Journal of Biomedical Informatics 58 (2015) 186–197

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data

Yiye Zhang [a,*], Rema Padman [b], Nirav Patel [c]

[a] School of Information Systems Management, The H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA, United States
[b] The H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA, United States
[c] Teredesai, McCann & Associates, P.C., Pittsburgh, PA, United States

ARTICLE INFO

Article history:
Received 1 April 2015
Revised 18 June 2015
Accepted 10 September 2015
Available online 28 September 2015

Keywords:
Clinical pathway
Clinical practice guideline
Visualization
Chronic kidney disease

ABSTRACT

Objective: Clinical pathways translate best available evidence into practice, indicating the most widely applicable order of treatment interventions for particular treatment goals. We propose a practice-based clinical pathway development process and a data-driven methodology for extracting common clinical pathways from electronic health record (EHR) data that is patient-centered, consistent with clinical workflow, and facilitates evidence-based care.

Materials and methods: Visit data of 1,576 chronic kidney disease (CKD) patients who developed acute kidney injury (AKI) from 2009 to 2013 are extracted from the EHR. We model each patient's multi-dimensional clinical records into one-dimensional sequences using novel constructs designed to capture information on each visit's purpose, procedures, medications and diagnoses. Analysis and clustering on visit sequences identify distinct types of patient subgroups. Characterizing visit sequences as Markov chains, significant transitions are extracted and visualized into clinical pathways across subgroups.

Results: We identified 31 patient subgroups whose extracted clinical pathways provide insights on how patients' conditions and medication prescriptions may progress over time. We identify pathways that show typical disease progression, practices that are consistent with guidelines, and sustainable improvements in patients' health conditions. Visualization of pathways depicts the likelihood and direction of disease progression under varied contexts.

Discussion and conclusions: Accuracy of EHR data and diversity in patients' conditions and practice patterns are critical challenges in learning insightful practice-based clinical pathways. Learning and visualizing clinical pathways from actual practice data captured in the EHR may facilitate efficient practice review by healthcare providers and support patient engagement in shared decision making.

1. Introduction

Evidence-based medicine is widely acknowledged as a systematic approach for delivering consistent, credible, and safe healthcare [1]. It is a critical assumption in the Affordable Care Act of 2010 to achieve coordinated, patient-centered, and effective healthcare [2]. However, validated models, methods and tools required to apply evidence-based medicine at the point of care, particularly those that accommodate insights from current practice, are lacking. An important source for medical evidence is clinical pathways, indicating the most widely applicable order of treatment interventions for particular patient groups [3]. Clinical pathways translate best available evidence into practice, are known to reduce in-hospital complications, and reduce length of stay and medical expenses [3]. In the US, more than 80% of the hospitals use clinical pathways for at least one intervention [4]. Clinical pathways are developed based on clinical practice guidelines (CPGs), which list recommendations for various treatments based on evidence from randomized clinical trials (RCTs). However, the strength of such evidence differs by clinical areas. For example, while recommendations in CPGs for conditions such as hypertension are frequently based on consistent and good quality patient-oriented evidence [5], those for chronic kidney disease (CKD) are known to be mostly consensus-based due to the difficulties in patient recruitment and high cost of RCTs [6]. Hence, CPGs often face concerns about oversimplification, inadequate applicability to the average patient, and resistance from practicing physicians [7,8].

* Corresponding author.
E-mail addresses: yiyez@andrew.cmu.edu (Y. Zhang), rpadman@andrew.cmu.edu (R. Padman), ndpatel219@gmail.com (N. Patel).

http://dx.doi.org/10.1016/j.jbi.2015.09.009
1532-0464/

As health information technology (IT), such as electronic health records (EHRs), gain widespread adoption and use in healthcare industry, thereby accumulating vast amounts of real-time patient care data, there is tremendous opportunity to develop data-driven models, methods and tools to facilitate review of practice workflows and improve evidence-based care delivery by learning practice-based pathways of care [7]. In this study, we aim to leverage this opportunity using a 'paving the COWpaths' approach [9], where 'COWpaths' refer to the most common clinical 'paths' or treatment patterns followed by patients, and oftentimes recorded using 'computer on wheels' (COW), and identified via retrospective analysis of patients' EHR data. They reflect the multitude of day-to-day clinical decisions made by many clinicians across many patients over a significant duration, and is an important and effective first step in managing change resulting from technological innovation. Upon thorough evaluation and validation by medical experts, we anticipate that these 'COWpaths' may be utilized as 'practice-based clinical pathways', to overcome the challenges faced by current CPGs, and also serve as individualized treatment guidelines.

In this paper, we aim to learn practice-based clinical pathways for chronic kidney disease (CKD), a chronic condition where patients gradually lose their kidney's functions, progressing from stage 1 to stage 5, and end stage renal disease (ESRD) [10]. It is a costly, complex and high mortality health condition affecting 26 million US adults, with another 73 million at increased risk for the disease [11]. In 2011, CKD patients above 65 years old made up 9.2% of US Medicare population (2.3 million), but incurred 18.2% of Medicare costs ($45.5 billion) [12]. Typically, the management of CKD is focused on delaying the progression of the condition, such as maintaining patients in their current disease stage and delaying the progression from stage 5 to dialysis [13]. Two widely known CPGs for CKD are Kidney Disease Outcomes Quality Initiative (KDOQI) and Kidney Disease: Improving Global Outcomes (KDIGO), which summarize global CPGs for various complications of kidney disease [14,15]. Yet, many recommendations are consensus-based, or based on evidence from small-scale randomized clinical trials (RCTs) [14,15]. Recent studies suggest that care delivery changes in CKD management may improve clinical outcomes, enhance quality of patient experience, and reduce annual total per capita health spending [16]. This is of critical importance in the special case of CKD patients who have developed acute kidney injury (AKI) whose consequences are serious and even fatal at times. Yet, few standard preventive and therapeutic options exist for AKI, and there is a growing movement among nephrologists to develop a CPG for AKI [6]. In this study, we specifically focus on AKI as a likely condition that can benefit from insights provided by practice-based clinical pathways that may be learned from EHR data.

We propose a practice-based clinical pathway development process, shown in Fig. 1, that integrates health IT and domain knowledge, including representation of multidimensional and longitudinal EHR data, identification of distinct patient subgroups, and extraction of common treatment patterns as candidate clinical pathways that constitute the focus of this paper. Subsequently, medical experts need to evaluate candidate clinical pathways and their outcomes, and make modifications and redesign, when necessary, to complete the process. Furthermore, we provide visualizations of the learned pathways that can be used by healthcare providers for practice review and decision support, and by patients to engage in shared-decision making, communication, and education. This paper is organized as follows. In Section 2 we describe the data representation, clinical pathway extraction methods, as well as our study data. Analytical results and evaluation are presented in Section 3. We discuss limitations and future work in Section 4 and Section 5 concludes the paper.

2. Materials and methods

2.1. Related work

Previous studies have identified clinical pathways and care patterns from EHR using statistical models, process mining, and machine learning [17–21]. Among studies that used statistical models, Lin et al. and Poelmans et al. modeled clinical pathways as a hidden Markov model (HMM), a stochastic model for randomly changing processes, that includes sequences of hidden states and observations [22,23]. Hidden states in the HMM are assumed to be states in the clinical pathway, and observations are actual interventions that occurred in practice. Generally, a large amount of data is necessary to train a HMM. On the other hand, process mining approaches take process logs as inputs, and in the case of clinical pathway mining, clinical activities recorded in EHR are used as process logs [20]. Many of the process mining approaches, such as Heuristic Miner and Fuzzy Miner, assume that event logs contain sufficient information and minor noise [24,25]. This assumption leads to spaghetti-like workflow models that are hard to interpret, as diversity is innate to most of the health data [19]. Recently, while still a process mining approach, Huang et al. presented sequence mining algorithms for clinical pathway patterns that did not look for an entire clinical pathway from start to end, but rather patterns of clinical pathways including time differences between events [21]. Also, machine learning techniques provide a potential solution to spaghetti-like workflow models by segmenting patients into relatively homogeneous groups before learning clinical pathways [18,26,27]. For example, Greco et al. used hierarchical clustering to cluster patients' sequences of visits [27], and Lakshmanan et al. segmented patients by their outcomes, followed by further clustering using DBScan and frequent pattern mining using SPAM [18].

Models from these studies have primarily been applied to clinical processes where each point in the pathway is a clinical event that bears, or is assumed to bear, clear temporal relationship with another [19,21]. However, in the outpatient setting especially, time stamps associated with clinical activities are typically *per day* [18]. Such lack of accurate temporal ordering presents a mining challenge for previous methods to accurately learn the co-progression of interventions and outcomes. For example, CKD patients visit clinics and hospitals every few months depending on the severity of their conditions. During each visit, they

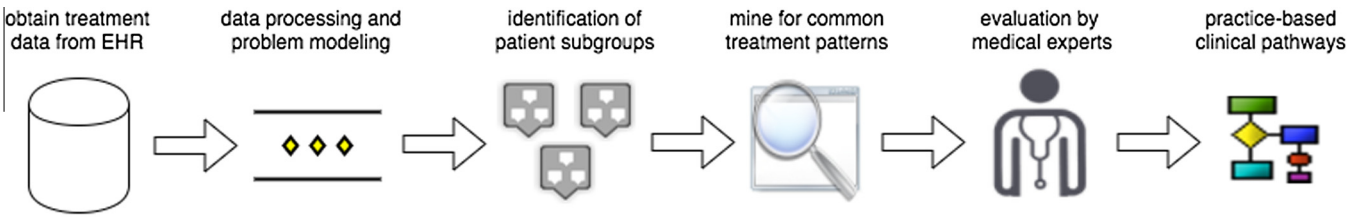

Fig. 1. Practice-based clinical pathway development process.

participate in various clinical activities such as receiving procedures and medication prescriptions. Table 1 is a sample summary of EHR data associated with a patient's visits. As the table shows, no clear temporal relationship exists between each clinical activity in the EHR data within one visit, because their time stamps are all on the same *day*. Furthermore, patients with chronic conditions commonly suffer from multiple comorbidities [12], resulting in multiple diagnoses noted and medications prescribed during a visit. To solve this challenge, we propose a novel data representation method to enable frequent pattern mining on such multidimensional and longitudinal treatment data. Furthermore, we expect our clinical pathway learning algorithm to be consistent with clinicians' decision process, and allow it to adapt to different levels of data complexity, to enable the efficient learning of interpretable clinical pathways.

2.2. Data transformation

In this paper, we consider each visit as a point in the clinical pathway, and track the co-progression of interventions and outcomes over time. Below we illustrate the data representation steps to establish correct temporal relationships among multiple clinical events of different types.

Def 1. e^l_k, k = 1, …, K_l, l = 1,…, L, represents a set of K_l events of L specific types, referred to as 'node' hereafter, that occurs during a patient's medical visit.

In Table 1, 'Office' event is of encounter type, 'CKD Stage 3' event of diagnosis type, 'Diuretics' event of medication type, and 'Renal ultrasound' of procedure type.

For example, in Table 1 patient 1's record can be transformed using node as below in Table 2. Procedure, medication, and diagnosis nodes are named as *Pj*, *Mj*, and *Dj*, respectively, $j \in Z^+$. The numbering of the nodes only serves to distinguish nodes, and does not represent any temporal ordering. Medications from 7/1/12 and 12/1/12 are represented as '*M2*' because the items contained in these 3 nodes are identical: {ACE inhibitors, diuretics}.

Our goal is to identify common sequences from the data that can constitute as clinical pathways. To do so efficiently, we introduce a modeling element called super node to represent unique visit content. Each super node captures a unique combination of visit purpose, procedure, medication and diagnosis. Super nodes collapse multi-dimensional records of visit such that they can be represented as a sequence of visits (super nodes) ordered by visit dates. Each patient has one and only one sequence, starting with the first visit recorded in the EHR and ending with the last visit.

Def 2 (*Super node (SN)*). Event sequence V is a set of supernodes {V_1, V_2, …, V_K}, where each node V_i is a nonempty subset of e^l_k, where K = all possible combinations of K_1, K_2, K_3 and K_4. Each visit of each patient can be represented as an element of the event sequence. Visits are ordered chronologically such that n^{th} visit occurs before $(n+1)^{st}$ visit.

Table 2
Transformation using nodes.

Patient	Date	Purpose	Procedure	Medication	Diagnosis
1	3/1/12	Education	N/A	M1	D1
1	7/1/12	Office	P1	M2	D2
1	12/1/12	Office	N/A	M2	D1
1	2/1/13	Hospital	N/A	M3	D3

Table 3
EHR extract illustrating modeling elements, post-data transformation.

Patient	Visit date	Super node
1	3/1/12	V3
1	7/1/12	V1
1	12/1/12	V5
1	2/1/13	V2

For example, patient 1's record can be transformed using super nodes as shown in Table 3. Each super node is labeled as Vj, $j \in \boldsymbol{Z}^+$. Similar to nodes, the numbering of each super node only distinguishes one super node from another and does not represent its chronological occurrence in the sequence.

Def 3 (*Visit sequence*). The n^{th} visit of patient p, p = 1, …P, with N_p visits, can be represented as

$$\text{Visit}^p_n\{V_i\}, i = 1, \ldots K,\ n = 1, \ldots N_p$$

For example, in Table 1, Visit^1_1 = {Education, CKD Stage 4, hypertension, ACE inhibitors} $\in$ {V_i}, i = 1, …, K, and Visit^1_2 = {Office, renal ultrasound, AKI, CKD Stage 4, Hypertension, ACE inhibitors} $\in$ {V_i}, i = 1, …, K. Hence, patient p's visit sequence can be represented as:

Q_p = {Visit^p_n}, n = 1, …, N_p, where $\text{Visit}^p_n \in \{V_i\}$, i = 1, …, K, and $|Q_p| = N_p$. For example, Q_1 = {V3, V1, V5, V2}, where patient 1 has 4 visits.

2.3. Identification of patient subgroups

We expect considerable diversity across patients undergoing various interventions. Hence, prior to clinical pathway extraction, we cluster patients' sequences of visits into subgroups such that we can extract clinical pathways for distinct types of patients. Since each patient has one and only one sequence, this is essentially the same as clustering of patients. To measure similarity among sequences, we use longest common subsequence (LCS), the maximum number of items that 2 sequences have in common, while preserving the order of occurrence but possibly separated [28].

$$\text{LCS}(x, y) = \max\{|u| : u \in S(x, y)\}$$

where $|u|$ is the length of the common subsequence for the pair of sequences (x, y), and $S(x, y)$ is the nonempty set of common subsequences of sequences x and y. A distance measure based on LCS is defined as:

Table 1
EHR summary, pre-data transformation.

Patient	Visit date	Description of visit			
		Visit purpose	Procedure	Medication	Diagnosis
1	3/1/12	Education	N/A	ACE[a] inhibitors	CKD stage 4, hypertension
1	7/1/12	Office	Renal ultrasound	ACE inhibitors, diuretics	AKI, CKD stage 4, hypertension
1	12/1/12	Office	N/A	ACE inhibitors, diuretics	CKD stage 4, hypertension
1	2/1/13	Hospital	N/A	ACE inhibitors, diuretics, statins	AKI, CKD stage 5, hypertension

[a] Angiotensin converting enzyme.

$$\mathrm{dLCS}(x,y) = |x| + |y| - 2\mathrm{LCS}(x,y)$$

where $|x|$ is the length of sequence x and $|y|$ is the length of sequence y. Comparison of each pair of visit sequences generates a dLCS based distance matrix for the entire sample of patients. For example, 100 visit sequences will yield a 100 by 100 distance matrix of dLCS. Since patients' clinical pathways follow clinical decisions made at each visit that is dependent on earlier decisions, we expect recurrent splittings of patient population [29], resulting in a hierarchical data structure. Hence, we use hierarchical clustering, a well-known cluster analysis technique that has been applied to biomedical data analysis extensively [30–32], with the dLCS matrix, to cluster patients into subgroups.

2.4. Clinical pathway extraction

When making treatment plans during a visit, clinicians from our study site report that they commonly review 2 previous visit records. Our clinical pathway extraction algorithm mimics this decision making process for CKD management [26,33], and models patients' visit sequences as discrete, time homogeneous Markov chains. Markov chains exhibit memoryless property; for example, in a first-order Markov chain the current state only depends on the previous state:

$$P(X_t = Y_j | X_{t-1} = Y_i, X_{t-2} = Y_k, \ldots, X_1 = Y_l) = P(X_t = Y_j | X_{t-1} = Y_i)$$

The assumption of memoryless property in representing sequential clinical workflow has been used in Meier et al. [34], Vankipuram et al. [35], Bouarfa et al. [36], and Li et al. [37]. In addition, we assume time-homogeneity, such that $(X_{s+t} = Y_j | X_s = Y_i)$ is independent of s. Time-homogeneity gives rise to a state transition probability distribution, which is denoted as $A = \{\alpha_{ij}\}$ where $\alpha_{ij} = P(X_{t+1} = Y_j | X_t = Y_i\}, 1 \le i,\ j \le N, \alpha_{ij} \ge 0$, and $\sum_{j=1}^{N} \alpha_{ij} = 1$. Through the Markov model process, we want to represent a clinical practice that each treatment decision from a patient's visit is dependent on information from 2 previous visits. Below we show the modeling process and define super pairs *SP* as the state elements of the Markov chain.

Def 4 (*Super pair (SP)*). A superpair $\mathrm{VV_m} = \{\mathrm{V_i}, \mathrm{V_j}\}$, i ≠ j, i, j = 1, . . ., K consolidates pairs of supernodes into a single superpair node.

In the above example, patient 1's visit sequence will transform from {V3, V1, V5, V2} to form a Markov chain of {VV1, VV4, VV3}, where VV1 represents transition from V3 to V1, VV4 represents transition from V1 to V5, and VV3 represents transition from V5 to V2. In this Markov chain, the occurrence of VV4 depends only on VV1, and VV3 depends only on VV4. In terms of actual visits, this means that V5 depends on V3 and V1, and V2 depends on V1 and V5, as in the clinicians' practices. Fig. 2 summarizes the transformation of EHR data elements to generate Markov chains.

The transition matrix of Markov chain includes all transitions from all patients' visit sequences of super pairs within each subgroup. Clinical pathways are built by connecting transitions of super pairs in the Markov chain that are above a desirable threshold in probability, *minPr*, and frequency, *minC*, such that we capture significant transitions in the patient population. A pseudocode of the algorithm is shown in Fig. 3.

To illustrate the algorithm, Fig. 4 shows a 5 by 5 transition matrix from a Markov chain. We can observe the transition probability, P_{ij}, and transition frequency, C_{ij}, of each transition, where i, j = 1–5. If we set the threshold for transition probability to be 0.3 and frequency to 3, then we have 2 pathways: VV1–VV2–VV4–VV3–VV5–VV2 (in green arrows), and VV1–VV2–VV3–VV5–VV2 (in blue arrows). Alternatively, if the threshold is raised to 0.5 for probability and 4 for frequency, the extracted pathway is narrowed down to VV1–VV2–VV4–VV3–VV5 (in orange arrows). In our analysis, the thresholds are determined to facilitate clinical interpretation.

2.5. Study data

EHR extract of 1,576 CKD patients from 1/1/2009 to 6/30/2013 was obtained from a community nephrology practice in Western Pennsylvania. All patients had at least one instance of AKI (ICD9 code = '584.xx' or doubling of creatinine [38]), and at least 5 visits. There are 726 female patients and 850 male patients in the dataset. Caucasian, African american, and other races make up 90%, 7%, and 3% of the study patients, respectively. Patients are categorized by age; 70%, 25%, and 5% of the patients are 70 years and older, between 50 and 70 years, and below 50 years old, respectively. A total of 407 patients are reported to have deceased during the study period, but this may be an underestimate due to loss to follow up. Table 3 lists the clinical components studied in the paper. Visit purpose is categorized into office, hospital and CKD education visits. We include in this study only the major conditions suffered by CKD patients: CKD stage 1 to stage 5, AKI, hypertension, diabetes, end stage renal disease (ESRD); and top 4 common drug classes: Angiotensin converting enzyme (ACE) inhibitors, Angiotensin receptor blockers (ARB), diuretics, and statins. According to KDOQI guideline on blood pressure [14], there are few RCTs that have compared the use of ACE against ARB, as well as ACE/ARB combined with diuretics, so the potential to generate new insights regarding the use of these medications motivated their inclusion in our analysis. All procedures specified using Current Procedural Terminology (CPT) codes in the data are included in the pathway analysis [39]. Given the number of visit types, procedures, diagnoses and medications, the total possible number of super nodes is 3,276.

3. Evaluation

3.1. Data transformation

A total of 17,358 visits are included in the dataset. Table 4 lists the percentage of each visit purpose, diagnosis and drug class in the data, and in the super nodes representing unique combinations of visit contents. Our analysis found limited number of procedures in the data so we omit them in the table. A total of 804 super nodes were generated from the data. Fig. 5 shows the distribution of super nodes, whose shape exhibits a power law distribution [40]. Roughly 80% of the observations are accounted for by the top 20% of the super nodes, in accordance with the Pareto principle [41].

The common diagnoses and medications in this patient population across all visits are also indicated in the table. For example, 83.9% of patients are, or were at some point, in CKD stage 3, and CKD stage 3 was noted as a diagnosis in 45.4% of the visits. The table also reveals the diversity level that each pathway component, such as a diagnosis, contributes by displaying its percentage in the set of all unique super nodes. Relatively low percentage in the super node set suggests that the pathway component occurs with a relatively stable group of components such that their combination maps into the same super node, whereas relatively high percentage suggests that the component tends to occur with varying groups of other components, resulting in larger variability. For example, hypertension is present in only 10.4% of the unique set of super nodes, though clinicians marked hypertension as a diagnosis in 97.7% of the visits. On the other hand, ARB shows up in 36.9% of the super node sets, while it was part of the prescription in only 26.5% of the visits.

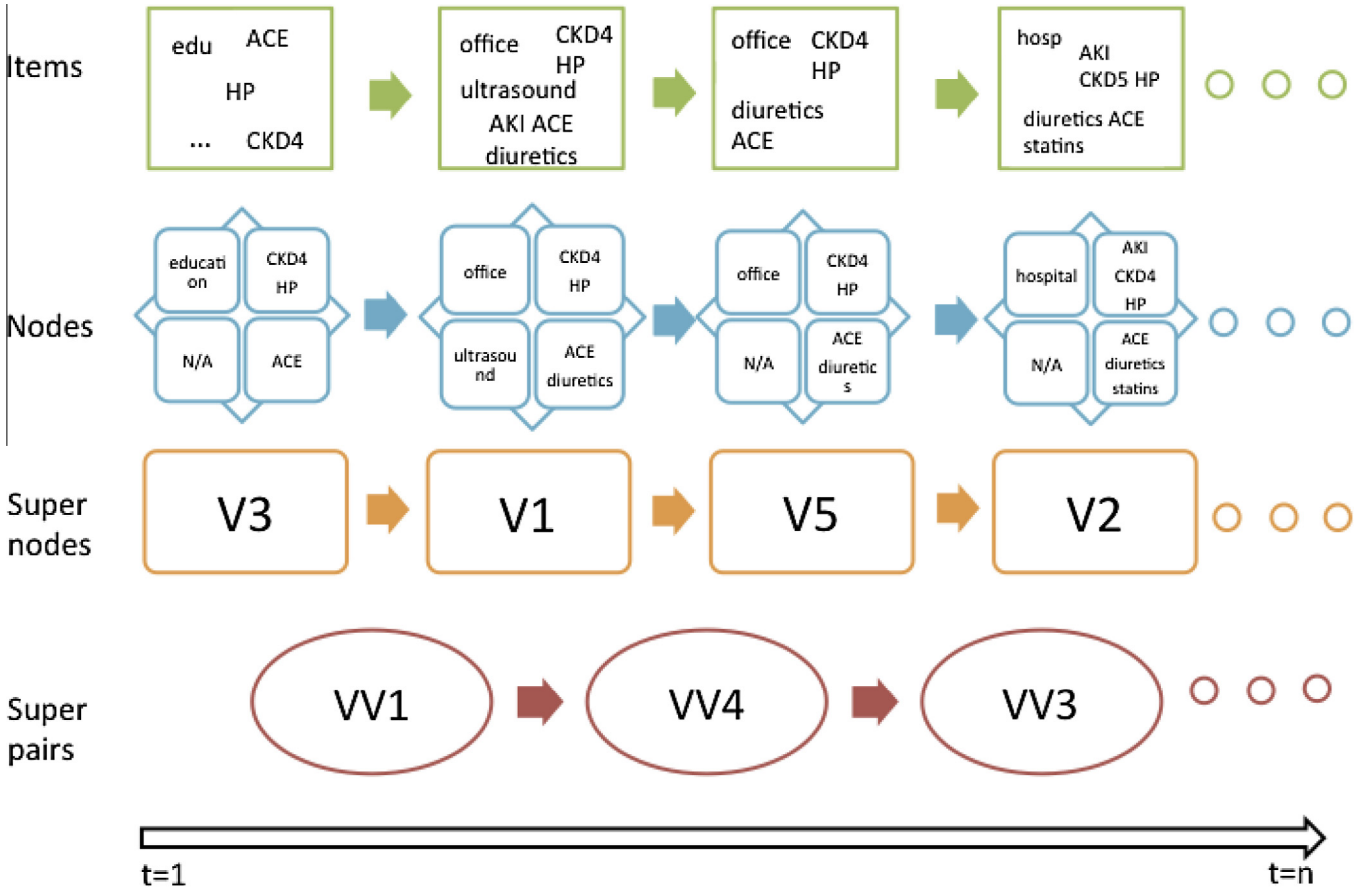

Fig. 2. Modeling visit history as Markov chain.

Algorithm 2 Clinical pathway extraction

Input: Q of SPs, $minPr$, $minC$
Output: Clinical pathways CP

for $p = 1$ to P **do**
  for $n = 1$ to $N - 1$ **do**
    Scan Q_p and save n^{th} and $n + 1^{th}$ SPs in $Transition$ as *parent* and *child*
  end for
end for
for $i = 1$ to I **do**
  for $j = 1$ to J **do**
    for i^{th} element in *parent*, and j^{th} element in *child*, calculate b_{ij} and c_{ij}
    if $b_{ij} \geq minPr$ and $c_{ij} \geq minC$ **then**
      pass
    else
      delete from $Transition$
    end if
  end for
end for
Scan $Transition$ and save into CP all possible pathways that can be connected

Fig. 3. Pseudocode of the clinical pathway extraction algorithm.

3.2. Subgroup clustering

Cluster analysis identified 31 patient subgroups from the data on 1,576 patients. Table 5 lists summary statistics and the most frequent visit content across subgroups. No procedures were among the most frequent visit contents, so we omit it from the table. We calculated the support for the most frequent visit content using Sequential Pattern Discovery using Equivalence classes (SPADE), a type of frequent sequence mining technique [42]. Support indicates the percentage of patients who experienced the specific visit at least once in their visit sequences, and therefore high support reflects similarity among patients' visit sequences. Broadly, subgroups separate by diagnoses, then further by the combination of medication prescription. Support values across

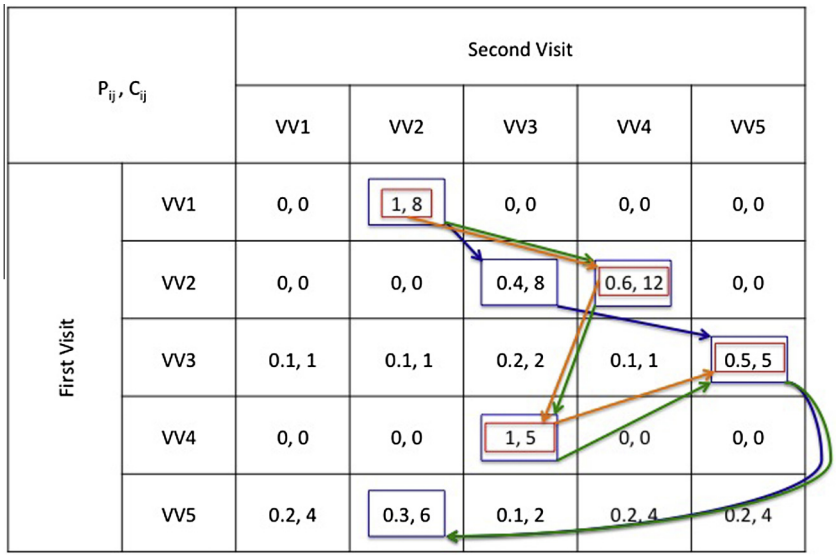

Fig. 4. Extraction of clinical pathways using Markov chain transition matrix.

Table 4
Summary statistics of visit purpose, diagnosis and drug class.

Description	% Patients affected (out of 1576)	% in visits (out of 17,358)	% in all unique SNs (out of 804)
Visit purpose			
Office	98.4	62.3	49.5
Hospital	82.3	35.3	42.3
Education	19.8	2.4	8.2
Diagnosis			
CKD stage 1	2.3	0.5	3.9
CKD stage 2	14.1	3.7	12.9
CKD stage 3	83.9	45.4	27.5
CKD stage 4	63.7	35.7	28.5
CKD stage 5	15.6	4.6	14.2
AKI	100.0	28.3	38.6
Hypertension	97.0	97.7	10.4
Diabetes	51.1	54.5	3.9
ESRD	16.9	8.5	12.9
Drug class			
ACE inhibitors	42.0	31.5	40.5
ARB	33.8	26.5	36.9
Diuretics	69.9	63.5	56.7
Statins	63.4	60.5	52.5

subgroups are fairly high, except for subgroup 31 with 78 patients, where no commonality was found in terms of visit content among patients.

3.3. Extracted clinical pathways

Clinical pathway extraction and visualization was performed for all subgroups to identify common pathways of disease progression and treatment. Fig. 6 displays the time to extract clinical pathways across subgroups in seconds, from the longest to shortest. While most pathways depict expected progression patterns, such as progression of CKD via development of AKI, others contained insightful information such as the potential association of education sessions with improvement, which needs further investigation using larger sample of patients. The state space of the Markov chain contains 3,505 super pairs. We illustrate the clinical pathway extraction and visualization step after subgroup clustering using two examples below.

Fig. 7 plots clinical pathways extracted from all 14 patients' sequences in subgroup 29, where most patients are in CKD stage 4 and hypertensive and take ACE inhibitors and statins. Each node represents a super pair, and the edges represent the transitions

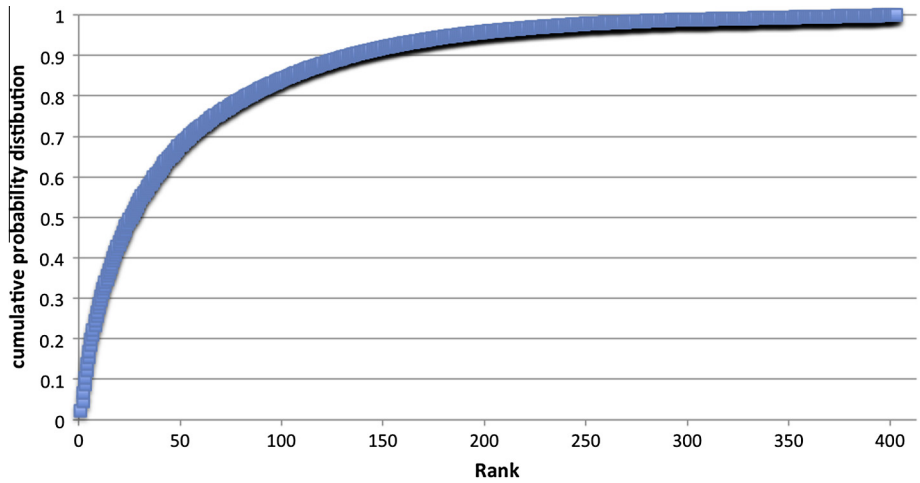

Fig. 5. Distributions of super nodes.

Table 5
Summary statistics across patient subgroups.

Sub group	# Patients	Visit content with the highest support			
		Purpose	Diagnoses	Drug Class	Support
1	80	Office	CKD stage 3, diabetes, hypertension	–	0.54
2	16			ACE	1
3	55			ACE, ARB, diuretics, statins	0.78
4	122			ACE, diuretics, statins	0.7
5	21			ACE, statins	1
6	10			ARB	1
7	36			ARB, diuretics	0.75
8	22			ARB, statins	0.95
9	74			Diuretics	0.69
10	83			Diuretics, statins	0.84
11	75			Statins	0.63
12	158		CKD stage 3, hypertension	–	0.52
13	29			ACE	0.72
14	66			ACE, diuretics, statins	0.77
15	14			ACE, ARB, diuretics	0.86
16	32			ACE, diuretics	0.69
17	26			ACE, statins	0.96
18	14			ARB	0.93
19	19			ARB, diuretics	0.95
20	20			ARB, statins	0.95
21	86			Diuretics	0.57
22	100			Diuretics, statins	0.59
23	68			Statins	0.71
24	90		CKD stage 3/4, diabetes, hypertension	ARB, diuretics, statins	0.67
25	38		CKD stage 3/4, hypertension	ARB, diuretics, statins	0.6
26	18		CKD stage 4, diabetes, hypertension	ACE, diuretics	0.67
27	14			ACE, statins	1
28	69			Diuretics, statins	0.94
29	14		CKD stage 4, hypertension	ACE, statins	1
30	29	Hospital	AKI, CKD stage 3	–	0.55
31	78	Deceased			0.15

among super pairs. The number of occurrences of a super pair is shown by its relative size, and the thickness of the edges represents frequency of connections. The threshold for *minPr* is 0.1 and for *minC* is 1, and the average length of the sub-pathway is 4.3. Fig. 8 expands the sub-pathway made up of red nodes, which is one of the most dominant patterns seen in subgroup 29. Out of the 14 patients in subgroup 29, 11 patients experience at least a part of this sub-pathway. This sub-pathway illustrates a process of patients repeating office visits and education sessions to prepare for progression to ESRD. Patients later experienced AKI-related hospitalizations, which led their conditions to worsen to ESRD, and subsequently passed away. These patients were on ACE inhibitors and statins throughout the pathway duration. The KDOQI guideline for CKD recommends ACE inhibitors for lowering blood pressure [14].

Fig. 9 displays clinical pathways from subgroup 4 of 122 CKD stage 3, diabetic and hypertensive patients taking ACE inhibitors, diuretics and statins. The threshold for *minPr* is 0.05 and for *minC* is 2, and the average path length is 4.4. Out of the 122 patients, 72 patients experienced at least a subset of the activities in these clinical pathways. A sub-pathway made up of red nodes is characterized by consecutive hospitalizations leading eventually to the demise of these patients. Fig. 10 displays the expansion of this sub-pathway, of which 36 patients experience at least a part. Another sub-pathway, in Fig. 11, shows that even though CKD patients rarely improve, patients showed improvement from CKD stage 4 to CKD stage 3, and receiving education sessions on CKD may have potentially contributed to such improvement. Fifty-four patients experience at least a part of this sub-pathway. In both sub-pathways, a diuretic was given to patients along with ACE inhibitors. The KDOQI guideline recommends combined use of both medications to increase patient adherence[14]. Further investigation of these patients' life style choices, health conditions, and medication usage may identify treatments that help patients achieve and maintain improvements in their health conditions.

3.4. Comparison against an existing method

We evaluate our algorithm against Heuristics Miner [25] from ProM [43] using subgroup 29. Clinical pathways generated by our

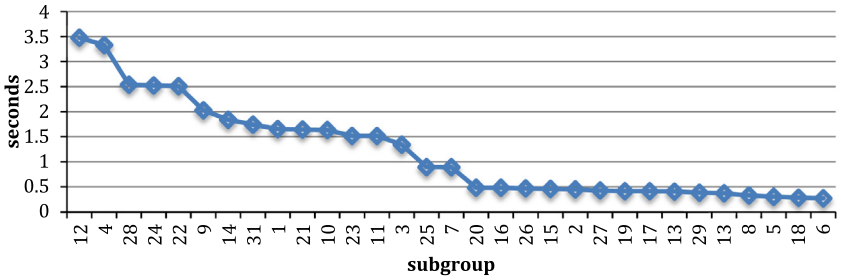

Fig. 6. Time to extract clinical pathways.

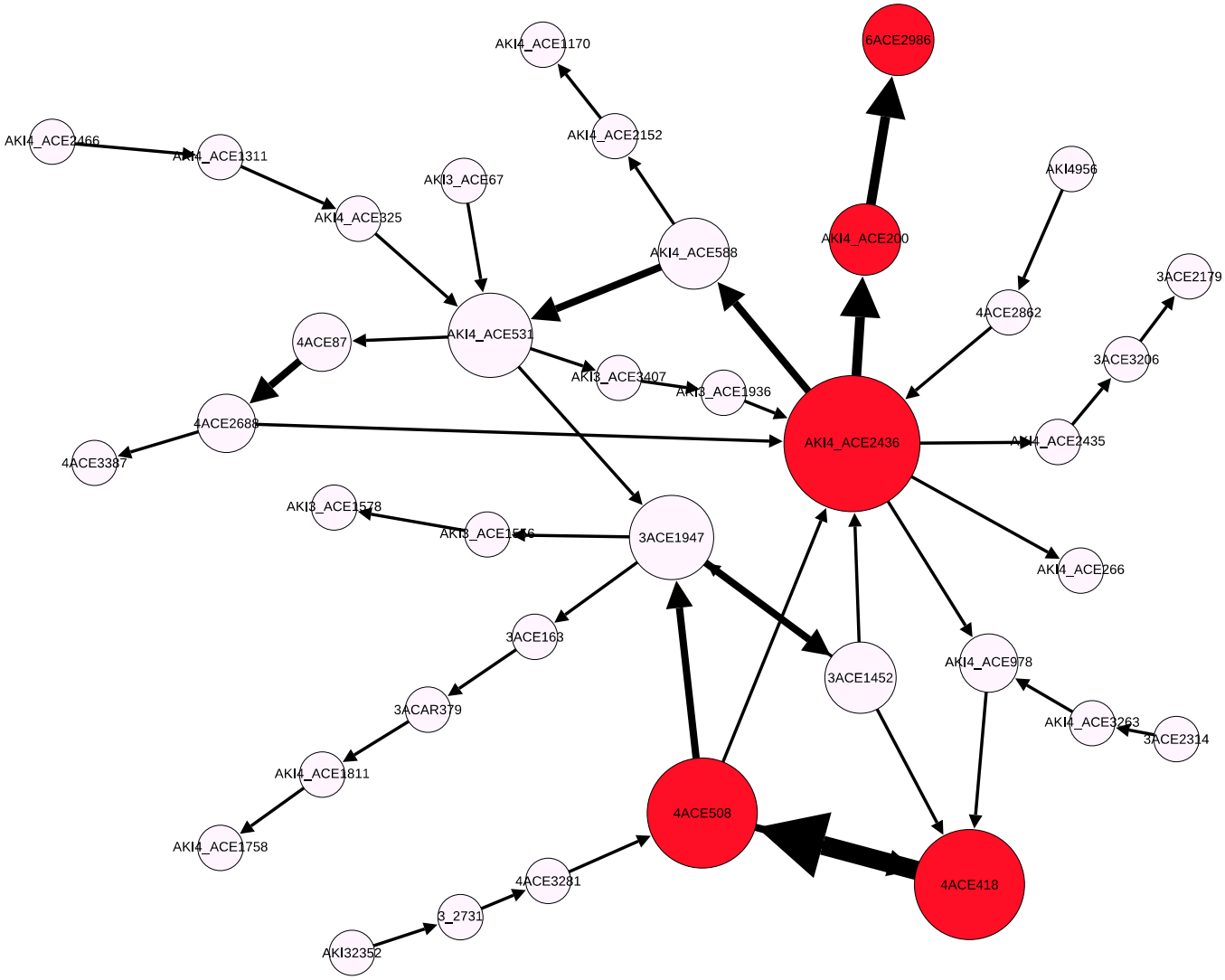

Fig. 7. Clinical pathway mined for subgroup 29.

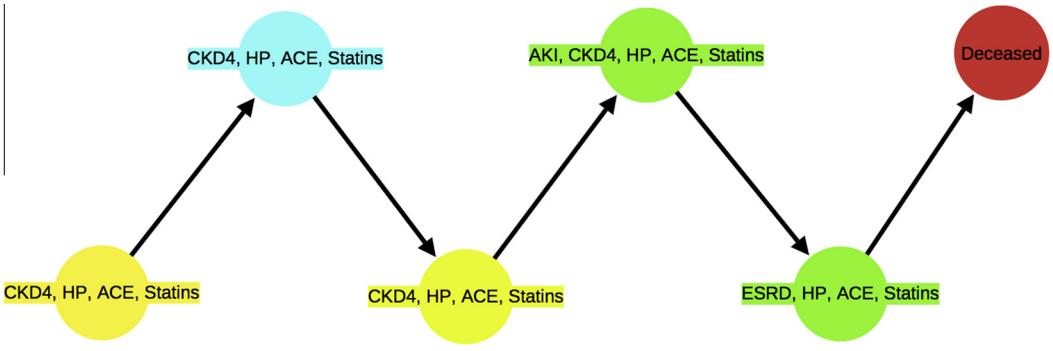

Fig. 8. Visualization of a clinical pathway for patients in subgroup 29. Yellow node: office visit, green node: hospitalization, blue: education visit, red: deceased, CKD4: CKD stage 4, HP: hypertension. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

algorithm and Heuristics Miner are shown in Figs. 7 and 12, respectively. The number in the boxes in Fig. 12 is the number of times that each item appeared in the EHR, and the numbers on the edges are the number of times the transitions took place [25,44]. The major advantage of our algorithm over Heuristics Miner is the ability to show the correct temporal associations and information encoding. For example, dependencies identified using Heuristics Miner suggest associations such as "prescription of statins precedes hospitalization" that are imprecise and distracting for decision makers. A more appropriate association should be that an office visit that includes diagnoses of CKD stage 3 and hypertension and prescription of statins is followed by a hospital visit, and so on,

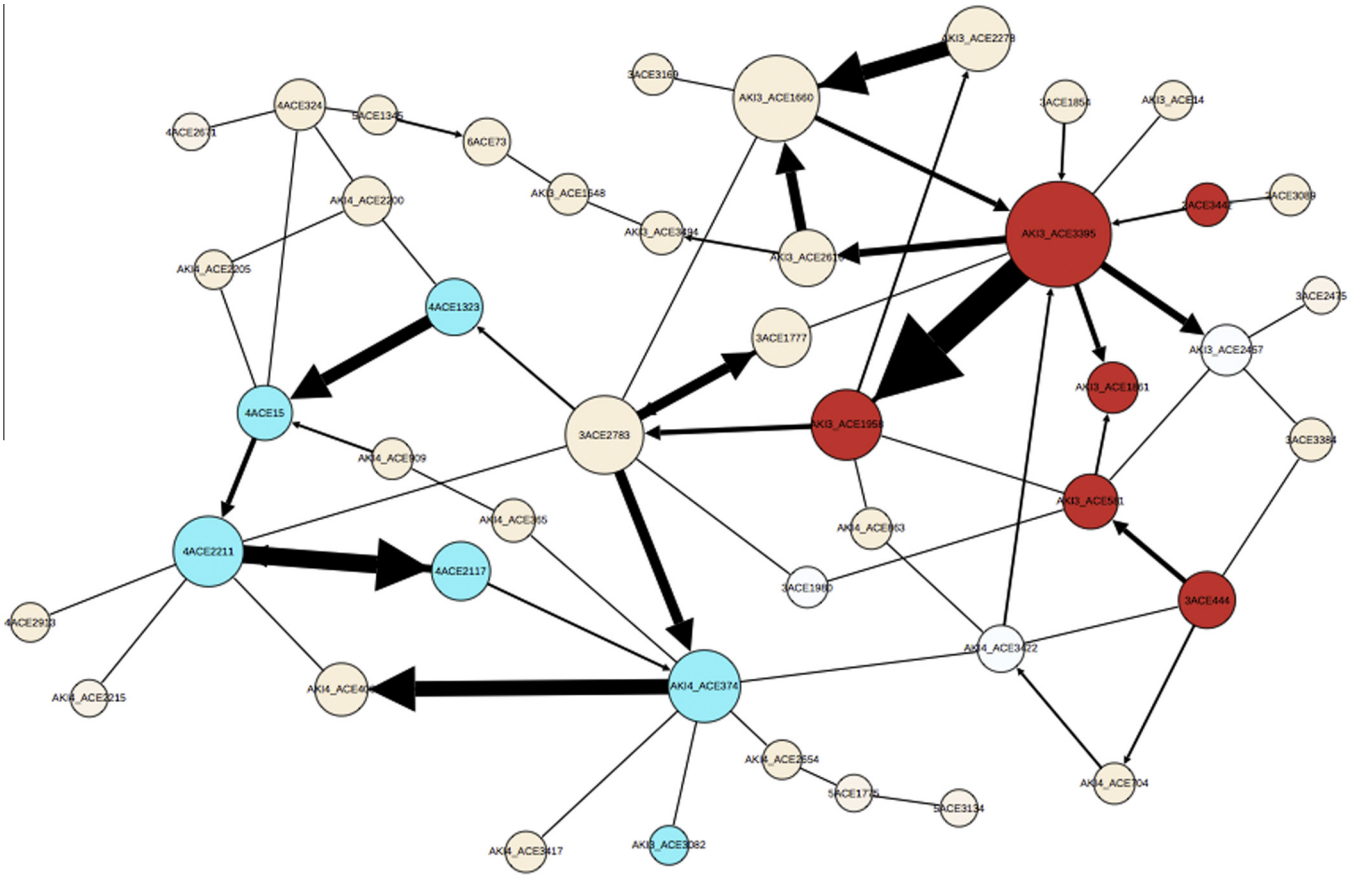

Fig. 9. Clinical pathway mined for subgroup 4.

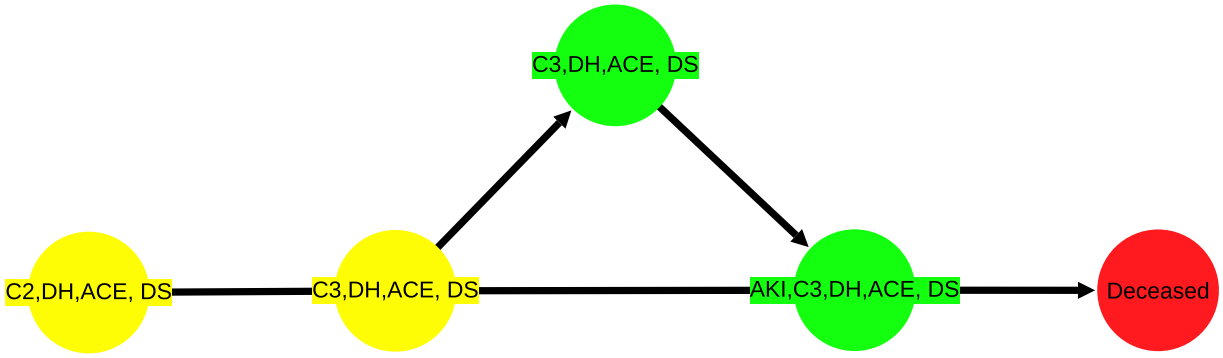

Fig. 10. Visualization of a sub-pathway for patients in subgroup 4. Yellow node: office visit, green node: hospitalization, red: deceased, C2/3: CKD stage 2/3, DH: diabetes and hypertension, DS: diuretics and statins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which is what our algorithm is capable of finding. Another advantage of our algorithm is its depiction of pathways with realistic complexity and associations. Fig. 12 presents an overly simplified process, which is rare in healthcare settings. On the other hand, our algorithm ensures that each visit in the clinical pathway is dependent on previous 2 visits by using super pairs and Markov chains, to reflect the actual decision making process in disease management. For example, we can elicit sub-pathways, such as Fig. 8, from the complete clinical pathway in Fig. 7, that present a comprehensive association between visits.

4. Discussion

4.1. Implications

The implications of learning clinical pathways from data can be two-fold. Healthcare providers can utilize them to compare local practice against consensus guidelines, and to identify common practice patterns, promising care delivery pathways, and unwanted care variations. The data representation method presented in this paper allows efficient comparison among patients'

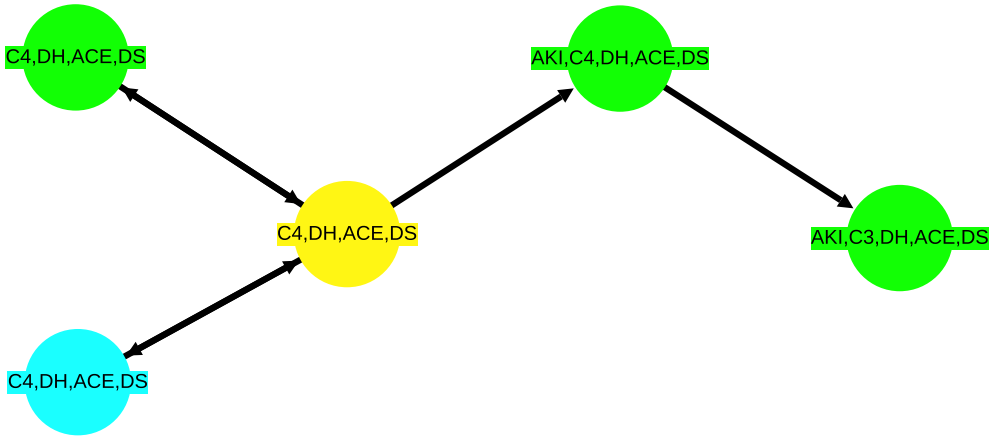

Fig. 11. Visualization of a sub-pathway for patients in subgroup 4. Yellow node: office visit, green node: hospitalization, red: deceased, C3/4: CKD stage 3/4, DH: diabetes and hypertension, DS: diuretics and statins. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

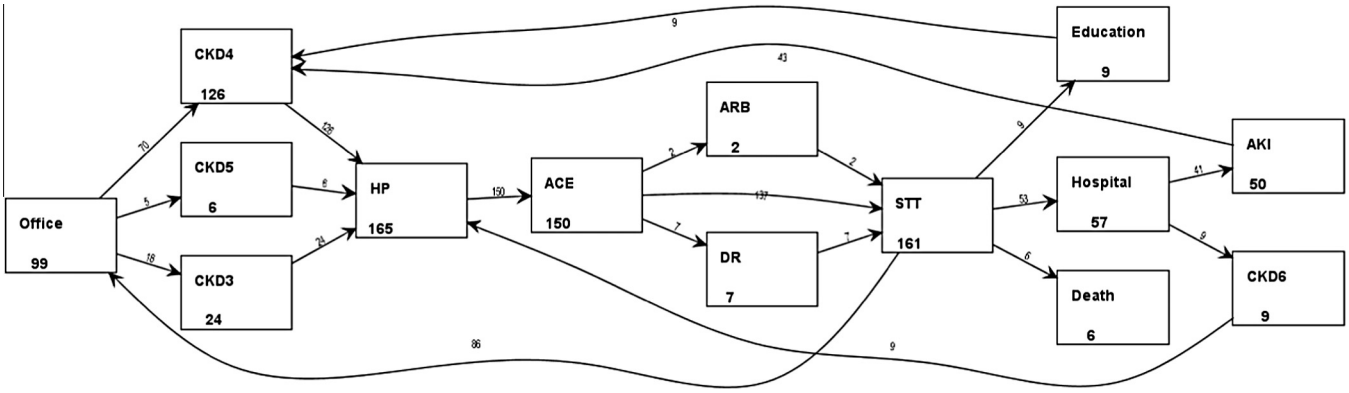

Fig. 12. Clinical pathway mined using Heuristic Miner in subgroup 29.

clinical histories and detection of variations in health conditions and medication usage that are otherwise hard to achieve. Examining variability in the whole patient sample or by diagnosis or drug class, and analyzing the resulting clinical pathways, may provide important insights to healthcare providers about treatment choices. In addition, learning clinical pathways across practices in defined geographic regions can also inform better population health management.

We have demonstrated that patients can be categorized into subgroups based on their conditions and medications. These are subject to change as patients continue their visits. Over time, patients and clinicians can learn about the changing subgroup assignment and clinical pathways for each patient, determine a patient's potential position on a given pathway, and facilitate discussions about personalized treatment decisions. According to a 2012 report by the Institute of Medicine, patients have considerable desire for shared decision making, but there is a significant gap between expected and actual engagement in the care they experience [45]. Visualizing practice-based clinical pathways and personalizing them for individual patients and their clinicians have the potential to support consistent and effective communication between patients and their healthcare providers, and improve satisfaction and health outcomes.

4.2. Limitations

Accuracy and completeness of the data is crucial for developing insightful practice-based clinical pathways. When data is captured in EHR to meet billing and reimbursement objectives, extensive time and effort are needed to clean the data and understand the underlying treatment processes so that clinical pathways mined from such data can produce meaningful information. For instance, chronic comorbidities such as hypertension are usually noted only during patients' initial visit or when there is a significant change in disease state, hence tracking and verifying this information at every visit is a challenge. Similarly, depending on the EHR, sometimes medication information is updated only when a change is made, and is not always recorded in between. Additionally, although the entire dataset dates back to 1994, the practice was using an older version of ICD9 codes for diagnosis until 2009 and did not have separate codes for CKD stages. Missing laboratory data compounded the problem and hindered manual calculation of the missing CKD stages. Finally, many of the patients in the dataset are active patients in the clinics; hence, the pathways presented here provide only a partial view of their entire clinical history.

The unique modeling approach for multidimensional data representation described in this paper supports efficient pattern search and provides flexibility in dealing with the diverse patient information. However, common patterns in patients' treatment data that constitute the clinical pathways are identified at the expense of missing rare events. In addition to learning clinical pathways, our methods can also be used to identify rare events of interest in the pattern of treatments, if appropriate patient filtering and selection of probability and frequency thresholds are applied in the algorithm.

4.3. Future work

In this study, the number of patients in each subgroup ranged from 10 to 158. High variability in disease progression, treatments and day-to-day management was a challenge in identifying interesting pathways supported by large numbers of patients. Larger patient samples in future studies may allow us to draw more conclusive comparisons between practice-based evidence and guideline recommendations. Evaluation of the practice-based clinical pathways developed in this paper is an important next stage. Analyzing pathways associated with specific outcomes, such as dialysis or mortality, and segmenting patients who have differing outcomes can provide useful insights on the effectiveness of treatments and patient management strategies. In terms of methodologies, the next challenge is the incorporation of additional factors such as laboratory values in the clinical pathway to predict patients' future visit content and develop probable clinical pathways for each individual. Availability of cost data may further facilitate the tracking of data-driven clinical and cost pathways simultaneously, utilizing the methodologies presented in this paper. These methods can be applied to not only CKD treatment data, but also to other datasets with similar data components and structures. Expanding the scalability of the method, such that it can handle more relevant components, causes and complications, is an ongoing area of research.

5. Conclusion

Developing practice-based clinical pathways has the potential to positively influence practice review, shared decision making between patients and healthcare providers, and cost reduction. In this study, we explored methods to efficiently summarize and learn clinical pathways from treatment data of 1,576 CKD patients in the outpatient care delivery environment. Selection of clinical pathway components was performed based on consultation with clinicians and existing CPGs with the goal of learning practice-based evidence from data that can help disambiguate guidelines. Patients' multidimensional and longitudinal clinical histories are represented as one-dimensional sequences of visits, capturing each visit's purpose, procedures, medications and diagnoses. We apply hierarchical clustering with LCS distance measure on patient sequences to separate patients into 31 distinct subgroups. Further modeling of visit sequences as Markov chains allows us to identify and visualize pathways that show typical disease progression, practices that may be consistent with guidelines, and sustainable improvements in patients' health conditions. Our data transformation and pathway extraction methods are efficient, consistent with actual clinical practice, and also generalizable to other clinical areas with similar data structures.

Competing Interests Statement

Authors have no competing interests to declare.

Contributorship Statement

Yiye Zhang and Dr. Rema Padman initiated the study and obtained data from Teredesai, McCann & Associates, P.C. Study design and implementation were conducted jointly by Yiye Zhang, Dr. Rema Padman, and Dr. Nirav Patel. Yiye Zhang performed the data analysis and drafted the paper, which was revised and confirmed carefully by Dr. Rema Padman and Dr. Nirav Patel. Dr. Rema Padman's revision of the paper mainly focused on the analytical components, whereas Dr. Nirav Patel reviewed the clinical aspects of the paper. Data interpretation was performed jointly by the three authors.

Acknowledgments

We are grateful to the physicians and staff of the community nephrology practice who generously provided data from their Electronic Health Record system for this study. We particularly thank the physicians, Dr. P. Teredesai and Dr. Q. Xie, and staff, L. Smith and A. Barletta, who gave us important clinical and technical information about the data and the key characteristics of CKD and its treatments. We also thank Dr. L. Wasserman at Carnegie Mellon University who provided valuable feedback while developing the methodology.

This study was designated as Exempt by the Institutional Review Board at Carnegie Mellon University.

References

[1] D.L. Sackett, Evidence-based Medicine: How to Practice and Teach EBM, vol. xiv, second ed., Churchill Livingstone, Edinburgh; New York, 2000. 261 p.
[2] Patient Protection and Affordable Care Act, Stat. 42, 2010.
[3] T. Rotter, L. Kinsman, E. James, A. Machotta, H. Gothe, J. Willis, et al., Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs, Cochrane Database Syst. Rev. (3) (2010) CD006632. PubMed PMID: 20238347.
[4] S. Saint, T.P. Hofer, J.S. Rose, S.R. Kaufman, L.F. McMahon Jr., Use of critical pathways to improve efficiency: a cautionary tale, Am. J. Manag. Care 9 (11) (2003) 758–765. PubMed PMID: 14626473.
[5] Physicians, AAoF, Rating the Strength of Evidence 2014 [cited 2014 09/01], Available from: <http://www.aafp.org/journals/afp/authors/ebm-toolkit/strength.html>.
[6] K. Uhlig, E.M. Balk, J. Lau, A.S. Levey, Clinical practice guidelines in nephrology – for worse or for better, Nephrol., Dialysis, Transplantat.: Off. Publ. Eur. Dialysis Transplant Assoc. – Eur. Renal Assoc. 21 (5) (2006) 1145–1153. PubMed PMID: 16520349.
[7] M.A. Fauman, How do physicians use practice guidelines?, Psychiat Times (2006).
[8] S. Saria, A $3 trillion challenge to computational scientists: transforming healthcare delivery, IEEE Intell. Syst. 29 (4) (2014) 82–87.
[9] Y. Zhang, R. Padman, J.E. Levin, Paving the COWpath: data-driven design of pediatric order sets, J. Am. Med. Inform. Assoc.: JAMIA 21 (e2) (2014) e304–e311. PubMed PMID: 24674844. Pubmed Central PMCID: 4173172.
[10] V. Jha, G. Garcia-Garcia, K. Iseki, Z. Li, S. Naicker, B. Plattner, et al., Chronic kidney disease: global dimension and perspectives, Lancet 382 (9888) (2013) 260–272. PubMed PMID: 23727169.
[11] About Chronic Kidney Disease: The National Kidney Foundation, Inc.; 2013 [cited 2014 8/22], Available from: <http://www.kidney.org/kidneydisease/aboutckd.cfm>.
[12] A.J. Collins, R.N. Foley, B. Chavers, D. Gilbertson, C. Herzog, K. Johansen, et al., 'United States renal data system 2011 annual data report: Atlas of chronic kidney disease & end-stage renal disease in the United States, Am. J. Kidney Dis.: Off. J. Natl. Kidney Found., 59 (Suppl 1) (2012) A7, e1-420. PubMed PMID: 22177944.
[13] Chronic Kidney Disease (CKD) and Diet: Assessment, Management, and Treatment, National Kidney Disease Education Program, 2014.
[14] Kdoqi, KDOQI clinical practice guidelines and clinical practice recommendations for diabetes and chronic kidney disease, Am. J. Kidney Dis.: Off. J. Natl. Kidney Found. 49 (Suppl. 2) (2007) S12–S154. PubMed PMID: 17276798.
[15] P.E. Stevens, A. Levin, Kidney disease: improving global outcomes chronic kidney disease guideline development work group M. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline, Ann. Intern. Med. 158 (11) (2013) 825–830. PubMed PMID: 23732715.
[16] G. Abra, M. Patel, D. Moore, Trend-bearing Chronic Kidney Disease Care Model, 2013.
[17] F. Caron, J. Vanthienen, K. Vanhaecht, E. Van Limbergen, J. De Weerdt, B. Baesens, A process mining based investigation of adverse events in care processes, Health Inform. Manage. J. (2013) 310–317.
[18] G.T. Lakshmanan, S. Rozsnyai, F. Wang, Investigating clinical care pathways correlated with outcomes, Lect. Notes Comput. Sci. 8094 (2013) 323–338.
[19] W. Van der Aalst, T. Weijters, L. Maruster, Workflow mining: discovering process models from event logs, IEEE Trans. Knowledge Data Eng. 16 (9) (2004) 1128–1142.
[20] W.M.P. van der Aalst SpringerLink (Online service), Process Mining Discovery, Conformance and Enhancement of Business Processes, Springer-Verlag, Berlin, Heidelberg, 2011, http://dx.doi.org/10.1007/978-3-642-19345-3.
[21] Z. Huang, X. Lu, H. Duan, On mining clinical pathway patterns from medical behaviors, Artif. Intell. Med. 56 (1) (2012) 35–50. PubMed PMID: 22809825.

[22] F.-R. Lin, L.-S. Hsieh, S.-M. Pan, Learning clinical pathway patterns by hidden Markov model, In: The 38th Annual Hawaii International Conference on System Sciences, IEEE, 2005.
[23] J. Poelmans, G. Dedene, G. Verheyden, H.V.D. Mussele, S. Viaene, E. Peters, Combining business process and data discovery techniques for analyzing and improving integrated care pathways, Lect. Notes Comput. Sci. 6171 (2010) 505–517.
[24] C.W. Günther, W.M.P. Van Der Aalst, Fuzzy mining – adaptive process simplification based on multi-perspective metrics, Lect. Notes Comput. Sci. 4714 (2007) 328–343.
[25] A.J.M.M. Weijters, W.M.P. van Der Aalst, A.K.A. De Medeiros, Process mining with the heuristics miner-algorithm, BETA Working Paper Series: Eindhoven University of Technology, 2006.
[26] Y. Zhang, R. Padman, L Wasserman, On learning and visualizing practice-based clinical pathways for chronic kidney disease. In: American Medical Informatics Association 2014 Annual Symposium, Washington, DC, 2014.
[27] Gianluigi Greco, Antonella Guzzo, L. Pontieri, Mining hierarchies of models: from abstract views to concrete specifications, Lect. Notes Comput. Sci. 3649 (2005) 32–47.
[28] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, Inc., Hoboken, NJ, 2008.
[29] S. Dlugosz, Clustering Life Trajectories: A New Divisive Hierarchical Clustering Algorithm for Discrete-valued Discrete Time Series, ZEW, Mannheim, 2011.
[30] H. Wao, J.W. Beckstead, J. Beal, M. Aluoch, T.C. Skipper, J.J. Orrick, Identifying subgroups of care providers participating in a Telehealth educational intervention: hierarchical cluster analysis of evaluation data, J. Int. Assoc. Providers AIDS Care June (2013). PubMed PMID: 23771871.
[31] T. Sakagami, T. Hasegawa, T. Koya, T. Furukawa, H. Kawakami, Y. Kimura, et al., Cluster analysis identifies characteristic phenotypes of asthma with accelerated lung function decline, J. Asthma: Off. J. Assoc. Care Asthma 51 (2) (2014) 113–118. PubMed PMID: 24102534.
[32] H.W. Bae, S. Rho, H.S. Lee, N. Lee, S. Hong, G.J. Seong, et al., Hierarchical cluster analysis of progression patterns in open-angle glaucoma patients with medical treatment, Investigat. Ophthalmol. Vis. Sci. 55 (5) (2014) 3231–3236. PubMed PMID: 24781944.
[33] Y. Zhang, R. Padman, L. Wasserman, N. Patel, P. Teredesai, Q. Xie, On clinical pathway discovery from electronic health record data, IEEE Intell. Syst. 30 (1) (2015) 70–75.
[34] J. Meier, A. Dietz, A. Boehm, T. Neumuth, Predicting treatment process steps from events, J. Biomed. Informat. 53 (Feb) (2015) 308–319. PubMed PMID: 25510607.
[35] M. Vankipuram, K. Kahol, T. Cohen, V.L. Patel, Toward automated workflow analysis and visualization in clinical environments, J. Biomed. Informat. 44 (3) (2011 Jun) 432–440. PubMed PMID: 20685315.
[36] L. Bouarfa, P.P. Jonker, J. Dankelman, Discovery of high-level tasks in the operating room, J. Biomed. Informat. 44 (3) (2011 Jun) 455–462. PubMed PMID: 20060495.
[37] Y. Li, S. Swift, A. Tucker, Modelling and analysing the dynamics of disease progression from cross-sectional studies, J. Biomed. Informat. 46 (2) (2013) 266–274. PubMed PMID: 23200810.
[38] W. Van Biesen, R. Vanholder, N. Lameire, Defining acute renal failure: RIFLE and beyond, Clin. J. Am. Soc. Nephrol.: CJASN 1 (6) (2006) 1314–1319. PubMed PMID: 17699363.
[39] Association AM, AMA – About CPT [Internet], Chicago: American Medical Association; [updated 2015; cited 2015 1/11], Available from: <http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.shtml>.
[40] A. Wright, D.W. Bates, Distribution of problems, medications and lab results in electronic health records: the Pareto principle at work, Appl. Clin. Informat. 1 (1) (2010) 32–37. PubMed PMID: 21991298. Pubmed Central PMCID: 3189502.
[41] M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351.
[42] M.J. Zaki, SPADE: an efficient algorithm for mining frequent sequences, Mach. Learn. 42 (1/2) (2001) 31–60.
[43] Verbeek HMWE, Bose RPJC, ProM 6 Tutorial, 2010.
[44] A.J.M.M.T. Weijters, J.T.S.J. Ribeiro, HeuristicsMiner 6.0: Users Guide.
[45] C. Alston, Institute of Medicine (U.S.), Institute of Medicine (U.S.), Roundtable on Value & Science-Driven Health Care, Best Practices Innovation Collaborative, Communicating with Patients on Health Care Evidence, 14 pages.