HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2017 January 12.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2015 ; 2015: 2530–2533. doi:10.1109/EMBC.2015.7318907.

Early Detection of Heart Failure with Varying Prediction Windows by Structured and Unstructured Data in Electronic Health Records*

Yajuan Wang,
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA

Kenney Ng,
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA

Roy J. Byrd,
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA

Jianying Hu,
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA

Shahram Ebadollahi,
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA

Zahra Daar,
Geisinger Health System, Danville, PA 17822

Christopher deFilippi,
Univ. of Maryland School of Medicine, Baltimore, MD 21201

Steven R. Steinhubl, and
Scripps Health, San Diego, CA 92121

Walter F. Stewart
Sutter Health Research, Walnut Creek, CA 94520

Abstract

Heart failure (HF) prevalence is increasing and is among the most costly diseases to society. Early detection of HF would provide the means to test lifestyle and pharmacologic interventions that may slow disease progression and improve patient outcomes. This study used structured and unstructured data from electronic health records (EHR) to predict onset of HF with a particular focus on how prediction accuracy varied in relation to time before diagnosis. EHR data were extracted from a single health care system and used to identify incident HF among primary care patients who received care between 2001 and 2010. A total of 1,684 incident HF cases were identified and 13,525 controls were selected from the same primary care practices. Models were compared by varying the beginning of the prediction window from 60 to 720 days before HF diagnosis. As the prediction window decreased, the performance [AUC (95% CIs)] of the

*Research supported by NIH R01HL116832.

Corresponding author; phone: 914-945-2075; wangya@us.ibm.com.

predictive HF models increased from 65% (63%–66%) to 74% (73%–75%) for the unstructured, from 73% (72%–75%) to 81% (80%–83%) for the structured, and from 76% (74%–77%) to 83% (77%–85%) for the combined data.

I. Introduction

Heart failure (HF) prevalence is increasing and is among the most costly diseases to Medicare [1]. HF affects approximately 5.7 million people in the United States, and about 825,000 new cases per year with ~$33 billion total annual cost [2–4]. The lifetime risk of developing HF is 20% at 40 years of age [2]. HF has a high mortality rate: ~50% within 5 years of diagnosis [5] and causes or contributes to approximately 280,000 deaths every year [6]. There has been relatively little progress in slowing progression of HF severity largely because there are no effective means of early detection of HF to test interventions.

Early detection of HF offers the opportunity to test and ultimately develop effective lifestyle and pharmacologic interventions. However, HF is a clinically complex and heterogeneous disease that is challenging to detect in routine care because of the diversity of alternative explanations for symptoms [7].

The rapid adoption of electronic health records (EHRs) and advances in machine learning and natural language processing (NLP) opens new opportunities to develop novel and cost-effective methods to detect HF before it is too late to modify the natural history of the disease. EHRs provide rich, longitudinal patient records containing structured and unstructured data (e.g., progress notes). Structured data are relatively easy to extract and incorporate into a predictive model. But, unstructured data may contain potentially valuable information on Framingham HF signs and symptoms and other relevant indicators of disease progression or explanations for health status that are documented long before specific diagnoses emerge. In this study, we evaluated the independent and combined predictive power of EHR structured data (diagnoses, clinical measures, labs, medication orders, image orders and hospitalizations) and unstructured data from progress notes that was specific to Framingham HF signs and symptoms and to left ventricular ejection fraction (LVEF). The prediction window was assessed from 60 to 720 days prior to the diagnosis date and the predictive ability of clinical factors was examined to identify factors that are more effective for early detection.

II. Methods

A. Study Subjects

We used a case-control study design where the study was nested within a cohort of ~400,000 primary care patients who received care between 2001 and 2010 from the Geisinger Clinic. A total of 1,684 incident HF cases were identified who met the following criteria: 1) HF diagnosis (or an associated ICD-9 diagnosis of HF) appeared in a minimum of three clinical encounters within 1.5 years of each other. If the time span between the first and second HF diagnoses was less than 1 year, the date of the first encounter was used as the HF diagnosis date; otherwise the date of the second encounter was used. 2) Age was between 50 and 85 at

the time of HF diagnosis. 3) The time between the first primary care physician visit and the HF diagnosis date was greater than 1.5 years so that sufficient follow-up time before the time window was available.

Up to 10 eligible clinic-, sex-, and age-matched (in 5-year age intervals) controls were selected for each incident HF case for a total of 13,525 subjects. Patients were eligible as controls if they had no HF diagnosis before the date – 1 year post-HF diagnosis of the matched case. Control subjects were required to have their first office encounter within 1 year of the incident HF patient's first office visit and have ≥1 office encounter 30 days before or any time after the case's HF diagnosis date to ensure similar duration of observations among cases and controls.

B. Extracting Features from EHR

A feature vector representation for each patient was generated based on the patient's EHR data. EHR data can be viewed as event sequences over time (e.g. a patient can have multiple diagnoses of hypertension on different dates). To convert such event sequences into feature variables, an observation window (e.g., two years) was specified. Then all events of the same feature within the window were represented by one or more relevant summary statistics. The aggregation function can produce simple feature values like counts and averages or complex feature values that take into account temporal information (e.g., trend and temporal variation). In this study, only basic aggregation functions were applied: counts for categorical variables and means for numeric variables. Table I summarizes the structured variables extracted from the EHR records of HF case and controls that were evaluated for predictive modeling.

An NLP application was developed to extract Framingham HF signs and symptoms from the unstructured text in EHR progress notes. Table II lists the 15 (out of 17) Framingham criteria that were extracted. Instances of affirmations (positive mentions) and denials (negative mentions) were identified. Details of the NLP extraction of Framingham HF signs and symptoms have been previously described [8] with an overall precision (or positive predictive value) of 0.925 and recall (or sensitivity) of 0.896 relative to manual chart review. Since LVEF is an important indicator of ventricular function, LVEF values and categories were also extracted from clinical notes. LVEF was classified into three categories: reduced (≤40%), moderate (40%–50%) and preserved (≥50%). Count feature aggregation was used for the Framingham criteria and LVEF categories while mean aggregation was used for the numeric LVEF variable.

C. Predictive Analytics

We examined how the performance of a predictive model varied in relation to the duration of the prediction window. For example, a prediction window of 365 days means that no EHR data was used from cases in the one year period before the assigned diagnosis date and the comparable date assigned to group-matched controls (see Figure 1). The observation window refers to the time period before the prediction window from which EHR data were used to predict HF. The "index date/prediction date" refers to the time point prior to the diagnosis date for each case and control where EHR data are used to discriminate who will

be a HF case versus not. The index date separates the prediction window from the observation window. In the current study, we varied the prediction window size from 60 days to 720 days. A fixed observation window size of 720 days was used for all models.

To build predictive models we first used the information gain [9] measure to select the most discriminating subset of features. Next, random forest classifiers [10] (with 100 trees) were trained and evaluated. The Area under the ROC Curve (AUC) was used to measure the predictive performance. 10-fold cross validation (CV) was used in all model iterations and the AUC (95% CI) was the primary outcome of interest. An initial set of experiments was performed to identify the optimal number of features to use for the unstructured and structured feature types. Predictive models using random forest classifiers were built with fixed observation window (720 days) and predictive window (180 days) sizes, where the number of features selected by the information gain measure was allowed to vary. Cross validation was used to determine the optimal number of features for each feature type. Predictive models were then built using the unstructured and structured feature types, the optimal number of features and a fixed observation window size of 720 days but varying predictive window sizes (ranging from 60 to 720 days). These models were evaluated to characterize HF prediction performance and the predictive ability of clinical factors as a function of time before diagnosis.

III. Results

A. Impact of Prediction Window Size on HF Prediction

The optimal number of features for each feature type was identified for unstructured feature types: $n=15$ with 71% (70%–73%) AUC; for structured feature types: $n=100$ with80% (79%–81%) AUC; and for combined structured and unstructured feature types: $n = 500$ with 80% (79%–82%) AUC. Figure 2 shows the impact of varying the prediction window size on HF prediction performance for classifiers built with the different feature types. As expected, by increasing the prediction window size from 60 days to 720 days (so the prediction is made earlier in time), the AUCs decrease from 74% (73%–75%) to 65% (63%–66%) for unstructured data, from 81% (80%–83%) to 73% (72%–75%) for structured data, and from 83% (82%–85%) to 76% (74%–77%) for the combined feature types. Using combined unstructured and structured features, predictive performance is significantly but only slightly improved across all prediction window sizes over using just the unstructured or structured features alone ($p<0.05$), except at prediction window = 180 days. This observation is further confirmed by using $n=100$ combined unstructured and structured feature types. As seen in Figure 2, with the same number of features ($n=100$) selected for structured and combined feature types, classifiers built with the combined feature types achieved slightly better AUC performance over just the structured features: AUC 75% (74%–76%) ~ 83% (81%–84%) versus 73% (72%–75%) ~ 81% (80%–83%), $p<0.05$ except at prediction window = 120, 180 days.

B. Predictive Ability of Features

The predictive ability of the top features selected from each of the two feature types as a function of the predictive window size was also investigated. Predictive ability was

measured by the number of times a feature was selected in a 10-fold cross validation (CV) experiment using an information gain feature selection criterion. Figure 3(a) shows the 19 unstructured features that were selected at least once in the top m=15 for different predictive window sizes. Among these, 10 features were consistently selected in all 10 folds across all predictive window sizes: 4 positive Framingham criteria (APED, ANKED, DOE and RC), 4 negative Framingham criteria (APED, ANKED, PLE and RALE) and 2 LVEF features (value and reduced LVEF). The predictive ability of the remaining features shows interesting patterns that depend on the size of the predictive window. For example, RALE (positive) shows up in the top features only for predictive windows shorter than 480 days while HEP (negative) shows up in the top features only for longer predictive windows (600 days to 720 days). Similarly, Figure 3(b) and 3(c) show the predictive patterns of the top structured and combined features, respectively. For structured data, 168 features were selected at least once, and 60 were consistently selected in all 10 fold cross-validation models across all predictive window sizes. Twenty-three of the 60 features (38%) were related to hospitalization (e.g. Legally Blind, Cardiac Disorders, etc.), 16 (27%) were Lab measures (e.g. creatinine, B-type Natriuretic Peptide [BNP], etc.), 13 (22%) were for medication orders (e.g. digoxin, furosemide, etc.) and 8 (13%) were in reference to specific diagnoses (e.g., COPD, Diabetes, etc.). Some clinical variables showed increased predictive ability closer to the diagnosis date. For example, Hemoglobin (HGB) was selected for predictive window sizes less than 540 days, and Estimated Glomerular Filtration Rate (eGFR) and hematocrit (HCT) were selected only for predictive window sizes less than 300 days. In contrast, other clinical variables demonstrated increased predictive ability further away from the diagnosis date (e.g. Cholesterol was selected when the predictive window size was more than 420 days). For the combined feature type, 169 features were selected at least once. Forty-seven were consistently selected in all 10-fold cross-validation models across all predictive window sizes. Similar to the structured features, the majority – 21 (45%) out of the 47 features were related to hospitalization, 10 (21%) were for medication reconciliation, 8 (17%) were in reference to specific diagnoses, 6 (13%) were Framingham criteria (APED positive, ANKED positive and negative, DOE positive, RC positive and PLE negative), and 2 (4%) were LVEF (numeric and reduced category). Some clinical factors with increased early or late predictive ability are highlighted in Figure 3(c).

IV. Discussion

Treatment of HF has universally focused on post-diagnosis, after irreversible remodeling and functional impairment have occurred [11]. Early detection of HF would open opportunities to test interventions that may delay the progression of heart failure. This study characterized HF prediction performance and the predictive ability of structured and unstructured clinical factors extracted from EHRs as a function of time (from 60 days up to 2 years) before diagnosis. It utilized NLP to extract Framingham criteria from unstructured clinical notes and machine learning techniques to investigate separate unstructured and structured information for HF prediction. The prediction performance for unstructured, structured and combined data from EHR demonstrated consistent improvement as the predictive window decreased. The combined unstructured and structured data from EHR demonstrated significantly better predictive performance compared to using just the structured or

unstructured data alone across all prediction window sizes ($p<0.05$, except at prediction window = 180 days when compared to structured data only). While the gain in AUC was relatively small when using combined structured and unstructured data, unstructured data were over-represented in the combined model. Some Framingham criteria documented in clinical notes long before the diagnosis showed consistent predictive ability across all prediction window sizes (APED, DOE, RC positive, PLE negative, ANKED positive and negative). Our study also confirmed the predictive ability of certain labs (BNP, INR, creatinine, etc) when using structured data for HF prediction; however, these lab features were not consistently selected when combined structured and unstructured features were used. Furthermore, no labs were selected for predictive window size = 180 days (see Figure 3c). This implies the potential value and robustness of some Framingham criteria (mentioned above) over certain labs.

In this study, we used a very basic approach for feature construction, limiting the use of clinical knowledge and evidence to refine features and other information (e.g., missing values) to improve feature representations. In addition, we used a fixed observation window to investigate the impact of varying predictive windows; however, in clinical settings, a long patient history might not always be available to enable HF assessment. We will continue to explore how feature construction from data driven and knowledge driven approaches can enhance model performance, including development of more advanced feature construction methods to investigate the temporal information of clinical factors and investigate the combined impacts of varying observation windows and predictive windows for HF prediction. Furthermore, since the different types of HF (diastolic HF with preserved LVEF versus systolic HF) require different intervention strategies, we would like to apply and extend this approach to the prediction and identification of important clinical factors for different HF subtypes.

V. Conclusion

In conclusion, our study demonstrated that both unstructured and structured information in EHR can facilitate early detection of HF as early as two years prior to diagnosis. The combined data achieved superior performance compared to using structured or unstructured data alone across all the prediction window sizes. Unstructured and structured factors exhibited different patterns of predictive ability with some being more useful closer to and others farther from the diagnosis date.

Acknowledgments

The authors would like to thank Mr. Harry Stavropoulos from IBM Research for providing database support and Ms. Melody J. Chin, Ms. Heather Law from Sutter Health and Ms. Elise Blaese from IBM for project facilitation.

References

1. Sun J, Hu J, Luo D, et al. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc. 2012; 2012:901–10. [PubMed: 23304365]
2. Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics--2014 update: a report from the American Heart Association. Circulation. Jan 21; 2014 129(3):e28–e292. [PubMed: 24352519]

3. Roger VL, Go AS, Lloyd-Jones DM, et al. Heart disease and stroke statistics--2011 update: a report from the American Heart Association. Circulation. Feb 1; 2011 123(4):e18–e209. [PubMed: 21160056]
4. Roger VL, Go AS, Lloyd-Jones DM, et al. Heart disease and stroke statistics--2012 update: a report from the American Heart Association. Circulation. Jan 3; 2012 125(1):e2–e220. [PubMed: 22179539]
5. Roger VL, Weston SA, Redfield MM, et al. Trends in heart failure incidence and survival in a community-based population. JAMA. Jul 21; 2004 292(3):344–50. [PubMed: 15265849]
6. Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2010. Natl Vital Stat Rep. May 8; 2013 61(4):1–117.
7. King M, Kingery J, Casey B. Diagnosis and evaluation of heart failure. Am Fam Physician. Jun 15; 2012 85(12):1161–8. [PubMed: 22962896]
8. Byrd RJ, Steinhubl SR, Sun J, et al. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform. Jan 10.2013
9. Bishop, CM. Pattern Recognition and Machine Learning. Springer; 2006.
10. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32.
11. de Couto G, Ouzounian M, Liu PP. Early detection of myocardial dysfunction and heart failure. Nat Rev Cardiol. Jun; 2010 7(6):334–44. [PubMed: 20458341]

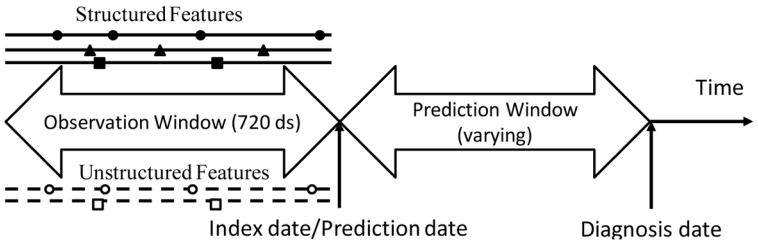

Figure 1.
Illustration of the timeline for predictive modeling.

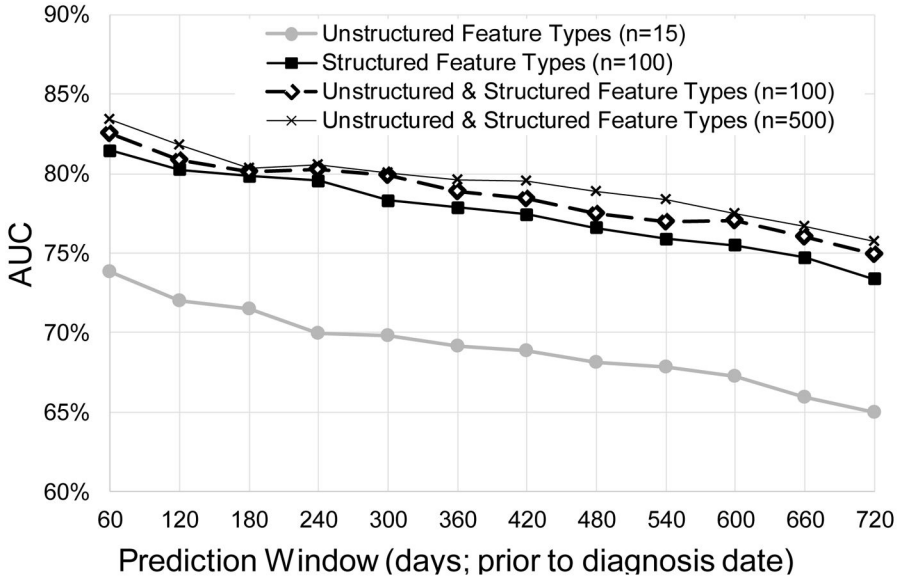

Figure 2.
Predictive modeling performance (AUC) with different feature types and predictive window sizes.

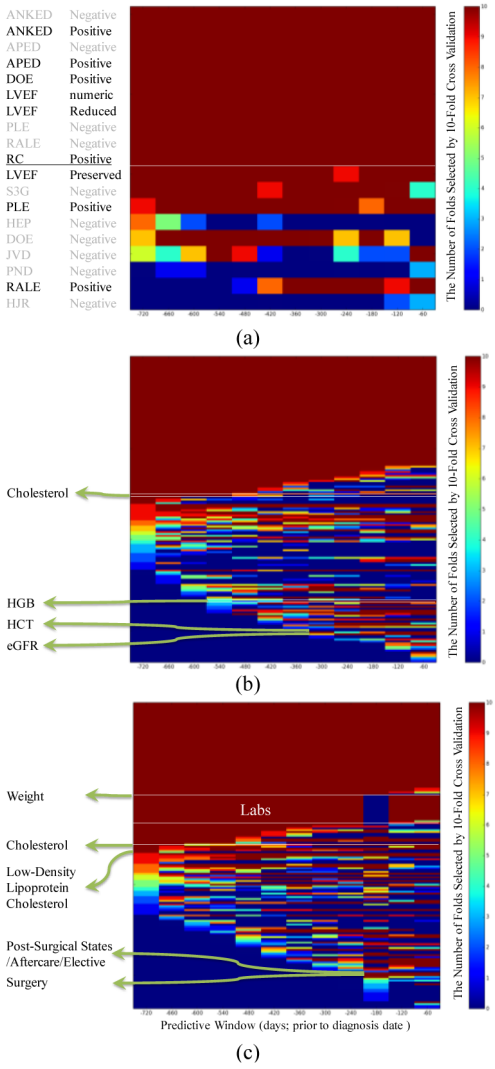

Figure 3.

Heat maps of the top m features selected by any fold in 10-fold cross validation (CV) for unstructured feature types (a; m=15), structured feature types (b; m=100) and combined feature types (c; m=100). Each row along the vertical axis is a feature. The horizontal axis indicates the prediction window size from 720 days to 60 days prior to the diagnosis date. The dark red indicates that the feature is selected in all of the 10 folds in 10-fold CV; the dark blue indicates that the feature is not selected in any of the 10 CV folds. Examples of clinical variables are labeled to highlight the different patterns of predictive ability as a function of predictive window size.

Table I

Structured Variable Categories, Examples, Number Of Unique Variables, And Aggregation Methods.

Variable Category	Example Variables	# of Variables	Aggregation Method
Clinical Diagnosis	Diabetes, Cardiac disorders, etc.	18,569	count
Clinical Measures	Pulse, Systolic blood pressure, etc.	6	mean
Labs	Cholesterol, Serum Glucose, etc.	2,336	mean
Medication Reconciliation	Beta Blockers, Loop Diuretics, etc.	1,250	count
Prescription Order	Furosemide, digoxin, etc.	3,952	count
Imaging Order	Echo orders, etc.	18	count
Hospitalization	The ICD-9 associated with the hospitalization.	7,304	count

Table II

Extracted Framingham HF Criteria, Code Names And The Feature Subtypes.

Extracted Framingham HF criteria –Code names	**Sub-types**[a]
Major Criteria	
Acute Pulmonary Edema – APEdema (APED)	P, N
Paroxysmal Nocturnal Dyspnea or orthopnea – PNDyspnea (PND)	P, N
Jugular Venous Distention – JVDistension (JVD)	P, N
Rales – Rales (RALE)	P, N
Radiographic Cardiomegaly – RCardiomegaly (RC)	P, N
S3 Gallop – S3Gallop (S3G)	P, N
Hepatojugular Reflux – HJReflux (HJR)	P, N
Central venous pressure > 16 cm H_2O – ICV Pressure (ICV)	P, N
Weight Loss of 4.5 kg in 5 days, due to HF treatment (WTL)	P
Minor Criteria	
Bilateral Ankle Edema – AnkleEdema (ANKED)	P, N
Dyspnea on Ordinary Exertion – DOExertion (DOE)	P, N
Hepatomegaly – Hepatomegaly (HEP)	P, N
Nocturnal Cough – NightCough (NC)	P, N
Pleural Effusion – PleuralEffusion (PLE)	P, N
Tachycardia (rate of ≥120 min^{-1}) – Tachycardia (TACH)	P

[a] P: Positive/Affirmation, N: Negative/Denial.