ORNL/TM-2018/1034

Advanced Analytics Studies Applied to US Department of Veterans Affairs' Corporate Data Warehouse

Hoony Park
Jason Laska
Hilda B. Klasky
Aileen Boone
Ozgur Ozmen
Karthik Rajasekar
Aneel Advani
Colby Cox
Mark Pleszkoch
Edmon Begoli

October 2018

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2018/1034

Computational Sciences and Engineering Division

ADVANCED ANALYTICS STUDIES APPLIED TO US DEPARTMENT OF VETERANS AFFAIRS' CORPORATE DATA WAREHOUSE—INITIAL DRAFT

October 2018

Hoony Park
Jason Laska
Hilda B. Klasky
Aileen Boone
Ozgur Ozmen
Karthik Rajasekar
Aneel Advani
Colby Cox
Mark Pleszkoch
Edmon Begoli

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

This page is intentionally blank.

TABLE OF CONTENTS

This page is intentionally blank.

LIST OF FIGURES

This page is intentionally blank.

LIST OF TABLES

This page is intentionally blank.

ACRONYMS

2D	two-dimensional
3D	three-dimensional
AI	artificial intelligence
CCS	Clinical Classifications Software
CDW	Corporate Data Warehouse
CM	clinical modification
CPT	Current Procedural Terminology
DL	deep learning
DSE	Data Science Environment
EHR	electronic health record
FY	fiscal year
HIT-AA	Health Information Technology–Advanced Analytics
LDA	latent Dirichlet allocation
ML	machine learning
MLA	machine learning algorithm
ORNL	Oak Ridge National Laboratory
PCA	principal component analysis
PCS	Procedure Coding System
PTSD	post-traumatic stress disorder
RBM	restricted Boltzmann machine
Rx	Prescription
SIHD	stable ischemic heart disease
TF	Term Frequency
t-SNE	T-distributed Stochastic Neighbor Embedding
VA	US Department of Veterans Affairs

This page is intentionally blank.

ABSTRACT

In this report, we describe our experience of applying several advanced analytics algorithms to the US Department of Veterans Affairs' (VA's) Corporate Data Warehouse (CDW) electronic health records datasets during FY2017-18. While various algorithms were applied to the CDW data, the main goal of this effort was to provide useful insights into the *operational* aspects (as opposed to the purely clinical aspect) of this specific implementation. Since there are not many reports in the public literature on advanced analytics applied to this data, this report provides unique insight in this regard. We focused on two machine learning (ML) applications: (1) clinical pathway inference (patterns dominating clinical pathways applied to a male dataset of Stable Ischemic Heart Disease patients as use case) and (2) medical-concept representation learning's capability to inform and refine cohort membership based on probable patient outcomes.

During our study on clinical pathway inference, we used two main modeling techniques: (a) topic probabilistic modeling (latent Dirichlet allocation [LDA]) and (b) feature imbedding. We observed that the applicability of LDA to the pathway inference remains in question. LDA provides results that are intuitive and easy for humans to comprehend; however, when applied to the pathway inference, it also generates pathway patterns that consist of a few dominant pathway components. In addition, LDA provides no information regarding the relationships among components in the same pathway. The outputs of LDA are also found to categorize patients based on their trace data of clinical procedures. However, the results suggest that LDA is biased toward statistically dominant components. This makes it particularly hard to discriminate pathway subbranches. To address the issue, we designed a new pathway inference methodology that integrates temporal ordering of pathway components into LDA results, as existing feature embedding tools such as *word2vec* are not readily applicable to our task. In addition, we developed our own feature embedding tool customized for patient traces. We conclude this study by suggesting two approaches to use embedding representation of a component into LDA and briefly list our lessons learned and recommendations for future work.

For our study on medical-concept representation learning ability to inform and refine cohort membership based on probable patient outcomes, during our empirical evaluation we compared methods from medical-concept learning to standard one of a kind (one-of-K) encoding to evaluate the change in effectiveness as done in Choi, Schuetz, Stewart, and Sun (2016). We performed two primary empirical evaluations. The first evaluation was on a curated collection of 60,000 patients with no more than 1 year of medical history included. The second was a collection of patients with no restriction to the amount of medical history included. The hypothesis across both evaluations is that representation learning is broadly useful independent of downstream processing models; thus, for each experiment, we trained three models—a logistic regression model, a two-layer neural network model, and a nearest-neighbor model—and then averaged the results. Each model was trained using fivefold cross validation. We provided evidence that medical-representation learning improves predictions of the primary diagnosis category of a short patient history through 12 experiments. In addition, we provided evidence that medical-representation learning fails to improve prediction of the primary diagnosis category of an arbitrarily long patient history through 12 experiments. We conclude this study with a brief list of our lessons learned and recommendations for future work.

We conclude the report with a discussion of recent technology trends as they relate to our artificial intelligence and, more specifically, to our ML research and approaches as described previously.

1. INTRODUCTION

In fiscal year (FY) 2017–2018 of the Health Information Technology-Advanced Analytics (HIT-AA) Project, Oak Ridge National Laboratory (ORNL) responded to the US Department of Veterans Affairs' (VA's) drive to assess the quality of patient care, measure value, and improve safety in health care delivery by designing and building advanced analytic software systems that leverage the Data Science Environment (DSE), one of three principal components in ORNL's advanced analytics architecture. Through the DSE, ORNL has successfully begun to (1) assess the quality of treatment protocol guidelines and (2) quantify the value of guideline-recommended clinical pathways and treatment protocols by determining how variants in patient characteristics (risk factors), testing, and timing of treatments might lead to better patient outcomes.

Meeting the VA's advanced analytic needs for business intelligence and clinical decision support, ORNL has researched and developed analytic models and algorithms that support workflow analytics using artificial intelligence (AI) and, more specifically, machine learning (ML) methods, which enable analytics for complex event processing. In the following sections, ORNL provides the FY 2017–2018 report for the three VA HIT-AA research tasks. In Section 1, we present an overview of the two research plans. Section 2 describes the problem domain study and models by research task, Section 3 details the development of analytic models and methods by research task, Section 4 discusses the application of advanced analytics (models and methods) in the DSE, and Section 5 discusses future work and opens issues.

1.1 HIT-AA RESEARCH OVERVIEW

Using advanced analytics on big, heterogeneous health data is an answer to VA's ambition to monitor and assess quality, safety, and value to measurably improve its delivery of health care to veterans. Through novel methods and research, ORNL's HIT-AA effort supports the VA's aim to ensure veterans receive high-quality health care. The HIT-AA project uses big health data in conjunction with specific clinical workflows and extensions for cohorts of patient cases in real time. It is a big advance to have specific information that is just a few clicks away to measure quality and outcomes of care, not only for traditional stable ischemic heart disease (SIHD) patient cohorts for which there are published guidelines (Fihn et al., 2012) but also for more focused sub-cohorts such as those patients who have SIDH and who are over 80 years of age, have been diagnosed with post-traumatic stress disorder (PTSD), have had diabetes for over 10 years, and have had a myocardial infarction in the last year.

The HIT-AA project has engendered research that aims to investigate various areas of advanced analytics such as ML and neural networks, specifically focusing on (1) clinical pathway inference (patterns dominating clinical pathways from the VA's Corporate Data Warehouse [CDW]), (2) medical-concept representation learning ability to inform and refine cohort membership based on probable patient outcomes, and (3) game theoretic approaches to structure an inference model for guideline-based clinical cohort analytics and quality measurement (see *HIT-AA: Game-Theoretic Approach for Understanding and Modeling Stable Ischemic Heart Disease*). This HIT-AA research could potentially be utilized to improve the quality of care to veterans as well as enable metrics regarding safety and value.

1.1.1 Clinical Pathway Inference

Specific Aims: This research aims to infer patterns dominating clinical pathways from CDW and then to use the obtained insights to further analyze cohorts from the perspective of clinical pathway implementations. More specifically, this research is (1) investigating different methodologies using probabilistic modeling (latent Dirichlet allocation [LDA]), word embedding, and in the future, a restricted Boltzmann machine (RBM), (2) performing a comparative study on pathways inferred from different methodologies looking at the evaluation strength and limitation of each methodology, and (3) exploring

opportunities with inferred pathways such as the evaluation and refinement of pathways and cohort analysis from a pathway perspective.

Significance: As integrated multidisciplinary care maps, clinical pathways aim to improve outcomes of patients' health and clinical efficiency by standardizing care processes and have demonstrated their value in support of patient care management. However, there remains a lack of consensus regarding what constitutes a clinical pathway. As a result, a clinical pathway has numerous variations in practice and often reflects no relation to the ideal pathways elaborated by pathway designers. This research aims to fill the gap by computationally modeling clinical pathways.

Inferred patterns allow us to define a distance metric between patients from the perspective of clinical activities performed on them. Because this essentially enables identification of subgroups within a cohort by mapping the patients into the defined metric space, a comparative cohort analysis can be performed between different subgroups. Also, patients whose clinical activities are unusual from those of rest of the patients with the same diagnosis are easily detected and evaluated.

Innovation: By performing a comparative study of the results from the three approaches (LDA for inferring clinical pathway patterns from the CDW database, distributed representations that capture semantics of words, and in the future, a RBM and its variations), this research not only assesses strengths and weaknesses of each methodology but also generate non-trivial insights regarding the practice of clinical activities hidden in CDW, which might not be feasible if a single methodology is applied, potentially improving the quality of care.

1.1.2 Medical-Concept Representation Learning

Specific Aims: This research uses medical-concept representation learning to predict cohort membership specializations based on patient and historical outcome information. We investigate the ability of medical-concept representation to inform and refine cohort membership based on probable outcomes. Broadly, the goal is to provide some data-driven flexibility into how data are represented to ML models to improve the quality of analysis. This flexibility in representation can enable learning relationships between codes, prescription information, and outcomes independently and jointly. (Reasons why similar information might be represented in different ways in an electronic health record include variability in coding, free text entry, etc.) Medical-concept learning has demonstrated the ability to relate International Classification of Diseases, ninth revision (ICD-9), codes related to eye problems even though the Clinical Classifications Software groups these codes into different clinical categories (Choi, Schuetz, Stewart, & Sun, 2016).

This research produces a software capability operating on CDW data to drill down into SIHD cohort information and provide more context into probable outcomes using medical-concept representation learning.

Significance: Providing more finely grained refinements to cohort-based guidelines, informed by historical patient outcomes, could improve outcomes by ensuring a more obvious mapping of individual needs to the cohort. This method could provide additional context to help inform how guidelines, devised for cohorts, can best aid the patient. Potentially, this research could lessen the tension between the generality of guidelines and the specifics of treating an individual patient.

Medical-concept representation learning is an area of active research and has created cohorts and increased prediction accuracy. Medical-concept representation learning has been found to improve both the predictive capability of ML tools for specific tasks (Choi, Schuetz, Stewart, & Sun, 2016; Zhu et al., 2016) and to improve the creation of cohorts for cohort analysis (Zhu et al., 2016). The broad intuition

behind medical-concept learning is to provide some data-driven flexibility in how the data are presented to ML models to improve the quality of analysis. Specifically, this is achieved by embedding data into a vector space. Unlike standard categorical one-of-K encoded representations, medical-concept representation learning updates during model training. Part of the research objective is to determine and evaluate different embedding methodologies and quantify their relative contribution to improving downstream analysis.

Innovation: Focusing more on representation as opposed to algorithm development, this tool makes available the most fine-grained analysis of cohort information, informing treatment in a more precise manner and potentially improving the quality of care.

2. DEVELOPMENT OF PATHWAY INFERENCE MODELS AND METHODS

An intermediary step, developing analytics models and methods, is the core of the solution. We built models and developed methods for inference of clinical pathway patterns.

The following sections present conceptual descriptions of the advanced data analytics algorithms used in this project to infer clinical pathway patterns from the CDW database. We plan to use two main modeling techniques: (1) topic modeling and (2) feature imbedding. In the future, we would like to also apply the RBM to see how it compares to topic modeling LDA.

2.1 TOPIC MODELING

Topic modeling is part of the field of information retrieval. It provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives. Topic modeling is helpful to

1. uncover hidden topical patterns that pervade the collection;
2. annotate documents according to topical patterns; and
3. use the annotations to organize, summarize, and search texts.

There are several types of topic modeling algorithms. During this project, we use the LDA model. In the future, we would like to also apply the RBM model to compare, and validate, the LDA model.
Topic modeling uses the following terms to define the modeling items: (1) vocabulary, (2) topic, (3) document, (4) word, and (5) corpus. In topic modeling, a vocabulary is a set of unique words. These terms are presented in

Table 1.

In order to apply topic modeling to the field of health care informatics, we need to map the terms using the modeling of the items to the clinical terminology. In the context of the clinical pathways for this study, the following terms are defined as:

- A vocabulary consists of the pathway components (i.e., a set of unique treatment activities).
- A topic is a clinical pathway. A sample of a clinical pathway is shown in Figure 1, which presents a redraw of Figure 2 for diagnosis of SIHD taken from the *2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease* (Fihn et al., 2012*)*. A clinical pathway is the flow of clinical events to diagnose or treat a disease or a health condition.

- A document is the equivalent of a collection of clinical activities performed on a particular patient during a period of time.

- A word, an item, or a token is a particular clinical treatment activity. In Figure 1, a word is each of the shapes (either a box or a diamond shape) identified by a unique identifier number.
- Finally, a corpus is a cohort of patients.

Table 1. Topic modeling terms mapped to pathway terminology

Topic modeling term	Definition	Topic modeling term mapped to clinical pathways terminology
Vocabulary	A set of (unique) words	Pathway components (set of unique treatment activities)
Topic	A pattern of words	Clinical pathway
Document	A sequence of words	Collection of clinical activities performed on a particular patient
Word	A basic unit of discrete data. An item, term, or token in the vocabulary	Clinical activity
Corpus	A collection of documents	A cohort

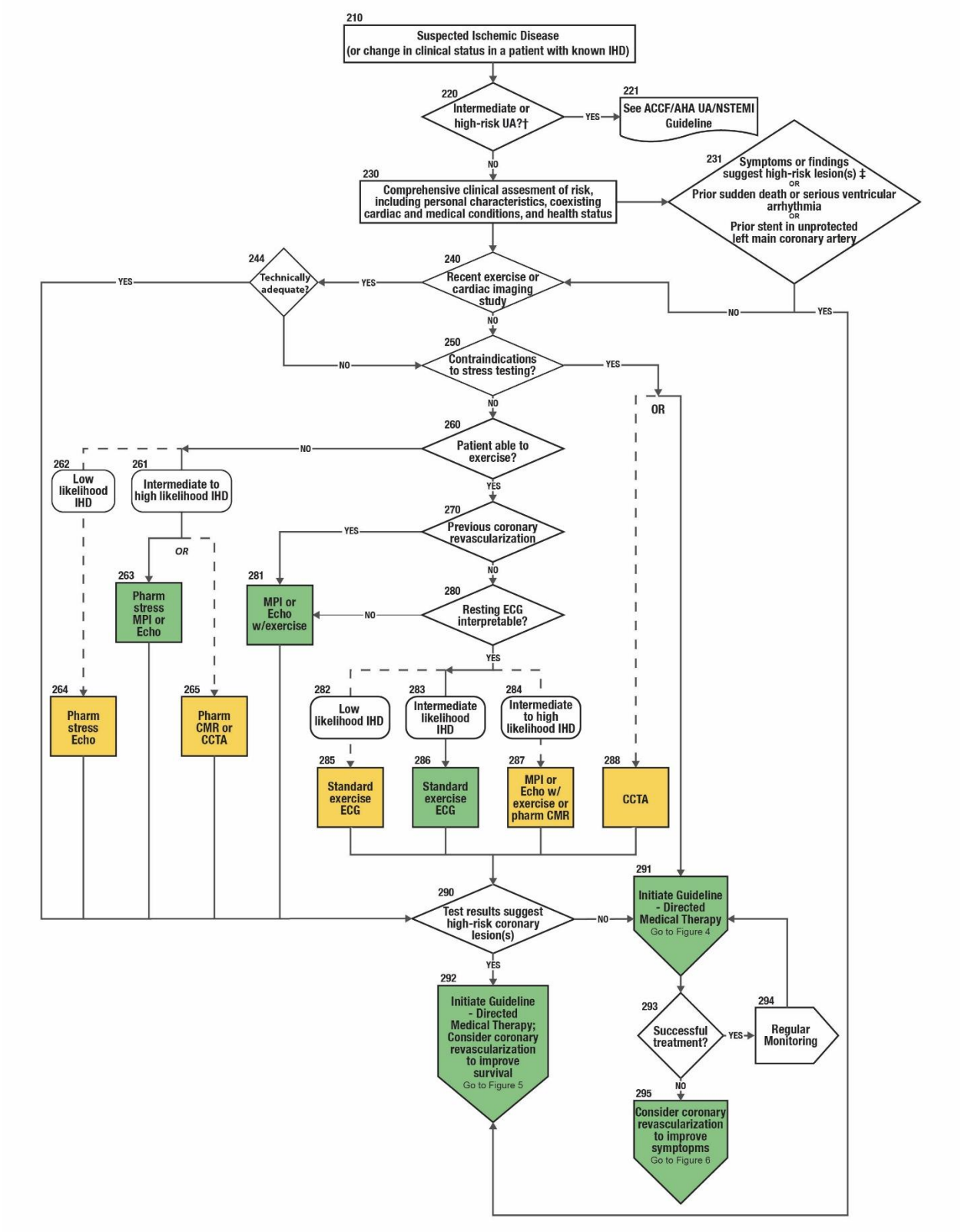


Figure 1. Clinical pathway for the diagnosis of SIHD (source: Fig. 2 Diagnosis of SIHD found in the 2012 ACCF SIHD guideline).

2.1.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a part of the field of information retrieval. Specifically, LDA is a generative probabilistic model of a corpus (Blei, 2003). LDA works over documents that are generated randomly and are unordered. Each document is assumed to be generated by this (simple) process. Documents are represented as a random mixture over latent topics. Documents exhibit multiple topics (but typically not many), and they have different distributions over topics.

A topic is a distribution over a fixed vocabulary. These topics are assumed to be generated first, before the documents. The number of topics should be specified in advance.

2.1.2 LDA Generative Process for a Document

The following section presents the process to generate documents in LDA (Figure 2). For this report, we rely on the work of Blei (2003) and Huang et al. (2014, 2013). We do not attempt to formally define LDA in this document. Rather, we present the referenced descriptions to illustrate our work and refer the reader to the references for the details. We follow as an example Huang's work using LDA to discover treatment patterns as a probabilistic combination of clinical activities.

- Choose N ~ Poisson($\boldsymbol{\xi}$)
- Choose N words or the topics

Generative process:

- Choose $\boldsymbol{\theta}$ ~ Dir($\boldsymbol{\alpha}$)
- For each of N words w_n:
  - Choose a topic z_n ~ Multinomial($\boldsymbol{\theta}$)
  - Choose a word w_n from $p(w_n|z_n, \beta)$

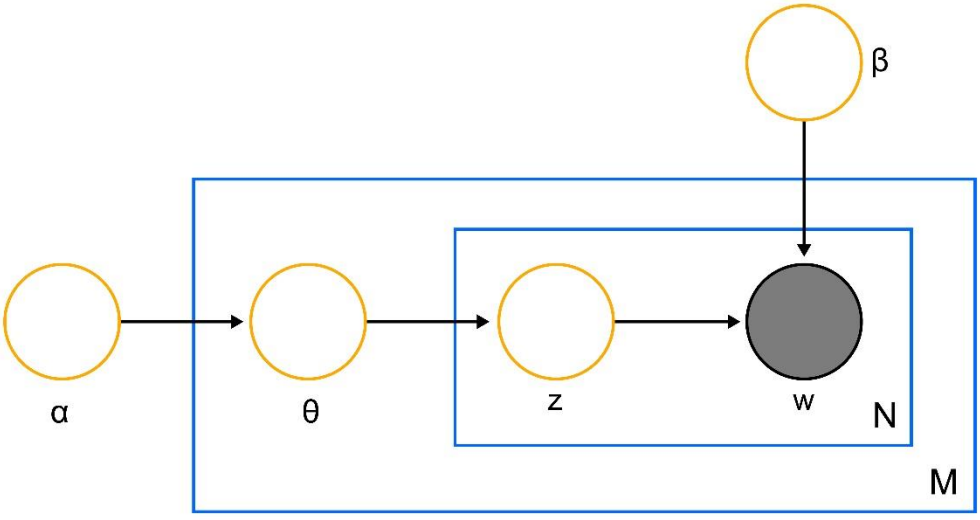


Figure 2. LDA generative process for a document.

Where:

α, β = parameters of the Dirichlet priors on topic (per document) and word (per topic) distributions, respectively

w = word

N = number of words

M = number of topics

Z = topic

2.1.3 Issues with LDA for Pathway Pattern Inference

The following are some of the drawbacks of LDA applied to clinical pathway pattern inference.

- LDA was originally designed for topic modeling of text documents.
- LDA requires prior distributions, which are always difficult to obtain and often intuitively initialized.
- LDA considers a document as a bag of words (or equivalently as a bag of clinical activities for the purpose of this text).
  - Spatial/temporal structures in occurrences of treatment activities are ignored.
  - Pathway subbranches are hard to discriminate.
- LDA tends to generate *dominant* patterns based on frequencies of occurrences.
  - The results might be too obvious.

2.2 FEATURE EMBEDDING

Feature embedding, also known as word or concept embedding, is a neural network implementation that aims to learn distributed representations of words. Unlike LDA and RBM, which learn the distribution of mere occurrences of input words, feature embedding learns occurrences of words against occurrences of other words that neighbor them in documents. In data mining, a word can be associated or predicted based on the words that surrounds it (Figure 3).

Figure 3. Different embedding strategies means different mappings.

Similarly, feature embedding is used to represent a health care event over time per patient in his medical history. In this context, a word, a feature, or a concept is a health care event or clinical feature. Each medical event, per patient, is represented in a vector, as shown in Figure 4. Then, the vectors are stacked to form a matrix of events per patient (Figure 4). What this means is that for each patient we concatenate medical events in a period of time in a sequence ordered by time that represents the patient's history. By creating this matrix per patient, we can identify cohorts, compare patient treatments, and identify similarities and differences.

Identifying patients' similarities in electronic health record (EHR) data is a very challenging process. We do not have control over how data are collected. The data used in this study may be incomplete,

inaccurate, or inconsistent. For this work, we follow the work presented in Zhu et al. (2016). During this study, we are using the Gensim Python package.

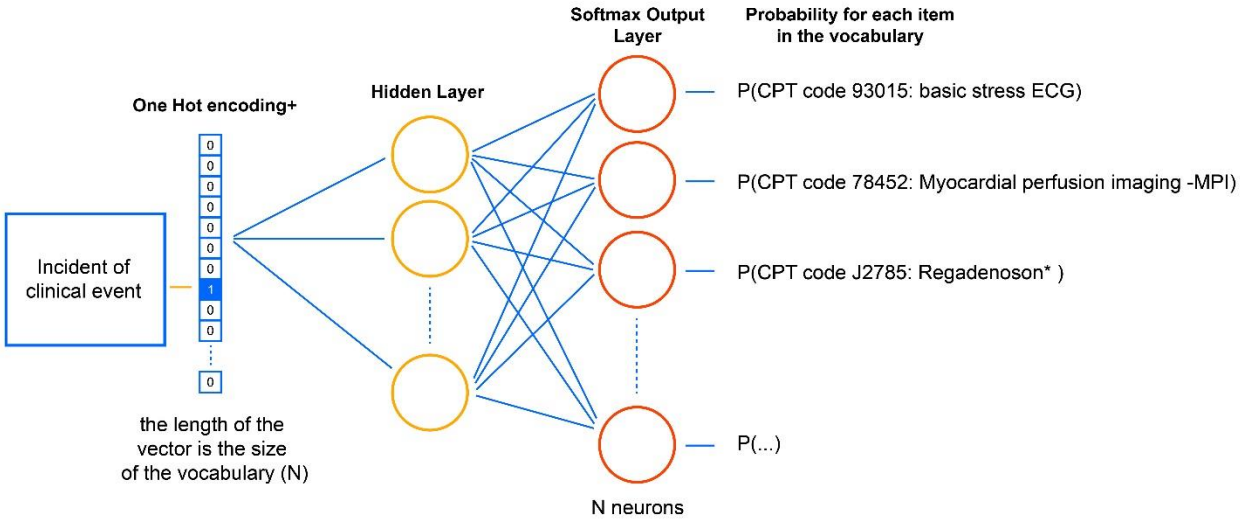


Figure 4. Feature embedding.

2.2.1 LDA vs. Feature Embedding

LDA explains the relationship between patient and pathway component. It is intuitive and easy to understand. In LDA, there are no relationships between components. Whereas humans cannot comprehend the embedding of a component, feature embedding explains the relationship between components.

Having these two different advanced algorithms, we wonder if we can combine the two to increase our understanding of clinical pathways. Combining LDA and feature embedding could be implemented in a future extension of this project.

3. TESTING OF PATHWAY INFERENCE MODELS AND METHODS

After building models and developing methods and algorithms, we tested both ML approaches in the DSE. The pathway inference approach is discussed below.

3.1 PATHWAY INFERENCE AND CDW DATA

In this section, we describe our efforts in pathway inference from CDW data. We first report an empirical evaluation of LDA applied to a small cohort of SIHD patients retrieved from the CDW data. Then we give status updates on the ongoing effort of integrating LDA with the feature embedding technique, a neural network approach that captures a degree of associations between pathway components in their distributed representations.

Clinical pathways are integrated multidisciplinary care maps used to manage and guide implementation of evidence-based clinical procedures and activities. They seek to improve the outcomes of patient health

and clinical efficiency by standardizing care processes. However, there remains a lack of consensus regarding what constitutes a clinical pathway. As a result, a clinical pathway has numerous variations in practice and often reflects no relation to the ideal pathways elaborated by pathway designers. This research aims to fill the gap by computationally modeling clinical pathways. Clinical pathway modeling is an effort to derive unique patterns that constitute each clinical pathway by sifting through EHRs. Once extracted, such patterns disclose not only consensus clinical activities practiced in VA hospitals but also non-trivial knowledge with regard to specific diseases. The patterns can thus be used to measure the gap between the pathways suggested and the actual practices in VA hospitals and can also be further examined to refine the suggested pathways.

Inference of clinical pathway patterns from EHRs is a challenging task. Therapies and treatment procedures are too diverse and complex to be represented by a simple model. Among many methodologies, which have been proposed in the recent past, most notable works are based on LDA, a probabilistic approach originally designed for topic modeling of text corpus. We observed, however, the applicability of LDA to the pathway inference remains in question. Most notably, LDA considers a document as a bag of words disregarding the order of their appearances in sentences and tends to generate dominant patterns based on frequencies of occurrences. Therefore, spatial (or temporal) structures of treatment activities are ignored in generating clinical pathways. This makes it particularly hard to discriminate pathway sub-branches. To address the issue, we designed a new pathway inference methodology that integrates temporal ordering of pathway components into LDA results.

3.1.1 LDA and Pathway Inference

Latent Dirichlet allocation (LDA) is a probabilistic approach for automatically organizing, understanding, searching, and summarizing a large electronic text archive. Given a set of text documents, LDA uncovers hidden topical patterns that pervade the collection and annotate each document by relevancy to topics. Formally, LDA produces topics in terms of probability distributions over the entire set of words (or vocabulary) in the *corpus*. Each document is then represented as a probability distribution over these topics. Here the number of topics is assumed to be known a priori.

To apply LDA to inference of a clinical pathway, we need to rename the terms used in the topic modeling to those used in a clinical pathway (Table 1). A component of a pathway replaces *word*, a patient trace (treatment activities performed on the patient arranged by the time of execution) replaces *document*, a pathway replaces a *topic*, and the set of the entire pathway components (or treatment activities) replaces *vocabulary*. LDA, when applied to pathway inference, produces a number of clinical pathways, where each pathway is defined as a probability distribution over treatment activities and each patient trace as a probability distribution over the pathways.

LDA requires a dataset in the form of a list of bags of words. More specifically, all the words in a document are presented together to LDA as if placed in a bag. LDA then computes occurrences of each word in each bag to produce the result. To apply LDA to pathway inference, we need to provide a patient trace as a bag of treatment activities. In fact, transformation of CDW data into suitable data is the key to the success of the approach. This process mainly consists of three steps:

1. Identification of pathway components in CDW
2. Development of procedures to extract occurrences of each component
3. Grouping of highly temporarily correlated components

Here, step 3 is the focus of our current effort to address the issue of applying LDA to pathway inference. This is discussed in greater detail in the subsequent sections.

3.2 EMPIRICAL STUDY: LDA OVER CDW DATA

The overall procedure of applying LDA to pathway inference is illustrated in Figure 5.

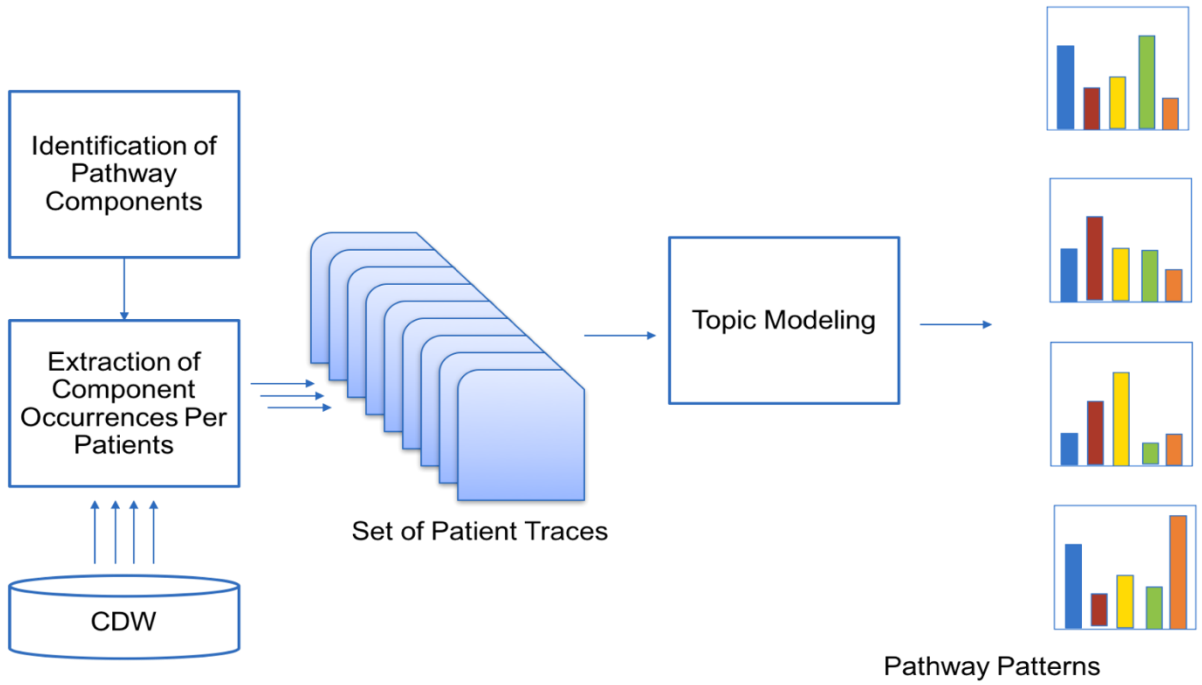


Figure 5. Overall procedure of topic modeling over CDW data for inference of pathways.

To evaluate LDA as an approach to inferring clinical pathways, we identified 20,458 records of patients who underwent cardiovascular stress tests (current procedural terminology [CPT] code 93015) in 2017. For each patient record, a trace is generated by including data that spanned a month following the day of the stress test. This generated

- 149,218 unique pathway components
- 2,490,724 occurrences of the components

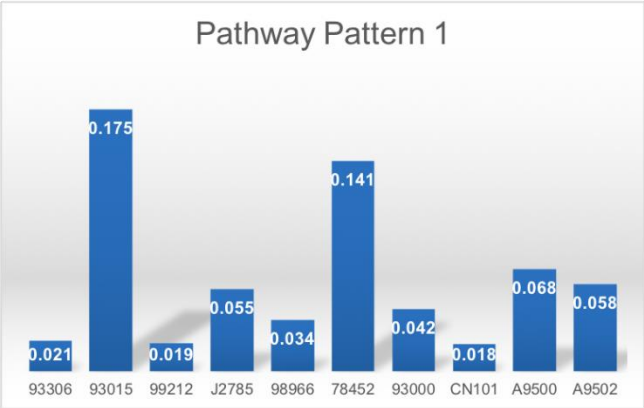
For the initial investigation, we decided to down select the data by focusing on inpatients and considered the following component types only.

- *InpsurgicalICDpcs*
- *Intcpt*
- *IntDischargeDRG*
- *RxFill*
- *OutpCPT*
- *InptICPpcs*

This reduced the data to

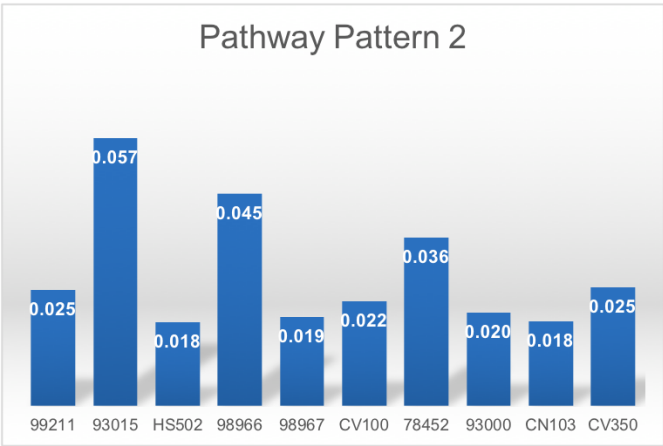
- 20,458 traces
- 6,203 unique components
- 415,232 occurrences of the components

We then transformed the patient traces into bags of components and applied LDA. Figure 6(a) through (j) show 10 pathway patterns LDA identified, where for each pathway, 10 components of the highest probabilities are shown. Also, descriptions of the components in each pattern and a proposed clinical interpretation of the pattern are provided. Note, components labeled with CPT and ICD-10 Procedure Coding System (PCS) are included, but not prescription drugs, which will be available in the near future.



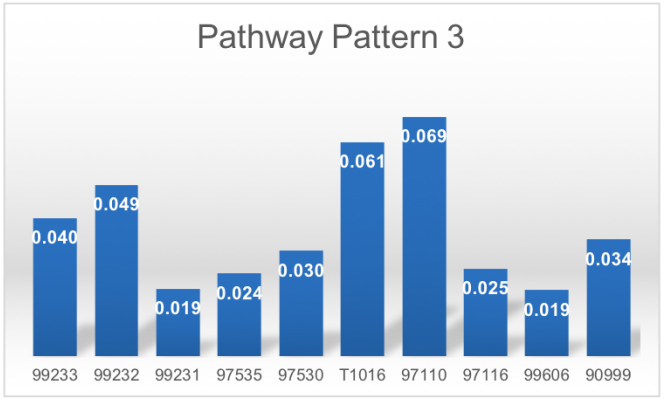
93306	Echocardiography Procedures
93015	Cardiovascular stress test using maximal or submaximal
99212	Established Patient Office or Other Outpatient Services
J2785	Injection, regadenoson, 0.1 mg
98966	Telephone assessment and management service
78452	Myocardial perfusion imaging, tomographic (SPECT)
93000	Electrocardiogram, routine ECG with at least 12 leads
CN101	Opioid Analgesics
A9500	Technetium tc-99m sestamibi, diagnostic, per study dose
A9502	Technetium tc-99m tetrofosmin, diagnostic, per study dose

Figure 6(a). Pathway pattern 1 represented as a probability distribution (left) and descriptions of components (right). This cohort consists of patients with heart disease with events related to diagnostic stress tests for SIHD. Only the top 10 components are shown.



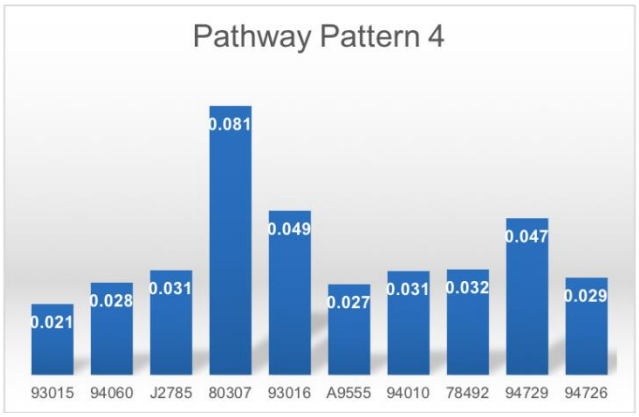
99211	Established Patient Office or Other Outpatient Services
93015	Cardiovascular stress test using maximal or submaxima
HS502	Oral Hypoglycemic Agents, Oral
98966	Telephone assessment and management service
98967	Telephone assessment and management service
CV100	Analgesics
78452	Myocardial perfusion imaging, tomographic (SPECT)
93000	Electrocardiogram, routine ECG with at least 12 leads
CN103	Non-Opioid Analgesics
CV350	Antilipemic Agents

Figure 6(b). Pathway pattern 2 represented as a probability distribution (left) and descriptions of components (right). This cohort represents patients with heart disease and diabetes, likely a sub-cohort of pattern 1 with added pharmacology therapy for diabetes. Only the top 10 components are shown.



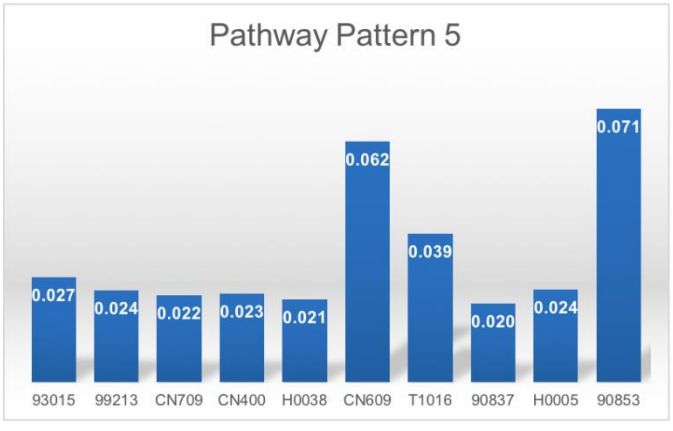
99233	Subsequent Hospital Care
99232	Subsequent Hospital Care
99231	Subsequent Hospital Care
97535	Physical Medicine and Rehabilitation Therapeutic Procedures
97530	Physical Medicine and Rehabilitation Therapeutic Procedures
T1016	Case management, each 15 minutes
97110	Therapeutic procedure, 1 or more areas, each 15 minutes
97116	Therapeutic procedure, 1 or more areas, each 15 minutes
99606	Medication therapy management service(s) provided by a pharmacis
90999	Other Dialysis Procedures

Figure 6(c). Pathway pattern 3 represented as a probability distribution (left) and descriptions of components (right). This cohort may consist of patients who are recovering from acute cardiac events such as heart attacks, with diagnostic tests to gauge continued recovery. Only the top 10 components are shown.



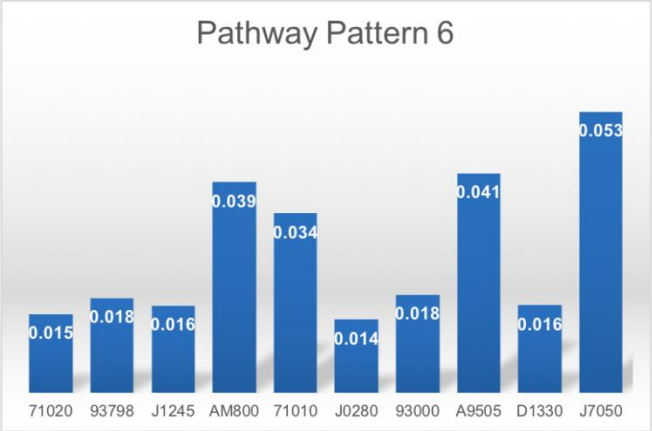
93015	Cardiovascular stress test using maximal or submaximal
94060	Pulmonary Diagnostic Testing and Therapies
J2785	Injection, regadenoson, 0.1 mg
80307	Presumptive Drug Class Screening
93016	Cardiovascular stress test using maximal or submaximal
A9555	Rubidium rb-82, diagnostic, per study dose, up to 60 millicuries
94010	Pulmonary Diagnostic Testing and Therapies
78492	Myocardial imaging, positron emission tomography (PET), perfusion
94729	Pulmonary Diagnostic Testing and Therapies
94726	Pulmonary Diagnostic Testing and Therapies

Figure 6(d). Pathway pattern 4 represented as a probability distribution (left) and descriptions of components (right). This cohort may consist of patients who have pulmonary embolism or are going through lung cancer screening. Only the top 10 components are shown.



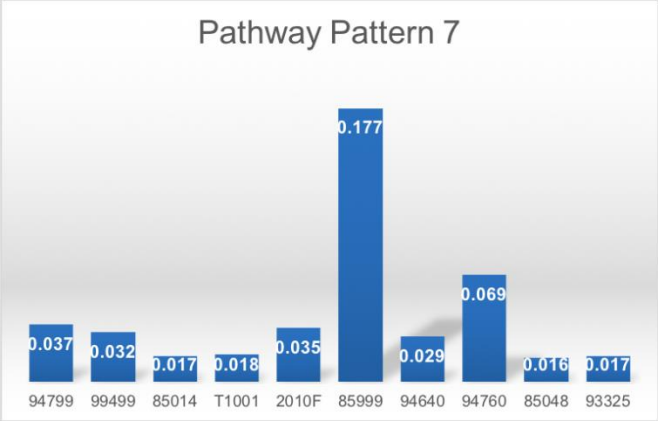
93015	Cardiovascular stress test using maximal or submaximal
99213	Established Patient Office or Other Outpatient Services
CN709	Antipsychotics, Other
CN400	Anticonvulsants
H0038	Self-help/peer services, per 15 minutes
CN609	Antidepressants, other
T1016	Case management, each 15 minutes
90837	Psychotherapy
H0005	Alcohol and/or drug services; group counseling by a clinician
90853	Other Psychotherapy

Figure 6(e). Pathway pattern 5 represented as a probability distribution (left) and descriptions of components (right). This cluster may consist of patients with psychiatric conditions, who are also being counseled for alcoholism, or a cluster of VA patients who are a homeless population with social support and mental health treatment. Only the top 10 components are shown.



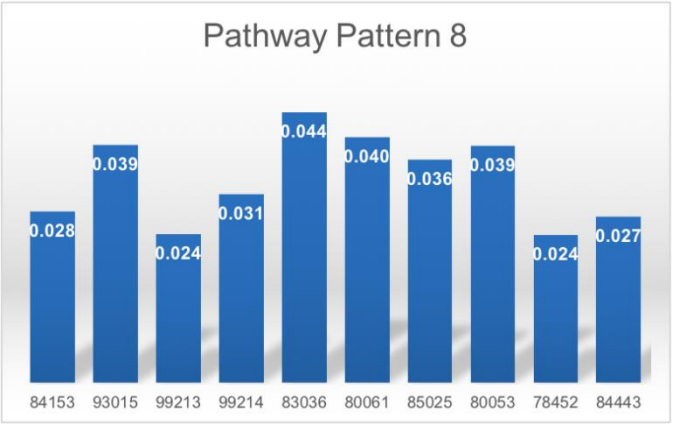
71020	Under Diagnostic Radiology (Diagnostic Imaging) Procedures of the Spine and Pelvis
93798	Physician services for outpatient cardiac rehabilitation
J1245	Injection, dipyridamole, per 10 mg
AM800	Antivirals
71010	Radiologic examination, chest
J0280	Injection, aminophyllin, up to 250 mg
93000	Electrocardiogram, routine ECG with at least 12 leads
A9505	Thallium tl-201 thallous chloride, diagnostic, per millicurie
D1330	Dental related
J7050	Infusion, normal saline solution , 250 cc

Figure 6(f). Pathway pattern 6 represented as a probability distribution (left) and descriptions of components (right). This cohort of patients could represent prostate cancer staging and therapy, with associated prophylaxis post-dental procedures or for septicemia. Only the top 10 components are shown.



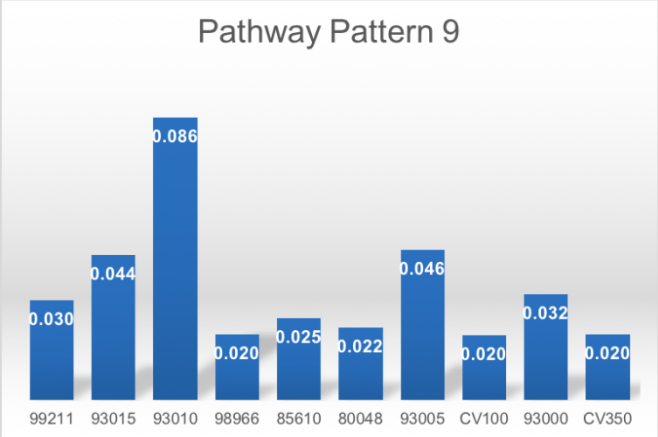
94799	Pulmonary Diagnostic Testing and Therapies
99499	Other Evaluation and Management Services
85014	Blood count
T1001	Nursing assessment / evaluation
2010F	Physical Examination
85999	Hematology and Coagulation Procedures
94640	Pulmonary Diagnostic Testing and Therapies
94760	Noninvasive ear or pulse oximetry for oxygen saturation
85048	Blood count
93325	Echocardiography Procedures

Figure 6(g). Pathway pattern 7 represented as a probability distribution (left) and descriptions of components (right). This cluster could represent a cohort of elderly patients who have regular outpatient visits or standard post-hospitalization outpatient follow-up visits. Only the top 10 components are shown.



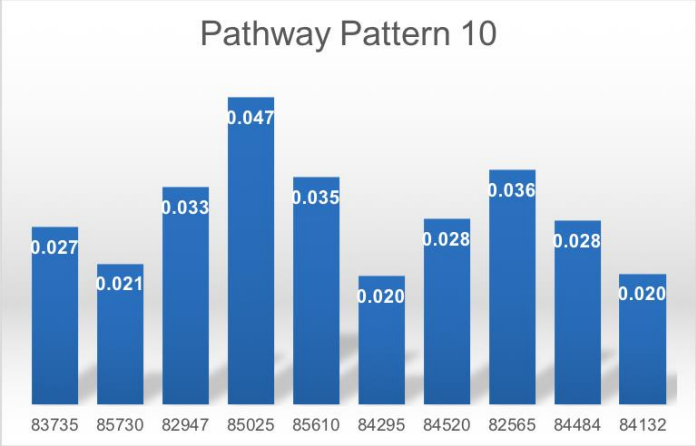
84153	Prostate specific antigen (PSA)
93015	Cardiovascular stress test using maximal or submaximal
99213	Established Patient Office or Other Outpatient Services
99214	Established Patient Office or Other Outpatient Services
83036	Hemoglobin
80061	Organ or Disease Oriented Panels
85025	Blood count
80053	Organ or Disease Oriented Panels
78452	Myocardial perfusion imaging, tomographic (SPECT)
84443	Chemistry Procedures

Figure 6(h). Pathway pattern 8 represented as a probability distribution (left) and descriptions of components (right). This cohort may represent elderly patients who have regular outpatient visits. Only the top 10 components are shown.



99211	Established Patient Office or Other Outpatient Services
93015	Cardiovascular stress test using maximal or submaximal
93010	Electrocardiogram, routine ECG with at least 12 leads
98966	Telephone assessment and management service
85610	Prothrombin time
80048	Organ or Disease Oriented Panels
93005	Electrocardiogram, routine ECG with at least 12 leads
CV100	Analgesics
93000	Electrocardiogram, routine ECG with at least 12 leads
CV350	Antilipemic Agents

Figure 6(i). Pathway pattern 9 represented as a probability distribution (left) and descriptions of components (right). This cohort could consist of patients on aspirin and blood thinners, who are post-cardiac patients and have outpatient angina treatment. Only the top 10 components are shown.



83735	Chemistry Procedures
85730	Thromboplastin time, partial (PTT)
82947	Glucose
85025	Blood count
85610	Prothrombin time
84295	Sodium
84520	Urea nitrogen
82565	Creatinine
84484	Chemistry Procedures
84132	Potassium

Figure 6(j). Pathway pattern 10 represented as a probability distribution (left) and descriptions of components (right). These patients could be cohorts of outpatient diabetic patients or those who are followed for other renal conditions. Only the top 10 components are shown.

We examined the results in order to evaluate whether LDA can capture clinical pathway patterns without reference to a candidate pathway and can discriminate patients in terms of conducted treatment activities. For this, we compared the 10 pathways, clustered patients based on their probability distributions over pathway patterns, and mapped components of each pathway into days by estimating their expected occurrences.

As shown in Figure 7, LDA disclosed pathways that mostly consist of cardiovascular-related components. However, LDA does not provide an explanation for how different components are organized into different pathways. More specifically, information about how the components in the same pathway are related is not available, which is a serious drawback in reconstructing the actual pathways.

Next, we clustered patient traces. With the 10 pathways LDA identified, each patient trace can be represented as a point in a ten-dimensional vector space. That is, for each patient trace, the probabilities associated with the 10 pathways are coordinates in the vector space. We applied a k-means clustering (method of vector quantization) algorithm that produced 10 clusters and identified patient traces. We visualized the result by reducing the dimension from 10 to 2 and 3 using the T-distributed Stochastic

Neighbor Embedding (t-SNE), a machine learning algorithm (MLA) for visualization, as shown in Figures 8 and 9. In the figures, the 10 clusters of patients are displaced in different colors. The results show that most clusters are found to be localized and spatially characterized even in lower dimensions. The result demonstrates that clustering of patient traces with respect to LDA outputs can characterize patient groups based on their treatment procedures. Outlier detection needs further study. As shown in Figures 8 and 9, patient traces that are spatially dispersed cannot be distinguished visually in the present two-dimensional (2D) or three dimensional (3D) representations.

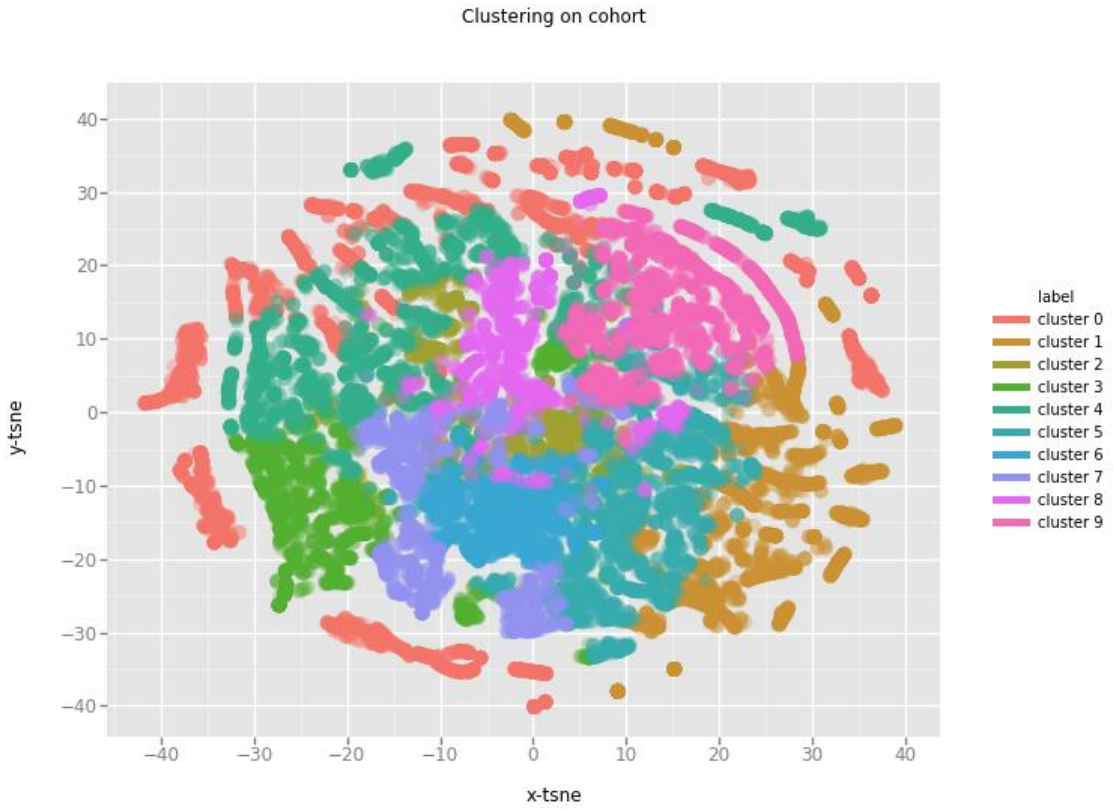


Figure 7. Clustering of patient traces using LDA results. K-means algorithm produced 10 clusters of the patient traces. The clusters are shown after projected into 2D space using t-SNE.

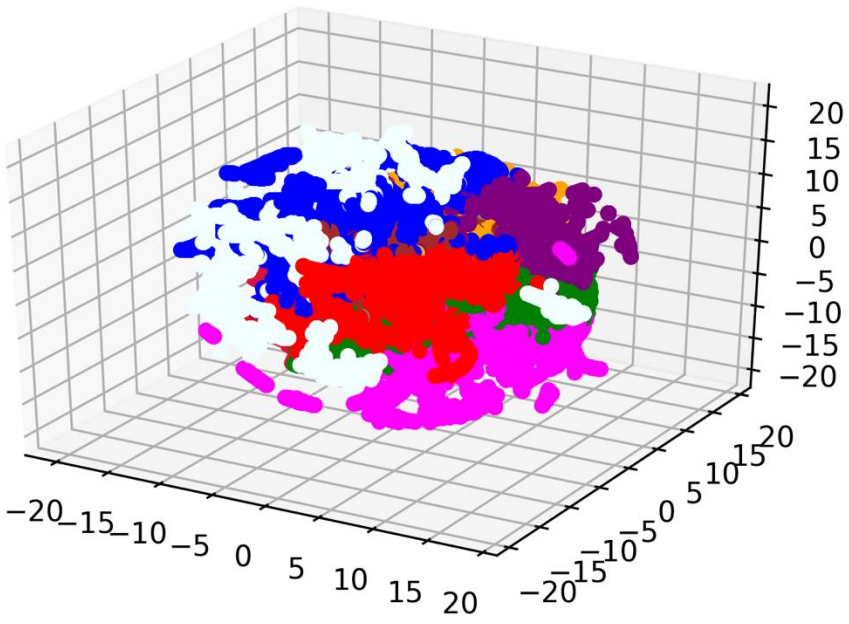


Figure 8. Clustering of patient traces using LDA results in a 3D space. K-means algorithm produced 10 clusters of the patient traces. The clusters are shown after projected into 3D space using t-SNE. (Note that colors assigned to the clusters are not the same as those used in Figure 7.)

Finally, we studied how each pathway component is likely to occur on each day (i.e., between day 1 and day 31). Here, "day" is relative days from the day when the initial cardiovascular stress test was given. This is intended to obtain an empirical pattern of each pathway. For this we first mapped all the pathway components into the days of their occurrences (Figure 9). This is a complementary view to Figure 7. The intent of this study is to evaluate whether LDA can capture how pathways are practiced in VA hospitals. Although the result is from a small explorative study, it captures clear patterns in occurrences of dominant components in each pathway pattern. For more comprehensive future studies, we plan to perform a more thorough investigation in analyzing results to confirm this observation. (See Appendix A, Figures A-1(a) through A-1(j) for expected occurrences of components in each pathway pattern computed using posterior probabilities [Gibbs sample].)

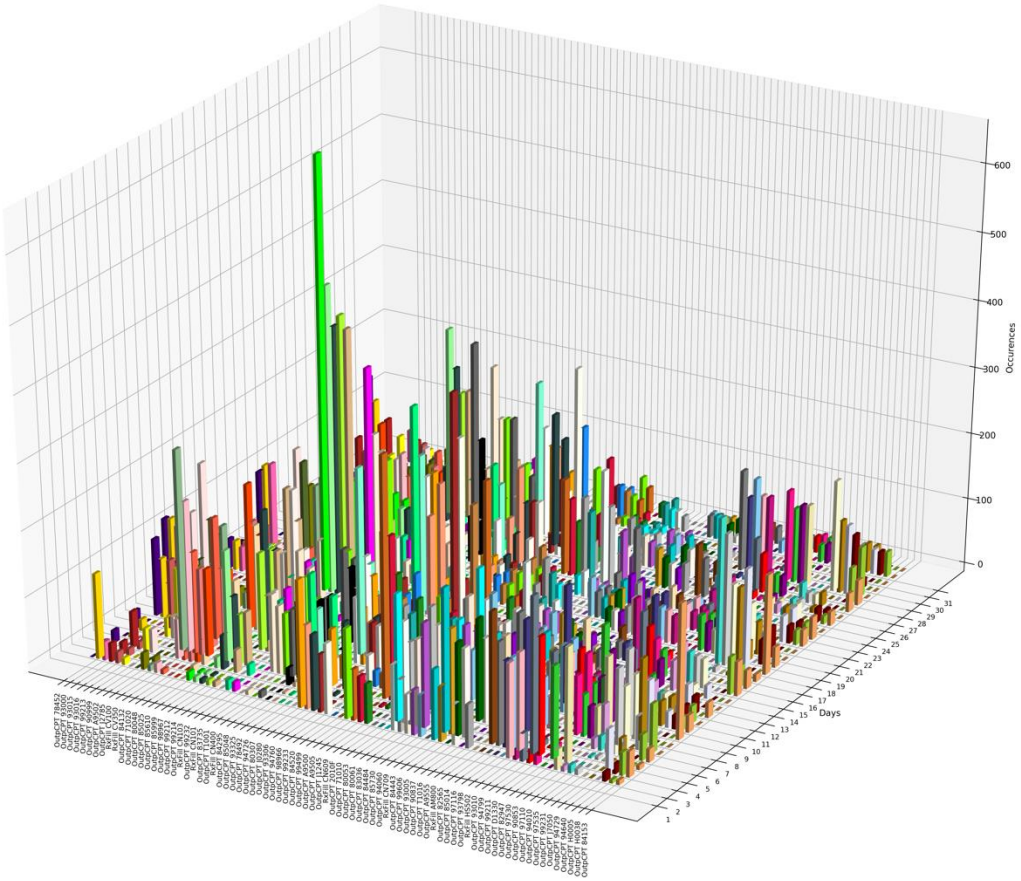


Figure 9. The number of occurrences of pathway components over the period of one month. Day 1 is the day when the first cardiovascular stress test was given.

3.2.1 Feature Embedding

Feature embedding is a neural network implementation that aims to learn distributed representations of words. It maps a word into a high dimensional space—as high as a 300 dimension, in practice—where the word is placed close to semantically similar words. Here "semantically similar" denotes two words that tend to appear in vicinity. Feature embedding, thus, incorporates occurrences of words against occurrences of other words that neighbor them in documents into the learning process.

When applied to patient trace data where pathway components are placed on the day (or the exact timestamp, for an investigation of finer resolution), pathway components that co-occur within the same time window of interest map to a high dimensional vector space where they are placed close to one another. Figure 10 illustrates placements of a time window over to six patient traces where pathway components are aligned by the day of their occurrences. Note each time window slides to right to include new sets of co-occurring components.

We found existing feature embedding tools such as *word2vec* are not readily applicable to our task. Those tools were developed to include pairs of words that are located within a fixed distance in the same sentence when computing the embedding. Here, a distance is simply the difference of word positions in a sentence. In practice, this is done by placing a window into a sentence and considering all pairs of words in the window. The window then slides to the right to include new sets of pairs. In summary, for the

existing tools, only the positions of the words within a sentence are relevant information to reflect temporal correlations on the embedding.

In contrast, for our task, we need to consider pairs of pathway components that occur within a fixed time window—that is, not the positions of components in the trace, but the exact time difference between the occurrences should be taken into consideration when computing the feature embedding. This suggests defining a sliding window in terms of time and considering the pairs of pathway components within the window. Consequently, the number of components within a window differs to a great extent.

To address the difference, we developed our own feature embedding tool customized for patient traces. In particular, we

- implemented "Efficient Estimation of Word Representations in Vector Space" by Mikolov, Sutskever, Chen, Corrado, & Dean (2013) from Google,
- adopted the skip-gram with negative-sampling for efficient estimation,
- implemented the tool using PyTorch, a deep neural network package from Facebook, and
- developed and tested the tool in a machine with four Nvidia Volta GPUs that we instantiated in the enclave.

Unlike existing tools, it accepts sequences of time-ordered items with window parameter in terms of seconds.

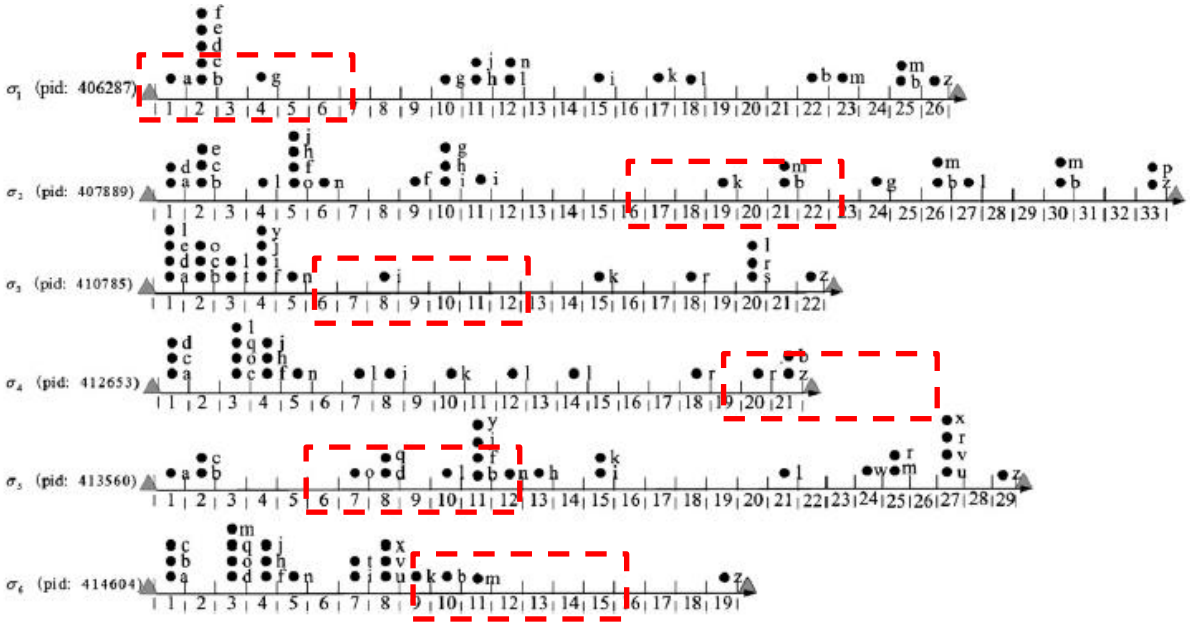


Figure 10. Illustrative example of placing sliding window over to patient traces for feature embedding. For each patient trace, a window of the same temporal size is placed to consider pairs of components in the window for embedding. The window slides into the next position until it passes day 31 (Huang, Lu, Duan, & Fan, 2013).

3.2.2 LDA and Feature Embedding

As illustrated in a previous section, LDA provides results that are intuitive and easy for humans to comprehend. When applied to the pathway inference, it also generates pathway patterns that consist of a few dominant pathway components. Also, for each trace, it assigns a few pathways with high probability,

leaving the rest of the pathways with very low probabilities. This essentially helps to characterize each pathway and each patient by their dominant components and pathways, respectively.

However, LDA provides no information regarding the relationships among components in the same pathway. This imposes a serious drawback when it is applied to the inference of sequential aspects of a pathway, where temporal correlation and ordering of occurrences are important factors. On the other hand, feature embedding provides a relationship between two components in terms of distance in a vector space. However, incorporation of such an embedded representation of a component into the LDA framework is an open question. In this report we describe the two approaches that we are currently exploring.

The first approach to integrate feature embedding and LDA is to apply a clustering algorithm over pathway components utilizing their pairwise distances in the embedded space. Once the component groups are obtained through the clustering, the subsequent procedure is the same as a regular LDA processing. The only difference is the final LDA outputs are represented in terms of the clustered component groups rather than individual components. This approach is intuitive, and results are easy to comprehend. However, since components are expected to belong to multiple groups in practice, a clustering algorithm that permits nonexclusive clustering should be applied. We are evaluating fuzzy-clustering (a form of clustering in which each data point can belong to more than one cluster) algorithms to this end. The overall procedure is illustrated in Figure 11.

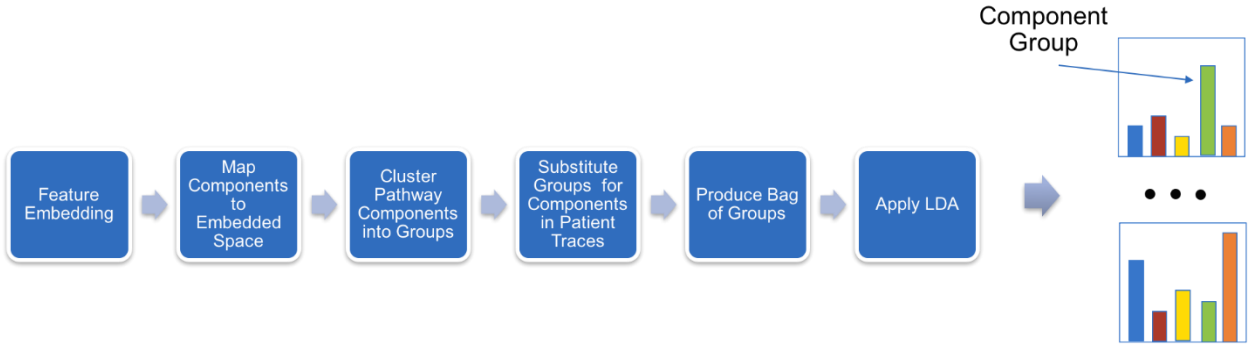


Figure 11. Clustering approach to integrate LDA and feature embedding. The final LDA results are presented in terms of component groups.

The second approach is to transfer feature embedding framework to learn LDA weights (Figure 12). More specifically, this approach borrows LDA output structures (sparse document and topic representations), but not LDA itself. With this approach, a pathway is defined as a point in the same embedded space of the components. As in LDA, a pathway is then represented as a probability distribution over the entire components. Here, we convert the distance from the pathway point to a component into probability. Formally, the probability of the *i*-th component in a pathway is

$$p_i = \frac{e^{-d_i^2/\sigma^2}}{\sum_j e^{-d_j^2/\sigma^2}},$$

where d_i is distance between the i-th component and the pathway, and σ^2 is variance of distances (following Gaussian distribution likelihood). A patient trace is represented similarly by computing distances to each pathway point from the components found in the trace.

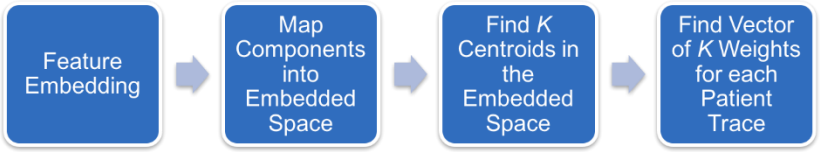


Figure 12. Transfer of feature embedding to learn LDA weights.

3.3 LESSONS LEARNED AND NEXT STEPS

In this study, we applied and developed a topic modeling approach to infer clinical pathway patterns for an SIHD cohort by sifting through electronic health records (EHRs). Once extracted, such patterns disclose not only consensus clinical activities practiced in VA hospitals but also nontrivial knowledge with regard to SIHD. In particular, we

A. demonstrated how consensus pathway patterns are practiced in VA hospitals by Latent Dirichlet Allocation (LDA) and showed they can also be used to further cluster cohorts
B. introduced the feature embedding concept (e.g., word2vec) to address the weakness of conventional topic modeling methods such as LDA for pathway inference and implemented an in-house feature embedding software customized for clinical pathway inference and
C. implemented the preceding procedures using highly portable and easily reusable open source programming environments including Jupyter and Apache Spark.

According to the empirical study presented in the previous sections, LDA, when applied to modeling of clinical pathways, extracts a given number of unique salient patterns, each of which characterizes certain aspects of clinical pathways. The outputs of LDA are also found to categorize patients based on their trace data of clinical procedures. However, the results also suggest that LDA is biased toward statistically dominant components. In addition to the aforementioned weakness of LDA (i.e., ignorance to temporally correlated components), this issue should be further studied. We will investigate whether the feature embedding approach can also mitigate the issue.

4. DEVELOPMENT OF REPRESENTATION LEARNING MODELS AND METHODS

This section discusses the development and testing of medical-concept representation-learning models and methods that apply representation learning to cohort clustering, thereby improving our understanding of the process of predicting cohort membership. The objective of this research is to give some data-driven flexibility to how data are presented to machine learning models, thus improving the quality of analysis.

4.1 MOTIVATION AND APPROACH

Data analytics for clinical decision making can be challenging. First, there is tension between the medical guidelines determined from studying groups of patients and the act of treating a specific patient. Medical professionals have deep expertise in the current treatment protocol guidelines at multiple levels of granularity. To ensure relevant and useful results, these guidelines are an analysis of cohorts of patients; however, professionals treat a single patient at a time. This fact creates strain between the generality of the guidelines and the specifics of the individual. More fine-grained refinements to cohort-based guidelines informed by historical patient outcomes could improve outcomes by ensuring a data-driven

mapping of the individual needs to the cohort. When treating an individual, it may be challenging to determine

- how a person fits with previous cohort analysis and
- how a cohort may be refined to better inform outcomes.

Additional context can help inform how guidelines can best assist the individual. To that end, we investigate two methods to transform how data are presented to algorithms to improve performance:

1. aggregate data using medical groupers
2. apply representation learning methods on aggregated data

Secondly, health care analytics has unique challenges associated with the data of health care. Indeed, "It is widely held that 80% of the effort in an analytic model is preprocessing, merging, customizing, and cleaning datasets, not analyzing them for insights" (Rajkomar et al., 2018). There are many challenges associated with deriving health care insights from health care data. Health care data are designed to address many different objectives across many stakeholders such as including documentation for compliance purposes, maintaining privacy, ensuring accurate billing information, and informing health care decisions. As such, similar information may be included multiple times in different formats in multiple places. This redundancy can impair performance due to overrepresentation. As such, it is difficult to answer questions relating to how information should be presented to algorithms in the service of health care analytics.

Targeting treatment to individuals is a challenging process. Scientific studies capture phenomena at a general level, but medical practitioners treat individuals. How do we address this gap? We can do this by using big data analytics, which is a challenging endeavor. The aim is to **confidently** map an individual to the **smallest** appropriate group or cohort. In this study, we apply representation learning to cohort clustering. Through cohort clustering, we understand the process of predicting cohort membership.

Work has been done for cohort clustering, but there is often a piece missing: how does this drive better health care outcomes? An example in the literature is separating widely disparate diagnoses: COPD, diabetes, obesity, heart failure (Verberne et al., 2017). Also, can we use the same techniques to perform sub-cohort clustering? In Figure 13, we present a sample set of 60 K CDW patients to facilitate iterating quickly, gaining intuition about what relationships to amplify and where the noise is. This study drives our interactions with the data refinement team to better direct analysis going forward.

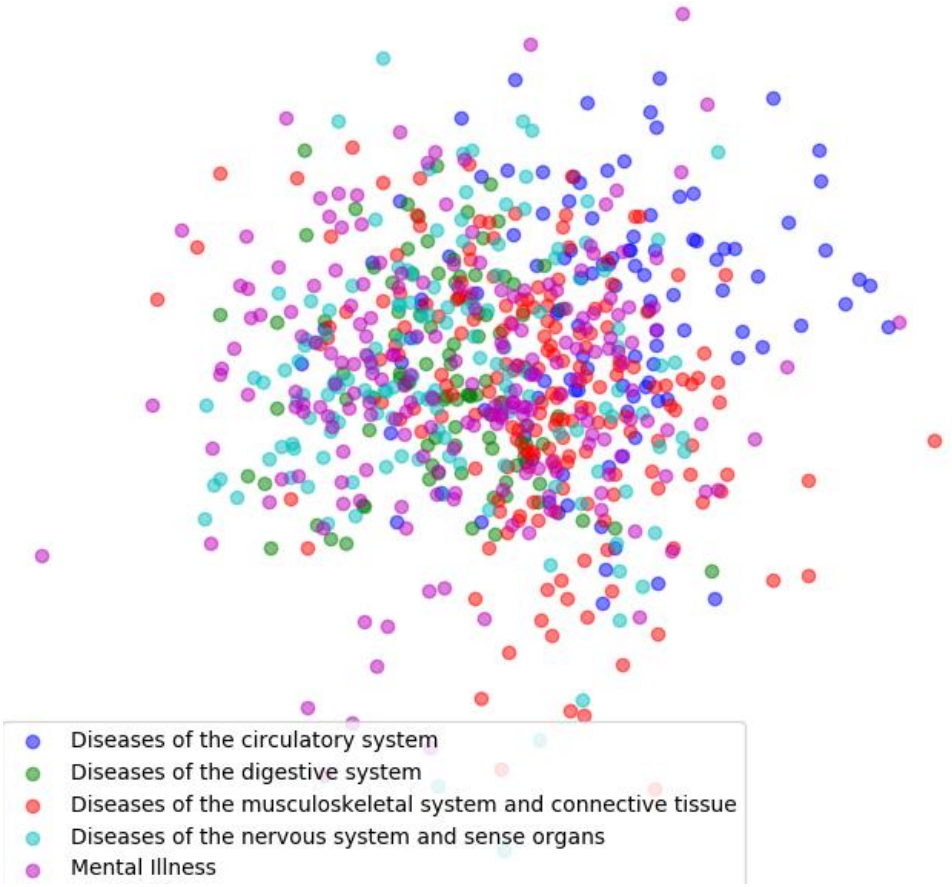


Figure 13. Visual relationship between codes from dense representation (60 K patients using VA CDW data).

4.2 MEDICAL-CONCEPT REPRESENTATION LEARNING FOR COHORT CLUSTERING

To understand representation learning, we need to understand ML. ML is the capability of AI algorithms to acquire new knowledge by identifying data patterns in raw data; ML also includes the capability of making decisions based on the knowledge acquired. Further, the MLA can make predictions. The performance of MLAs heavily, and literally, depends on the "representation" of the raw data to be analyzed. It also depends on the information passed to the MLA. Each piece of information passed to the MLA is called a "feature." The MLA's dependence on representation is a very common dependence in computer science problems. This dependence is what can make an algorithm perform better or worse. Thus, to make an MLA efficient, it is important to carefully select the set of features that will help to perform better on each case or task given. However, sometimes, it is difficult to identify what are the appropriate features that should be given to the MLA. In these cases, it is helpful to use representation learning.

Representation learning is a class of unsupervised learning methods to assist in the selection of the presentation of data to algorithms to facilitate analysis. Representation learning helps to discover not only the mapping from representation to output but also the representation itself (Goodfellow, Bengio, & Courville, 2016). Learned representations often result in much better performance than what can be obtained (Zhu et al., 2016) with hand-designed representations. Representation learning also allows AI systems to rapidly adapt to new tasks, with little help from people or from other systems. In addition, representation learning enhances performance enabling the discovery of new features**.** Representation

learning is tightly related to feature embedding; both techniques have been utilized in natural language processing and ML. When it is not possible to obtain a good representation to solve a problem, we need to use deep learning (DL). DL introduces representations that are expressed in terms of other simpler representations.

"Different data representations can entangle and hide more or less the different explanatory factors of variation behind the data" (Bengio, Courville, & Vincent, 2013). While different data representations can obscure certain factors, medical-concept representation learning has been found to improve both the predictive capability of ML tools for specific tasks (Choi, Schuetz, Stewart, & Sun, 2016; Zhu et al., 2016) and the creation of cohorts for cohort analysis (Zhu et al., 2016). Representation learning applied to health care has improved the ML tasks of diagnosis prediction, cohort creation, cohort membership, refinement of cohort membership, and outcome prediction. However, there is a need to improve outcomes by informing patient cohort information using probable outcomes.

In this research, we investigate the ability of medical-concept representation to inform and refine cohort membership based on patient information. Medical-concept learning is an area of active research and has shown promise. The broad intuition is to provide some data-driven flexibility into how the data are represented to ML models to improve the quality of analysis. This flexibility in how data are communicated using representation learning has resulted in learning relationships both **within** a field, such as codes, prescription information, and outcomes independently, and **across fields** (Rajkomar et al., 2018)**.** One example is that medical-concept learning has demonstrated the ability to relate ICD-9 codes related to eye problems 224.4 and 370.00 even though the Clinical Classifications Software (CCS) groups these codes into different clinical categories (Choi, Schuetz, Stewart, & Sun, 2016).

Our study is based on the work described in Choi et al. (2016); Miotto, Li, Kidd, and Dudley (2016), and Choi, Schuetz, Stewart and Sun (2016). Each of the previous works cited in this area focuses on the ability of predicting information concerning a visit (encounter). As such, this work hypothesizes an extension of previous visit-based prediction methods to a patient-based prediction.

For the implementation of representation learning, we use the skip-gram algorithm (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Skip-gram is a neural network-based algorithm capable of capturing relationships and predicting neighboring words between a sequence of words based on the co-occurrence of words inside a context. Skip-gram is a standard method for learning word representations and to capture relations between words by training on large amounts of text. In the same way that a text is a sequence of words, a health care record can be seen as a sequence of medical events for diagnosis, treatment, medication, and outcomes in the patient's life span. There is evidence that learned representations allow for a type of reasoning by analogy determined by vector addition. Indeed, some such analogies have been reported in the popular scientific press such as king - man + woman = queen. In this example, if E denotes the result of encoding, the researchers determined that the vector in the encodings of all words in the vocabulary closest to E(king) - E(man) + E(woman) was E(queen).

During our representation learning study, we analyzed patient visit records and their associated diagnosis codes (ICD-9 and ICD-10), procedures (CPT codes), and drug usage (National Drug Code [NDC]).

5. TESTING OF REPRESENTATION LEARNING MODELS AND METHODS

In this section, we describe (1) our method of employing representation learning to CDW data, (2) our work plan to evaluate the effectiveness in employing representation learning for clustering cohorts by primary diagnosis category, and (3) our empirical evaluations that compared methods from medical-concept learning to standard one-of-K coding to evaluate the change in effectiveness.

5.1 EMPIRICAL STUDY: REPRESENTATION LEARNING OVER CDW DATA

One standard way to represent raw data is through medical groupers. As part of this effort, we have developed tools operating on real CDW data to assist in the preprocessing, normalization, and standardization of datasets. The first is a suite of grouping tools that aggregate fine-grained code information into higher-level semantic collections. These tools reduce the number of distinct diagnoses or procedures to be considered by an algorithm, which improves learning efficiency. These tools benefit analysis in three fundamental ways.

1. These tools amplify the commonality between codes. For broad-level analysis, fine-grained distinction in minutes between the lengths of administrative visits may be less helpful than the knowledge that a patient is participating in an administrative visit.
2. The resulting multiset of aggregated codes often have a higher entropy than the original, so the raw codes with few data are often removed.
3. Since the amount of the data is the same, but the number distinct aggregated codes is smaller, there are more repetitions of the same signal for the algorithm to learn from.

First, we list the groupers that rely on external data. These include

1. Group ICD-10 and ICD-9 codes specifically for PCS,
2. Group ICD-10 and ICD-9 codes specifically for Clinical Modification (CM),
3. Group CPT code using CCS information.

These groupers, commonly referenced in publications, are a critical step to measuring how helpful some exploratory research can be in assisting the VA in its mission.

Secondly, we list the groupers than rely only on internal data. These cross reference existing fields in the CDW database to leverage institutional knowledge and best practices. (An opportunity exists to evaluate the relative benefit of comparing both the external grouper information and internal grouper information. This evaluation could lead to discussions about the value of future investments in grouping technology.) These groupers include

1. Group prescription (Rx) information by aggregating Drug Name information into Drug Category,
2. Group CPT code level information into “MajorCPTCategory” information.

In the service of cohort clustering, the level of aggregation provided by the final grouper may be too coarse. Specifically, the CPT codes associated with *hearing aids*, *wheelchair seats*, and *distilled water* are grouped into the same category. Moreover, codes associated with emergency room visits and short consultations concerning smoking cessation are grouped into the same category. This suggests that further comparative analysis may be needed to justify the inclusion of this CPT grouper over the CCS grouper.

Each of the previous works cited in this area focuses on the ability to predict information concerning a visit (encounter). As such, this work hypothesizes an extension of previous visit-based prediction methods to a patient-based prediction.

5.2 WORK PLAN: EVALUATING EFFECTIVENESS OF REPRESENTATION LEARNING

For the purpose of evaluating the effectiveness in employing representation learning for clustering cohorts by primary diagnosis category, we implement the following work plan. The goal is a software capability

operating on CDW data that can drill down into SIHD cohort information and provide more context into probable outcomes using medical-concept representation learning. The work plan steps are as follows.

1. Obtain access to broad-based cohort information across multiple ICD-9/ICD-10 codes, CPT codes, and Rx codes.
2. Obtain access to outcome information (previous literature has focused around the 3–6 month time period).
3. Develop methods to standardize coding before representation learning using groupers

    A. for ICD diagnostic codes (CM codes),
    B. for ICD hospital inpatient codes (PCS codes),
    C. for CPT codes, and
    D. possibly include Logical Observation Identifiers Names and Codes for laboratory results (Shivade et al., 2013).

4. Construct categorical, one of K, encoding of grouped and normalized data.
5. Using the Gensim Python package, apply a medical-concept representation learning method (Zhu et al., 2016; Choi, Schuetz, Stewart, & Sun, 2016; Miotto, Li, Kidd, & Dudley, 2016) to identify relationships within CPT and Rx fields.

    A. Over patient histories of up to 1 year
    B. Over patient histories of longer duration

6. Train cohort clustering methods on steps 4 and 5.
7. Visualize results of steps 5 and 6.

Due to recent successes in the literature, we think this is likely to assist in the delivery of higher quality health care by improving diagnosis cohort-clustering. Medical representation learning has improved predictions of heart failure up to 23% (Choi, Schuetz, Stewart, & Sun, 2016) and improved predictions of new disease onset by 15% (Miotto, Li, Kidd, & Dudley, 2016). This cohort analysis improves clinical practice informing outcome-informed guideline information about the patient at the time of care. From a research standpoint, medical-concept representation learning has the ability to flexibly include subject matter knowledge into algorithmic tools, improving the quality of analysis. There needs to be a historical focus on algorithm choice and subsequent performance, but recent work has highlighted the importance of data representation on results (Figure 14).

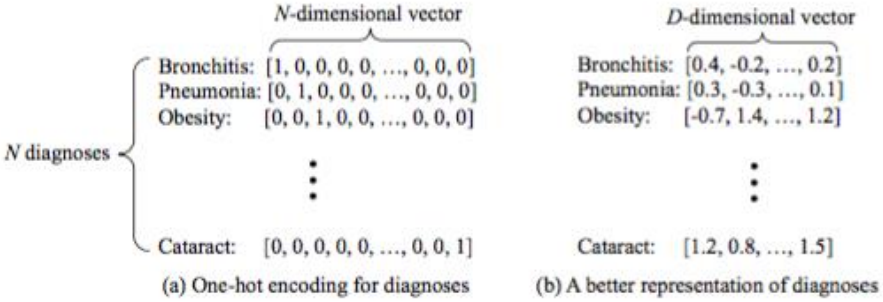


Figure 14. Example of representation learning for diagnosis. (Choi, Schuetz, Stewart, & Sun, 2016).

Specifically, this is achieved by embedding data into a vector space. Unlike standard categorical one-of-K encoded representations, medical-concept representation learning updates during model training. For

example, diagnosis could be represented as an indicator vector such that all entries except one contain the value zero (Choi, Schuetz, Stewart, & Sun, 2016). This representation clearly communicates which diagnosis a patient has, but it obscures the relationships between diagnoses. Indeed, the distance between any two distinct diagnoses represented by indicator vectors is equal. To the extent that downstream processing relies on distance for dimensionality reduction, such as principal component analysis (PCA) or clustering, distance between individual indicators may not be useful as an input. Representation learning is a method for capturing additional context between data to present to downstream analytics.

We have examined the impact of applying a skip-gram representation learning method adopted from Mikolov, Sutskever, Chen, Corrado, & Dean (2013) to add the context associated with temporal co-occurrence to the input data (Figure 15). We have encouraged further work in this area by considering other representation learning methods such as auto-encoder-based methods, whether it be a denoising autoencoder such as in Miotto, Li, Kidd, & Dudley (2016) (Figure 16) or a variational auto-encoder (Kingma & Wellington, 2013).

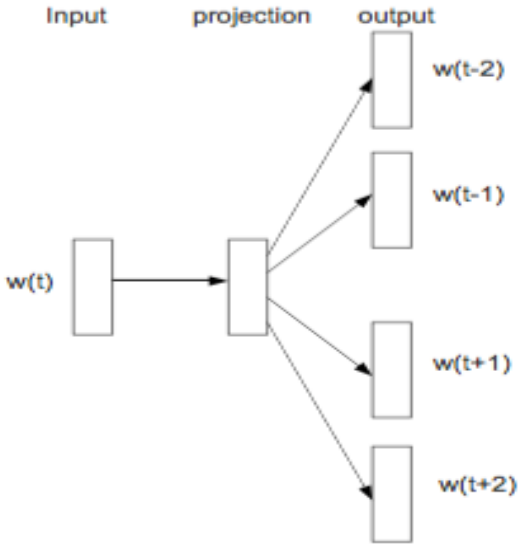


Figure 15. Skip-gram model objective is to learn word vector representations that are good at predicting nearby codes (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).

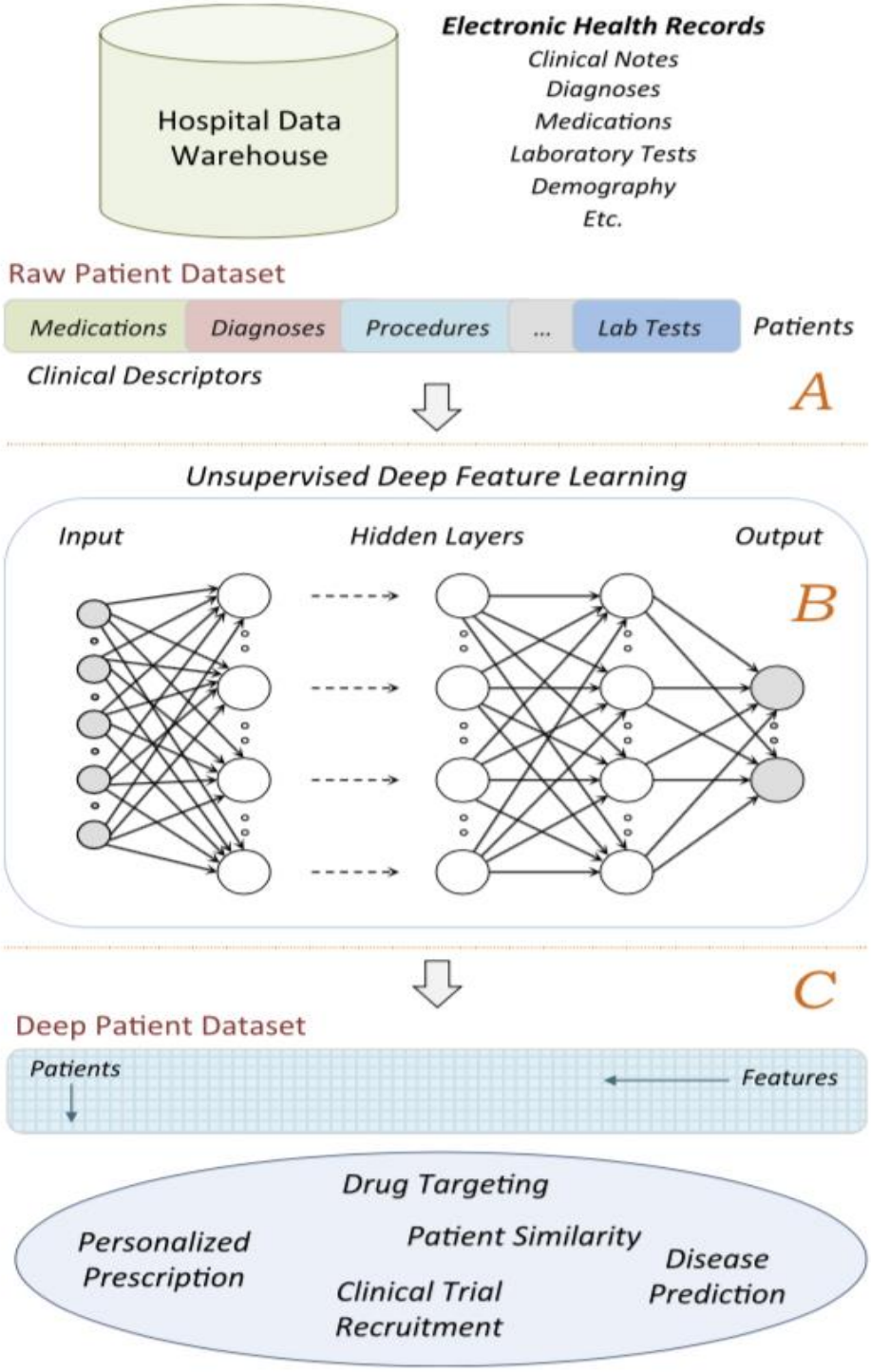


Figure 16. "Deep patient" uses auto-encoders to preform dimensionality reduction with some robustness in the representation(Miotto, Li, Kidd, & Dudley, 2016). This conceptual framework shows the following: A. preprocessing stage, B. modeling of raw representations, and C. application of deep features to database.

While medical-concept representation learning has clear benefits in existing literature, it is a data-driven method with drawbacks. There is an opportunity for unrepresentative data to highlight existing but medically spurious relationships. There are some best practices from the literature (Miotto, Li, Kidd, & Dudley, 2016; Choi, Schuetz, Stewart, & Sun, 2016; Zhu et al., 2016; Rajkomar et al., 2018) that attempt to address some possible pitfalls associated with these methods. These include

- using expert feedback to curate labels (Miotto, Li, Kidd, & Dudley, 2016; Zhu et al., 2016),
- limiting analysis to codes with not too few occurrences,
- limiting analysis to patients with not too few codes,
- using a visit as the unit of analysis, not a patient, and
- inspecting the intermediate results to sanity check the intermediate results.

There has been evidence from interpreting the results of representation learning to provide feedback on the quality of representation (Choi, Schuetz, Stewart, & Sun, 2016). A mitigation strategy employed by previous methods is to visualize or otherwise interpret the relationships learned in representation learning to provide evidence against spurious relationships (Figure 17). A common example of this approach has been chronicled in Mikolov, Yih, & Zweig (2013) with popular press *MIT Technology Review* referencing the arXiv work (MIT, 2015). We show below two examples of such inspection of intermediate results. Figure 17 gives four examples from the published literature of representation learning, clustering similar codes together by using co-occurrence relationships. Again, similar to Miotto, Li, Kidd, & Dudley (2016) and Choi, Schuetz, Stewart, & Sun (2016), we may visualize learned relationships between CPT codes learned by skip-gram. This collects the results of four queries. For a given code, each query asks what the most similar code and its similarity score to the input code is. The results (Table 2) demonstrate some expected relationships between hearing aid codes, EKG codes, medical discussion codes, and physical therapy codes. The fact that these relationships are anticipated provides a useful sanity check of the representation's general usefulness.

Table 2. Representation learning result on expanded CDW dataset.

code	sim score	code description
V5014		repair modification of hearing aid
92593	0.92	hearing aid check
93005		EKG
93010	0.84	EKG interpretation
98966		5-10 minute medical discussion by telephone
98967	0.72	11-20 minute medical discussion by telephone
97530		bending lifting reaching to improve functional performance
97116	0.86	gait training

Coordinate 112	Coordinate 152
Kidney replaced by transplant (V42.0) Hb-SS disease without crisis (282.61) Heart replaced by transplant (V42.1) RBC antibody screening (P) Complications of transplanted bone marrow (996.85) Sickle-cell disease (282.60) Liver replaced by transplant (V42.7) Hb-SS disease with crisis (282.62) Prograf PO (R) Complications of transplanted heart (996.83)	X-ray, knee (P) X-ray, thoracolumbar (P) Accidents in public building (E849.6) Activities involving gymnastics (E005.2) Struck by objects/persons in sports (E917.0) Encounter for removal of sutures (V58.32) Struck by object in sports (E917.5) Unspecified fracture of ankle (824.8) Accidents occurring in place for recreation and sport (E849.4) Activities involving basketball (E007.6)
Coordinate 184	**Coordinate 190**
Pain in joint, shoulder region (719.41) Pain in joint, lower leg (719.46) Pain in joint, ankle and foot (719.47) Pain in joint, multiple sites (719.49) Generalized convulsive epilepsy (345.10) Pain in joint, upper arm (719.42) Cerebral artery occlusion (434.91) MRI, brain (780.59) Other joint derangement (718.81) Fecal occult blood (790.6)	Down's syndrome (758.0) Congenital anomalies (759.89) Tuberous sclerosis (759.5) Anomalies of larynx, trachea, and bronchus (748.3) Autosomal deletions (758.39) Conditions due to anomaly of unspecified chromosome (758.9) Acquired hypothyroidism (244.9) Conditions due to chromosome anomalies (758.89) Anomalies of spleen (759.0) Conditions due to autosomal anomalies (758.5)

Figure 17. Example of visualization of the relationships learned from skip-gram (Choi, Schuetz, Stewart, & Sun, 2016).

5.3 EVALUATIONS

For our empirical evaluation, we compared methods from medical-concept learning to standard one-of-K coding to evaluate the change in effectiveness as done in Choi, Schuetz, Stewart, & Sun (2016). We performed two primary empirical evaluations. The first evaluation was on a curated collection of 60 K patients with no more than 1 year of medical history included. The second was a collection of patients with no restriction to the amount of medical history included. This section evaluates the impact of these methods to inform clustering at a patient level rather than a visit level.

We briefly discuss below some experimental design commonalities between the two evaluations. The first attempt to consider the sub-clustering of SIHD patients was redirected due to a lack of representativeness in the initial data cohort. In particular, the task of predicting individual SIHD diagnosis codes suffered from an extremely unbalanced data prediction problem as I20 is by far the most likely ICD-10 code associated with SIHD in the cohort. As such, we considered a more general problem of predicting the most likely primary diagnosis category. We consider the most frequently occurring of the top 10 primary diagnosis categories of each patient to be its label. The diagnosis categories were determined by grouping the ICD-10 using the multi-level CCS category Level I. The hypothesis across both evaluations is that representation learning is broadly useful independent of downstream processing models; thus, for each experiment, we trained three models—a logistic regression model, a two-layer neural network model, and a nearest-neighbor model—and then averaged the results. Each model was trained using (five) fold cross

validation. Specialized parameter settings can be found in the delivered code on our project's internal GitLab repository.[1]

5.3.1 Short History Evaluation

This evaluation provided evidence that medical-representation learning improves predictions of primary diagnosis category of a short patient history through 12 experiments. We report the accuracy of predicting the label using the following inputs in performing this task: the multiset of CPT codes, the multiset of Rx codes, and the multiset of both procedure and prescriptions (CPT + Rx) (Table 3).

Table 3. Accuracy of primary diagnosis category prediction using short patient histories.

Data	One-hot	Skip-gram	TF + skip-gram	PCA
CPT	35%	45%	52%	41%
Increase over baseline		28%	48%	17%
Rx	32%	33%	34%	34%
		0	0	0
CPT + Rx	39%	48%	48%	41%
		23%	23%	5%

The first column of Table 3 details the accuracy determined by predicting using only the multiset given by the row title. This is the baseline method without representation learning. The second column demonstrates that training a skip-gram representation learning model increases predictive accuracy by 28% for procedures. One challenge with using the multisets of codes is that variation introduced by different amounts of codes in their history. To explore this distinction, we combined a skip-gram model with a Term Frequency (TF) model. The TF model first represents the multiset of inputs as a probability distribution over the inputs; then the trained skip-gram model augments the probabilities using the interrelationships between categories. This combination method resulted in the best empirical result for a 48% improvement over baseline. For comparison purposes, as was done in both Miotto, Li, Kidd, & Dudley (2016) and Choi, Schuetz, Stewart, & Sun (2016), we also report the result of principal component analysis (PCA) as a representation learning method for completeness. The results of PCA were worse in than skip-gram in the procedure case.

5.3.2 Long History Evaluation

This evaluation provided evidence that medical representation learning fails to improve prediction of primary diagnosis category of an arbitrarily long patient history through 12 experiments. We report the accuracy of predicting the label using the following inputs in performing this task: the multiset of CPT codes, the multiset of Rx codes, and the multiset of both procedure and prescription (CPT + Rx) codes. In each of the 12 experiments, the results were identical. The clustering algorithms failed to predict beyond the baseline. Subsequent analysis demonstrated a difference in "label purity" between the two evaluations, particularly since a patient was given a singular label associated with the most occurring diagnosis category. The data demonstrated that the assignment from patient to category became less definitive with the addition of more history. This seems intuitively clear. Over short durations a patient's medical visits may focus on few categories, whereas over all time periods, they may be more spread out

[1] In addition, we evaluated a Support Vector Machine model as well but found its performance was similar to the two-layer neural network model; however, it took much longer to evaluate.

over the categories. Empirically, we witnessed that for over 20% of long history patients the probabilities between the top two occurring categories was less than 5%. Furthermore, 50% of patients had a difference of less than 15%. This provides evidence that clustering methods struggle not due to any specific deficiency of representation learning but due to the changing strength of the label assignment per patient.

5.4 LESSONS LEARNED AND NEXT STEPS

We summarize the lessons learned in the following items:

- The task of diagnosis clustering can be improved 40% by using representation learning.
- Standard methods from medical literature might struggle to find meaningful relationships with raw prescription fields.
- Representation learning might capture medically meaningful relationships in both prescription and procedure data.

We summarize the next steps in this research with the following three items.

1. Clustering by primary diagnosis category at scale over a long duration is challenging since a patient's primary diagnosis category might be mixed. The variability introduced at this stage can drown out subsequent analysis. As such, we hypothesize that we should focus on clear, unambiguous signals to inform the analysis such as detecting transitions in diagnosis, detecting worsening chronic conditions such as diabetes, or predicting hospitalizations. These will have clear binary indications of activity over shorter time horizons, both of which have improved performance.
2. We suggest demonstrating the extent to which the sub-cohort clustering of a subtype of SIHD behaves similarly to the current cohort clustering by primary diagnosis category. In particular, does performance degrade over long histories? We propose pursuing mitigation strategies for increasing performance over longer time horizons such as build hierarchies over short time periods, add demographic information, include more domain knowledge in data cleaning, and leverage sequential models to consider changing membership over time.
3. Proposed future work includes prototyping, testing, and evaluating existing "interpretable models" such as Variational Autoencoder and comparing them with "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," a paper by published in *Scientific Reports* by Miotto, Li, Kidd, and Dudley (2015). This is an evaluation task meant to assess which model might better suit the needs of the VA.

6. TECHNOLOGY TRENDS IN HEALTH CARE

In this section we discuss recent technology trends in health care as they relate to our AI and, more specifically, our ML research and approaches as described above.

6.1 FORECAST FOR HEALTH DATA

As noted by the medical and research communities, the volume of health care data is increasing at an unprecedented rate. In 2013, 153 exabytes of patient data were generated, and by 2020 that number is expected to grow to 2,314 exabytes, which equates to about 48% growth annually (Stanford Medicine, 2017). This growth could be partially attributed to proposals to shift data collection from human transcription, which is highly error prone, to systems of sensor-collected and cloud-stored data, allowing for continuous real-time data collection (Rolim et al., 2010), which is the current most promising technological trend. Datasets of this magnitude cannot be efficiently analyzed by conventional means; however, through advances in AI, ML, and advanced analytics, these data can provide improved

individualized patient care. Through the implementation of continuous monitoring and interpretation of patient-generated data, clinical pathways could be generated that are self-adaptive and able to enhance both the efficiency and quality of the physician-patient interaction (Alexandrou, Skitsas, & Mentzas, 2011) Future implementations of advanced analytical methods can be enhanced by the use of these techniques in cloud computing, lowering costs and increasing scalability.

Below, we further discuss trends related to the AI/ML methods used in our research as well as trends in some related areas in medicine which would benefit from AI, including genetic research, medical-image analysis, real-time clinical decision support, business intelligence for hospital administration, appointment scheduling, diagnoses, and population health management.

6.2 TRENDS IN HEALTH CARE USING AI

As demonstrated in our research, AI/ML methods can provide tailored, precise clinical pathways for very specific sub-cohorts. Foundational medical guidelines such as Fihn et al. (2012) offer guidelines for diseases such as SIHD but are not adapted to specialized patient cohorts. Using AI and ML techniques, we can assess the efficacy of treatment protocol guidelines as they are applied to sub-cohorts of patients with divergent clinical characteristics. Through clinical pathway inference, we can adapt clinical pathways to reflect a more cohort-specific pathway, and using medical-concept representation learning, we can refine clinical pathway guidelines, thereby providing improved specificity and accuracy in clinical and population outcomes. For individual patients, using AI/ML, we can analyze a cohort member's EHR to better target treatment protocols that more precisely address the patient's therapeutic needs.

A well-written clinical pathway, which is flexible enough to accommodate individual patients and needs, can lead to real-time clinical decision support when paired with an AI-driven user interface. A clinical pathway can recognize a series of events that it has seen before and offer real-time suggestions on the next potential step according to a statistical analysis of the most probable path to lead to a successful outcome for the patient. An AI-driven user interface could adapt the clinical pathway to lead to individualized patient care while offering guidance that enables physicians to be more effective with their time (Alexandrou, Skitsas, & Mentzas, 2011).

As discussed in our research above, one method of modeling clinical pathways is pathway inference, which uses LDA, an RBM, and word embedding. In terms of topic modeling, LDA is a methodology for determining what a document is saying based on the frequency of terms it uses, while an RBM is a neural network system with two layers where no node on the same layer is connected, as will be demonstrated in ORNL's future work. The idea is that the documentation from previous patients is analyzed, patterns are inferred, and those patterns are used to implement new clinical pathways. These patterns allow patient cohorts to be further broken down into sub-cohorts, which allows for more refined clinical pathways to be applied. Additionally, the further refinement of patient groups allows for any aberrant treatments to be easily identified and rectified. Even among patients with the same illness and treatment, such as colorectal cancer, there are significant variations in outcomes. These "unwarranted variations" result from genetic background, tumor micro-environment, and response to treatment. Aside from biological contributions, other factors, such as socioeconomic status and geographical location, factor into clinical outcome. By further stratifying patient cohorts beyond traditional methods, better risk profiles can be adapted, leading to better comparisons of facility efficacy as well (Menon, Cunningham, & Kerr, 2016).

The medical-concept representation learning method is also used in our research. The goal of this method is to go from a clinical pathway that was written for a broad group of patients and adapt it to provide individualized, precision patient care. By learning the concepts of medical information, this method provides flexibility by obtaining a pathway and allowing multiple variations of similar data to pass through the system. This tailored clinical pathway is accomplished by interpreting free-text entries in

EHRs as well as different naming methods for drugs, diseases, and labs (Choi, Schuetz, Stewart, & Sun, 2016). For example, other methods of interpreting EHRs would view ischemic heart disease, coronary artery disease, and coronary heart disease as separate conditions even though these terms are often used interchangeably. The same logic follows for using the name brand for drugs vs. the generic name. This technique allows modeling systems to be more adaptive to different regions and populations. This methodology should provide the most granularity and thus the most individualized treatment path leading to the highest quality of care.

Through the ORNL advanced analytics architecture, which is also supporting the Million Veterans Program, we are given the opportunity to use AI/ML methods for genomics science. Particularly advantageous uses of ML techniques in genomics are to recognize patterns that can be used to annotate genes by mapping untranslated regions, introns, and exons along entire chromosomes. In addition to annotating chromosomes, ML can be used to distinguish various disease phenotypes in DNA microarrays. (Libbrecht & Noble, 2015) Both of these uses dramatically reduce the amount of time it takes for a researcher to accomplish the given task, thereby increasing efficiency. In DNA, where the vast majority of data is unexpressed, predictive algorithms, which determine the likelihood of any given sequence being expressed, aid researchers greatly and will lead to a dramatic increase in the output of work seen in this field.

Medical image analysis, aided by AI, is already helping radiologists analyze two-dimensional medical images such as radiographs and ultrasounds, and three-dimensional convolutional neural networks designed to help with MRIs are currently under research (Tang et al., 2018). As AI-assisted radiological diagnosis continues progressing, this technology is predicted to help radiologists by triaging and providing preliminary diagnoses to radiologists, which will greatly speed up their workflow (Tang et al., 2018).

The use of AI in clinical administration and health policy has the potential to improve patient care through ensuring effective medication strategies and reducing costs. US hospitals have experienced between 174 and 320 drug shortages on the last day of each quarter since the first quarter of 2013 (ASHP, 2018), which are strongly associated with a decrease in the number of suppliers, failure to comply with manufacturing standards, and a number of drugs having sales of generic versions (GAO, 2016). These massive shortages leave patients either waiting for a drug that could improve their quality of life or possibly paying an inflated rate for an alternative. One way to mitigate this situation is to use a database like AHFS to find alternative drugs with similar pharmacological properties. AI can be applied during the billing process to cross reference codes of drugs with known shortages to a list of known alternatives and offer suggestions.

Another potential use of AI is the optimization of hospital appointment scheduling. Hospitals often face times with surges of patient admittances. During these times, patients face extended wait times and physicians face extended work hours. AI could improve the efficiency by identifying bottlenecks in the patient pathways and optimize routes of treatment for peak efficiency. Some procedures take minimal time to perform but a substantial amount of time to get to. This efficiency could be improved by use of AI patient routing.

AI methods have also been developed to improve interpretation of waveforms that can be gathered by a simple Holter monitor. ST-segment deviation can be analyzed as quickly as a visual inspection with accuracy that measures exact amplitudes that cannot be detected by eye (Myers, Scirica, & Stultz, 2017). This indicator is extremely helpful in determining a patient's risk factors, when coupled with demographics (Kaul et al., 2001). Additionally, AI methods can be used to perform large-scale population health management such as epidemiology simulations of the spread of infectious disease and

demographics-driven health care analysis. These studies can improve the quality of research from public health groups and recognize issues that were unnoticed before.

6.3 HEALTH CARE'S FUTURE ANALYTIC NEEDS

To achieve all the benefits of real-time, clinical AI/ML approaches to patient cohorts and individual clinical analytics, the computing infrastructure needs to support very large-scale data storage and highly scalable, intensive computing platforms. The main conceptual solution to these needs is the use of large-scale cloud computing and shared access to these resources in an extensible way. As cloud systems are typically outsourced, there is also increased compatibility with existing infrastructure, which lowers the overall cost and improves accessibility for smaller health care systems. These systems are also highly scalable for increased use with future growth. In edge computing, data are collected and analyzed in a geographically local vicinity. This system is much more expensive to set up as the health care facility has to buy the computational power to suit their needs as well as upkeep; however, the benefit is a lower latency period (on the order of milliseconds) as the data are processed closer to its origin. In summation, the future of health care technology is found in large-scale cloud computing, using AI methods on big health data, which can meet health care's analytic needs for clinical decision support and business intelligence.

7. ACKNOWLEDGEMENTS

The authors are most grateful to the staff of the US Department of Veterans Affairs for their support of Oak Ridge National Laboratory as part of the FY 17–18 ORNL VICTOR Part A, Part B: Health Information Technology-Advanced Analytics project. Special thanks are extended to Dr. Jonathan Nebeker, Dr. Tamara Box, and Dr. Merry Ward and their teams of subject matter experts at the VA for their technical expertise and helpful discussions they provided in support of this effort.

8. REFERENCES

Alexandrou, D. A., Skitsas, I. E., & Mentzas, G. N. (2011). A holistic environment for the design and execution of self-adaptive clinical pathways. *IEEE Transactions on Information Technology in Biomedicine, 15*(1), 108–118. Retrieved from http://imu.ntua.gr/sites/default/files/biblio/Papers/a-holistic-environment-for-the-design-and-execution-of-self-adaptive-clinical-pathways.pdf

American Society of Health-System Pharmacists (ASHP). (2018, March 31). "Drug Shortages Statistics." Retrieved from https://www.ashp.org/Drug-Shortages/Shortage-Resources/Drug-Shortages-Statistics

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Choi, E., Bahadori, M. T., Searles, E., Coffey, C., Thompson, M., Bost, J., ..., & Sun, J. (2016). Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1495–1504). ACM.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*.

Fihn, S. D., Gardin, J. M., Abrams, J., Berra, K., Blankenship, J. C., Douglas, P. S., ..., & Kligfield, P. D. (2012). 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *Journal of the American College of Cardiology*, *60*(24), e44–e164.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (Vol. 1). Cambridge: MIT press.

Huang, Z., Dong, W., Ji, L., Gan, C., Lu, X., & Duan, H. (2014). Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, *47*, 39–57.

Huang, Z., Lu, X., & Duan, H. (2013). Latent treatment pattern discovery for clinical processes. *Journal of Medical Systems*, *37*(2), 9915.

Huang, Z., Lu, X., & Duan, H. (2013). Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, *46*, 111–127.

Kaul, P., Fu, Y., Chang, W. C., Harrington, R. A., Wagner, G. S., Goodman, S. G., ..., & Topol, E. J. (2001). Prognostic value of ST segment depression in acute coronary syndromes: insights from PARAGON-A applied to GUSTO-IIb. *Journal of the American College of Cardiology*, *38*(1), 64–71. Retrieved from http://www.onlinejacc.org/content/accj/38/1/64.full.pdf

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv*:1312.6114.

Libbrecht, M. W. & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics, 16*(6), 321. Retrieved from https://www.nature.com/articles/nrg3920

Menon, M., Cunningham, C., & Kerr, D. (2016). Addressing unwarranted variations in colorectal cancer outcomes: A conceptual approach. *Nature Reviews Clinical Oncology*, *13*(11), 706.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751).

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*, 26094.

MIT. (2015, September). King - Man + Woman = Queen: The Marvelous Mathematics of Computational Linguistics. A View from Emerging Technology from the arXiv. *MIT Technology Review*.

Myers, P. D., Scirica, B. M., & Stultz, C. M. (2017). Machine learning improves risk stratification after acute coronary syndrome. *Scientific Reports. 7*(1). Retrieved from https://www.nature.com/articles/s41598-017-12951-x.pdf

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ..., & Sundberg, P. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, *1*(1), 18.

Rolim, C. O., Koch, F. L., Westphall, C. B., Werner, J., Fracalossi, A., & Salvador, G. S. (2010, February). A Cloud Computing Solution for Patient's Data Collection in Health Care Institutions. In *eHealth, Telemedicine, and Social Medicine, 2010. ETELEMED'10. Second International Conference on* (pp. 95–99). Retrieved from https://ieeexplore.ieee.org/document/5432853/#full-text-section

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., & Lai, A. M. (2013). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, *21*(2), 221–230.

Stanford Medicine. (2017, June). Harnessing the power of data in health. *Stanford Medicine 2017 Health Trend Report.* Retrieved from: https://med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf

Tang, A., Tam, R., Cadrin-Chenevert, A., Guest, W., Chong, J., Barfett, J., ..., & Poudrette, M. G. (2018). Canadian Association of Radiologists white paper on artificial intelligence in radiology. *Canadian Association of Radiologists Journal.*

U.S. Government Accountability Office (GAO). (2016). "Drug shortages: certain factors are strongly associated with this persistent challenge." Retrieved from https://www.gao.gov/assets/680/678281.pdf

Verberne, L. D., Leemrijse, C. J., Swinkels, I. C., van Dijk, C. E., de Bakker, D. H., & Nielen, M. M. (2017). Overweight in patients with chronic obstructive pulmonary disease needs more attention: a cross-sectional study in general practice. *NPJ Primary Care Respiratory Medicine*, *27*(1), 63.

Zhu, Z., Yin, C., Qian, B., Cheng, Y., Wei, J., & Wang, F. (2016). Measuring patient similarities via a deep architecture with medical concept embedding. In *Data Mining* (ICDM), *2016 IEEE 16th International Conference* on (pp. 749–758). IEEE.

This page is intentionally left blank.

APPENDIX A. OCCURENCES OF COMPONENTS IN PATHWAY COMPONENTS

Appendix A, Figures A-1-(a) through A-1-(j) show expected occurrences of components in each pathway pattern computed using posterior probabilities (Gibbs sample).

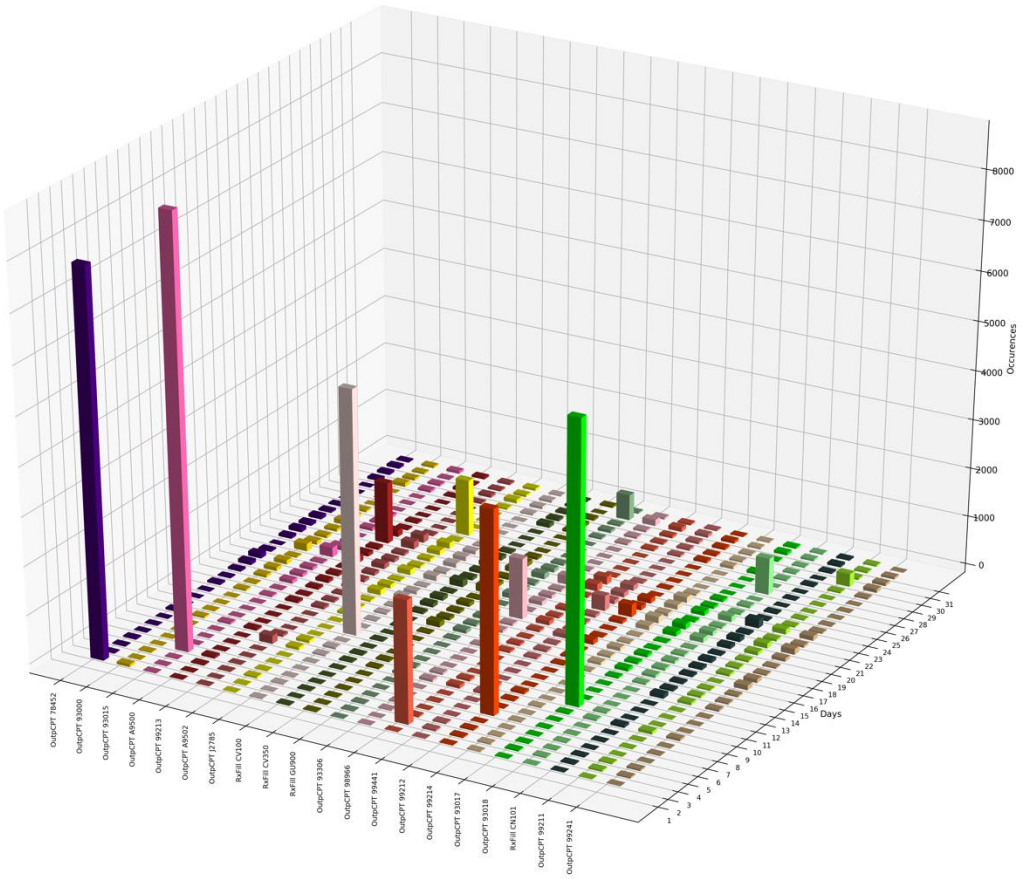


Figure A-1(a). The expected occurrences of components in pathway component 1. The expected amount is computed by Gibbs sample.

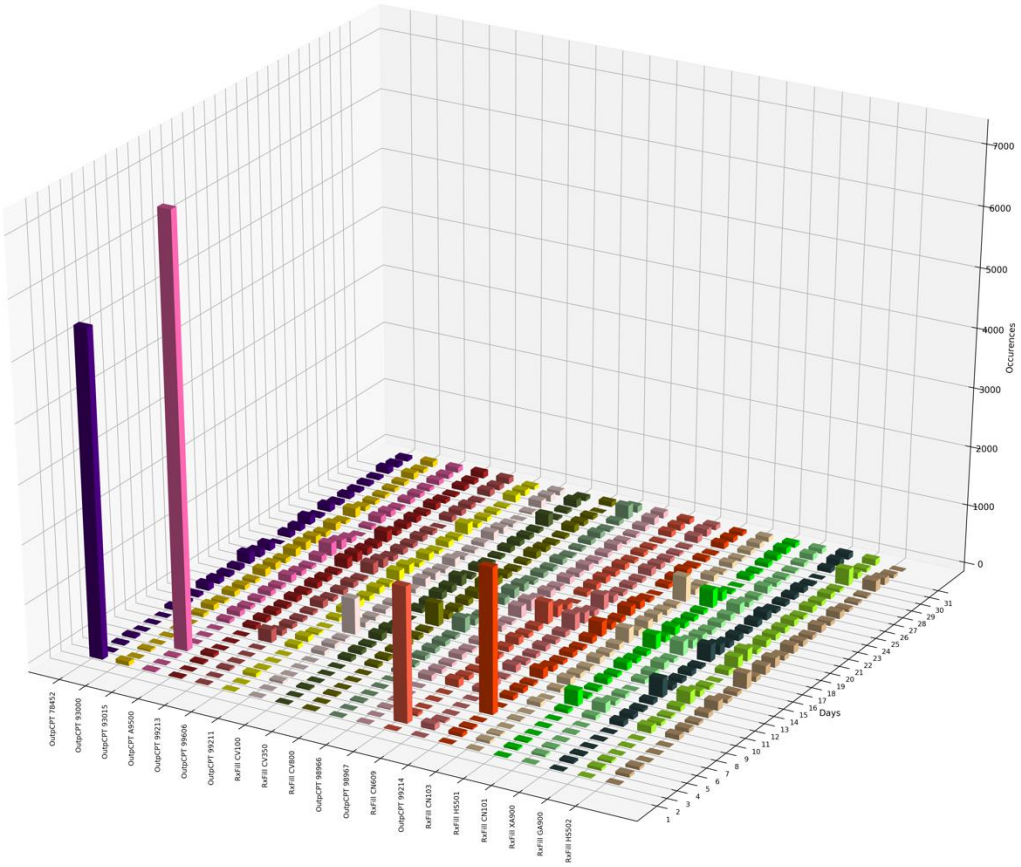


Figure A-1(b). The expected occurrences of components in pathway component 2. The expected amount is computed by Gibbs sample.

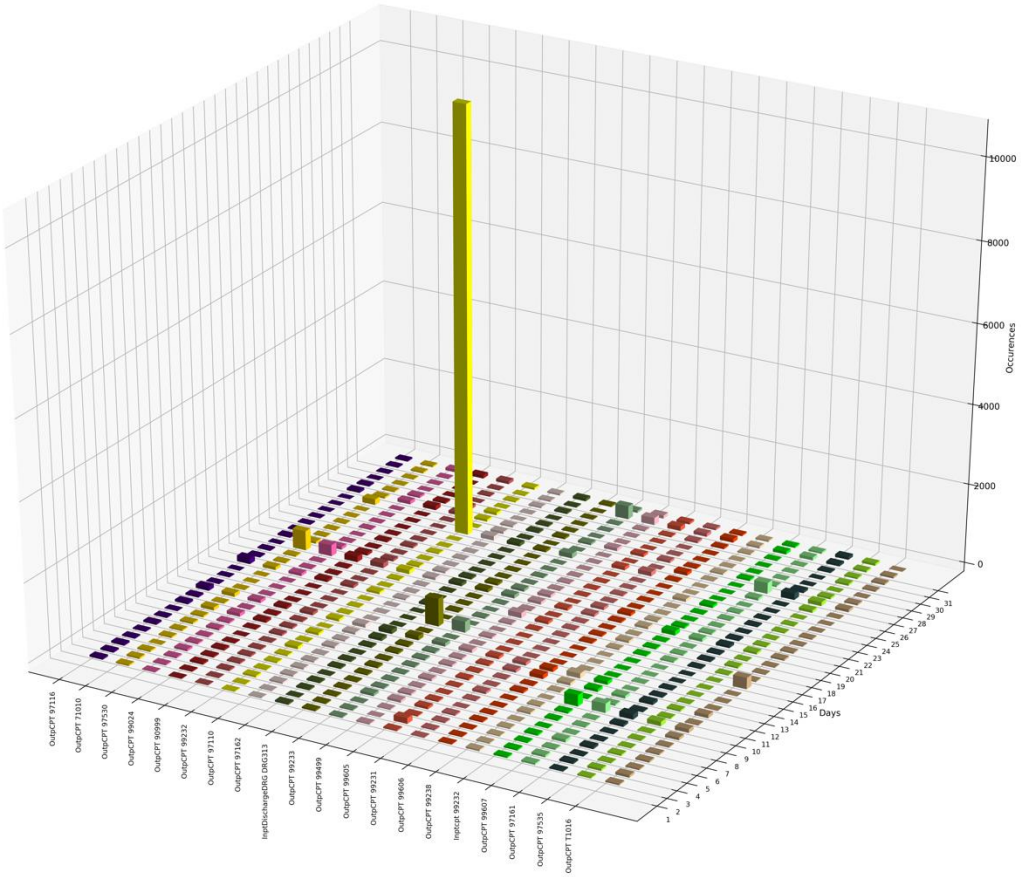


Figure A-1(c). The expected occurrences of components in pathway component 3. The expected amount is computed by Gibbs sample.

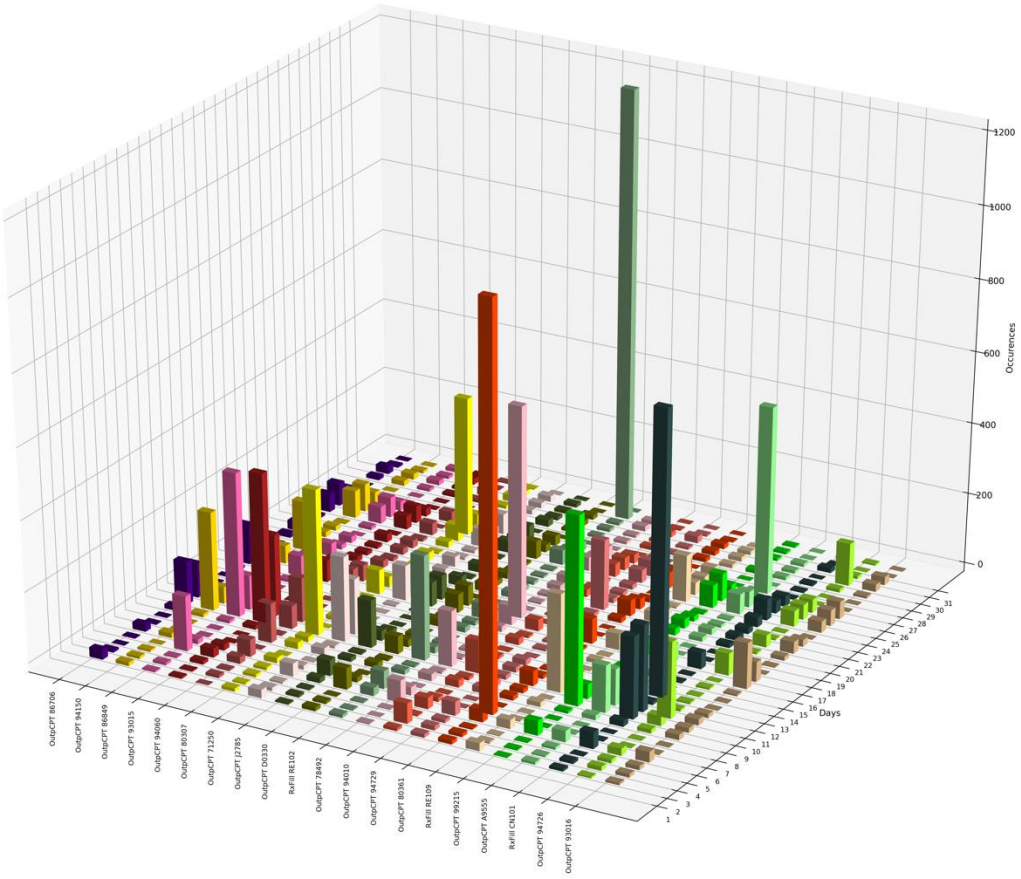


Figure A-1(d). The expected occurrences of components in pathway component 4. The expected amount is computed by Gibbs sample.

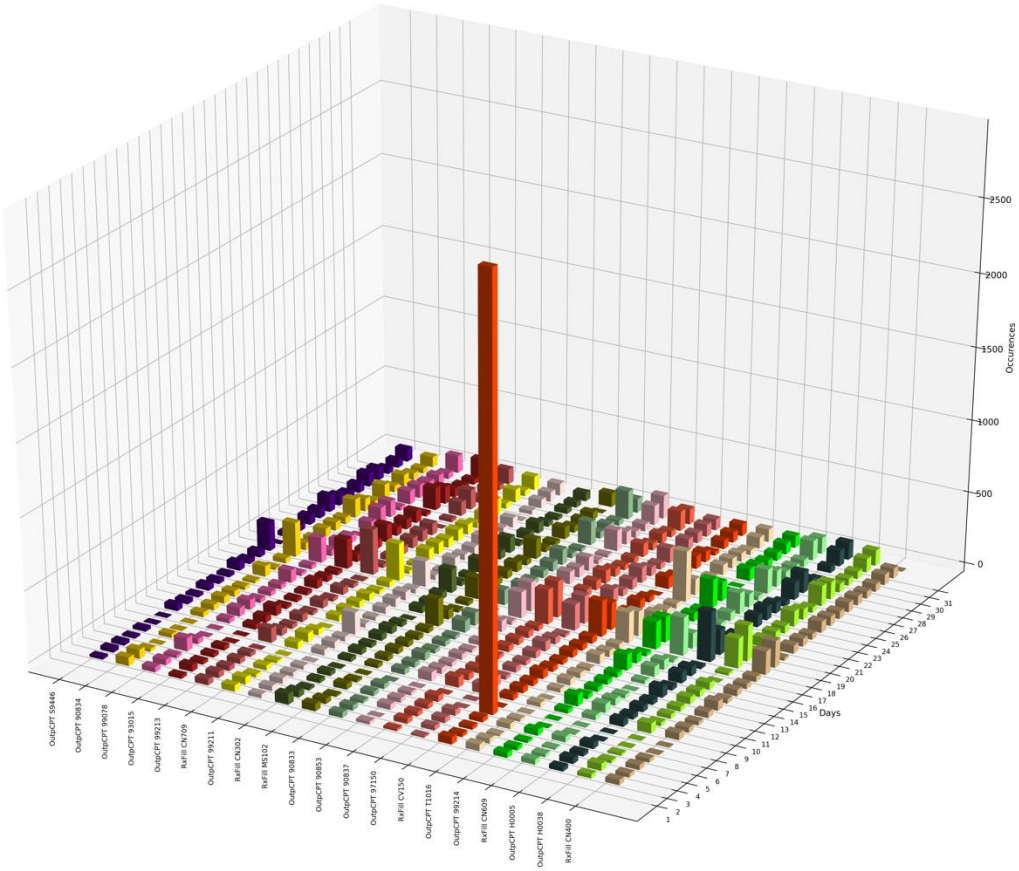


Figure A-1(e). The expected occurrences of components in pathway component 5. The expected amount is computed by Gibbs sample.

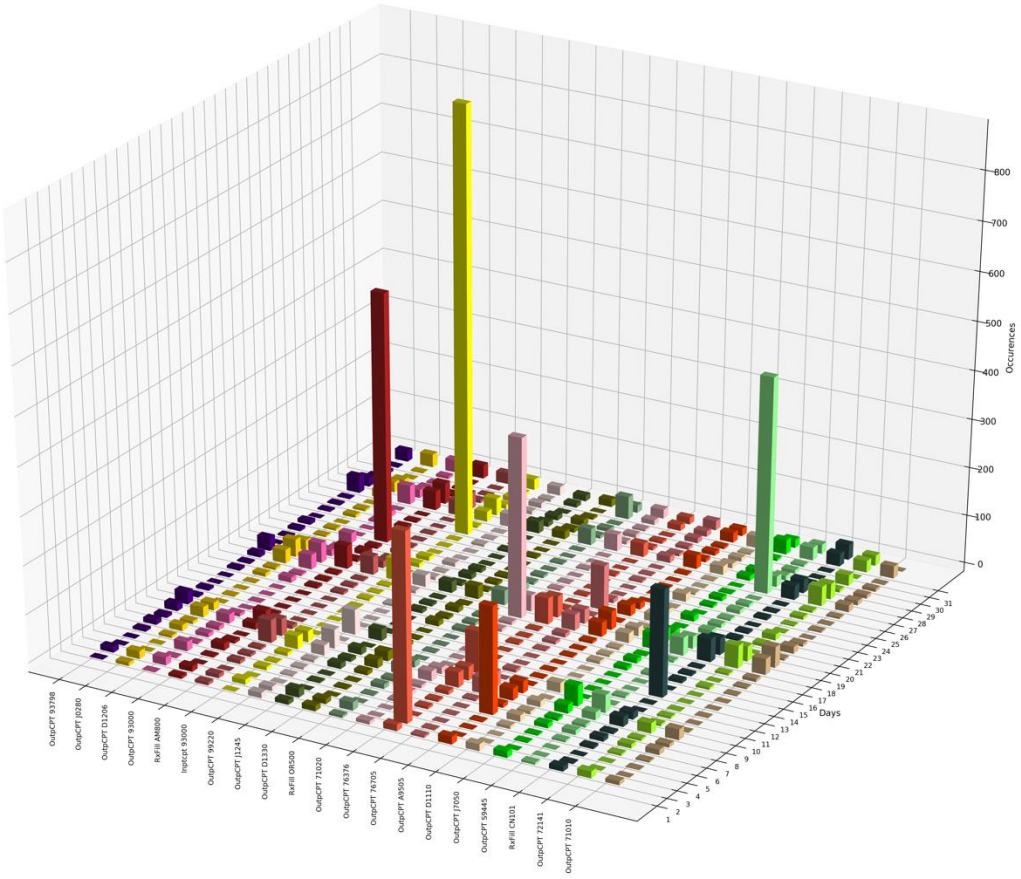


Figure A-1(f). The expected occurrences of components in pathway component 6. The expected amount is computed by Gibbs sample.

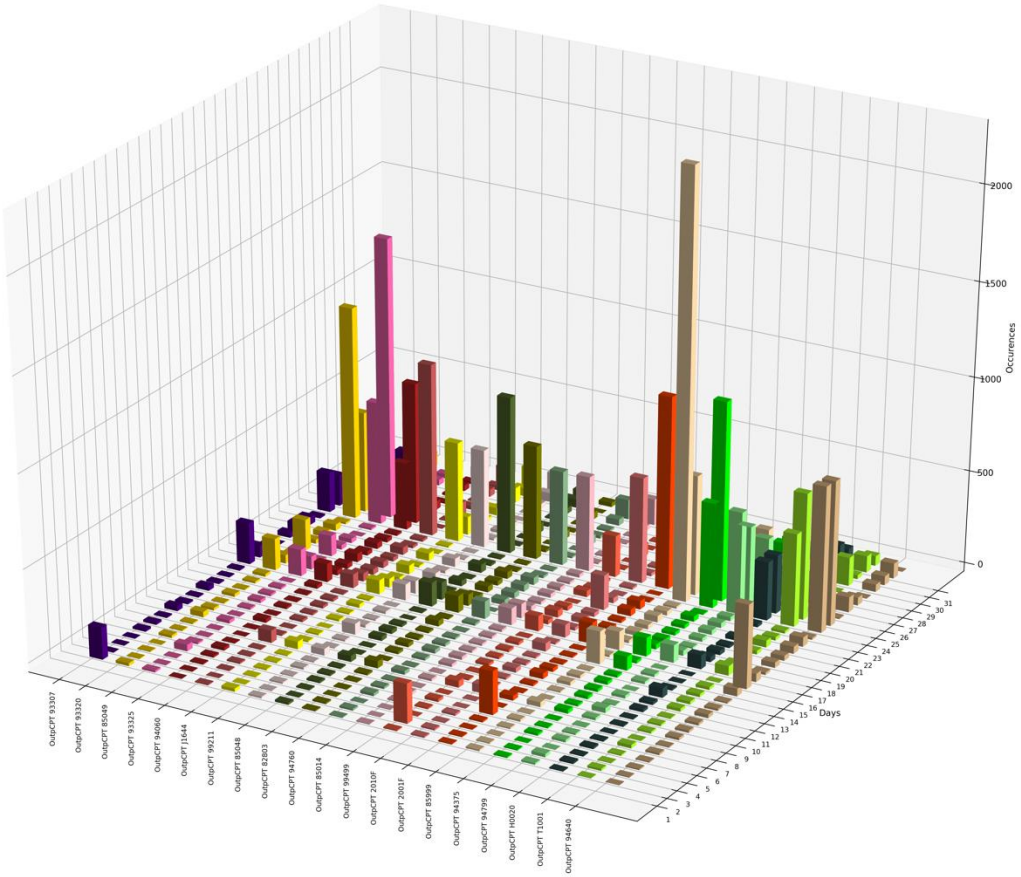


Figure A-1(g). The expected occurrences of components in pathway component 7. The expected amount is computed by Gibbs sample.

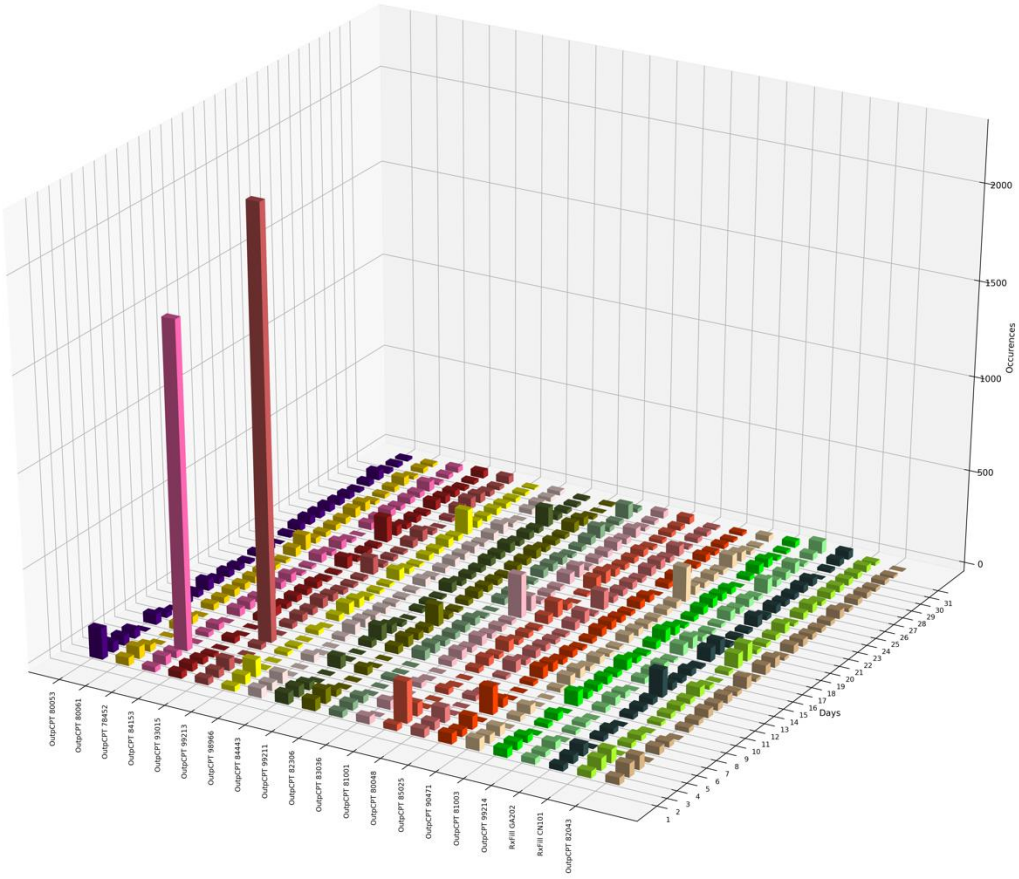


Figure A-1(h). The expected occurrences of components in pathway component 8. The expected amount is computed by Gibbs sample.

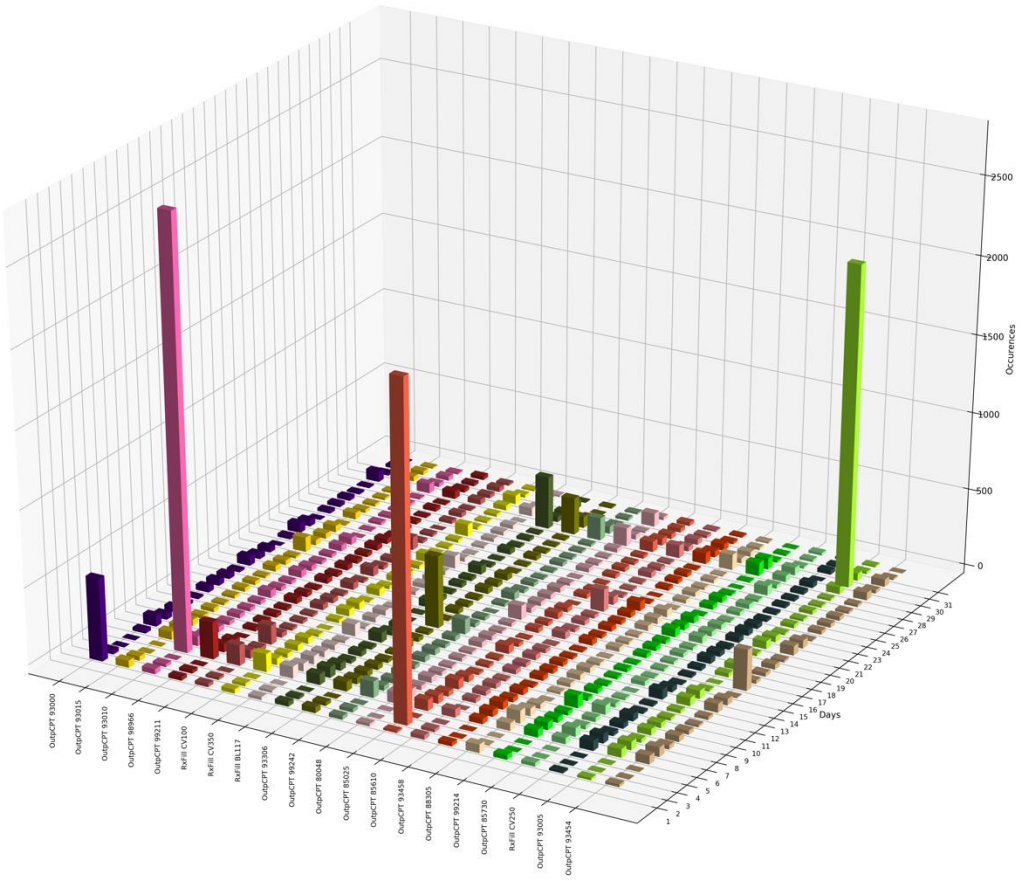


Figure A-1(i). The expected occurrences of components in pathway component 9. The expected amount is computed by Gibbs sample.

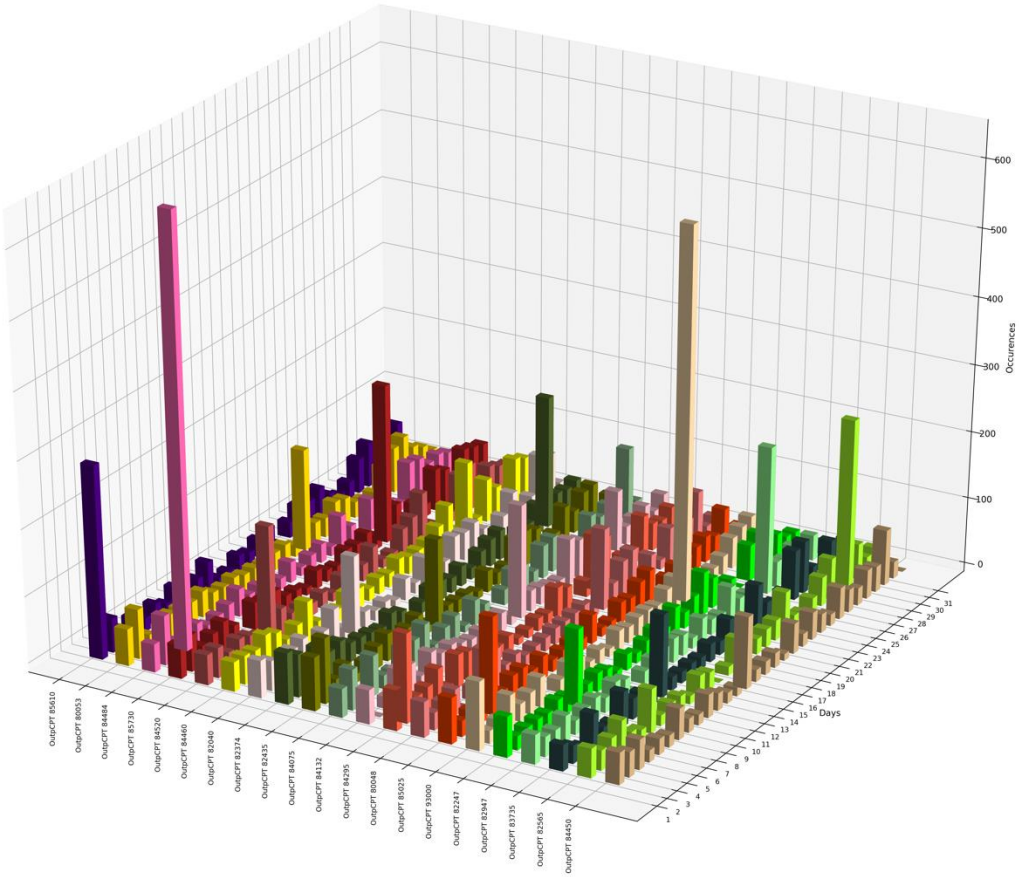


Figure A-1(j). The expected occurrences of components in pathway component 10. The expected amount is computed by Gibbs sample.