Predicting Chronic Heart Failure Using Diagnoses Graphs

Saurabh Nagrecha, Pamela Bilo Thomas, Keith Feldman, Nitesh V. Chawla

Predicting Chronic Heart Failure Using Diagnoses Graphs

Saurabh Nagrecha[1], Pamela Bilo Thomas[1,2], Keith Feldman[1], and Nitesh V. Chawla[1,2]

[1] Department of Computer Science and Engineering,
Interdisciplinary Center for Network Science and Applications (iCeNSA),
University of Notre Dame, Notre Dame IN 46556, USA,
{snagrech,pthomas4,kfeldman,nchawla}@nd.edu
[2] Indiana Biosciences Research Institute,
1345 W 16th St #300, Indianapolis, IN 46202, USA

Abstract. Predicting the onset of heart disease is of obvious importance as doctors try to improve the general health of their patients. If it were possible to identify high-risk patients before their heart failure diagnosis, doctors could use that information to implement preventative measures to keep a heart failure diagnosis from becoming a reality. Integration of Electronic Medical Records (EMRs) into clinical practice has enabled the use of computational techniques for personalized healthcare at scale. The larger goal of such modeling is to pivot from reactive medicine to preventative care and early detection of adverse conditions. In this paper, we present a trajectory-based disease progression model to detect chronic heart failure. We validate our work on a database of Medicare records of 1.1 million elderly US patients. Our supervised approach allows us to assign likelihood of chronic heart failure for an unseen patient's disease history and identify key disease progression trajectories that intensify or diminish said likelihood. This information will be a tremendous help as patients and doctors try to understand what are the most dangerous diagnoses for those who are susceptible to heart failure. Using our model, we demonstrate some of the most common disease trajectories that eventually result in the development of heart failure.

Keywords: heart failure, cardiovascular disease, directed acyclic graph, Medicare, EMR, health care

1 Introduction

Today the healthcare industry finds itself at the precipice of a significant change, as the past decade has seen the adaption and integration of electronic medical records (EMR) into clinical practice. Beyond the logistical benefits of maintaining and organizing patients' medical data, clinicians and researchers can perform novel research using these secondary data sources [1–3]. In fact EMRs ability to provide a computationally accessible set of structured data representing the expansive healthcare feature space has fueled the emergence of a sundry

of informatics tools ranging from early clinical decision support systems, to the statistical analysis of, to predictive analytics aimed at identifying patients at risk for readmission [4, 5].

Building on the success of these tools, many researchers have seen healthcare informatics as the junction between another line of parallel clinical research, the shift from reactive to preventative medicine. Medical research is itself an evolving field, and has advanced in parallel with the emergence of EMR. Clinicians have put forth a strong effort in advancing the care paradigm from reactive medicine, where clinicians treat the conditions currently afflicting a patient, to preventative care where clinicians undertake courses of action "for the purpose of preventing disease or detecting it in an asymptomatic stage" [6]. As such, the early detection and treatment of adverse health conditions represents an exciting opportunity for the informatics community. Others have found that a combination of research areas, including, but not limited to, graph-based data mining, entropy-based data mining, and topological-based data mining, work best for knowledge discovery and towards an end goal of supplementing human learning with machine learning [7]. Eventually the goal is to have P4-medicine (predictive, preventative, participatory, personalized) available for all patients by using big data and the combined human computer interaction and knowledge discovery/data mining approach [7].

A number of works have built on this foundation, focusing predictive tasks from disease prediction, to the prediction of breast cancer survivability [8, 9]. However, these tools suffer from a fundamental flaw, they identify patients' health conditions as isolated events, i.e. a disease will occur in a patient's future medical chart, or a patient will recover from early stage breast cancer. One must remember that an individuals' health condition does not only consist of when doctors measure them in a clinical environment. Although the rate of onset may vary, the progression of disease represents a highly fluid state. As such, it may be more valuable to view these patients' conditions as trajectories, rather than binary events.

While this seems like a significant shift in thinking, medical subfields have already established the concept of a disease trajectory, sometimes denoted as disease 'progression'. In particular, research in relation to neurodegenerative disorders such as Parkinson's and Alzheimer's have quite well established this concept [10, 11]. More recently, the trajectory concept has begun expanding into the general healthcare population. Many clinicians have long postulated that an underlying progression of related diagnoses may relate to diagnoses for which we do not explicitly relate a temporal aspect. Today, the data collected through the expanding EMR now allows for researchers to examine such hypotheses in detail. Perhaps Jensen et. al, have provided one of the best examples to date, where through their work they successfully extracted diagnosis trajectories by analyzing millions of longitudinal patient records and utilized a novel way of describing biological disease progression [12].

In this work, we build on this concept and present a novel graph-based diagnosis trajectory model. While recent advances have taken what effectively

represent an "unsupervised" approach to trajectory discovery, we aim to provide a target based "supervised" methodology. We will begin with a discussion of the underlying methodology used in constructing the underlying diagnosis graph. From here, we will discuss utilizing the temporal relations extracted from the graph, showing that we can identify paths that significant differentiate the occurrence of the target diagnosis. Finally, we will provide a case study of the methodology in relation to patients with congestive heart failure.

2 Data Description

Electronic Medical Records (EMRs) log information on patients in the form of diagnosis codes for each of their visits. This log effectively narrates a patient's medical history as identified by medical practitioners and can predict their future health outcomes. Here we describe our data source, the data structure, how chronic heart failure appears these diagnoses logs and how prevalent it is within our patients.

Provenance Our data comes from the Medicare records of 1,145,541 elderly patients in the United States. The accuracy and completeness of these records makes them invaluable to demographic and epidemiological research [13, 14, 9]. The data is completely anonymized — both in terms of the patients and the healthcare providers. For a given patient, we applied a threshold of a maximum of 5 in-patient visit, and each visit corresponds to a maximum of 10 diagnosis codes from the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM). These ICD-9-CM codes are designed to convey an intrinsic hierarchy of diagnosis detail— the full 5-digit code represents the specific condition, location and/or severity, and its leading 3 digits represent the medical diagnosis family. This "code collapse" [9] helps us identify the family of patients who develop heart failure in our data.

Identifying heart failure. We observe ground truth evidence of presence/absence of heart failure with ICD-9-CM diagnoses for individual patients. Diagnoses represented by the family of '428.xy' codes cover all diagnoses for heart failure. Specific examples of the 428 diagnosis family include *Systolic heart failure* (428.2); which breaks down into *Systolic heart failure, unspecified* (428.20), *Acute systolic heart failure* (428.21), *Chronic systolic heart failure* (428.22) and *Acute on chronic systolic heart failure* (428.23). We labeled as 'HF' all patients for whom we observed the '428' diagnosis family, and labeled the rest as 'NHF' for heart failure and non-heart failure respectively.

Table 1 shows a sample patient's medical history. Here we see the chronological history of the patient through each successive visit expressed in terms of full ICD-9-CM codes. For each visit, the first code is the principal diagnosis, followed by any secondary diagnoses made during that visit. The data presents these diagnoses in their full ICD-9-CM form where 733.00 represents *Osteoporosis, unspecified*. Some diagnoses, such as *Pathologic fracture* (733.1) and *Nutritional marasmus* (261), use fewer than the maximum 5 digits in ICD-9-CM.

Table 1: **Example Patient History**: Each row represents a distinct visit in chronological order. In each visit, the data shows multiple ICD-9-CM code diagnoses for our example patient. Note that the code for heart failure (428.0) appears in the fifth visit.

Visit	Vector of ICD-9-CM Disease Codes
1	7331 (Pathologic fracture, unspecified site), 73300 (Osteoporosis, unspecified), 2761 (Hyposmolality and/or hyponatremia), 4928 (Other emphysema), 73743 (Scoliosis associated with other conditions)
2	7331 (Pathologic fracture, unspecified site), 73300 (Osteoporosis, unspecified), 73741 (Kyphosis associated with other conditions), 73743 (Scoliosis associated with other conditions), 261 (Nutritional marasmus)
3	7331 (Pathologic fracture, unspecified site), 73300 (Osteoporosis, unspecified), 73741 (Kyphosis associated with other conditions), 73743 (Scoliosis associated with other conditions), 261 (Nutritional marasmus)
4	485 (Bronchopneumonia, organism unspecified), 2765 (Volume depletion disorder), 2769 (Electrolyte and fluid disorders not elsewhere classified), 496 (Chronic airway obstruction, not elsewhere classified),73300 (Osteoporosis, unspecified)
5	48230 (Pneumonia due to Streptococcus, unspecified) **4280 (Heart failure)**, 5119 (Unspecified pleural effusion), 2761 (Hyposmolality and/or hyponatremia), 2768 (Hypopotassemia), 73300 (Osteoporosis, unspecifie), 73741 (Kyphosis associated with other conditions), 7331 (Pathologic fracture, unspecified site)

Using Table 1 as an example, we see that in visit #5, the patient was diagnosed with *Congestive heart failure, unspecified* (428.0) and therefore belongs to the class 'HF'.

Summary Statistics. The EMR data used in this study covers 1,145,541 elderly Medicare patients over the course of 5,727,705 total visits. Over the course of these visits, the patients registered a total of 12,396 unique ICD-9-CM diagnoses codes, which represent 1,064 families of 3 digit collapsed codes. This set of patients exhibits a heart failure rate of 46.6%, which is extremely high compared to the United States, about 5.7 million (2.2%) adults have heart failure [15]. However, some have observed the overall prevalence of heart failure in elderly patients in the United States as high as 10.6 to 13.5% (Chart 20-2 [15]). Since our study focuses on patients on Medicare, this number is further amplified.

Experiments. Based on this EMR data, we group our analysis into two distinct phases— 1) building a representational model for heart failure and 2) predicting heart failure outcomes for unseen patients. First, we infer the nature of disease progression for patients with and without observed heart failure. Based on the learned model, we identify trajectories, individual diagnoses, and edges that give the best indication of heart failure. We then use this model on previously unseen

Table 2: **Preprocessed Example Patient History**. Using the same example data as Table 1, we derive a compact history of the patient. As a result of this preprocessing, we arrive at a set of input diagnoses (visits #1 to 4) to create a trajectory and decouple it from the labeled outcome (visit #5). Note that the diagnosis has been collapsed to its family

Visit #	ICD-9-CM codes
1	733 (Other disorders of bone and cartilage), 276 (Disorders of fluid electrolyte and acid-base balance), 492 (Emphysema), 737 (Curvature of spine)
2	261 (Nutritional marasmus)
3	NA
4	485 (Bronchopneumonia, organism unspecified), 496 (Chronic airway obstruction, not elsewhere classified)
5	**428 heart failure**

patients and predict whether they will develop heart failure and validate this against observed ground truth data about these test patients.

3 Building a Representational Predictive Model

Researchers have built contemporary disease progression models using patient data with already known target outcomes. [10, 11]. In contrast, our supervised approach helps contextualize disease progression trajectories against an in-situ control set of patients, i.e. our data contains trajectories followed by patients who eventually were diagnosed with Heart Failure and those who were not. This approach highlights the diagnosis trajectories that intensify or diminish likelihood of heart failure in patients. The identification of such divergence in diagnoses helps pinpoint signals for heart failure from overall population trends. In this section, we describe how we restructure Medicare EMR data to obtain supervised disease progression trajectories. We then merge individual trajectories in the form of a compact Directed Acyclic Graph to model class-aware patient population-wide trends in diagnoses. Using this model, we identify key differentiating diagnoses and trajectories that help separate patients who are likely to develop heart failure from those who do not.

Preprocessing. We transform the data from raw medical histories shown in Table 1 to extract class-aware trajectories for patients using the following steps. We first collapse the diagnosis codes to their 3 digit counterparts, then eliminate duplicate families of diagnoses and then decouple the diagnosis history used for prediction from the observed outcome. Table 2 shows the result of applying this preprocessing to the example history from Table 1.

1. *Removing patients who receive a heart disease diagnosis on their first visit* Out of the 46.6% of the patients in our dataset who develop heart failure,

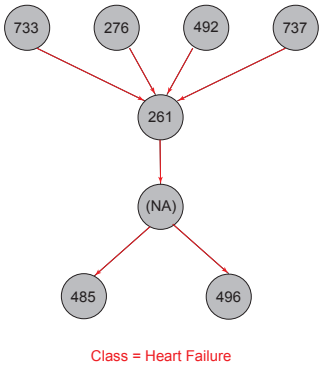


Fig. 1: **Example Patient Disease Progression** A directed acyclic graph can model a single patient's disease progression as shown above. Each layer represents a distinct visit, and each child node belongs to the next visit. In each visit layer, we identify several unique diagnosis codes. Each individual node in the graph represents these diagnosis codes.

18.0% receive a heart failure diagnosis in their very first visit. Since this study revolves around the concept of diagnoses leading up to heart failure, we consider these patients out of scope for our training and testing data. This removal of heart failure cases reduces the rate of observed heart failure in the rest of the patients down to 34.8% from the original 46.6%.

2. *Decoupling input data and target labels*— In patients with heart failure, we right-censor the diagnosis data when the first '428' code appears. This ensures that there is no "data leakage", i.e. we do not predict heart failure based on an observed diagnosis of heart failure since it is a chronic condition.
3. *Pre-pruning diagnoses and pathways*— To mitigate the impact of spurious/noisy disease trajectories in our analysis, we set a minimum support threshold of 100 for the nodes and edges in our graph. By imposing this threshold, we ensure that none of the diagnoses or the pathways between them draws conclusions from a set of fewer than 100 patients out of a total sample size of 1.1M patients.
4. *Code Collapse*— The original data contains 12,396 "Minor Category" diagnosis codes, whereas our analysis targets the "Major Category" outcome (heart failure). Collapsing the 5-digit diagnoses codes down to their respective 3-digit major categories helps reduce the complexity of the problem and matches the granularity of the observed outcome. As a result, we now use 1,064 diagnosis families to chart patient trajectories, which is 8.6% of the original complexity.
5. *Removing duplicate diagnoses*— We only consider new and previously unobserved diagnoses in our analyses. In Table 1, this means that we consider diagnoses for *Other disorders of bone and cartilage* (733) only for their first

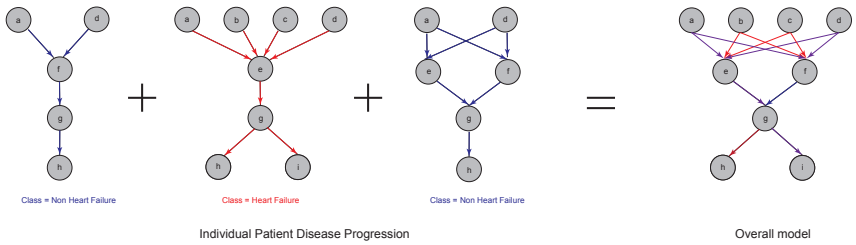


Fig. 2: **Combining individual Patient Histories**

visit. We hope to address the trade-off of dropping duplicate diagnosis in Future Work.

6. *Removing superfluous diagnoses*— ICD-9-CM diagnosis codes starting with V (Supplementary classification of factors influencing health status and contact with health services) and E (External causes of injury) reveal little about the progression of disease and were taken out of the graph.

Disease Progression for Individual Patients. For the example patient in Table 2, we can now create a disease progression history based on their diagnoses Figure 1. Each node represents a diagnosis and each edge ($e(i, j)$) represents a potential transition from diagnosis i to j across successive patient visits. Each of these edges is strictly directed from a diagnosis in visit $(t-1)$ to a diagnosis in visit t and nodes in the same visit do not have edges between them. This makes the graph a Directed Acyclic Graph (DAG). Each patient in our training data has an outcome label associated with them, which we use as a label for each of these patient-centric DAGs.

A Class-Aware Heart Failure Model. While each patient's disease progression DAG contains signals for their eventual outcome, but a domain expert aided causal analysis for each patient would not scale to the 1.1M patients in our dataset. Instead, we aggregate these histories into a unified model across all patients to get a consensus on which diagnoses and pathways signal which outcome for an unseen patient. Within our patient population, we perform a 5-fold cross-validation, training on 80% of the patients and testing the model on the other 20% over 5 folds of the data. As shown in Figure 2, we simply combine observed nodes and edges across all training set patients and create a composite DAG. The nodes at each level represent the superset of possible diagnoses at that visit and edges between each level represent the observed transitions between diagnoses.

The weight of each edge $e(i, j)$ corresponds to the observed confidence of heart failure among patients who were diagnosed with i and then j. Similarly, we assign a weight to each node n representing the observed confidence of heart failure for that node. A patient can have multiple diagnoses within the same visit, each of which adds to the support of the corresponding nodes and edges. This does

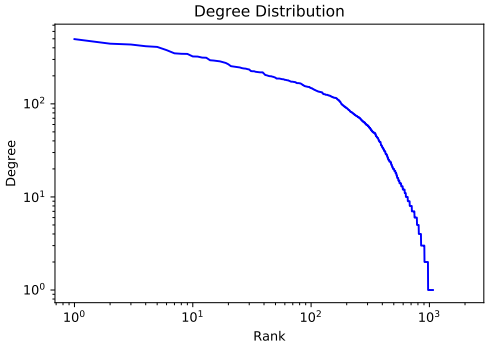
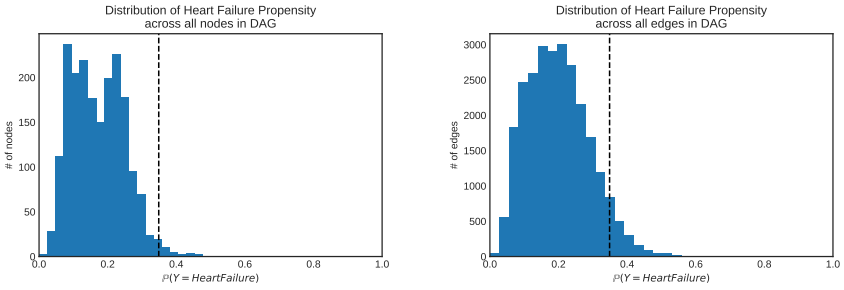


Fig. 3: **Degree Distribution of Trained Diagnosis Graph**

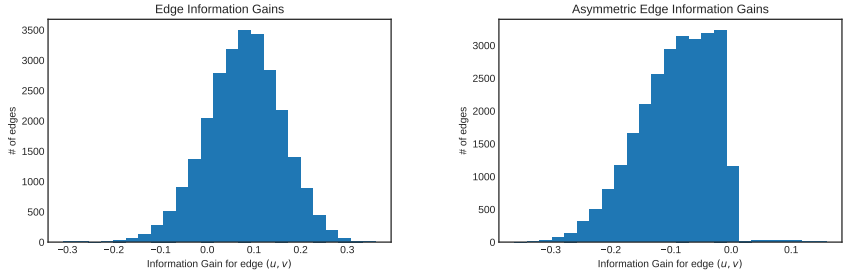
not guarantee the total incoming support into a node being equal to the total outgoing support. Another salient feature of this model is that it cane distinguish the same diagnosis code between visits. For example, if one observes code '261' in visit #1 and #2 for different patients, we create nodes labeled '261_1' and '261_2' to preserve their unique trajectory histories. By this definition, it is possible for '261_1' and '261_2' to have completely different weights. In our model, we label each node's diagnosis code according to the visit it was observed in.

Model Inference The overall trained model is an interconnected representation that contains 1,974 nodes and 26,229 edges. with an average in-degree and out-degree of 13.29. A relatively few nodes and edges contain a high likelihood of heart failure as seen in Figure 4. These high-confidence nodes and edges indicate underlying diagnoses and trajectories that lead to high rates of heart failure.



(a) **Imbalance in Node Weights** (b) **Imbalance in Edge Weights**

Fig. 4: **Heart failure trajectories are highly imbalanced**: We observe that the progression of heart failure follows a minor set of nodes and edges in the learned model.



(a) **Information Gain** (b) **Asymmetric Information Gain**

Fig. 5: **Information Gain in Edges**: We compute the information gain for each edge and use it as an edge attribute. By making information gain class aware, we treat non-heart failure intensifying edges as negative information gains. This makes it easier to isolate signals for heart failure intensifying paths.

We identify nodes and edges with a high propensity for heart failure in Table 3 and Table 4 respectively. These nodes and edges describe diagnostic pathways that indicate heart failure. In addition to extremely high likelihood of heart failure, we also identify diagnoses that effectively discern heart failure. For this, we use information gain (or InfoGain, for short) for successive diagnoses in patients. A higher information gain indicates a higher class polarization between heart failure and non-heart-failure. Information Gain (IG) is the *reduction* in entropy for a given edge in our model. Specifically, an edge $e(i, j)$ with high information gain indicates that the diagnosis of j after i leads to a higher confidence of arriving at either class. We compute each node's entropy from its class distribution using

$$H(i) = \sum_{k} -p_k(i)log(p_k(i)) \qquad H(j) = \sum_{k'} -p_{k'}(j)log(p_{k'}(j))$$
$$IG(i, j) = H(i) - H(j)$$

where i and j are the respective source and destination diagnoses, $k \in \{HF, NHF\}$, and $p_k(.)$ represents the probability of observing class k in a given node. Information Gain for $e(i, j)$ is simply $IG(i, j) = H(i) - H(j)$, where higher values of IG are more helpful in our search for heart failure propensity intensifying markers.

The sheer abundance of pathways towards non-heart-failure outcomes eclipses the relatively low InfoGain of individual edges which intensify heart failure. These are the majority of the edges which form the positive side of Figure 5a. However, we are primarily interested in edges which intensify likelihood of heart failure. To achieve this, we artificially penalize InfoGain in edges which have a higher likelihood of non heart failure by simply making them negative. This isolates and highlights heart-failure intensifying pathways in the network. Fig-

ure 5b shows how this transformation affects the edge attributes and isolates the relatively few edges which exhibit a high InfoGain favoring heart failure.

The above steps outline how we process raw patient records into a supervised representational model. This graphical model not only amalgamates patient disease trajectories, but it also highlights key pathways leading to heart failure.

4 Predicting heart failure for an Unseen Patient

Now that we have a representational and interpretable model to predict heart failure, we use it to predict outcomes for our held-out test dataset. We describe how we convert a new patient's diagnosis history into predicted probabilities and how we evaluate these predicted outcomes.

How to Predict Given a test patient's diagnosis history, we replicate the steps in the training section to arrive at a graph similar to Figure 1. Here, we make an important assumption about the nature of our model— we assume that the probabilities at each stage obey a Bayesian model. Using the probabilities from our trained model, we can predict relative odds of heart failure and non-heart failure by simply multiplying the class-wise probabilities for each edge and normalizing them. Given a test patient with a disease progression graph G_{test}, the unnormalized value of $\mathbb{P}(Y = HF)^* = \prod_e p(e(i,j))$, $\forall e(i,j) \in G_{test}$ and $p(e(i,j)) \in G_{trained}$. We then similarly compute the unnormalized value of $\mathbb{P}(Y = NHF)^*$ and finally output the normalized value of $\mathbb{P}(Y = HF)$.

In this work, we assume a Markovian model when using the Bayesian network structure to model disease progression. This means that dependencies and graph attributes (for instance, support and confidence) do not extend beyond *immediate* descendants directly in our model, i.e. $A|B$ and $B|C$ can model disease progression, but not $A|B, C$ directly. In future work, the model can extend to include higher-order dependencies [16]. This would enable us to model dependencies of the form $A|B, C$ and beyond.

Evaluation & Baselines We compute the above probabilities for all patients in the test set and evaluated against their true observed outcomes. We then compute the Receiver Operating Characteristics in terms of False Positive Rate and True Positive Rate for these predictions. Our key prediction metric is the area under the ROC curve (AUROC), a higher AUROC indicating superior predictive performance.

How soon can we predict Heart Failure? Each visit in our trained model is represented by a layer of nodes in the DAG. A prediction made using the first i layers of the DAG corresponds to a prediction made on i visits of an unseen test patient. Deliberately pruning the number of layers in the trained DAG is equivalent to reducing the complexity of our trained model and being able to predict our target outcome earlier. In order to see if this trade-off negatively influences predictive performance, we test the unseen patient histories on DAGs

Table 3: **Top 20 confidence nodes for heart failure**. These nodes represent the diagnoses with the highest likelihood of heart failure in our model.

Visit #	Diagnosis	Confidence	Support
2	Rheumatic fever with heart involvement	0.5947	882
1	Rheumatic fever with heart involvement	0.5685	690
2	Pulmonary congestion and hypostasis	0.5117	2088
1	Cardiomyopathy	0.4923	7338
1	Diseases of mitral valve	0.4746	1872
2	Cardiomyopathy	0.466	11864
1	Pulmonary congestion and hypostasis	0.4552	1046
1	Poisoning by agents primarily affecting the cardiovascular system	0.4448	679
1	Hypertensive heart and renal disease	0.4383	3784
2	Hypertensive heart and renal disease	0.4366	7624
1	Diseases of mitral and aortic valves	0.4312	8724
1	Diseases of aortic valve	0.4232	199
2	Diseases of mitral valve	0.4188	2764
2	Diseases of aortic valve	0.4097	487
1	Diseases of other endocardial structures	0.4032	3524
1	Chronic pulmonary heart disease	0.4016	4924
1	Nephrotic syndrome	0.3963	1676
2	Diseases of mitral and aortic valves	0.3949	15168
2	Chronic pulmonary heart disease	0.3918	9477
1	Hypertensive heart disease	0.375	41999

pruned to predict heart failure from 1, 2, 3 and 4 (maximum number of test visits in our data) visits and evaluate their area under the ROC curve.

5 Results

The techniques described above cover three key aspects of our research. First, we train a class-aware model of heart failure from patient history data. Second, we interpret the model to identify key diagnoses and disease progression pathways which intensify or mitigate the chances of a given patient developing heart failure. Third, we show how this model performs when predicting heart failure outcomes for a completely unseen set of patients.

Model Inference Looking at the highest confidence nodes for heart failure in Table 3, several common themes appear in diagnoses that tend to proceed heart failure - namely, rheumatic fever, pulmonary congestion, cardiomyopathy, blood poisoning, kidney disease, hypertension, and aortic and mitral valve disease. For these diagnoses, it does not appear to matter much if one diagnoses a patient on 1 or 2 - the progression to heart disease seems to occur at about the same confidence levels. By absolute numbers in this data set, the diagnoses that lead to heart failure the most are cardiomyopathy and aortic mitral valve diseases.

Rheumatic heart diseases and pulmonary congestion patients appear less than the former three in the data set, but have a higher probability of leading to heart failure.

Referring to the highest confidence edges for heart failure given in Table 4, we can see that may of the same destination diagnoses match the diagnoses given in the high confidence nodes in Table 3. The edges give us some idea about the diagnoses that come first that may lead to heart disease given another diagnosis. For instance, the high confidence nodes in Table 3 told us that diagnoses such as rheumatic fever, pulmonary congestion, and cardiomyopathy lead to heart failure. 14 out of the 20 top confidence edges involve cardiomyopathy, which indicates that cardiomyopathy is a strong component in leading to heart failure. Cardiac dysrhythmia is a diagnosis that is particularly deadly when combined with further diagnoses.

The findings of this graph seem to confirm the results of other studies. Others have identified cardiomyopathy and valve dysfunction as precursors for heart disease [17], [18], [19]. The American Heart Association has recommended that patients who have chronic kidney disease are in the highest risk group for development of cardiovascular disease [18]. Researchers have associated nephrotic syndrome with cardiovascular disease [18]. Pulmonary congestion is very common in patients with heart failure due to its relation to high pressure in the left ventricle of the heart. Many patients who have heart failure are also found to have fluid overload which is a common result of pulmonary congestion. Detection and treatment of pulmonary congestion can help prevent progression of heart failure [20]. The result of blood poisoning and sepsis is often multiple organ failure, including septic cardiomyopathy, which can lead to heart failure [21]. Other studies have found that chronic pulmonary heart disease is a predictor of chronic heart failure in China [22]. For many years, doctors have known that rheumatic fever can contribute to heart failure occurring later in life [23]. Hypertension is also a major contributing factor in congestive heart failure [24].

Almost all the top confidence edges involved rheumatic heart disease, pulmonary congestion and hypostasis, mitral and aortic valve disease, or cardiomyopathy, which therefore accentuates the importance of those diseases in the diagnosis of heart failure. The high confidence edges given in Table 4 let us know that diseases such as cardiac dysrhythmia, diabetes, ischemic heart disease, and lung diseases, diagnosed beforehand, can ultimately result in heart failure.

Table 5 shows us the highest information gain nodes tend to come from source diagnoses that are mental or noncardiac in nature (Affective psychoses, cerebral degeneration, malignant neoplasm of bladder, Parkinson's disease, etc.) followed by an acute myocardial infarction, endocardium diseases, or cardiomyopathy. This seems to suggest that these diagnoses are the first cardiac problems that might occur in patients with other mental or noncardiac issues. This model of information gain suggests that screening for those three diseases, since they appear as some of the first cardiac diagnosis on a trajectory that leads to heart failure.

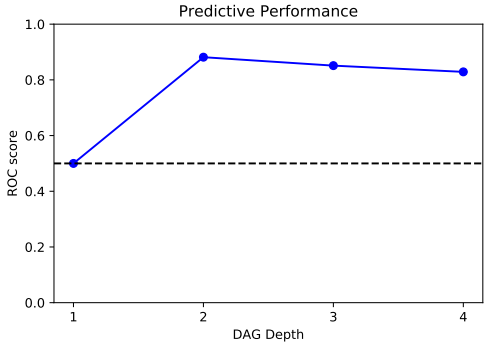


Fig. 6: **Predictive performance for various number of visits**. Area under the ROC curve plateaus with an increased amount of visits.

Heart Failure Prediction Our model can predict heart failure in patients from diagnoses from their second visit (i.e.: their first disease progression) as seen in Figure 6. Adding diagnoses from subsequent visits makes the predictive performance plateau in comparison to the second visit. As discussed in our *Data Preprocessing* stage in Section 3, we only have a maximum of 4 patient visits for our prediction task.

6 Discussion

Which disease progression trajectories lead to heart failure? A diagnosis of cardiomyopathy is a very common theme that appears in many high confidence edges and high information gains that lead to heart failure. Cardiomyopathy appears in 14 out of the 20 high confidence edges and 5 of the top 20 high information gains. We can therefore conclude that cardiomyopathy is an important factor in the progression of a heart failure trajectory. Monitoring patients for cardiomyopathy and intervening early is therefore important in limiting heart disease.

Besides cardiomyopathy, most of the other high information gain edges had a destination edge of acute myocardial infarction or endocardium diseases. These three appear as "gateway" diagnoses that eventually results in heart failure later in their medical record for patients who do not currently have a diagnosis of heart disease.

While cardiomyopathy is very common in the high confidence edge nodes, it does not occupy the top four high confidence edges. Cardiac dysrhythmia appears as a source diagnosis in the two top confidence edges, indicating that those with cardiac dysrhythmia should watch out for rheumatic heart disease or pulmonary congestion. Additionally, pulmonary congestion appears as a destination node for three out of the top four confidence edges, indicating that pulmonary

congestion is a complication that, given other diagnoses such as cardiac dysrhythmia, diabetes, or chronic ischemic heart disease, could eventually result in heart failure.

Can we predict heart failure? Using this model, we observe we can predict heart failure using the conclusions found from this data. The ROC curve given in Figure 6 indicates that diagnoses given in the first visit contains most of the information that leads to heart failure. We receive diminishing returns from subsequent diagnoses after that first visit. One reason could be that most of the diseases that eventually result in heart failure have already appeared by their first visit to a doctor, and rarely do patients not have diseases that are indicative of heart failure at their first visit, and then they go on to get heart failure later. Table 5 gives some examples of patients that have the highest jump in the probability of developing heart failure after their first visit. Certain cardiac events put those who were originally being treated for mental diagnoses in particular (affective psychoses, cerebral degenerations, Parkinson's disease, neurotic disorder) on a path to heart failure beginning in Visit 2. In general, the data tells us that the disease progression from Visit 1 to Visit 2 gives the most indication that a patient will eventually become a heart failure patient.

7 Conclusion

By constructing a DAG of Medicare patients and their visits, we found trends in diseases that result in an ultimate diagnosis of heart failure. We conclude that cardiomyopathy is a condition that is commonly associated with heart failure such that screening for cardiomyopathy should be a common part of preventative treatment. Additionally, we know that many patients' first diagnoses on a heart failure path are acute myocardial infarctions, endocardium diseases, and cardiomyopathy. Doctors who see patients for other medical issues, especially mental issues as observed, should know of these complications since they are often the first that show up in diagnoses that do not otherwise lead to heart failure. We also found that rheumatic heart disease, pulmonary congestion and hypostasis, cardiomyopathy, blood poisoning, and valve and aortic diseases are common comorbidities that occur before doctors diagnose patients with heart failure. Because the highest information gains in our DAG are on paths that concern mental disorders such as psychosis, cerebral degeneration, and Parkinson's, the conclusion can be made that patients being seen for these disorders should also be monitored for heart disease.

The ultimate goal of such a system is to be able to effectively predict likelihood of heart failure, which we demonstrate using our trained DAG. We show that the most indicative diagnoses belong to the first disease progression in terms of their information gain and area under the ROC curve. This underscores the usefulness of our model in extracting signals which can be used for early detection of heart failure.

Table 4: **Highest Confidence disease progression edges in trained DAG.** Top 20 diagnostic edges with extremely high confidence of heart failure. These edges represent those at the extreme right of the distribution in Figure 4b. Visit number corresponds to source diagnosis.

Source		Destination		Conf	Supp
Visit #	Diagnosis	Visit #	Diagnosis		
1	Cardiac dysrhythmias	2	Rheumatic fever with heart involvement	0.6995	183
1	Cardiac dysrhythmias	2	Pulmonary congestion and hypostasis	0.6870	131
1	Diabetes mellitus	2	Pulmonary congestion and hypostasis	0.6798	178
1	Other forms of chronic ischemic heart disease	2	Pulmonary congestion and hypostasis	0.6486	148
1	Other diseases of lung	2	Cardiomyopathy	0.6400	150
1	Cardiomyopathy	2	Pneumonia, organism unspecified	0.6209	153
1	Cardiomyopathy	2	Acute myocardial infarction	0.6198	121
1	Conduction disorders	2	Cardiomyopathy	0.6123	325
1	Acute myocardial infarction	2	Cardiomyopathy	0.6000	180
1	Old myocardial infarction	2	Cardiomyopathy	0.5986	147
1	Cholelithiasis	2	Cardiomyopathy	0.5943	106
1	Cardiomyopathy	2	Other diseases of lung	0.5882	170
1	Diverticula of intestine	2	Cardiomyopathy	0.5847	118
1	Cardiomyopathy	2	Conduction disorders	0.5738	237
1	Cardiac dysrhythmias	2	Cardiomyopathy	0.5724	1277
1	Cardiomyopathy	2	Transcient cerebral ischemia	0.5714	105
1	Ill-defined descriptions and complications of heart disease	2	Cardiomyopathy	0.5708	212
1	Diabetes mellitus	2	Chronic pulmonary heart disease	0.5706	340
1	Essential hypertension	2	Pulmonary congestion and hypostasis	0.5610	164
1	Iron deficiency anemias	2	Cardiomyopathy	0.5577	104

Table 5: **Top 20 Information Gain edges in trained DAG.** These edges represent diagnoses which go from seemingly benign to high likelihood of heart failure.

Source			Destination			InfoGain	Supp
Visit #	Diagnosis	Conf.	Visit #	Diagnosis	Conf.		
1	Affective psychoses	0.1541	2	Acute myocardial infarction	0.3462	0.1597	153
1	Affective psychoses	0.1541	2	Other diseases of endocardium	0.3355	0.1544	225
1	Affective psychoses	0.1541	2	Hypertensive heart disease	0.3279	0.1533	210
1	Malignant neoplasm of bladder	0.1769	2	Acute myocardial infarction	0.3462	0.1489	106
1	Other cerebral degenerations	0.1717	2	Acute myocardial infarction	0.3462	0.1324	105
1	Other cerebral degenerations	0.1717	2	Other diseases of endocardium	0.3355	0.1272	194
1	Other cerebral degenerations	0.1717	2	Hypertensive heart disease	0.3279	0.1261	141
1	Parkinson's disease	0.1795	2	Acute myocardial infarction	0.3462	0.1252	151
1	Parkinson's disease	0.1795	2	Other diseases of endocardium	0.3355	0.1200	167
1	Parkinson's disease	0.1795	2	Hypertensive heart disease	0.3279	0.1189	161
1	Neurotic disorder	0.1874	2	Acute myocardial infarction	0.3462	0.1068	249
1	Neurotic disorder	0.1874	2	Other diseases of endocardium	0.3355	0.1016	286
1	Neurotic disorder	0.1874	2	Hypertensive heart disease	0.3279	0.1005	247
2	General symptoms	0.2176	3	Cardiomyopathy	0.2980	0.1004	152
2	Other disorders of urethra and urinary tract	0.2208	3	Hypertensive heart and renal disease	0.3109	0.1004	146
2	Other disorders of urethra and urinary tract	0.2208	3	Cardiomyopathy	0.2980	0.0980	199
1	Senile and presenile organic psychotic conditions	0.1898	2	Acute myocardial infarction	0.3462	0.0979	179
1	Senile and presenile organic psychotic conditions	0.1898	2	Other diseases of endocardium	0.3355	0.0926	221
2	Other and unspecified anemias	0.2282	3	Hypertensive heart and renal disease	0.3109	0.0917	207
1	Senile and presenile organic psychotic conditions	0.1898	2	Hypertensive heart disease	0.3279	0.0915	225

References

1. Hans-Ulrich Prokosch, T Ganslandt, et al. Perspectives for medical informatics. *Methods Inf Med*, 48(1):38–44, 2009.
2. Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.
3. Ashwin Belle, Raghuram Thiagarajan, SM Soroushmehr, Fatemeh Navidi, Daniel A Beard, and Kayvan Najarian. Big data analytics in healthcare. *BioMed research international*, 2015, 2015.
4. Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005.
5. Oanh Kieu Nguyen, Anil N Makam, Christopher Clark, Song Zhang, Bin Xie, Ferdinand Velasco, Ruben Amarasingham, and Ethan A Halm. Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison. *Journal of hospital medicine*, 2016.
6. Stanislav V Kasl and Sidney Cobb. Health behavior, illness behavior and sick role behavior: I. health and illness behavior. *Archives of Environmental Health: An International Journal*, 12(2):246–266, 1966.
7. Andreas Holzinger. Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. *IEEE Intelligent Informatics Bulletin*, 15:6–14, 2014.
8. Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.
9. Darcy A Davis, Nitesh V Chawla, Nicholas A Christakis, and Albert-László Barabási. Time to care: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery*, 20(3):388–415, 2010.
10. Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.
11. Patricia A Wilkosz, Howard J Seltman, Bernie Devlin, Elise A Weamer, Oscar L Lopez, Steven T DeKosky, and Robert A Sweet. Trajectories of cognitive decline in alzheimer's disease. *International Psychogeriatrics*, 22(02):281–290, 2010.
12. Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5, 2014.
13. Diane S Lauderdale, Sylvia E Furner, Toni P Miles, and Jack Goldberg. Epidemiologic uses of medicare data. *Epidemiologic Reviews*, 15(2):319–327, 1993.
14. Janet B Mitchell, Thomas Bubolz, John E Paul, Chris L Pashos, José J Escarce, Lawrence H Muhlbaier, John M Wiesman, Wanda W Young, Roberts Epstein, and Jonathan C Javitt. Using medicare claims for outcomes research. *Medical care*, 32(7):JS38, 1994.
15. Dariush Mozaffarian, Emelia J Benjamin, Alan S Go, Donna K Arnett, Michael J Blaha, Mary Cushman, Sandeep R Das, Sarah de Ferranti, Jean-Pierre Després, Heather J Fullerton, et al. Heart disease and stroke statistics2016 update. *Circulation*, 133(4):e38–e360, 2016.

16. Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.
17. John J McMurray and Simon Stewart. Epidemiology, aetiology, and prognosis of heart failure. *Heart*, 83(5):596–602, 2000.
18. Mark J. Sarnak, Andrew S. Levey, Anton C. Schoolwerth, Josef Coresh, Bruce Culleton, L. Lee Hamm, Peter A. McCullough, Bertram L. Kasiske, Ellie Kelepouris, Michael J. Klag, Patrick Parfrey, Marc Pfeffer, Leopoldo Raij, David J. Spinosa, and Peter W. Wilson. Kidney disease as a risk factor for development of cardiovascular disease. *Hypertension*, 42(5):1050–1065, 2003.
19. Donald Lloyd-Jones, Robert J. Adams, Todd M. Brown, Mercedes Carnethon, Shifan Dai, Giovanni De Simone, T. Bruce Ferguson, Earl Ford, Karen Furie, Cathleen Gillespie, Alan Go, Kurt Greenlund, Nancy Haase, Susan Hailpern, P. Michael Ho, Virginia Howard, Brett Kissela, Steven Kittner, Daniel Lackland, Lynda Lisabeth, Ariane Marelli, Mary M. McDermott, James Meigs, Dariush Mozaffarian, Michael Mussolino, Graham Nichol, Véronique L. Roger, Wayne Rosamond, Ralph Sacco, Paul Sorlie, Randall Stafford, Thomas Thom, Sylvia Wasserthiel-Smoller, Nathan D. Wong, and Judith Wylie-Rosett. Heart disease and stroke statistics—2010 update. *Circulation*, 121(7):e46–e215, 2010.
20. Eugenio Picano, Luna Gargani, and Mihai Gheorghiade. Why, when, and how to assess pulmonary congestion in heart failure: pathophysiological, clinical, and methodological implications. *Heart Failure Reviews*, 15(1):63–72, 2010.
21. Laszlo M. Hoesel, Andreas D. Niederbichler, and Peter A. Ward. Complement-related molecular events in sepsis leading to heart failure. *Molecular Immunology*, 44(1):95 – 102, 2007. XXI International Complement Workshop Beijing, China, October 22-26, 2006.
22. YM Cao, DY Hu, Y Wu, and HY Wang. A pilot survey of the main causes of chronic heart failure in patients treated in primary hospitals in china. *Zhonghua nei ke za zhi*, 44(7):487–489, 2005.
23. Edward F Bland and Duckett Jones. Rheumatic fever and rheumatic heart disease. *Circulation*, 4(6):836–843, 1951.
24. Daniel Levy, Martin G Larson, Ramachandran S Vasan, William B Kannel, and Kalon KL Ho. The progression from hypertension to congestive heart failure. *Jama*, 275(20):1557–1562, 1996.