Learning Clinical Workflows to Identify Subgroups of Heart Failure Patients

Chao Yan, MS[1], You Chen, PhD[1], Bo Li, PhD[1], David Liebovitz, MD[2], Bradley Malin, PhD[1]
[1]Vanderbilt University, Nashville, TN; [2]University of Chicago, Chicago, IL

Abstract

Heart Failure (HF) is one of the most common indications for readmission to the hospital among elderly patients. This is due to the progressive nature of the disease, as well as its association with complex comorbidities (e.g., anemia, chronic kidney disease, chronic obstructive pulmonary disease, hyper- and hypothyroidism), which contribute to increased morbidity and mortality, as well as a reduced quality of life. Healthcare organizations (HCOs) have established diverse treatment plans for HF patients, but such routines are not always formalized and may, in fact, arise organically as a patient's management evolves over time. This investigation was motivated by the hypothesis that patients associated with a certain subgroup of HF should follow a similar workflow that, once made explicit, could be leveraged by an HCO to more effectively allocate resources and manage HF patients. Thus, in this paper, we introduce a method to identify subgroups of HF through a similarity analysis of event sequences documented in the clinical setting. Specifically, we 1) structure event sequences for HF patients based on the patterns of electronic medical record (EMR) system utilization, 2) identify subgroups of HF patients by applying a k-means clustering algorithm on utilization patterns, 3) learn clinical workflows for each subgroup, and 4) label each subgroup with diagnosis and procedure codes that are distinguishing in the set of all subgroups. To demonstrate its potential, we applied our method to EMR event logs for 785 HF inpatient stays over a 4 month period at a large academic medical center. Our method identified 8 subgroups of HF, each of which was found to associate with a canonical workflow inferred through an inductive mining algorithm. Each subgroup was further confirmed to be affiliated with specific comorbidities, such as hyperthyroidism and hypothyroidism.

Introduction

Heart failure (HF) is one of the most common indications for admission to the hospital among older adults[1]. HF manifests in a clinically detectable manner when the heart is unable to supply an adequate flow of blood to meet the body's needs. HF is an important contributor to both the burden and cost of national healthcare expenditures. Over five million people in the United States are estimated to exhibit HF to some degree and management of the disease costs the nation an estimated $32 billion annually[1,2]. In 2001, the American Heart Association and American College of Cardiology refined the HF phenotype into four gross stages, which has led to the development and deployment of a wide array of management options for HF treatments. Yet, management of the disease is complicated by the fact that it often associates with a diverse collection of comorbidities (e.g., anemia, type 2 diabetes, various infections, and thyroid problems), which can manifest in different manners and combinations across the evolving stages of the disease. As a consequence, HF is also one of the conditions leading to high readmission rates in hospitals. Healthcare organizations (HCOs) have established treatment protocols and workflows for HF patients with different comorbidities and stages of progression[3]. Additionally, to improve the definition and management of HF, various investigations[4-6] have been conducted to computationally specify the phenotype[7] and workflows affiliated with its management[8].

Traditionally, research has aimed to refine HF into several clinical subphenotypes based on heart-related issues, such as systolic or diastolic heart failure[9]. These are natural subtypes that HCOs can rely upon to design specialized treatment plans or clinical workflows[8,10]. While such research, and subsequent clinical designations, can assist HCOs to more effectively manage HF patients, they often rely on an expert informed perspective and experience. As a consequence, they involve a substantial amount of human effort[4] and focus on clinical phenomena that are expected to categorize the HF population. To reduce human effort and learn clinical concepts (or management pathways) that are not necessarily anticipated, several studies have shown that data-informed methodologies can be invoked to infer complex comorbidities[11-13], clinical workflows[14], and care teams[15]. Many of these studies rely on the co-accesses to patients' electronic medical records (EMRs) to infer collaborative care teams or workflows for specific diseases[11,15-16], however, such studies have focused on all possible diseases and the workings of an HCO in general. In doing so, they have neglected how such views are influenced by conditioning the investigation on a specific complex disease, such as HF.

The investigation communicated in this paper is motivated by the expectation that complex diseases, like HF, are associated with a range of workflows in an HCO. These workflows are unlikely to be explicitly documented because

they are associated with subtypes of the disease and/or comorbidities that lead to *ad hoc* coordination. If such workflows can be detected through a data-informed method, they may be refined and resourced by an HCO to more effectively manage patients of a certain HF subtype.

Thus, in this paper we study four-months of EMR data, collected in 2010, from Northwestern Memorial Hospital for over 750 HF inpatient episodes. We introduce a data-informed framework to infer the underlying workflows that transpire in the clinical enterprise. We then map the learned workflows into a similarity measure to characterize patients into different subgroups. Finally, we show that these subgroups have a strong correlation with a range of diagnoses (e.g., hyper- and hypothyroidism) and procedures. Specially, our investigation suggests there are a minimum of 8 subgroups of HF patients, each of which is associated with a canonical workflow. It should be recognized from the outset that a subgroup does not indicate that they are a distinct population in their phenotype per se, but that they are a subgroup in the manner by which they are managed.

Background

This paper introduces a framework to identify subgroups of HF via inferred clinical workflows. Since this work involves workflow subgroup identification via inference methods, we take a moment to review related work in 1) workflow modeling and 2) subgroup discovery. When possible, we show how these methods have specifically been applied to HF populations.

Workflow Modeling

Workflow modeling and analysis has shown promise in a wide array of settings, ranging from general business management to specific clinical domain domains. For instance, van der Aalst and colleagues demonstrated how high-level Petri nets can model the workflows in an office environment, with a particular focus on how information systems support the control of office work[17]. To support the effort, they developed a workflow management system (WMS) prototype based on such formalizations. They subsequently extended the notion of a WMS to support dynamic changes[18]. Workflow modeling from a clinical perspective is more complex than many office settings because HCOs are composed of a large number of interacting departments and individuals who coordinate as availability and need dictates.

Still, there has been some success in this domain. In particular, Chen and colleagues introduced a method to infer clinical workflows and measure their efficiency via the utilization of EMR event logs. It was shown that these workflows naturally partition into four general types according to their average and variance in their efficiency[11]. They posited that certain inefficiencies were likely due to the complexity of the patients. While the methods introduced in their investigation enabled the evaluation of workflow efficiency, it did not condition the workflows on specific patient phenotypes or determine if subgroups for the management of a specific disorder led to the manifestation of disparate workflows. EMR access logs were also used by Li and colleagues to infer workflows through a method based on hidden Markov models (HMMs)[19]. These HMMs were utilized to characterize the behavior of EMR users, as well as detect anomalous activities. However, this investigation did not study the clinical meaning of the workflow or how they could be specialized to certain patient subgroups.

Subgroup Identification

There is evidence to suggest that identifying subgroups of patients based on their clinical conditions can be applied to design personalized treatment plans. For instance, Mugge and colleagues identified a subgroup of patients with nonrheumatic atrial fibrillation (Afib) with an increased risk for cardiogenic embolism by assessing left atrial appendage function[20]. They identified two distinct patient groups according to appendage flow patterns: 1) well-defined peak filling with visible fibrillatory contractions of the appendage wall and 2) irregular, very low, peak filling with almost no visible appendage contractions. While such subgroup identification is notable from a descriptive perspective, it focuses on clinical diagnosis and, thus, neglects how to design management routines for Afib patients in a healthcare environment that is subject to communication challenges and resource limitations. Soulakis and colleagues[16] utilized the EMR access logs of over 500 HF patients to identify seven networks of around 5000 care providers. However, this work is limited in that it neglects the association between the network of care providers and the clinical conditions of the patients.

Methods

We designed a data-informed framework to identify HF subgroups to consist of three steps: 1) infer patient subgroups through event sequences, 2) learn a workflow for each subgroup and 3) assign diagnosis and procedure codes as

phenotype labels for each subgroup. To gain intuition into how this framework works, we begin with an introduction of the dataset used in this investigation.

Dataset

The dataset for this study is summarized in Table 1. The records are drawn from comprehensive access logs and billing information derived from the EMR system at Northwestern Memorial Hospital (NMH) over a four month period during 2010. This dataset, which we refer to as NMH-D, consists of 1,138,555 access events distributed over 16,567 unique inpatient stays. We extracted all patients diagnosed with heart failure by selecting any patient with an ICD-9 code in the range 428.0 through 428.9 and we refer to this specific HF extract as the NMH-HF dataset. Each access event is affiliated with the following attributes: 1) pseudonym of the inpatient, 2) ID of the EMR user, 3) reason for the access event (as designated by the user according to a pull-down list), 4) date and timestamp for the access, 5) general physical location in NMH where the patient is located, and 6) the clinical service on which the patient is managed (e.g., general medicine vs. obstetrics). Each inpatient episode is affiliated with its ICD-9 billing codes, which were assigned after discharge.

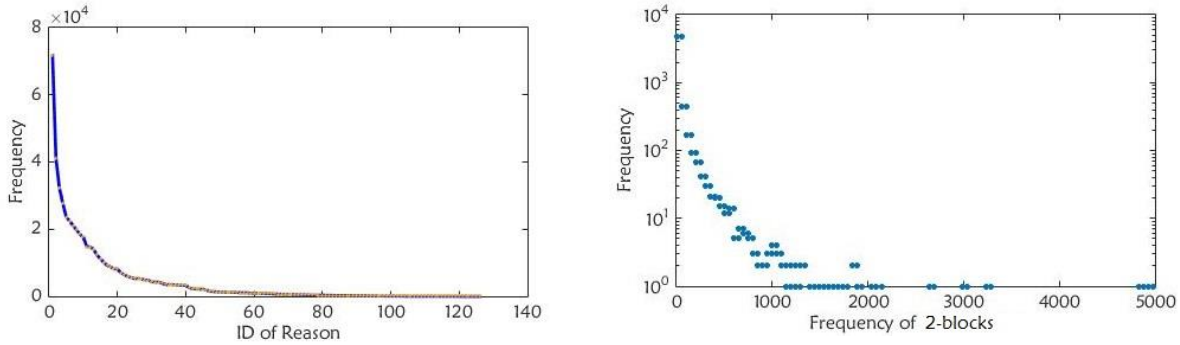
Table 1. A summary of the datasets in this study.

Dataset	Accesses	Patients	Reasons	Number of 2-blocks
NMH-D	1,138,555	16,567	142	(*not computed*)
NMH-HF	272,685	785	126	5823

We use the *Reason for Access* in the dataset, as opposed to the EMR user, as the smallest level of granularity associated with a workflow to mitigate noise in the analysis. Specifically, for each inpatient i, the corresponding event sequence R^i (defined in Table 2) is a series of ordered reasons for access. To orient the reader, the following is an example of a sequence of reasons one might encounter for a certain inpatient:

$\ldots \rightarrow$ *Attending Physician/Provider* $\rightarrow$ *Primary Staff Nurse* $\rightarrow$ *Resident-Inpatient Primary Service* $\rightarrow$ *Pharmacist* $\rightarrow \ldots$

In this work, we define an *n-block* as a sequence of n consecutive access events (with $n \geq 2$). Thus, a 2-*block* is defined as two consecutive access events. For instance, if an access event *Attending Physician/Provider* has a consecutive access event *Primary Staff Nurse*, then *Attending Physician/Provider* $\rightarrow$ *Primary Staff Nurse* is the corresponding 2-block. In fact, 2-blocks reflect the order relations between two neighbor events. If 2-blocks appears infrequently in event sequences, then the order relation between its containing two events are weak, which could be utilized by our framework to filter weak order relations, most of which are noise in EMR system functionality[9,22].



(a) *Reasons* (b) *Blocks with 2 Reasons*

Figure 1. Frequency distribution of a) Reasons and b) 2-blocks in NMH-HF dataset.

Figure 1 illustrates the frequency distribution of all reasons and the distribution of the frequency of 2-blocks. As can be observed in Figure 1(a), around 46% of the reasons (58 of 126) have a frequency of 1,000 or larger, which suggests that a reason subset is frequently invoked by care providers for different purposes. At the same time, in Figure 1(b), it can be seen that that around 90% of 2-blocks (5,246 of 5,823) appear no more than 100 times in the system. This suggests that the system is not dominated by several frequent 2-blocks, which indicates that our framework has an opportunity to identify subgroups through such event sequences.

Subgroups Identification Framework

The framework is composed of three main components: 1) generate patient subgroups via refined event sequences, 2) learn workflows for the subgroup population, and 3) extract diagnosis and procedure labels for subgroups. Here, refined event sequences are the transformation of original event sequences through applying filtering based on frequency. To help the reader understand our framework, we provide a legend of the common notation invoked throughout this paper in Table 2. The specific algorithmic process associated with the framework is communicated in Figure 2.

Table 2. A legend of the notation used in this paper.

Symbols	Interpretation
D	A set of event sequences, where an event sequence is affiliated with one patient episode
D_f	A set of event sequences filtering high frequency reasons in D
D'	A set of refined event sequences
ϕ	A tuple set of reasons and corresponding frequency
B_(2_block)	A tuple set of 2-block and corresponding frequency
B_(n_block)	A tuple set of blocks and corresponding frequency
R^i	An event sequence in D charactering inpatient *i's* episode
R'^i	A refined event sequence in D' characterizing inpatient *i*'s episode
r_t	A reason appearing at time stamp t
$b_{t,t+1}$	A 2-block with time stamp t and t+1
D_(2_block)	A 2-block set extracted from D_f
D'_(2_block)	A refined 2-block set from *D_(2_block)* by filtering lowest frequency 2-blocks
D_(n_block)	A set of blocks extracted from D', where each block can have a different length
D'_(n_block)	A set of blocks obtained by filtering lowest frequency blocks in *D_(n_block)*
m_r	The number of filtered highest frequency reasons in ϕ
m_b	The number of filtered lowest frequency 2-blocks in *D_(2_block)*
m_g	The number of filtered lowest frequency blocks in *D_(n_block)*
M_{PB}	A binary matrix characterizing the relationship between patient episodes and blocks in *D'_(n_block)*
M_{PB}'	A matrix transformed from M_{PB} using a polynomial kernel

1) Generate Patient Subgroups by Refined Access Sequences

The subgroup identification process is partitioned into five steps: 1) filter high frequency reasons contained in each event sequence in D into D_f, 2) generate 2-blocks in D_f and create a set *D_(2_block)*, 3) filter low frequency 2-blocks in *D_(2_block)* into *D'_(2_block)*, 4) use 2-blocks in *D'_(2_block)* to refine sequences in D_f to form a new sequence set D' and 5) clustering subgroup patients by using similarity of block elements in refined sequences in D'. Each of these steps is detailed in the following descriptions:

Filter high frequency reasons. Each raw event sequence in D can contain high frequency event reasons. We remove high frequency reasons because they correspond to the most general aspects of the workflow. These are unlikely to communicate clinical context that is critical to modeling a specific workflow. For example, both *Primary Staff Nurse* and *Primary Assistive Staff* appear the greatest number of times in event sequences. While it is anticipated that nurses provide support to patient care, they are critical to almost all aspects of the inpatient setting. In many respects, general nursing staff is akin to the stop words (e.g., prepositions or articles) in natural language text. And, as many investigations in natural language processing have illustrated, such information, can be triaged to improve pattern discovery. The new event sequence set D_f is generated by filtering the high frequent reasons.

Generate 2-blocks and filter low frequency (or noise). These two steps are incorporated because the event sequences are extracted from the access logs of EMR system, which contain noise in the order of relations[22]. We assume that noisy relations exist in low frequency blocks and, thus, filter weak relations from the event sequences D_f.

To do so, each event sequence in D_f is segmented into blocks by invoking the 2-blocks in $D'\_(2\_block)$. Thus, each event sequence is represented by varying sized blocks. With this transformation and linkage, D_f is transformed into D'. Thus, D' is the set of sequences where high frequent reasons and low frequent 2-blocks are both removed. The processing details are described from lines 6 through 17 in the algorithm in Figure 2. For example, after the process above, a patient in our dataset has a new sequence formed by two blocks $b_1 \rightarrow b_2$, where b_1 is a reason sequence:

Resident - Outpatient/ED/Proc Primary → *Patient Care* → *Radiology Technologist* → *Registration*

and b_2 is

Rehab Assigned Therapist → *Consultant* → *Rehab Assigned Therapist* → *Charging/Orders* → *Med Rec Coding*

D' is represented by linkage of new blocks and we extract all of the blocks in D' into $D\_(n\_block)$. At this point, we filter out low frequent blocks in $D\_(n\_block)$ to generate a new set $D'\_(n\_block)$.

Next, we cluster patients into subgroups using their associated bocks in $D'\_(n\_block)$. To do so, we generate a patient-by-block matrix M_{PB} to represent relations between patients and blocks. We then apply a polynomial kernel on M_{PB} to transform it to a new matrix M_{PB}' and perform k-means clustering.[1]

Input: D, a set of event sequences; m_r, the number of filtered highest frequency reasons; m_b, the number of filtered lowest frequency 2-blocks; m_g, the number of filtered lowest frequency blocks;
Output: $C = \{C_1, C_2, \ldots, C_k\}$, subgroups of patients

1: Let $\phi = \{(\varphi_i, f_i)\} \leftarrow$ (reason ID, frequency) tuple set from D
2: $D_f \leftarrow D \setminus$ (top m_r reasons with high frequency in ϕ)
3: $D\_(2\_block) = \{\tau_j\} \leftarrow$ 2-block set from D_f
4: Let $B\_(2\_block) = \{(\tau_j, f_j)\} \leftarrow$ (2-block, frequency) tuple set from D_f
5: $D'\_(2\_block) \leftarrow D\_(2\_block) \setminus$ (top m_b 2-blocks with low frequency in $B\_(2\_block)$)
// Where "\" indicates set exclusion
6: $D' \leftarrow \emptyset$;
7: **for each** R^i:
8: $R'^i \leftarrow \emptyset$, $c \leftarrow 1$, $B_c \leftarrow \emptyset$
9: **for each** consecutive 2-block $b_{t,t+1} = r_t \rightarrow r_{t+1}$ in R^i:
10: **if** $B_c \equiv \emptyset$ and $b_{t,t+1} \in D'\_(2\_block)$: $B_c \leftarrow r_t \rightarrow r_{t+1}$; **break**;
11: **end if**
12: **if** $b_{t,t+1} \in D'\_(2\_block)$: $B_c \leftarrow B_c \oplus r_{t+1}$; // Expand reason blocks at the tail
13: **else:** $R'^i \leftarrow R'^i \oplus B_c$; c ++; $B_c \leftarrow \emptyset$; // Form new representation of R^i
14: **end if**
15: **end for**
16: $D' \leftarrow D' \cup R'^i$;
17: **end for**
18: $D\_(n\_block) = \{\mu_k\} \leftarrow$ block set with different sizes from D'
19: Let $B\_(n\_block) = \{(\mu_k, f_k)\} \leftarrow$ (reason block, frequency) tuple set from D'
20: $D'\_(n\_block) \leftarrow D\_(n\_block) \setminus$ (top m_g blocks with the low frequency in $B\_(n\_block)$)
21: $M_{PB} \leftarrow$ binary matrix indicating the relationship between patients and blocks in $D'\_(n\_block)$
22: $M_{PB}' \leftarrow$ apply polynomial kernel on M_{PB}
23: $C = \{C_1, C_2, \ldots, C_k\} \leftarrow k$-means cluster of M_{PB}'
24: **return** C

Figure 2. Pseudocode of the cluster generation algorithm.

2) Learn workflows for each subgroup

Each subgroup is clustered using blocks in $D'\_(n\_block)$. Each event sequence (a patient episode) is characterized by these blocks. We group all blocks characterizing a subgroup into a workflow to represent the clinical process for this

[1] Applying the kernel makes it easier to separate groups by projecting the data into a higher set of dimensions.

type of HF patient. We then invoke an inductive mining algorithm (as implemented in ProM[23]) to infer and visualize workflows.

3) Extract Diagnosis and Procedure Labels

Care must be taken when learning workflows through a data-informed strategy, as they may not have labels that readily translate into administrative applications. This is important because the clinical workflows that are based on expert knowledge are associated with known semantics. Thus, we aim to relate inferred workflows to clinical context. To do so, we assign labels to each subgroup's workflow by discovering the billing codes that are the most discriminative for the workflow.

The labels for each workflow are derived from a series of processes. First, we extract the most frequent diagnosis and procedure codes from all inpatients in each subgroup. Next, we apply a *Z*-test (hypothesis testing method based on proportionality) to compute the *p*-value for each diagnosis and procedure codes in each subgroup[23]. For each subgroup C_a, the most distinguishable billing codes between pairs of subgroups (C_a, C_x), $x \neq a$, are extracted. Finally, we union these the distinguishable codes to characterize each subgroup and their affiliated workflows.

Results

This section reports on results from 1) a general view of the 10 identified distinct subgroups and 2) a case study of two representative subgroups - in terms of their affiliated workflows and the diagnosis and procedure codes that distinguish the subgroup from others.

Table 3. A summary of the number of patients and representative conditions for the HF subgroups.

Subgroup	Size	Billing Terms with High Frequency	Phenotype
C_1	447	401.9: Unspecified essential hypertension 416.8: Other chronic pulmonary heart diseases 425.4: Other primary cardiomyopathies 414.0: Coronary atherosclerosis of native coronary artery V45.81: Aortocoronary bypass status V58.61: Long-term (current) use of anticoagulants	General HF
C_2	14	276.2: Acidosis 280.0: Iron deficiency anemia secondary to blood loss 578.9: Hemorrhage of gastrointestinal tract V65.3: Dietary surveillance and counseling V10.04: Personal history of malignant neoplasm of stomach	HF associated with hemorrhage
C_3	15	428.4: Combined systolic and diastolic heart failure 276.8: Hypopotassemia V15.05: Allergy to other foods V15.06: Allergy to insects and arachnids	HF associated with anaphylactic reaction
C_4	22	263.9: Unspecified protein-calorie malnutrition 038.9: Unspecified septicemia 584.9: Acute kidney failure V45.11: Renal dialysis status	HF associated with renal failure and sepsis
C_5	13	729.5: Pain in limb 276.5: Dehydration 275.4: Hypocalcemia V42.0: Kidney replaced by transplant	HF associated with renal transplantation
C_6	58	276.7: Hyperpotassemia 584.9: Acute kidney failure 275.3: Disorders of phosphorus metabolism 428.2: Acute on chronic systolic heart failure 287.5: Thrombocytopenia V45.11: Renal dialysis status	Hyperthyroidism HF
C_7	5	(*too small to make a determination*)	(*too small to make a determination*)
C_8	6	(*too small to make a determination*)	(*too small to make a determination*)
C_9	34	244.9: Unspecified acquired hypothyroidism 733.0: Osteoporosis 401.9: Unspecified essential hypertension 425.4: Other primary cardiomyopathies	Hypothyroidism HF
C_{10}	171	403.9: Hypertensive chronic kidney disease	Acute or chronic renal failure HF

		585.9: Chronic kidney disease 584.9: Acute kidney failure 285.9: Anemia V45.01: Cardiac pacemaker in situ V45.82: Percutaneous transluminal coronary angioplasty status	

HF Subgroups Identified

By applying the framework to the 785 CHF inpatient episodes, we discovered 10 distinct subgroups as summarized in Table 3. For reference purposes, we use C_i to represent subgroup i. In this table, each subgroup is labeled using descriptions of the representative diagnosis and procedure billing codes. For instance, C_1 is a subgroup affiliated with general heart failure, C_3 is affiliated with allergies associated with HF, and C_6 is affiliated with hypothyroidism. We note that we neglect subgroups C_7 and C_8 because of their small size (only 5 and 6 patients, respectively). The representative billing codes come from two sources: 1) distinct codes between this subgroup and any other subgroup, and 2) high frequency codes associated with this subgroup.

Table 3 also reports the size and clinical label of each subgroup. It can be seen that each subgroup has a distinct specialized phenotype. For instance, C_1, the largest subgroup, is affiliated with the diagnosis of HF, which indicates that most of the patients went through a process associated with management of general comorbidities (e.g., hypertension, primary cardiomyopathies, and coronary atherosclerosis). C_{10}, the second largest subgroup, is affiliated with chronic kidney disease (CKD).

Case Study of Hyperthyroidism and Hypothyroidism in HF

To understand the intuitive nature of the identified subgroups, we report on a case study that compares two subgroups. Specifically, we focus on subgroups C_6 and C_9, which are similar in the number of patients they cover (58 and 34, respectively) and correspond to two typical HF subtypes. We illustrate the differences in these subgroups in terms of their clinical concepts and inferred workflows.

Clinical Differences. Figure 3 shows the concordance of the frequency distribution for the most significant codes (based on their p-values) affiliated with the C_6 and C_9 subgroups. Specifically, each (x,y) point corresponds to a specific code, where x and y is the proportion of patients in C_6 and C_9 who received the code, respectively. As such, codes that are close to the dashed diagonal line indicates they have a similar frequency in the two investigated subgroups. Clearly, these codes do not distinguish between the subgroups. By contrast, codes that are distant from the line can distinguish one subgroup from the other. In this figure, we marked diagnosis codes with a star and procedure codes as a circle.

We found that the patients in cluster C_6, were primarily diagnosed with i) *hyperpotassemia*, ii) *disorders of phosphorus metabolism*, iii) *acute kidney failure, unspecified*, and iv) *thrombocytopenia*. This combination of diagnoses makes sense intuitively. This is because excess potassium and phosphorus caused by hyperthyroidism are both associated with kidney failure[21]. Based on knowledge of these symptoms, it can be inferred that this subgroup suffers from **Hyperthyroidism HF**.

By contrast, the 34 patients in cluster C_9 have the following diagnosis labels: i) *unspecified acquired hypothyroidism*, ii) *osteoporosis*, and iii) *unspecified essential hypertension*. Osteoporosis is obviously associated with hypothyroidism because of a decreasing amount of calcium. And thus, we label C_9 with **Hypothyroidism HF**.

These HF subgroups demonstrate that differences in the inferred learned workflows may be associated with the clinical status, as well as procedures performed on the patients.

Workflows Differences. Though subgroup C_6 covers a larger number of patients than C_9, we find that it exhibits a simpler workflow structure (as shown in Figure 4). To illustrate, we first compare the main reasons of these two workflows. Subgroup C_6 contains: i) *Other Physician*, ii) *Covering Therapist*, iii) *Coordinator*, and iv) *Advanced Practice Nurse* as the main reasons (and corresponding roles), which appear to be affiliated with a generic healthcare process. By contrast, the workflow for C_9 has a more complicated structure, in which there is a greater diversity in the reasons. Specifically, this structure includes i) *Student Nurse*, ii) *Consultant*, iii) *Patient Care*, iv) *Assigned Staff*, v) *Advanced Practice Nurse*, vi) *Coordinator*, vii) *Anesthesiologist*, viii) *Dietary Clerk*, ix) *Radiologist*, and x) *Rehab Assigned Therapist*.

Moreover, this workflow contains some special reasons, such as xi) *Radiology Nurse/Resident/Technologist* and xii) *Radiology management*, which are likely linked with the diagnosis of *osteoporosis* induced by *hypothyroidism*. This

case study suggests that our initial hypothesis – that HF subtypes associate with different workflows and management processes – has standing. Moreover, we believe this case study is a clear illustration of how data-informed workflow mining can be leveraged to learn different subgroups of patients that can be subsequently labeled in a clinically meaningful manner.

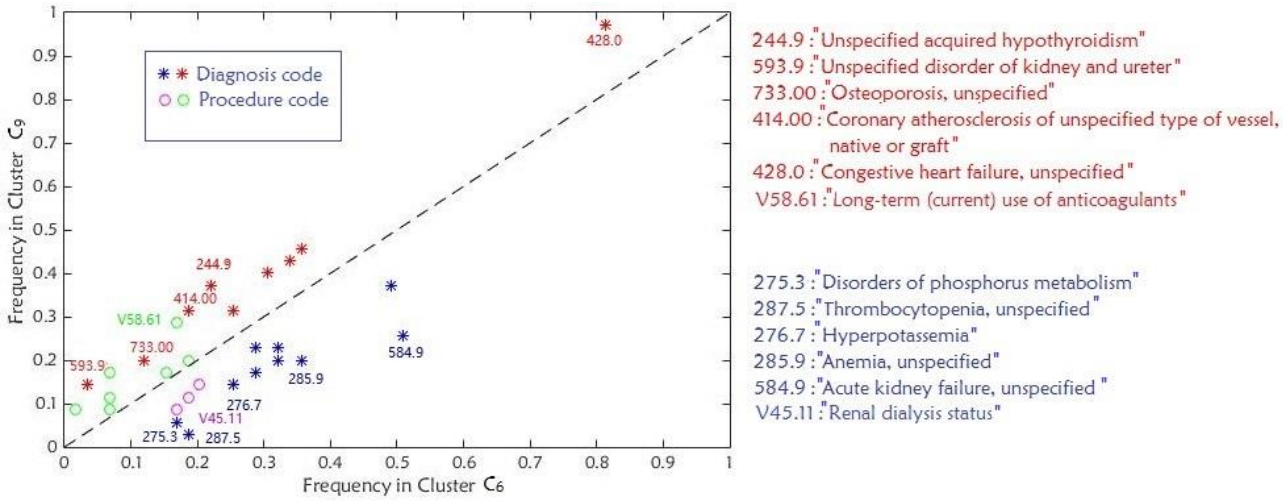
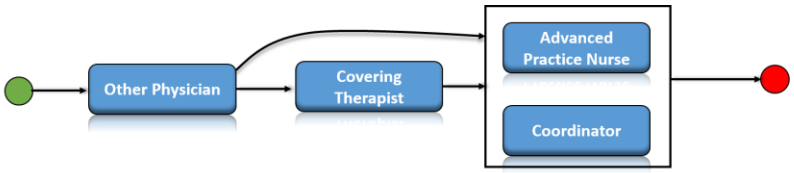
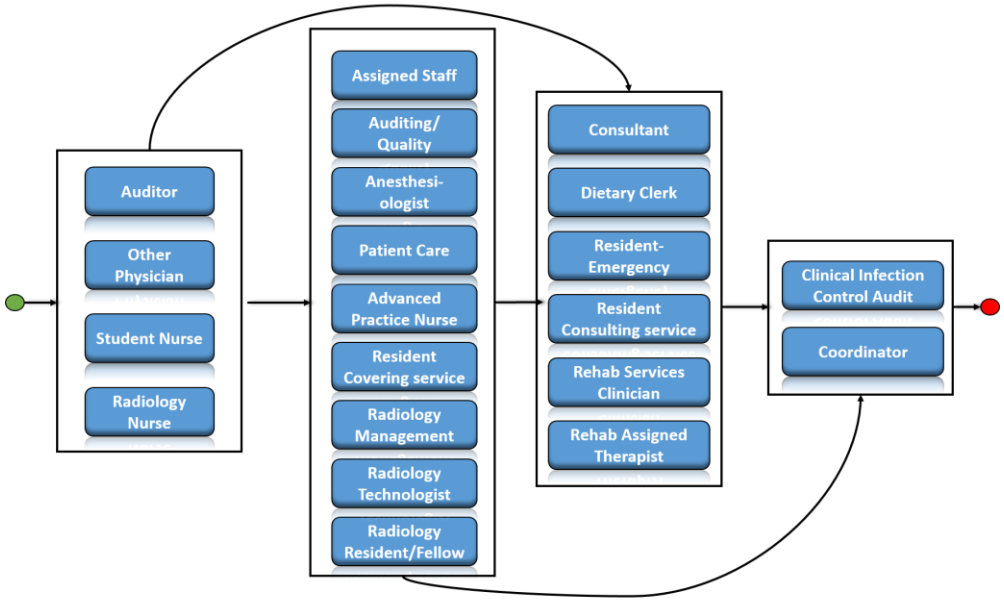


Figure 3. Concordance in the frequency of the diagnosis and procedural billing codes between HF subgroups C_6 and C_9.



(a) Workflow of C_6



(b) Workflow of C_9

Figure 4 A simplified view on the workflow structure of subgroups (a) C_6 and (b) C_9.

Discussion

We introduced a method to identify subgroups of HF through refined event sequences and subsequently infer workflow and phenotype for every subgroup. This approach is substantially different from the traditional definition of phenotypes rooted solely on metabolic and clinical observations. The subgroups identified in our framework share similar workflow patterns, suggesting they are managed in a similar way in clinical practice. To the best of our knowledge, this is the first investigation to show that subgroups of a complex disorder, such as HF, can be learned through workflows (in the form of event sequences) in the clinical enterprise. Our findings are further notable because they suggest that workflows stretching across departments and wards of an HCO can be learned from EMR utilization. We believe this methodology provides opportunities to make management strategies explicit and tune resource allocations accordingly. At the same time, we believe the learned workflows have standing because they were shown to correlate with diagnoses and procedures codes exhibited in the corresponding patient groups (information that was not used in the clustering process). As such, it may be possible to develop predictive models that assign a patient to a predefined and semi-personalized management regimen.

Despite our discoveries, we acknowledge that this is a pilot study on an HF population. There are several limitations of this project, which we wish to highlight for further refinement and future investigation. First, we relied on ICD-9 codes in the 428.* range, rather than rigorously validated computational phenotypes of HF[14,19,20] to define our cohort. Though the HF phenotype is considered a relatively well-defined diagnosis, it is a disease with multiple stages and confounding factors (as our hyperthyroidism and hypothyroidism subgroups illustrated). It is further conceivable that some of the patients who presented to Northwestern Memorial Hospital for a certain primary diagnosis (e.g., stroke) might have been treated for HF without its documentation in an ICD-9 billing code. As such, we believe that our selection criteria enable a highly precise investigation, but does not cover the gamut of HF patients.

Second, the workflows associated with the HF subgroups were only reviewed by one clinician. Rather, we mainly relied on identifying subgroups of HF in an information theoretic sense (e.g., similarity analysis of sequences of access reasons). Our method identified 10 subgroups and it appeared as though 8 had a clear clinical context differentiating from other subgroups (while 2 were too underpowered due to a small number of patients to make any judgement about). Still, before inferred workflows can be relied upon, they will require review by additional administrative and clinical experts to determine if they can be translated into decision support tools for an HCO.

Finally, we recognize that the size of the HF cohort is relatively small. During the four months of documented inpatient stays, we encountered less than a thousand patients treated for HF. Although our work yielded meaningful findings, we may not have captured all of the notable workflows or diagnostic labels for such workflows. For such an investigation to be useful for the HCO, we will need to investigate a large sample size over a longer period of time. It would also be ideal to compare the workflows, and the associated labels, with EMR data from other healthcare systems.

Conclusions

HF is a complex condition accompanied by diverse complications that progress across, a minimum of, four stages. HCOs have adopted various strategies (e.g., optional treatment plans for different developmental stages and complications) to improve quality of life for HF patients and reduce burden, as well as cost, for healthcare systems. Identifying subgroups of HF can assist in the management of patients with this disease. We introduced a data-informed framework to identify subgroups of HF patients through utilization of the EMR. For each subgroup, we provided external validation via the diagnosis and procedure codes of the corresponding patients. Our framework was evaluated on the event sequences of 785 HF inpatients from a large academic medical center. In doing so, we identified 8 HF subgroups, each of which was confirmed to be associated with a specific condition of HF (e.g., hyperthyroidism and hypothyroidism). Furthermore, each subgroup was characterized by different patterns of workflow. For instance, hypothyroidism associated HF involved more complex workflows than hyperthyroidism HF. We acknowledge that this investigation is a pilot study and further investigation is required with administrative review and validation across disparate healthcare enterprises.

Acknowledgements

This research was supported, in part, by the following grants from the National Library of Medicine at the National Institutes of Health K99LM011933, R00LM011933 and R01LM010207. The content in this work is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Go AS, Mozaffarian D, Roger VL, et al. Heart disease and stroke statistics—2013 update: a report from the American Heart Association. Circulation. 2013; 127(1): e6–245.
2. Heidenreich PA, Trogdon JG, Khavjou OA, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011; 123(8): 933–44.
3. Voigt J, John MS, Taylor A, Krucoff M, Reynolds MR, Gibson CM. A reevaluation of the costs of heart failure and its implications for allocation of health resources in the United States. Clin Cardiol. 2014; 37(5): 312–21.
4. Ruskoaho H. Cardiac hormones as diagnostic tools in heart failure. Endocr Rev. 2003; 24(3): 341-56.
5. Earnest MA, Ross SE, Wittevrongel L, Moore LA, Lin CT. Use of a patient-accessible electronic medical record in a practice for congestive heart failure: patient and physician experiences. J Am Med Inform Assoc. 2004; 11(5): 410-17.
6. Reingold S, Kulstad E. Impact of human factor design on the use of order sets in the treatment of congestive heart failure. Acad Emerg Med. 2007; 14(11): 1097-105.
7. Ahmad T, Pencina MJ, Schulte PJ, et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J Am Coll Cardiol. 2014; 64(17): 1765-74.
8. Dang J, HedayatiA, Hampel K, Toklu C. An ontological knowledge framework for adaptive medical workflow. J Biomed Inform. 2008; 41(5): 829-36.
9. Komamura K. Similarities and differences between the pathogenesis and pathophysiology of diastolic and systolic heart failure. Cardiol Res Pract. 2013: 824135.
10. Casper GR, Karsh BT, Or CKL, et al. Designing a technology enhanced practice for home nursing care of patients with congestive heart failure. AMIA Annu Symp Proc. 2005: 116-20.
11. Chen Y, Xie W, Gunter CA, et al. Inferring clinical workflow efficiency via electronic medical record utilization. AMIA Annual Symp Proc. 2015: 416-25.
12. Ho JC, Ghosh J, Steinhubl S, Stewart W, Denny JC, Malin B, Sun J. Limestone: high-throughput candidate phenotype generation via tensor factorization. J Biomed Inform. 2014; 52: 199–211.
13. Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium, J Am Med Inform Assoc. 2013; 20: e341–8.
14. Chen,Y Ghosh J, Bejan CA, et al. Building bridges across electronic health record systems through inferred phenotypic topics. J Biomed Inform. 2015; 55: 482-93.
15. Chen Y, Lorenzi N, Nyemba S, Schildcrout JS, Malin B. We work with them? Health workers interpretation of organizational relations mined from electronic health records. Int J Med Inform. 2014; 83(7): 495-506.
16. Soulakis ND, Carson MB, Lee YJ, et al. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. J Am Med Inform Assoc. 2015; 22(2): 299-311.
17. van der Aalst WMP. The application of Petri nets to workflow management. Journal of Circuits, Systems, and Computers. 1998; 8(1): 21-66.
18. van der Aalst WMP, ter Hofstede AHM, Kiepuszewski B, Barros AP. Workflow patterns. Distributed and Parallel Databases. 2003; 14(1): 5-51.
19. Li X, Xue Y, Malin B. Detecting anomalous user behaviors in workflow-driven web applications. Proc 31st IEEE International Symposium on Reliable Distributed Systems. 2012: 1-10.
20. Mügge A, Kühn H, Nikutta P, Grote J, Lopez JAG, Daniel WG. Assessment of left atrial appendage function by biplane transesophageal echocardiography in patients with nonrheumatic atrial fibrillation: identification of a subgroup of patients at increased embolic risk. J Am Coll Cardiol. 1994; 23(3): 599-607.
21. Chen Y, Nyemba S and Malin B. Auditing medical records accesses via healthcare interaction networks. AMIA Annual Symp Proc. 2012: 93-102.
22. Hartigan, JA, Wong MA. Algorithm AS 136: A *k*-means clustering algorithm. J Royal Statistical Society, Series C (Applied Statistics). 1979; 28(1): 100-8.
23. Leemans, Sander JJ, Dirk Fahland, van der Aalst WMP. Process and deviation exploration with inductive visual miner. Proc 12th International Conference on Business Process Management. 2014: 46.
24. Malin, BA. Correlating web usage of health information with patient medical data. AMIA Annual Symp Proc. 2002: 484-8.
25. Klein I, Danzi S. Thyroid disease and the heart. Circulation. 2007; 116(15): 1725-35.