Data-Driven Inference of Clinical Pathway Components for Identifying Basic Care Patterns from Electronic Health Records

Minsu Kim[1], Byung H. Park[1], Ozgur Ozmen[1], Everett Rush[1], Junghoon Chae[1], Makoto M. Jones[2], Randall W. Rupper[2], Jeffrey C. Humpherys[2], Merry Ward[2], Jonathan Nebeker[2]

[1] Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831, US
[2] Veterans Health Administration, Veterans Affairs, 810 Vermont Ave NW, Washington, DC 20571, US
kimm@ornl.gov

Abstract. Developing a clinical pathway is a labor-intensive process that requires the participation of experts in various areas, including clinical and informatics domains. This process will be greatly facilitated if the conformity of clinical guidelines is precisely measured and significant variations in adopting them are captured from electronic health records (EHRs). From a data analytics perspective, this requires the establishment of mapping between clinical concepts, their representations in terms of EHR data elements, and their temporal formation as clinical activities. This paper introduces a data-driven informatics framework that maps clinical concepts of EHR elements to an embedding space based on their temporal co-occurrences and groups them into cohesive clusters of clinical concepts called clinical pathway components (CPCs). The paper illustrates how a set of CPCs is discovered by applying the framework to a stable ischemic heart disease cohort of the US Department of Veterans Affairs.

clinical pathway, representation learning, clustering

1 Introduction

A clinical pathway (CP) is a guided care map to promote qualitative care for a specific cohort [1] in which both short- and long-term interventions are clearly defined by clinical professionals over the course of treatment. Electronic health records (EHRs) are patient-centric and real-time records, including treatment history of patients, such as laboratory services, procedures, inpatient/outpatient medications, radiology, nuclear medicine services, and consultations [5]. With the increasing availability of EHRs and advances in data analytics, opportunities to build data-driven approaches to infer evidence-based CPs have also grown [2]. However, developing a CP is still a labor-intensive process that requires

the participation of experts in various areas, including clinical and informatics domains. Moreover, the ability to capture the dynamics and knowledge that can be translated to actionable suggestions in enhancing CPs is greatly desired. From a data analytics perspective, this requires the establishment of compact and reusable representations of medical concepts from EHR data items that are represented as a coherent course of treatment as clinical activities.

Formally, the authors define a group of EHR data items (i.e., record values) to be contextually related if they tend to co-occur temporally and consistently in patients' clinical records. The authors further assume that such contextually related groups are building blocks for constructing a CP, and thus they are called *CP components (CPCs)*, as illustrated in Fig. 1. First, EHR items of a patient arranged by time stamps of their occurrences are defined as a *trace*, which essentially represents a course of treatments performed to the patient. Second, the traces of a given cohort into space are projected where elements of high-temporal affinities are closely placed. Then, a trace trajectory (i.e., a patient) is represented as a sequence of element groups. If the elements of such a group are consistently found as a group in trajectories of other traces, the elements are considered to be contextually related, thus forming a CPC. Hence, constructing an embedding space in which each clinical item is located based on their contextual similarity is key for identifying CPCs and CPs from the data. This paper proposes an embedding approach to construct an affinity map and a multistage clustering method to infer CPCs as basic care patterns in individual patient traces and identify widely used care patterns across all patients. The proposed method was applied to a stable ischemic heart disease (SIHD) cohort of the US Department of Veterans Affairs (VA).

2 Methods

Extracting treatment patterns from the traces of a given cohort is a multistep process that involves: (1) constructing clinical item embedding space, (2) identifying CPCs by mapping each trace into the space, and (3) consolidating CPCs into treatment patterns.

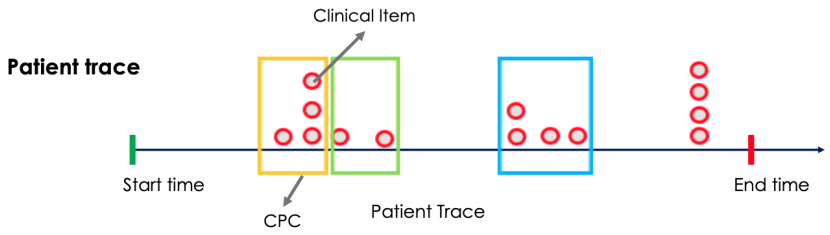


Fig. 1. Illustration of patient trace. Each patient trace is a temporal sequence of clinical items. By the proposed method, contextually related clinical items are identified and clustered as CPCs. Red circles indicate each clinical item, and colored rectangles indicate CPCs.

2.1 Definition of Terms

Key terms used in this paper are defined as follows:

- A *patient trace*, *trace*, is a collection of clinical items of each patient.
- A *clinical item*, or *item*, is a record of an individual clinical activity of a patient with a time stamp, such as a prescription order of a medication, order of a procedure, or order of a lab test.
- A *clinical item token*, or *item token* or token) is a unique identifier that refers to each clinical item without time stamp or patient specification.

2.2 Constructing Clinical Item Embedding Space

For the construction of the embedding space for clinical items, this work employed the global vector representation learning algorithm used by GloVe [7] to represent an item as a vector of pairwise temporal distances with all other items. Formally, $MTD_k(i,j)$ is defined as the minimum time distance between two clinical items i and j in patient trace k. This is formulated in Eq. (1a), where T_{ki} indicates the set of time points that clinical item i appears in patient trace k, and $|x-y|$ is the temporal distance between two time points x and y. Here, the unit of time is a day that can take a fractional form. Then, $AMD(i,j)$ denotes the average minimum temporal distance between i and j. Next, AMDs of all-pairwise items for trace constitute the global matrix GM, as shown in Eq. (1c).

$$MTD_k(i,j) = \begin{cases} 0, & \text{if } i = j \\ \min_{\forall x \in T_{ki}, \forall y \in T_{kj}} |x-y|, & \text{otherwise} \end{cases} \tag{1a}$$

$$AMD(i,j) = \frac{1}{\sum_{k=1}^{N} I_k} \sum_{l=1}^{N} MTD_l(i,j), \tag{1b}$$

$$GM[i,j] = AMD(i,j). \tag{1c}$$

2.3 Reduction of Global Matrix Dimension

There can be imbalances among EHR elements regarding their occurrences in the data. More specifically, some items are extremely rare and are found in as few as one patient trace. Because the global matrix is defined as an all-pairwise matrix, as in Eq. (1c), the imbalance in item occurrences and the sparsity caused by the rare tokens might cause poor embedding representation and inefficient usage of resources. To avoid this, the top items are selected, which account for 95% of the total occurrences in the data, and each row of the global matrix is constructed with respect to the top items only. For example, suppose that there are 1,000 unique items in a given EHR data and that the top 100 most frequently used unique items contain 95% of all item occurrences, then the dimension of the

constructed global matrix would be reduced from $1,000 \times 1000$ to $1,000 \times 100$, as shown in Eq. (2).

$$RGM[i,k] = AMD(i,k), \tag{2}$$

where i indicates each item, and k indicates each selected top item.

Then, a kernel principal component analysis (PCA) with the cosine kernel [9,12] is applied to the reduced global matrix (Eq. [2]) to generate a metric space in which contextually related items are located close. The distance between two items is measured as the cosine distance of the original vectors, as shown in Eq. (3).

$$D(i,j) \approx 1 - \frac{RGM_i \cdot RGM_j}{\|RGM_i\| \|RGM_j\|}, \tag{3}$$

where $D(i,j)$ is the cosine distance of two items i and j in the constructed metric space, which indicates the contextual distance between two items, and RGM_i and RGM_j indicate the row vector of RGM.

2.4 Identifying Clinical Pathway Component Candidates

To identify care patterns in each patient trace, clinical items of each trace were project into the constructed embedding space to capture contextually related items as clusters, which are the candidates for CPCs. The process is described as follows.

- First, clinical items in each patient trace are projected into the clinical item embedding space (Fig. 2).
- Next, the projected items are grouped into k clusters by using a hierarchical clustering algorithm with Ward's linkage [10], where k is an optimal number of clusters determined by the silhouette score. [8].

2.5 Identifying Clinical Pathway Components

Because CPC candidates are identified from individual traces, each candidate is essentially a variation of CPCs that is uniquely executed to a patient. Therefore, aggregating these CPC candidates across the entire traces is a way of producing CPCs as more cohesive and general treatment patterns. A heuristic method was devised to compare and merge similar CPC candidates into a single CPC (Fig. 3). The steps of aggregation are described as follows.

- First, a list of unique CPC candidates sorted by the number of occurrences is produced. Each CPC is defined as a set of items.
- Next, the top candidate from the list is selected as a CPC template and then merged with the remainder of the individual candidates only when the candidate is a super-set of the CPC template. If this occurs, the CPC is removed from the list.

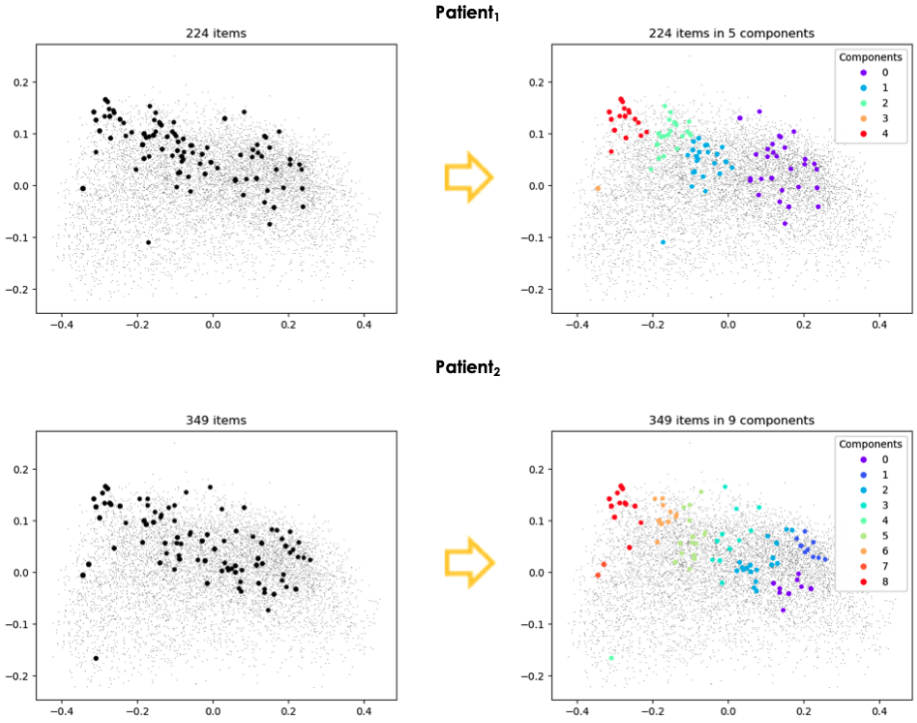


Fig. 2. An example of the identification process of CPC candidates in two patients. In both cases, the figures on the left indicate clinical items in each patient projected into the embedding space; the gray dots represent the whole embedding space, and the black dots represent clinical items of a patient. Figures on the right indicate the k clustering results, where each color represents identified CPC candidates.

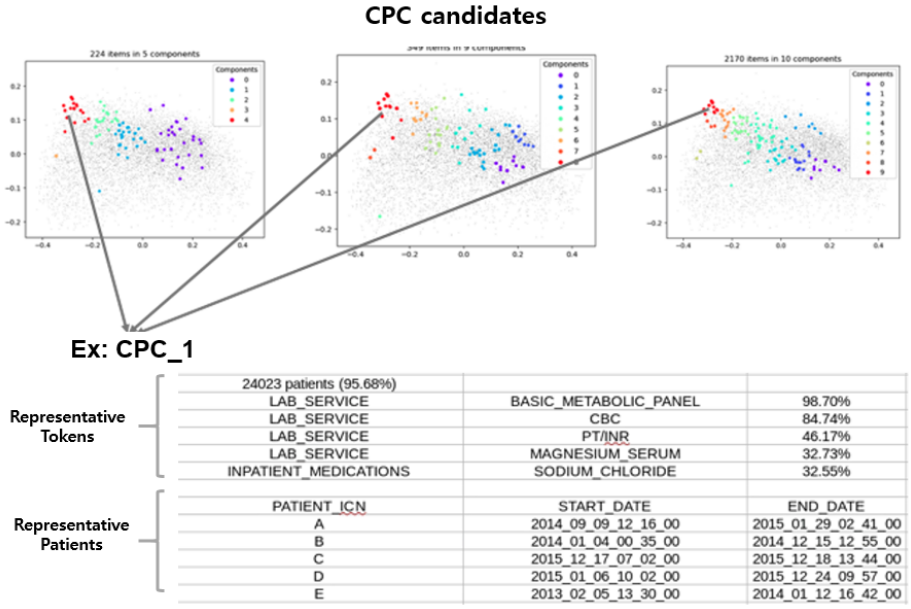
- Once all the candidates are processed, the next top candidate is selected as the next CPC template, and the process repeats until the candidate list is empty.

3 Results

This section presents the results of the proposed method applied to a cohort that consists of 25,345 SIHD patients. This cohort will be referred to as the *SIHD cohort*. The following sections describe the dataset and present the treatment patterns extracted from the data.

3.1 Data Description

A cohort of 25,345 patients was constructed by applying three rules in sequence. The first rule identifies the patients in the VA Corporate Data Warehouse diagnosed with SIHD (I25 codes as ICD-10s and 414 codes as ICD-9s [6]). The cohort



24023 patients (95.68%)		
LAB_SERVICE	BASIC_METABOLIC_PANEL	98.70%
LAB_SERVICE	CBC	84.74%
LAB_SERVICE	PT/INR	46.17%
LAB_SERVICE	MAGNESIUM_SERUM	32.73%
INPATIENT_MEDICATIONS	SODIUM_CHLORIDE	32.55%

PATIENT_ICN	START_DATE	END_DATE
A	2014_09_09_12_16_00	2015_01_29_02_41_00
B	2014_01_04_00_35_00	2014_12_15_12_55_00
C	2015_12_17_07_02_00	2015_12_18_13_44_00
D	2015_01_06_10_02_00	2015_12_24_09_57_00
E	2013_02_05_13_30_00	2014_01_12_16_42_00

Fig. 3. Example aggregation of CPC candidates into CPCs. By the proposed heuristic, each CPC candidate identified in each patient is aggregated into CPCs to capture universal care patterns among patients.

entry dates (i.e., first diagnosis) for patients are also critical, so the next rule identifies the patients who received cardiovascular stress test based on outpatient current procedural terminology and lists them with their dates. The stress test is widely used to diagnose SIHD. Assuming that the stress test would have triggered the first diagnosis, the next rule temporally aligns the first diagnosis and the first stress test, and if they are 1 week apart, then the diagnosis is assumed to be the first diagnosis. The temporal span of a patient trace was set to 36 months following the first diagnosis. Thus, the 25,345 patient traces contain 4,883,312 clinical items (i.e., tokens). These tokens were then mapped into the Unified Medical Language System (UMLS) [3] concept codes. Because some tokens could not be mapped to UMLS concepts, the final SIHD cohort data include 25,345 patient traces over 2,929,658 items.

3.2 Aggregation of Laboratory Panel Tests

Clinicians occasionally order a batch of lab tests called *laboratory panel tests*, or *lab panels*. The authors found that 651,631 of the 2,929,658 clinical items in the SIHD cohort were the lab panels, which took up more than 22% of all items. Among those 651,631 items, there are 3,595 unique lab panels, many of which are almost identical to one another. Some are only different by a few individual lab tests. For example, at least 32 lab panels are very similar to one another and

thus can all be annotated as a *complete blood count panel*. Hence, a heuristic method was devised that clusters panels based on their compositions, and 3,595 lab panels were reduced into 689 panel clusters. Lab panels were replaced with panel clusters. As a result, the final data include 2,314,855 clinical items of 25,345 patients with 7,876 unique tokens.

3.3 Construction of Clinical Item Embedding Space

A clinical item embedding space was constructed as described in Sections 2.2 and 2.3 to the SIHD cohort data. A clinical item embedding space was constructed by using SIHD cohort data. The authors found that 907 tokens account for more than 95% of all clinical items in the data. Therefore, with 7,876 unique tokens, the constructed RGM (Eq. [2]) has a dimension of $7,876 \times 907$. Then, a PCA with a cosine kernel was applied to the matrix, where the number of components used was 2, which explained more than 99% of variances [9]. Hence, the final embedding space has a dimension of $7,876 \times 2$.

3.4 Identification and Evaluation of Clinical Pathway Components

As described in Section2.4, clinical items were projected in each of the 25,345 patients into the constructed embedding space, and CPC candidates were identified (Fig. 2). As a result, 87,312 candidates were identified (i.e., an average of 3.44 candidates per patients). Next, as described in Section 2.5, 87,312 candidates were aggregated into 1,388 unique CPCs (Fig. 3). The authors consulted with physicians for an evaluation of the identified 1,388 CPCs. To facilitate their review process, for each CPC, a set of representative tokens and a set of representative patient traces were provided.

- Representative tokens of each CPC are the 10 tokens with the highest F1 scores. The list of the top 10 tokens for each CPC is listed in the Supplementary Table S1 provided on GitHub (https://github.com/mdy89/clinical_pathway_components).
- Ten representative patients were selected for each CPC according to their occurrences in patient records. For example, if patient A's record contains 1,000 clinical items, 900 of which are assigned to CPC_1, and patient B has 2,000 items, 1,200 of which are assigned to CPC_1, then patient A is considered more representative for CPC_1.

The clinical evaluation results of the top seven dominant CPCs (i.e., the most frequently observed CPCs) are listed in Table 1 in which a clinical annotation and fraction of occurrences of each group of patients are provided.

3.5 Identification of Care Patterns

In addition to the annotation results of the CPCs, the physicians also identified treatment patterns in terms of CPCs that consistently occur in patient traces.Some basic treatment patterns were introduced that comprised the top seven CPCs.

Table 1. Clinical evaluation results of the top seven CPCs.

	%_patients	Note
CPC_1	73.71%	Routine care. Mostly represented by basic labs and outpatient treatments.
CPC_2	47.77%	Urgent/emergent cardiac ischemia workup for symptomatic individuals.
CPC_3	35.09%	Scheduled or routine cardiac ischemia workup and follow-up visits.
CPC_4	29.34%	Establishing care that consists of drug screen, basic medications, and lab tests.
CPC_5	24.42%	Cardiac hospital care when cardiac workup is done elsewhere first (CPC_2 & 3 absent).
CPC_6	15.88%	Cardiac hospital care when cardiac workup is done locally first (coupled with CPC_2 & 3).
CPC_7	12.09%	Cardiac hospital care after CPC_3 or after transfer in.

- CPC_1 accompanied with either CPC_2 or CPC_3 is a cardiac workup followed by routine care or coronary artery bypass surgery (CABG).
- CPC_2 with occasional CPC_1 and CPC_4 is an initial workup with cardiac rehab.
- CPC_3 alone or occasionally with a CPC_1 is one of the following three cases: (1) an initial evaluation and straight to cardiac rehab, (2) a routine preoperative evaluation, or (3) an evaluation before antiarrhythmics.
- CPC_4 alone is an establishing care.
- CPC_5 accompanied with either CPC_1 or CPC_6 is a transferred-in situation in which a cardiac evaluation is first conducted, followed by CABG.
- CPC_2 or CPC_3 accompanied with either CPC_6 or CPC_7 is a cardiac workup followed by cardiac surgery and hospital care.

During the evaluation, consistent care patterns—such as routine examination, cardiac workup, lab tests, basic medications, and cardiac hospital cares—were identified, which are considered primary care patterns in SIHD patients [11]. This indicates that this method can capture basic care patterns from EHR data. However, because the proposed method focuses on the most widely used patterns during each step of the process, such as during CPC identification heuristics (Section 2.5), the resulting CPCs are also highly focused on the main patterns. There is still a possibility that some minor but clinically significant events are not explicitly represented in the results.

4 Conclusion

This work introduced a data-driven approach to infer CPCs from EHRs. The empirical evaluation on a VA SIHD cohort shows that the approach discloses CPCs that are consistently found in actual patient cases. However, because the proposed approach is strictly data-driven and unsupervised, it tends to favor frequent patterns, thus potentially insensitive to rare but clinically meaningful care patterns. To address this, the authors are considering the incorporation of additional criteria to evaluate the clinical outcome of each patient, such as 30 day mortality [4]. This could provide a way to improve the approach to capture patterns that are less frequent but could be clinically meaningful. For example, a patient cohort can be subdivided into subsets of patients by the criteria. Then, by applying the method to each subset of patients, the differences in patient care patterns can be identified with different clinical outcomes. This can also highlight minor but significant patterns that would be ignored before subdivision because they are significant only in a smaller group of patients.

5 Acknowledgements

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so, for the US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

1. Basse, L., Jakobsen, D.H., Billesbølle, P., Werner, M., Kehlet, H.: A clinical pathway to accelerate recovery after colonic resection. Annals of surgery **232**(1), 51 (2000)
2. Blumenthal, D., Tavenner, M.: The meaningful use regulation for electronic health records. New England Journal of Medicine **363**(6), 501–504 (2010)
3. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
4. Jamieson, W., Janusz, M., Miyagishima, R., Gerein, A.: Influence of ischemic heart disease on early and late mortality after surgery for peripheral occlusive vascular disease. Circulation **66**(2 Pt 2), I92–7 (1982)
5. Jha, A.K., DesRoches, C.M., Campbell, E.G., Donelan, K., Rao, S.R., Ferris, T.G., Shields, A., Rosenbaum, S., Blumenthal, D.: Use of electronic health records in us hospitals. New England Journal of Medicine **360**(16), 1628–1638 (2009)
6. Organization, W.H.: International statistical classification of diseases and related health problems: Tabular list, vol. 1. World Health Organization (2004)

7. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
8. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
9. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: International conference on artificial neural networks. pp. 583–588. Springer (1997)
10. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**(301), 236–244 (1963)
11. Wenger, N., Boden, W., Carabello, B., Carney, R., Cerqueira, M., Criqul, M.: Cardiovascular disability: Updating the social security listings. National Academy of Sciences (2010)
12. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)