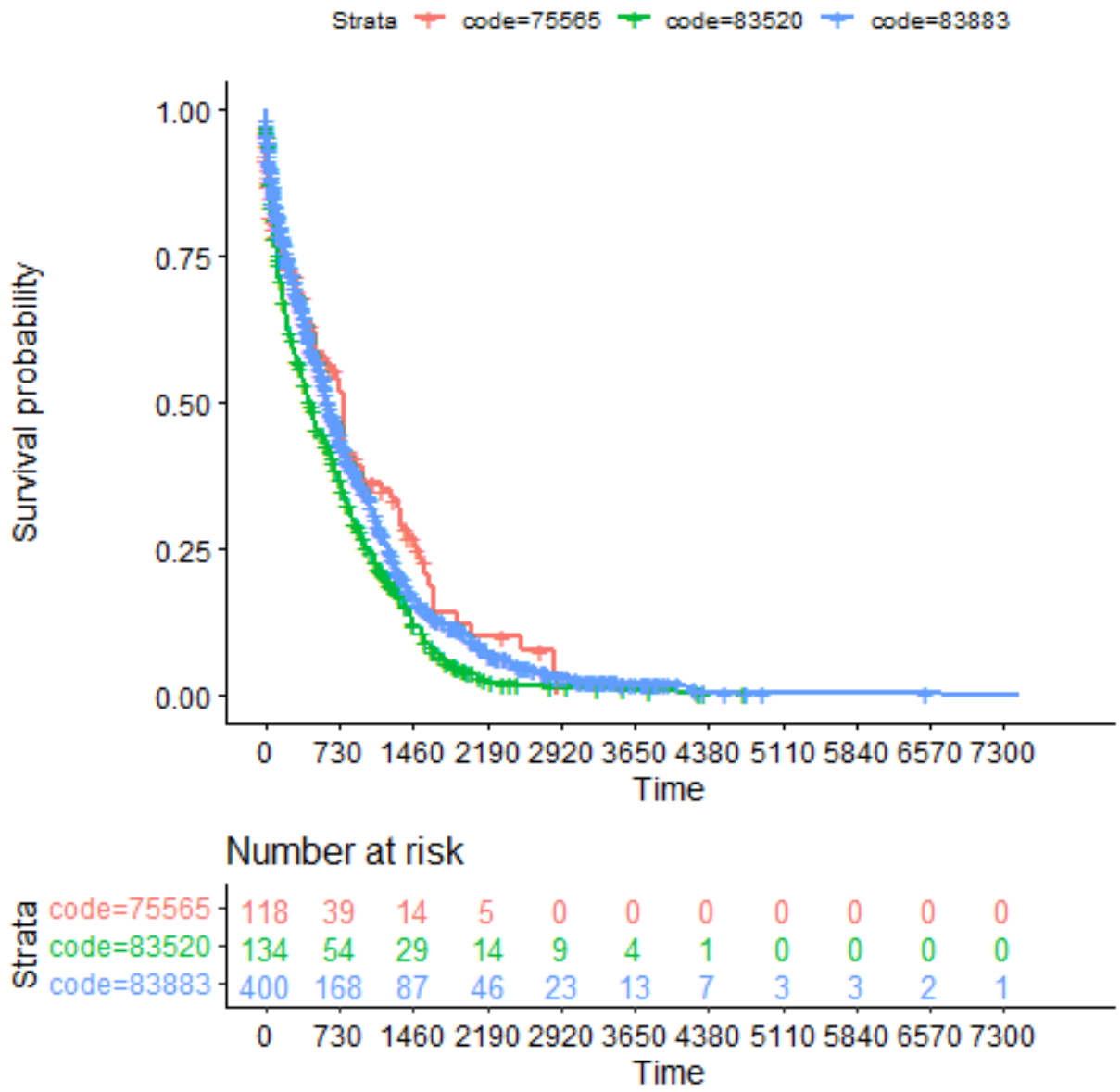

Figure 11: Survival analysis on the top procedures for early diagnosis

On the plot, the horizontal axis (x-axis) represents time in days, and the vertical axis (y-axis) shows the probability of being diagnosed with ATTRwt-CM. The lines represent survival curves of the three procedures. Each vertical drop in the curves indicates an event or a diagnosis. The vertical tick marks on the curves represents that a patient was censored at this time. At time zero, the survival probability is 1.0 that is no patient was diagnosed.

From the plot we can see that, at time 730 which is 2 years from the first heart failure diagnosis, the probability of diagnosing ATTRwt-CM with 75565 is 66.95%, 83520 is 59.71%, and 83883 is

58%. At time 1460 which is 4 years from the first heart failure diagnosis, the probability of diagnosing ATTRwt-CM with 75565 is 88.13%, 83520 is 78.35%, and 83883 is 78.25%.

n= 1429, number of events= 645
(80 observations deleted due to missingness)

	coef	**exp(coef)**	**se(coef)**	**z**	**Pr(>|z|)**
code83520	0.37015	1.447952	0.141699	2.612	0.009
code83883	0.10444	1.110091	0.129737	0.805	0.4208
Age	0.0109	1.010962	0.004015	2.715	0.00662
sexM	0.20593	1.228666	0.095908	2.147	0.03178

	exp(coef)	**exp(-coef)**	**lower .95**	**upper .95**	
code83520	1.448	0.6906	1.0968	1.911	
code83883	1.11	0.9008	0.8608	1.431	
Age	1.011	0.9892	1.003	1.019	
sexM	1.229	0.8139	1.0181	1.483	

Concordance= 0.522 (se = 0.016)
Likelihood ratio test= 20.91 on 4 df, p=3e-04
Wald test = 20.56 on 4 df, p=4e-04
Score (logrank) test = 20.62 on 4 df, p=4e-04

Table 9: Cox Proportional Hazard Model Summary

In the Cox proportional hazard model, the positive coefficients (coef) mean that the hazard (risk of ATTRwt-CM diagnosis) is higher. The R summary for the Cox model gives the hazard ratio (HR) for the subsequent groups relative to the first group, here the first group is patients with 75565 procedures. The exponentiated coefficients are known as *hazard ratios* which give the effect size of covariates. It can be observed that 83520, 83883, Sex: Male, and Age all are associated with the diagnosis of ATTRwt-CM.

Here, the p-values for the likelihood, Wald and score tests are significant This indicates that the model is significant. The test statistics are in close agreement in our model. In this multivariate analysis, the covariates Age and 83520 are significant as p-value < 0.05. However, the covariates sex and 83883 fail to be significant as p-value is greater than 0.05.

6.3 Prediction of ATTRwt-CM among Heart Failure Patients

The demographic and clinical characteristics of patients included were taken from the 2568764-patient dataset. Case (Patients with ATTR-CM after HF diagnosis and had the definitive procedures for ATTR-CM) and Control (Patients with HF diagnosis and no ATTRwt-CM diagnosis) patient cohorts were matched on age, sex, duration of medical history in the database, and number of healthcare visits. The mean age of patients across the cohorts was 76 years. There were similar proportions of male and female patients in both cohorts with 81.2 % male and 18.2 % female. The total number of healthcare encounters and total duration of diagnostic history information in the datasets had a mean of 143 and 6.29 years respectively. For this study the Heart Failure (HF) diagnosis was considered as the symptom onset and the prediction was done for the time after HF diagnosis for all the patients.

At first, we made our predictions using XGBoost, and Random Forest with 80:20 train and test split with default hyperparameters settings. Doing this we found, the Random Forest model had the highest accuracy of 80.14% (vs. 79.67% for XGBoost) as shown in Table 10. The model performed well in correctly predicting wild-type ATTR-CM HF vs. non-amyloid HF. The recall (sensitivity), and F1 score found were better for Random Forest model (Table 10), while the precision (positive predictive value [PPV]) was better for XGBoost.

The accuracy measures how well the model's predictions were made overall. It is the percentage of accurate forecasts or true positives as well as true negatives out of all predictions. For example, in the case of the XGBoost model, the accuracy is 79.67%, which means that on about 79.67% of the occasions, the model accurately anticipated the outcome. In general, an accuracy of 79.67% is regarded as good.

Precision is the percentage of positive instances that are accurately predicted compared to all the predicted positive instances of the model i.e., true positives and false positives. Greater precision means the model is producing fewer false positives. Recall is referred to as Sensitivity or True Positive Rate. It quantifies the percentage of positive cases that were correctly predicted to all positive instances actually observed that is true positives and false negatives. A greater recall value means that a bigger percentage of positive examples are being captured by the model. The harmonic mean of is the F1 score. It offers a balanced measurement that accounts for both recall and precision. A higher F1 score denotes a more favorable balance between recall and precision.

	XGBoost	Random Forest
Accuracy	79.67%	80.14%
Precision	78.14%	78.08%
Recall	80.77%	82.21%
F1 Score	79.43%	80.93%

Table 10: Performance evaluation of the prediction models

We also plotted the AUROC curve to see the performance of our models. Both the models had area under the ROC curve or AUC of 0.87. This is a good score for a binary classification model. AUC score ranges from 0 to 1, where 0 is for poor classifier which suggests random guessing, and 1 is for perfect classifier which suggests all the correct predictions were made. Our score of 0.87 suggests that our models had good discrimination power and is capable of distinguishing positive and negative cases with moderately high accuracy.

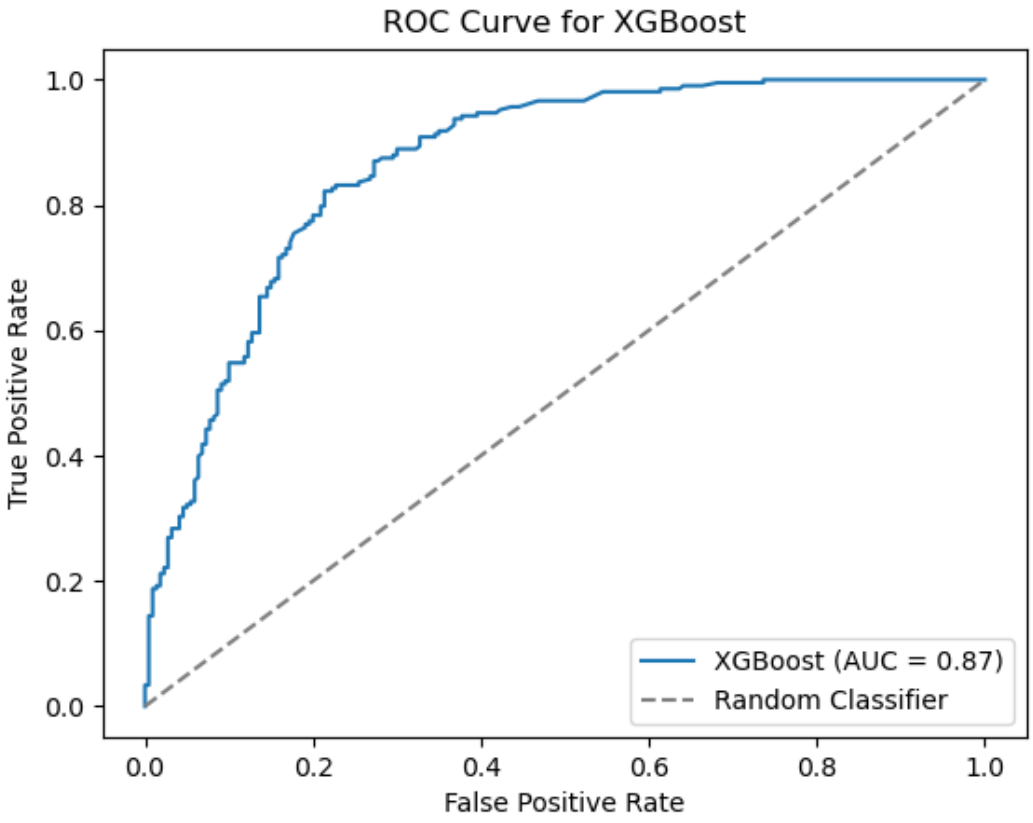

Figure 12: ROC Curve for XGBoost Model

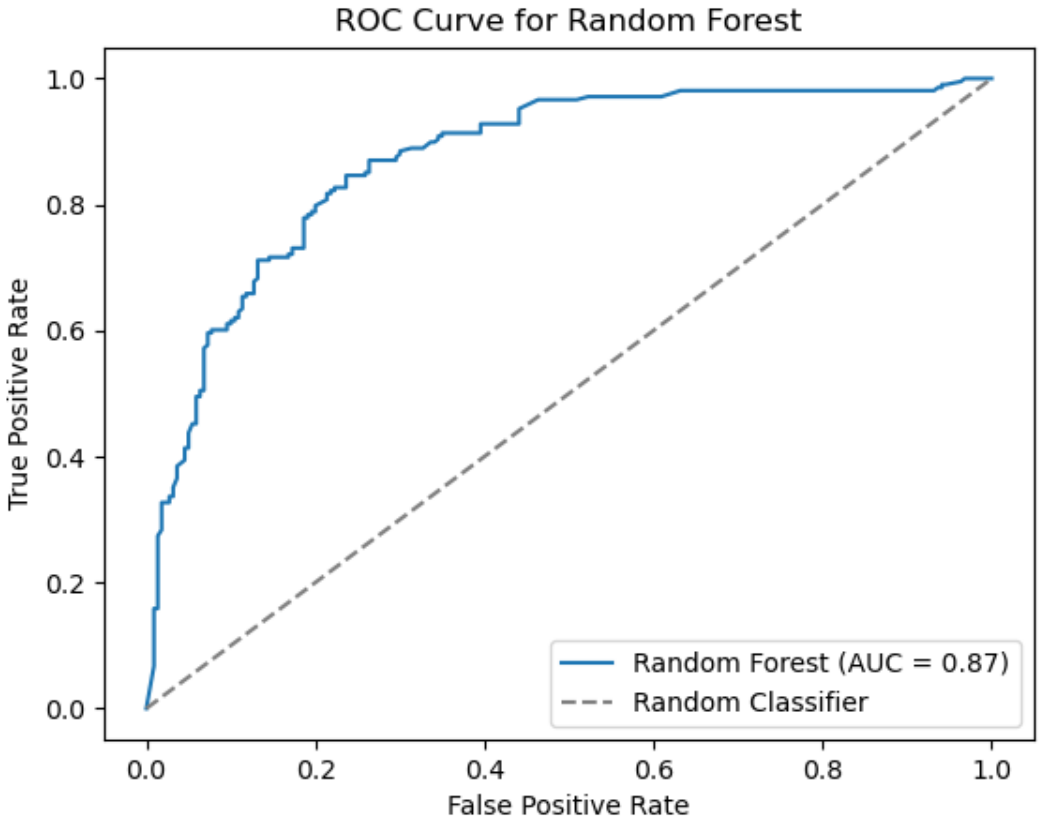

Figure 13: ROC Curve for Random Forest Model

The model outputs revealed the importance of the cardiac and non-cardiac clinical features that were associated with predicting ATTRwt-CM. Table 11 displays the feature importance for the International Classification of Disease (ICD) code-based cardiac and non-cardiac comorbidities and CPT code-based procedures used for predicting prevalence of ATTRwt-CM among heart failure patients. The strongest cardiac predictors included primary intrinsic cardiomyopathies, HFpEF, and First-degree AV block, whereas the strongest non-cardiac predictors included carpal tunnel syndrome, and synovitis/tenosynovitis, and ascites. The strongest combination of diagnoses included Combined systolic and diastolic Heart Failure + HFpEF, and Heart block + Cardiomegaly + HFpEF. The strongest procedure predictive of the disease was found to be Serum Test - Assay of free light chains; kappa and lambda with ratio or CPT 83883.

Features	XGBoost	Random Forest
Primary intrinsic cardiomyopathies	23.16%	19.61%
Carpal tunnel syndrome	19.48%	10.37%
Serum Test - Assay of free light chains; kappa and lambda with ratio	9.92%	6.22%
HFpEF	6.61%	7.40%
Cardiac MRI	3.76%	1.99%
Heart block,Cardiomegaly,HFpEF	2.98%	2.56%
Combined systolic and diastolic Heart Failure,HFpEF	2.71%	4.46%
HFrEF	2.41%	5.89%
First-degree AV block	2.40%	4.47%
Serum Test - Transthyretin (TTR) protein analysis	2.32%	2.08%
Secondary intrinsic cardiomyopathies	2.20%	6.19%
Atrial Fibrillation,Cardiomegaly,Joint disorders,Combined systolic and diastolic Heart Failure	2.11%	0.99%
Atrial Fibrillation,Joint disorders,HFpEF	2.10%	1.54%
Pericardial effusion/pericarditis	2.10%	4.43%
Synovitis and tenosynovitis	2.06%	2.74%
Atrial Fibrillation	2.03%	4.91%
Non-rheumatic heart valve disease	1.96%	5.07%
Heart block,Joint disorders,Combined systolic and diastolic Heart Failure	1.88%	0.78%
Heart block,Soft tissue disorders,HFpEF	1.83%	1.49%
Heart block,CKD,HFpEF	1.79%	2.04%
Atrial Fibrillation,Cardiomegaly,Soft tissue disorders,HFpEF	1.57%	1.96%
Cardiomegaly,Joint disorders,HFpEF	1.36%	1.66%
Heart block,Cardiomegaly,Joint disorders	1.23%	1.13%

Table 11: Feature importance for predicting ATTRwt-CM among HF patients

We then used nested cross validation for both machine learning models. In the outer loop of the nested cross validation, we used 5-fold cross validation to select the train and test set by shuffling the dataset. The then used 5-fold cross validation for a grid search algorithm to find out the best hyperparameters from our list. With the best set of hyperparameters the models were fit. Similar as before, in this case the Random Forest model had the highest mean accuracy of 79.14% (vs. 77.83% for XGBoost) as shown in Table 12. The recall, and F1 score found were better for Random Forest model (Table 12), while the precision (positive predictive value [PPV]) was better for XGBoost.

	XGBoost	Random Forest
Mean Accuracy:	77.83%	79.14%
Median Accuracy:	77.57%	79.39%
Mean Precision:	79.10%	78.12%
Median Precision:	79.43%	77.02%
Mean Recall:	77.01%	82.07%
Median Recall:	77.88%	82.11%
Mean F1 Score:	77.98%	80.00%
Median F1 Score:	79.22%	80.36%

Table 12: Performance Evaluation of The Prediction Models with Nested Cross Validation

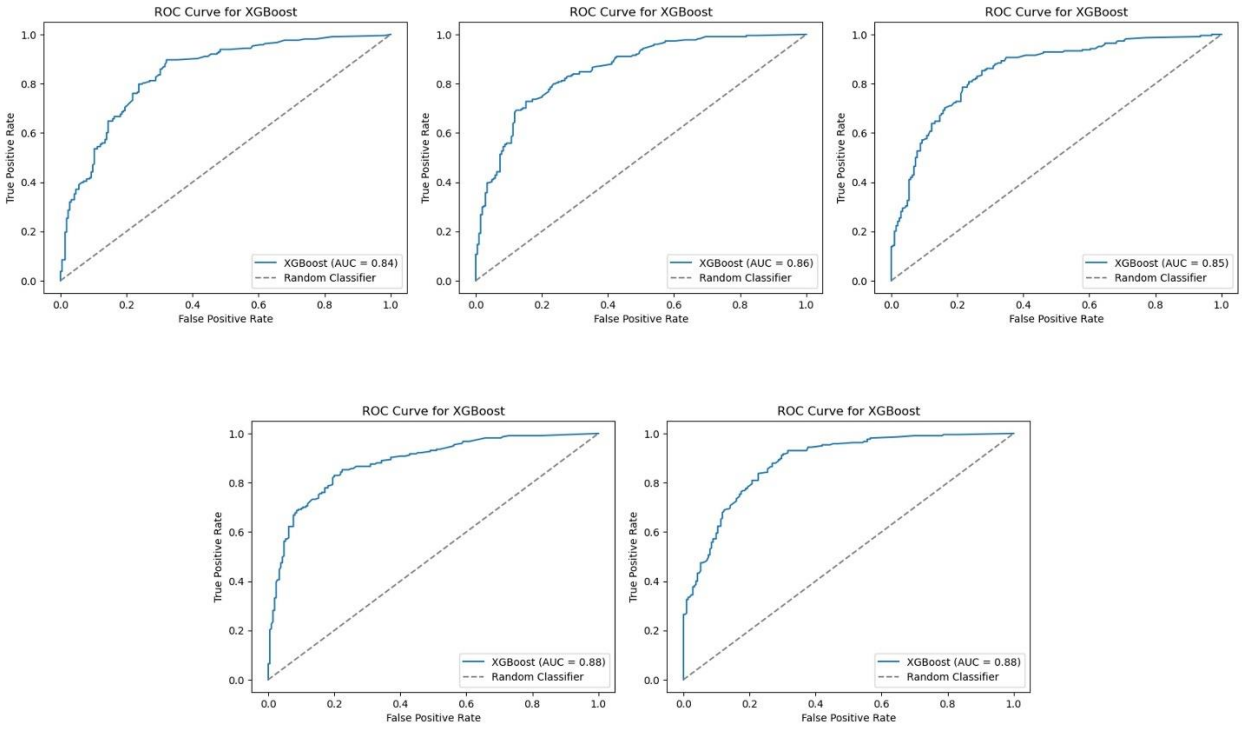

Figure 14: ROC Curves for XGBoost Model with 5-Fold Nested Cross Validation

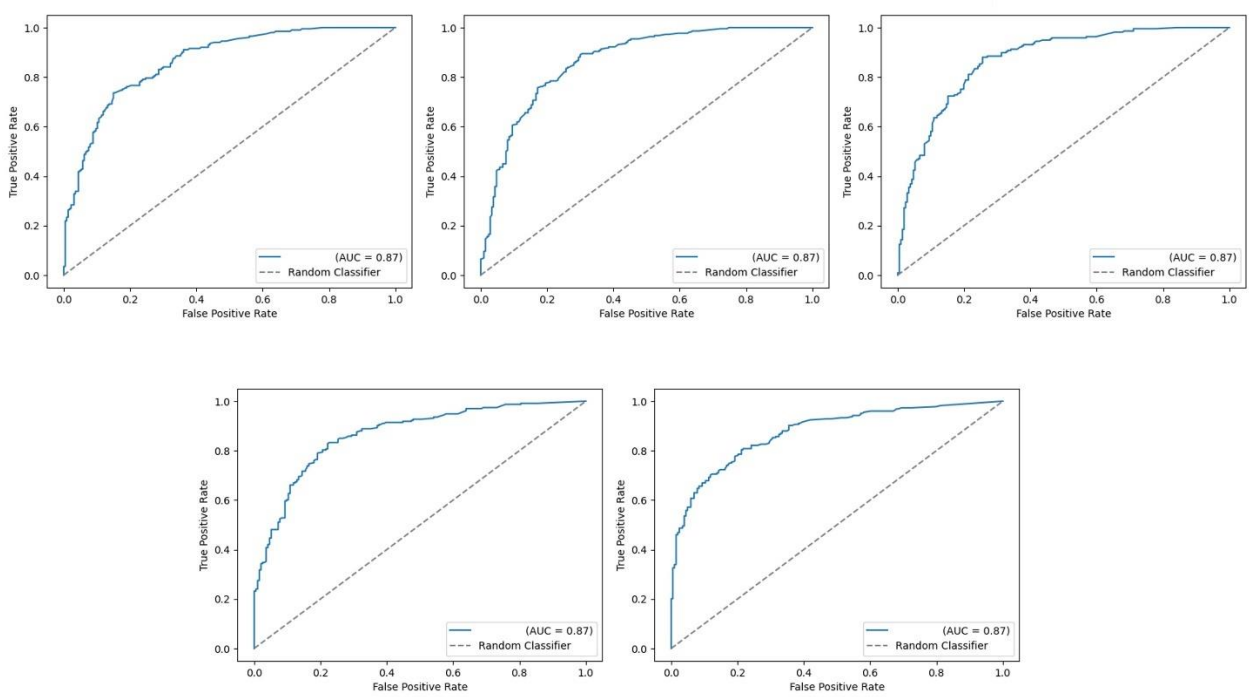

Figure 15: ROC Curves for Random Forest Model with 5-Fold Nested Cross Validation

Figures 14 and 15 are the ROC curve plots for the XGBoost and Random Forest models with 5 fold nested cross validation respectively. The AUROC for the XGBoost model had a mean of 0.862 and the AUROC for the Random Forest model had a mean of 0.87. From the plots it can be seen that both the models were good fit and had good discrimination ability with high accuracy.

Table 13 shows the mean of the feature importance of the cardiac and non-cardiac clinical features that were associated with predicting ATTRwt-CM. The strongest cardiac predictors included primary intrinsic cardiomyopathies, HFpEF, Secondary intrinsic cardiomyopathies, First-degree AV block, and HFrEF. he strongest non-cardiac predictors included carpal tunnel syndrome, and synovitis/tenosynovitis, and ascites. The strongest combination of diagnoses included Combined systolic and diastolic Heart Failure + HFpEF, and Heart block + Cardiomegaly + HFpEF. The strongest procedure predictive of the disease was found to be Serum Test - Assay of free light chains; kappa and lambda with ratio or CPT 83883.

Features	XGBoost	Random Forest
Primary intrinsic cardiomyopathies	33.98%	28.17%
Carpal tunnel syndrome	10.73%	13.77%
HFpEF	8.69%	11.04%
Serum Test - Assay of free light chains; kappa and lambda with ratio	7.43%	7.13%
Combined systolic and diastolic Heart Failure,HFpEF	5.18%	4.54%
Heart block,Cardiomegaly,HFpEF	4.90%	4.35%
HFrEF	4.50%	5.67%
Cardiac MRI	2.75%	1.06%
Atrial Fibrillation	2.56%	3.49%
First-degree AV block	2.34%	3.20%
Secondary intrinsic cardiomyopathies	2.08%	6.46%
Atrial Fibrillation,Cardiomegaly,Joint disorders,Combined systolic and diastolic Heart Failure	1.67%	0.53%
Synovitis and tenosynovitis	1.63%	1.10%
Pericardial effusion/pericarditis	1.55%	1.55%
Atrial Fibrillation,Joint disorders,HFpEF	1.51%	0.88%
Heart block,Soft tissue disorders,HFpEF	1.35%	0.73%
Serum Test - Transthyretin (TTR) protein analysis	1.35%	0.57%
Non-rheumatic heart valve disease	1.29%	1.54%
Cardiomegaly,Joint disorders,HFpEF	1.08%	0.90%
Atrial Fibrillation,Cardiomegaly,Soft tissue disorders,HFpEF	1.01%	1.11%
Heart block,CKD,HFpEF	0.93%	0.91%
Heart block,Joint disorders,Combined systolic and diastolic Heart Failure	0.76%	0.58%
Heart block,Cardiomegaly,Joint disorders	0.71%	0.74%

Table 13: Mean of feature importance for XGBoost and Random Forest models with nested cross validation

To further validate our model, we used a larger dataset for the prediction model with 1:2 ratio for the case and control cohort. We used nested cross validation for both machine learning models in this case as well to select the train and test set by shuffling the dataset and to find out the best hyperparameters. With the best set of hyperparameters the models were fit. In this case the XGBoost model had the highest mean accuracy of 81.11% (vs. 80.38% for Random Forest) as

shown in Table 14. The precision, recall, and F1 score found were better for XGBoost model (Table 14) compared to the Random Forest model.

	XGBoost	Random Forest
Mean Accuracy:	81.11%	80.38%
Median Accuracy:	81.42%	80.79%
Mean Precision:	75.15%	74.98%
Median Precision:	76.82%	74.47%
Mean Recall:	67.48%	64.93%
Median Recall:	67.27%	64.65%
Mean F1 Score:	70.98%	69.49%
Median F1 Score:	70.79%	70.00%

Table 14: Performance Evaluation of The Prediction Models with 1:2 case and control cohort ratio

Features	XGBoost	Random Forest
Primary intrinsic cardiomyopathies	45.50%	30.40%
Carpal tunnel syndrome	8.32%	16.06%
Serum Test - Assay of free light chains; kappa and lambda with ratio	7.74%	11.55%
HFpEF	7.46%	7.58%
Combined systolic and diastolic Heart Failure,HFpEF	6.62%	5.18%
HFrEF	3.89%	4.79%
Cardiac MRI	3.51%	1.84%
First-degree AV block	2.47%	1.96%
Secondary intrinsic cardiomyopathies	2.28%	5.06%
Non-rheumatic heart valve disease	1.43%	1.85%
Atrial Fibrillation	1.34%	1.71%
Heart block,Cardiomegaly,HFpEF	1.30%	2.48%
Pericardial effusion/pericarditis	1.15%	1.35%
Serum Test - Transthyretin (TTR) protein analysis	1.14%	0.53%
Atrial Fibrillation,Joint disorders,HFpEF	0.78%	0.75%
Heart block,Soft tissue disorders,HFpEF	0.75%	0.73%
Atrial Fibrillation,Cardiomegaly,Soft tissue disorders,HFpEF	0.73%	1.00%
Heart block,Joint disorders,Combined systolic and diastolic Heart Failure	0.68%	0.47%
Cardiomegaly,Joint disorders,HFpEF	0.66%	0.99%
Synovitis and tenosynovitis	0.65%	1.01%
Atrial Fibrillation,Cardiomegaly,Joint disorders,Combined systolic and diastolic Heart Failure	0.59%	0.63%
Heart block,Cardiomegaly,Joint disorders	0.54%	0.51%
Heart block,CKD,HFpEF	0.47%	1.55%

Table 15: Mean of feature importance for XGBoost and Random Forest models with 1:2 case and control cohort ratio

From the mean of feature importance, it can be observed that the rank of the importance features remained almost the same for the smaller and the larger dataset. This shows consistency of our model.

Table 16 and Table 17 shows the performance of the models when taking patient data until 1, 2, 3, and 4 years before their ATTRwt-CM diagnosis. 5-fold nested cross validation was used for the XGBoost and Random Forest models for the prediction in each previous year. It can be observed that for both models the performance dropped significantly in the previous years.

	1 Year Prior	**2 Year Prior**	**3 Year Prior**	**4 Year Prior**
Accuracy	65.21%	66.44%	66.06%	68.93%
Precision	66.98%	68.01%	64.40%	65.45%
Recall	46.98%	36.61%	26.78%	21.09%
F1 Score	55.10%	47.46%	37.58%	31.76%

Table 16: XGBoost model performance for predicting antecedent test group data

	1 Year Prior	**2 Year Prior**	**3 Year Prior**	**4 Year Prior**
Accuracy	65.78%	67.06%	67.12%	70.37%
Precision	68.23%	71.62%	74.65%	77.74%
Recall	47.44%	34.17%	22.17%	19.50%
F1 Score	55.72%	46.18%	33.99%	31.03%

Table 17: Random Forest model performance for predicting antecedent test group data

Procedures dropped for taking data up to 1 year before ATTRwt-CM diagnosis	779
Procedures dropped for taking data up to 2 years before ATTRwt-CM diagnosis	1010
Procedures dropped for taking data up to 3 years before ATTRwt-CM diagnosis	1148
Procedures dropped for taking data up to 4 years before ATTRwt-CM diagnosis	1245

Table 18: Numbers of procedures dropped from the dataset while going back in time

Table 18 shows the total number of the 3 procedures dropped while preparing the data for this analysis. From the table it is evident that the model performance dropped for the test data used after dropping the features. While the total number of occurrences of the procedures was 1436, more than 50% of those were done within 1 year before the ATTRwt-CM diagnosis of the patients. This relates to the drastic drop in the accuracy of the trained model on the new test set.

Features	1 Year Prior	2 Year Prior	3 Year Prior	4 Year Prior
Carpal tunnel syndrome	29.59%	27.55%	17.31%	9.91%
HFpEF	5.20%	5.71%	9.40%	17.62%
Synovitis and tenosynovitis	6.19%	6.74%	10.18%	8.05%
Primary intrinsic cardiomyopathies	8.54%	5.84%	5.32%	3.30%
HFrEF	2.81%	3.68%	7.48%	24.31%
Non-rheumatic heart valve disease	4.91%	4.65%	5.26%	3.27%
Pericardial effusion/pericarditis	2.02%	5.25%	4.87%	3.73%
Secondary intrinsic cardiomyopathies	2.77%	3.74%	5.20%	4.48%
Serum Test - Assay of free light chains; kappa and lambda with ratio	3.06%	3.71%	2.94%	2.02%
Cardiac MRI	4.12%	3.59%	2.37%	0.86%
Heart block,CKD,HFpEF	2.23%	2.93%	2.63%	2.92%
Combined systolic and diastolic Heart Failure,HFpEF	2.94%	2.97%	2.36%	2.57%
Heart block,Cardiomegaly,Joint disorders	3.00%	2.90%	2.39%	1.30%
Serum Test - Transthyretin (TTR) protein analysis	2.22%	2.28%	3.46%	2.88%
Heart block,Soft tissue disorders,HFpEF	2.70%	2.27%	2.85%	1.64%
Atrial Fibrillation,Joint disorders,HFpEF	2.04%	3.19%	2.49%	2.18%
First-degree AV block	3.10%	2.16%	2.37%	1.17%
Atrial Fibrillation	3.77%	2.25%	1.97%	1.72%
Atrial Fibrillation,Cardiomegaly,Soft tissue disorders,HFpEF	1.68%	2.19%	2.94%	1.27%
Cardiomegaly,Joint disorders,HFpEF	1.99%	1.85%	2.35%	1.61%
Heart block,Cardiomegaly,HFpEF	2.19%	1.62%	1.00%	1.67%
Atrial Fibrillation,Cardiomegaly,Joint disorders,Combined systolic and diastolic Heart Failure	2.00%	1.15%	1.63%	0.91%
Heart block,Joint disorders,Combined systolic and diastolic Heart Failure	0.92%	1.76%	1.22%	0.61%

Table 19: XGBoost model feature importance for predicting antecedent test group data

From the feature importance for the XGBoost model prediction on the previous years' data, changes in the top contributors of the full data can be observed. One of the major highlights of Table 19 is the important procedure for early analysis in the full data was "Serum Test - Assay of free light chains; kappa and lambda with ratio" has dropped down the list. The decrease in importance of this procedure indicates the decrease in number of this procedure in the previous years and it supports the fact that this procedure directly contributes to the early diagnosis of the disease. Table 20 shows the similar case for Random Forest model prediction. In both the tables similarity can be observed in the feature importance in every prior year.

Features	1 Year Prior	2 Year Prior	3 Year Prior	4 Year Prior
Carpal tunnel syndrome	27.39%	28.14%	18.90%	12.64%
Synovitis and tenosynovitis	8.69%	12.10%	13.77%	11.97%
HFrEF	4.31%	6.94%	15.98%	23.34%
HFpEF	5.66%	6.26%	9.52%	17.47%
Primary intrinsic cardiomyopathies	14.29%	7.09%	4.84%	3.52%
Non-rheumatic heart valve disease	5.11%	5.76%	6.93%	5.54%
Secondary intrinsic cardiomyopathies	4.54%	4.89%	5.27%	3.19%
Atrial Fibrillation	4.60%	3.20%	2.95%	4.23%
Pericardial effusion/pericarditis	1.98%	3.64%	3.88%	3.62%
First-degree AV block	3.27%	2.70%	2.53%	2.17%
Serum Test - Assay of free light chains; kappa and lambda with ratio	2.99%	3.65%	2.15%	0.95%
Cardiomegaly,Joint disorders,HFpEF	1.61%	1.92%	1.89%	1.91%
Combined systolic and diastolic Heart Failure,HFpEF	2.50%	1.94%	1.57%	1.73%
Atrial Fibrillation,Joint disorders,HFpEF	1.69%	1.84%	1.79%	1.03%
Heart block,CKD,HFpEF	1.07%	1.30%	1.79%	2.10%
Heart block,Cardiomegaly,Joint disorders	1.67%	1.85%	1.12%	0.46%
Atrial Fibrillation,Cardiomegaly,Soft tissue disorders,HFpEF	1.31%	1.25%	1.24%	0.81%
Heart block,Soft tissue disorders,HFpEF	1.32%	1.33%	1.10%	1.02%
Serum Test - Transthyretin (TTR) protein analysis	0.88%	1.09%	1.02%	1.23%
Heart block,Cardiomegaly,HFpEF	2.29%	1.17%	0.92%	0.81%
Cardiac MRI	1.38%	0.89%	0.21%	0.01%
Atrial Fibrillation,Cardiomegaly,Joint disorders,Combined systolic and diastolic Heart Failure	0.90%	0.46%	0.52%	0.20%
Heart block,Joint disorders,Combined systolic and diastolic Heart Failure	0.56%	0.60%	0.12%	0.07%

Table 20: Random Forest model feature importance for predicting antecedent test group data

7. DISCUSSION

We created a machine learning prediction model for ATTRwt-CM using EHR data from a large national database. We made use of ATTRwt-CM-related procedures, and combinations of cardiac and non-cardiac diagnoses that have been previously documented in the literatures. The results of this study can be used for early detection of ATTRwt-CM patients, and earlier therapy if successfully implemented inside the EHR of healthcare systems. For our model we picked the top 10 comorbidities and 10 combinations of the comorbidities from the literature of Huda, A., Castaño, A., Niyogi, A. et al. [6]. We also took all the procedures used in medical practices for diagnosis of ATTR-CM. Our goal was to establish a relation between these important features and diagnosis ATTRwt-CM earlier by using it. If the results are applied to the EHR system which contains all patient records, the patients with matching comorbidities and their combinations can be identified to be at risk for ATTRwt-CM and the procedures from this study can be done earlier to lead to definitive diagnosis of the disease.

The main prediction of this study matches with the results of previous literatures. In our study, we found all the top comorbidities and their combinations to be contributing factors in the prediction of the disease. We found primary intrinsic cardiomyopathy, carpal tunnel syndrome, HFpEF to be the morbidities with highest importance in the prediction models. Patients having these morbidities in their record can be considered to be at risk of ATTRwt-CM. One of the main reasons for the diagnosis delay for the disease is the requirement of invasive tests for the definitive diagnosis. But now a combination of non-invasive tests can be used to identify it. In our study we first identified the procedures which can detect the presence of this disease in less than 90 days, i.e., from test day to result. By using the top 3 findings from the statistical analyses we found that non-invasive tests Serum Test - Assay of free light chains; kappa and lambda with ratio, and Cardiac MRI are the important procedures for the prediction model. So, these procedures can be performed on the patients with the predictive comorbidities which can result in early diagnosis.

We also did prediction for the same train and test sets with dataset limited to 1, 2, 3, and 4 years prior to their ATTR-CM diagnosis to show the change in the important factors in predicting the outcome. The prediction accuracy and precision dropped significantly for years prior to the diagnosis. For prediction using XGBoost, though the comorbidities and their combinations had almost similar importance as the main prediction model, the importance of the non-invasive tests

dropped the most. The reason for this is the reduced number of the tests performed in the years prior to the diagnosis. From this observation we can establish a direct relation between the higher accuracy of prediction and the presence of the procedures for the patients. And thus, it can be realized that these procedures can contribute to the early diagnosis of the disease.

One of the strengths of this study was that it was based on patient data from a large national EHR system which consisted of more than 2.5 million heart failure patients from 63 health care organizations from North America, among which 1531 patients were diagnose with ATTRwt-CM after their heart failure diagnosis. Therefore, we chose heart failure to be the exact symptom of ATTRwt-CM in our study, which was not the case for many of the previous literatures. To choose the important features for our statistical and prediction models, we used previously established literatures and clinical findings. We also considered different code systems available for procedures and diagnoses for our analyses. This study can be used in the EHR systems to find out the patients who have the combinations of the comorbidities in their clinical history and can be tested earlier in the progress of the disease using the important non-invasive tests found here.

Our study also has some limitations. First, ICD codes were used to identify the case and control patients, and this imposes constraints on our investigation as HF and ATTRwt-CM can be incorrectly classified in the patient file. The model was built using an ICD code that only applies to ATTR-CM of the wild type. As a result of this diagnostic code's recent development and potential for inconsistent application, our model's applicability may be constrained by biases created by unique, institutional, or regional ICD coding practices. Additionally, control patients with HF who were not given a diagnosis can still have wild-type ATTR-CM. Having records from a large number of national health organizations in our datasets, and the long history of the patients it can be ruled that the selected diagnosis codes do not capture all the phenotypes. As a result, it is plausible that cardiac amyloidosis exists in some of the non-amyloid HF controls but has not yet been recognized. Secondly, we lack information on electrocardiographic voltage or echocardiographic markers as well as additional laboratory data which could contribute to the assignment of cases and controls. And lastly, some of the ATTRwt-CM diagnosed patients had missing or inconsistent data which forced us to drop them from the study. A larger cohort could give a more dependable analysis if used for our model.

8. CONCLUSION

ATTRwt-CM is a specific cause of HF which is associated with high morbidity and mortality. This disease is often misinterpreted as other cardiac illnesses and kinds of heart failure. Since ATTRwt-CM is more widespread than previously thought, it is crucial to identify it in time. Addressing the diagnosis delay has become essential because there is now a disease-modifying medication for ATTR-CM that has demonstrated treatment responsiveness for earlier treatment. Non-invasive investigations can also flag the potential presence of this condition and are essential to a thorough diagnosis.

In the medical field, an increasing number of machine learning (ML) models have been developed to predict diseases and phenotypes using data from medical claims. ML is particularly beneficial as it enables the identification of more individuals who may have diseases like wild-type transthyretin amyloid cardiomyopathy (ATTRwt-CM). This approach is advantageous over statistical methods because it can efficiently analyze the complex relationships among the diverse input predictors, whereas this task would be time-consuming if traditional statistical methods were used.

To create a prediction model for ATTRwt-CM, we utilized electronic health record (EHR) data from a large national database. EHR data provides more comprehensive information compared to claims data. Our model incorporated ATTRwt-CM-related diagnoses, combinations of cardiac and non-cardiac diagnoses, and techniques that have been previously documented in the literature. Implementing this model within healthcare systems' EHRs could potentially result in focused testing, early detection of ATTRwt-CM patients, and earlier initiation of therapy if successful.

Machine learning models tend to perform optimally when applied to the data they were trained on. In our case, when the model was applied to the entire EHR dataset, it demonstrated favorable overall performance, achieving an accuracy of 81.61% with the XGBoost algorithm. Notably, the Random Forest model performed even better within this cohort, achieving an accuracy of 81.07%.

Our research findings regarding the cardiac and non-cardiac phenotypes indicatory of wild-type ATTR-CM align with previous results in the literature and clinical practice related to this disease. The existence of comorbidities and their combinations for several years prior to the patients' diagnosis of ATTRwt-CM emphasizes the potential for earlier detection.

The findings presented here have clinical implications not only for wild-type ATTR-CM, but also for other diseases that are misdiagnosed or underdiagnosed. ICD being the worldwide language of medical diagnoses and easy to utilize in the EHR system, the methodology that we used has widespread applicability. It can give a systematic outline for realizing varied sets of signs and symptoms, particularly in the case of rare or under-recognized diseases. In our case, patients having these confirming findings in their regular clinical diagnoses should subsequently undertake confirmatory non-invasive diagnostic testing (for example, bone scintigraphy).

The generation of a statistical and prediction model for wild-type ATTR-CM from a sizeable EHR database is one of the study's strengths. Additionally, the use of ICD codes for phenotype mapping enabled us to look at the correlated symptoms and procedures that might suggest and precede wild-type ATTR-CM diagnosis. ML models have the capacity to identify patterns suggestive of the disease automatically. These patterns may not be obvious to the clinicians as ML models usually derive these using data from diverse diagnostic codes across different disease domains and organ systems.

When considering the results of our study, it is important to acknowledge several limitations. Firstly, the utilization of ICD codes to identify patients with heart failure (HF) imposes certain constraints on our investigation. HF is a clinical condition with a broad definition, and there is a possibility of misclassification in the patient records. Additionally, we lack crucial information such as electrocardiographic voltage, echocardiographic markers, and additional laboratory data, which could have aided in accurately assigning cases and controls. Given the nature of the datasets utilized in our analysis and the extensive medical history of the patients, it is evident that the diagnosis codes do not capture all the phenotypes. Consequently, it is plausible that some of the non-amyloid HF controls may actually have cardiac amyloidosis that has not yet been recognized.

The model was developed based on ICD code E85.82 which is specific to wild-type transthyretin amyloid cardiomyopathy (ATTR-CM). However, the recent development of this diagnostic code and the potential for inconsistent application may introduce biases stemming from unique organizational or regional ICD coding practices, thus restricting the applicability of our model. Furthermore, control patients with HF who were not given a diagnosis could still have wild-type ATTR-CM. The different forms of cardiac amyloidosis may exhibit overlapping cardiac and extracardiac symptoms.

This model could serve as an initial step in identifying individuals at risk who may require further evaluation using techniques such as cardiac magnetic resonance imaging (cardiac MRI), speckle-tracking echocardiography, bone scintigraphy, and blood tests for ATTRwt-CM. However, it is crucial not to interpret it as definitive evidence for the diagnosis of either ATTR-CM or cardiac amyloidosis.

9. REFERENCES

[1] Groft, Stephen C., Manuel Posada, and Domenica Taruscio. "Progress, challenges and global approaches to rare diseases." Acta Paediatrica 110.10 (2021): 2711-2716.

[2] Graf von der Schulenburg, J-Matthias, and Martin Frank. "Rare is frequent and frequent is costly: rare diseases as a challenge for health care systems." The European Journal of Health Economics 16 (2015): 113-118.

[3] Pande, Monu, and Ragini Srivastava. "Molecular and clinical insights into protein misfolding and associated amyloidosis." European Journal of Medicinal Chemistry 184 (2019): 111753.

[4] Rozenbaum, Mark H., et al. "Estimating the health benefits of timely diagnosis and treatment of transthyretin amyloid cardiomyopathy." Journal of Comparative Effectiveness Research 10.11 (2021): 927-938.

[5] AbouEzzeddine OF, Davies DR, Scott CG, et al. Prevalence of Transthyretin Amyloid Cardiomyopathy in Heart Failure with Preserved Ejection Fraction. *JAMA Cardiol.* 2021;6(11):1267–1274.

[6] 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 12, 2725 (2021).

[7] Lauppe RE, Liseth Hansen J, Gerdesköld C, et al. Nationwide prevalence and characteristics of transthyretin amyloid cardiomyopathy in Sweden. Open Heart 2021;8:e001755.

[8] Lane, Thirusha, et al. "Natural history, quality of life, and outcome in cardiac transthyretin amyloidosis." Circulation 140.1 (2019): 16-26.

[9] Rozenbaum, M.H., Large, S., Bhambri, R. et al. Impact of Delayed Diagnosis and Misdiagnosis for Patients with Transthyretin Amyloid Cardiomyopathy (ATTR-CM): A Targeted Literature Review. Cardiol Ther 10, 141–159 (2021).

[10] Witteles, Ronald M., et al. "Screening for transthyretin amyloid cardiomyopathy in everyday practice." JACC: Heart Failure 7.8 (2019): 709-716.

[11] Mitchell, Joshua D., et al. "Implementing a Machine-Learning-Adapted Algorithm to Identify Possible Transthyretin Amyloid Cardiomyopathy at an Academic Medical Center." Clinical Medicine Insights: Cardiology 16 (2022): 11795468221133608.

[12] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[13] 'Introduction to Logistic Regression' (2019). *Towards Data Science.* https://towardsdatascience.com/introduction-to-logistic-regression-66248243c14 Available at (Accessed at 21 April 2023)

[14] Walczak, Beata, and Desire Massart. "calibration in wavelet domai." In: Wavelets chemistry (Ed. Walczak), pg. 323-349. 2000.

[15] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-2808.

[16] Datta, Shounak, Vikrant A. Dev, and Mario R. Eden. "Developing QSPR for predicting DNA drug binding affinity of 9-Anilinoacridine derivatives using correlation-based adaptive LASSO algorithm." Computer Aided Chemical Engineering. Vol. 40. Elsevier, 2017. 2767-2772.

[17] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003 Jul 21;89(2):232-8. doi: 10.1038/sj.bjc.6601118. PMID: 12865907; PMCID: PMC2394262.

[18] Wang, Yuanchao, et al. "A hybrid ensemble method for pulsar candidate classification." *Astrophysics and Space Science* 364 (2019): 1-13.

[19] 'Understanding Random Forest' (2019). Towardsdatascience. Available at https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (Accessed at 10 July 2023)

[20] 'What is Random Forest' (2020). *TIBCO.* Available at https://www.tibco.com/reference-center/what-is-a-random-forest (Accessed at 10 July 2023)

[21] Maurer, Mathew S., et al. "Expert consensus recommendations for the suspicion and diagnosis of transthyretin cardiac amyloidosis." Circulation: Heart Failure 12.9 (2019): e006075.

[22] Gertz, Morie, et al. "Avoiding misdiagnosis: expert consensus recommendations for the suspicion and diagnosis of transthyretin amyloidosis for the general practitioner." BMC family practice 21 (2020): 1-12.