Journal of Cardiac Failure Vol. 20 No. 7 2014

Clinical Investigations

Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record

RAJAKRISHNAN VIJAYAKRISHNAN, MD,[1,2] STEVEN R. STEINHUBL, MD,[1,2,3] KENNEY NG, PhD,[4] JIMENG SUN, PhD,[4,5] ROY J. BYRD,[4] ZAHRA DAAR,[1,2] BRENT A. WILLIAMS, PhD,[1,2] CHRISTOPHER DEFILIPPI, MD,[6] SHAHRAM EBADOLLAHI, PhD,[4] AND WALTER F. STEWART, PhD[7]

Danville, Pennsylvania; La Jolla, California; Hawthorne, New York; Atlanta, Georgia; Baltimore, Maryland; and Concord, California

ABSTRACT

Background: The electronic health record (EHR) contains a tremendous amount of data that if appropriately detected can lead to earlier identification of disease states such as heart failure (HF). Using a novel text and data analytic tool we explored the longitudinal EHR of over 50,000 primary care patients to identify the documentation of the signs and symptoms of HF in the years preceding its diagnosis.

Methods and Results: Retrospective analysis consisted of 4,644 incident HF cases and 45,981 group-matched control subjects. Documentation of Framingham HF signs and symptoms within encounter notes were carried out with the use of a previously validated natural language processing procedure. A total of 892,805 affirmed criteria were documented over an average observation period of 3.4 years. Among eventual HF cases, 85% had ≥1 criterion within 1 year before their HF diagnosis, as did 55% of control subjects. Substantial variability in the prevalence of individual signs and symptoms were found in both case and control subjects.

Conclusions: HF signs and symptoms are frequently documented in a primary care population as identified through automated text and data mining of EHRs. Their frequent identification demonstrates the rich data available within EHRs that will allow for future work on automated criterion identification to help develop predictive models for HF. (*J Cardiac Fail 2014;20:459−464*)

Key Words: Heart failure, electronic health records, natural language processing.

The management of heart failure (HF) is one of the most critical challenges facing the health care system today. More than 5.8 million American adults have a diagnosis of HF and nearly 700,000 new cases are diagnosed each year.[1] Although the prevalence of HF is forecasted to increase by ∼25% over the next 20 years, it is already the leading cause of hospitalization for adults >65 years old and is responsible for almost $25 billion of direct medical costs annually. Those costs are expected to increase a staggering 215% by 2030.[2]

Early identification of those at risk of progressing to a HF diagnosis may provide an opportunity to improve both quality of life and reduce costs. However, the complexity and heterogeneity of the early clinical presentation of HF pose substantial challenges to its identification and diagnosis, especially in the primary care setting.[3]

The increasing presence of the electronic health record (EHR) in the primary care setting offers several advantages in our ability to explore early markers for the onset of conditions of interest, such as HF, in the years preceding clinical

From the [1]Department of Cardiology, Geisinger Medical Center, Danville, Pennsylvania; [2]Center for Health Research, Geisinger Medical Center, Danville, Pennsylvania; [3]Scripps Translational Science Institute, La Jolla, California; [4]T. J. Watson Research Center, IBM, Hawthorne, New York; [5]Georgia Institute of Technology, Atlanta, Georgia; [6]University of Maryland, Baltimore, Maryland and [7]Sutter Health, Concord, California.

Manuscript received September 17, 2013; revised manuscript received March 23, 2014; revised manuscript accepted March 27, 2014.

Reprint requests: Steven R. Steinhubl, MD, Director, Digital Medicine, Scripps Translational Science Institute, 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037. Tel: 858-554-5757; Fax: 858-546-9284. E-mail: steinhub@scripps.edu

See page 463 for disclosure information.

1071-9164/$ - see front matter

http://dx.doi.org/10.1016/j.cardfail.2014.03.008

diagnosis. They allow for access to longitudinal individual patient data in tens of thousands of patients, although the amount of data available makes manual review for pertinent findings almost impossible. Natural language processing (NLP) overcomes this limitation and allows for the automated extraction of pertinent data from free text information, such as found in medical encounter notes, which could potentially be used to provide clinical decision support.[4] We have previously shown that an NLP tool developed by our group to extract signs and symptoms potentially consistent with HF based on the Framingham criteria has good accuracy compared with expert human adjudication.[5,6]

In recent guidelines, the accuracy of signs and symptoms for the diagnosis of HF have been questioned.[7,8] The Framingham signs and symptoms criteria for HF diagnosis (Table 1), developed 40 years ago, are those classically associated with HF and are still used to diagnosis HF in research studies,[9,10] though their application to stratify risk for HF in a broad contemporary community based population has not been rigorously evaluated.

The primary aim of the present study was to use NLP of the EHR in a large primary care cohort to determine the prevalence of the Framingham criteria findings in both HF case and control subjects to lay the foundation for work to determine whether automated analytics of encounter notes in the EHR might enable the differentiation of subjects who would ultimately be diagnosed with HF.

Methods

Study Subjects

This was a case-control study using a retrospectively identified cohort of primary care patients who eventually developed HF and control subjects who did not. Patient EHRs dating from 2001 to 2010 within the Geisinger Health System were used to identify case and control subjects. The Geisinger Health System is an integrated health care system that provides health services in 31 counties of central and northeastern Pennsylvania and includes 41 community practice clinics that have been using the EPIC EHR since 2001. Data for this study were derived from the ~400,000 primary care patients served by these clinics.

Table 1. Framingham Diagnostic Criteria for Definite Heart Failure

Major Symptoms	Minor Symptoms
1. Paroxysmal nocturnal dyspnea (PND) or orthopnea	1. Bilateral ankle edema
2. Neck vein distension (JVD)	2. Nocturnal cough
3. Rales	3. Dyspnea on ordinary exertion
4. radiographic cardiomegaly	4. Hepatomegaly
5. acute pulmonary edema	5. Pleural effusion
6. s3 gallop	6. A decrease in vital capacity by 1/3 of the maximal value recorded*
7. Increased central venous pressure (>16 cm H_2O at RA)*	7. Tachycardia (>120 beats/min)
8. Circulation time of 25 s*	
9. Hepatojugular reflux (HJR)	
10. Weight loss of 4.5 kg in 5 days in response to treatment	

*Not used in the present analysis because not documented in routine clinical practice.

From these EHRs, we identified 4,644 incident HF cases with a clinical diagnosis based on meeting ≥1 of the following criteria: 1) HF diagnosis appearing on the problem list at least once; 2) HF diagnosis appearing in the EHR for 2 outpatient encounters; 3) ≥2 medications prescribed with the ordering provider associating that medication order with an ICD-9 diagnosis of HF; and 4) HF diagnosis appearing on ≥1 outpatient encounters and ≥1 medication prescribed with an associated ICD-9 diagnosis for HF. This operational diagnostic method for HF diagnosis has been previously validated.[11] The diagnosis date was defined as the first appearance of an HF diagnosis in the EHR. This means that the diagnosis of HF was based on EPIC EHR records, which comprise inpatient as well as outpatient diagnoses (more of the former than the latter as mentioned in criterion no. 1). Regarding criterion 2. (HF diagnosis appearing in the EHR for 2 outpatient encounters), once the HF diagnosis was confirmed based on this criterion, we considered the first of these visits to be the HF diagnosis date.

Approximately 10 eligible clinic-, sex-, and age-matched (in 5-year age intervals) control subjects were selected for each incident HF case (45,981 group-matched control subjects). Primary care patients were eligible as control subjects if they had no history of HF diagnosis before December 31, 2010. Control subjects were required to have had their first Geisinger Clinic office encounter within 1 year of the incident HF patient's first office visit and had ≥1 office encounter 30 days before or any time after the case subject's HF diagnosis date to ensure similar durations of observations among case and control subjects. In situations where 10 matches were not available, all available matches were selected. Therefore, for the purposes of this study we extracted the clinical notes portion of the EHRs for 51,625 patients. All patient encounters preceding the diagnosis of HF in case subjects and the matched date in control subjects were analyzed. In total, there were >3.3 million clinical notes, comprising >4 gigabytes of text data. The average number of clinical notes reviewed for cases was 25 (SD 19) and for controls 19 (SD 15).

EMR Data Extraction

An NLP application was developed and validated for identifying affirmations and denials of 14 of the 17 Framingham criteria for HF (Table 1). The remaining 3—circulation time of 25 seconds, increased central venous pressure, and a decrease in vital capacity by 1/3 of maximum on serial testing—were not identified, because they are not routinely documented in clinical practice. Details of the methods behind developing the automated text-mining tool have previously been published.[5] The program had a precision (or positive predictive value) of 0.925 and recall (or sensitivity) of 0.896 relative to manual chart review.[6]

Natural Language Processing

NLP systems recognize words or phrases as medical terms (in this case, Framingham criteria) that represent the domain concepts (named entity recognition), and they understand the relations between the identified concepts. We use unstructured information management architecture (UIMA) to identify clinically relevant entities in clinical notes mentioned as a part of the EHR. The entities are subsequently used for information retrieval and data mining of vast information available in the health records of patients.[4–6] Using this method can potentially identify any information available in the medical records of patients quickly and