Identifying Symptom Clusters in Breast Cancer and Colorectal Cancer Patients using EHR Data

Priyanka Gandhi
School of Science, IUPUI
Indianapolis, Indiana
prgandh@iu.edu

Xiao Luo
School of Engineering and Technology, IUPUI
Indianapolis, Indiana
luo25@iupui.edu

Susan Storey
Indiana University School of Nursing
Indianapolis, Indiana
sustorey@iu.edu

Zuoyi Zhang
Indiana University School of Medicine
Indianapolis, Indiana
zyizhang@indiana.edu

Zhi Han
Indiana University School of Medicine
Indianapolis, Indiana
zhihan@iu.edu

Kun Huang
Indiana University School of Medicine
Indianapolis, Indiana
kunhuang@iu.edu

ABSTRACT

Patients with chronic conditions such as breast cancer and colorectal cancer often present with different symptoms, such 'fatigue', 'pain' and 'depression'. These symptoms add to patients' distress and functional impairment if left untreated. In this research, we investigate a symptom clustering and association mining framework to firstly extract and cluster the symptoms from the Electronic Health Record (EHR) clinical reports, then secondly to analyze the associations between symptom clusters and clinical attributes. The universal sentence coder and a modified seed based k-means algorithm are used for symptom coding and clustering. The results show that the symptom clusters have different associations between breast cancer and colorectal cancer, as well as for different time frames after chemotherapy. The results also show that breast cancer patients have slightly more symptoms from these three symptom clusters compared to the colorectal cancer patients within 12 months after the chemotherapy. Whereas, the colorectal cancer patient cohort has slightly more depression on average between 48 months and 54 months after the chemotherapy. Through applying the association rule mining, we find some informative rules, such as 'if a patient is at a higher cancer stage of colorectal cancer (3B), but no fatigue symptom, he or she likely doesn't have depression and peripheral neuropathy'. Our methods can be generalized to analyze symptom clusters of other chronic diseases where symptom management is critical.

CCS CONCEPTS

• **Information systems → Data analytics**; • **Applied computing → Health informatics**.

KEYWORDS

Medical Term Embedding, Symptom Clustering, EHR data, Breast Cancer, Colorectal Cancer, Association Mining

ACM Reference Format:
Priyanka Gandhi, Xiao Luo, Susan Storey, Zuoyi Zhang, Zhi Han, and Kun Huang. 2019. Identifying Symptom Clusters in Breast Cancer and Colorectal Cancer Patients using EHR Data. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3307339.3342164

1 INTRODUCTION

Patients with chronic conditions often present with different symptoms, such as pain, fatigue, depression. These symptoms add to patient' distress and functional impairment if left untreated. Some symptoms co-occur, which are described as symptom clusters in the literature [3]. The symptom clusters could be gastrointestinal symptom cluster including nausea, vomiting, lack of appetite, or the psychoneurological symptom cluster including depressive symptoms, anxiety, or other types. In this research, we make use of data stored in the Electronic Health Record (EHR) systems to analyze the symptom clusters and their association with the clinical and demographic attributes. EHR systems have been adopted by many countries in the past years [13]. In the United States, there are three stages of meaningful use of EHRs. Stage III is to improve population health outcomes, improve clinical outcomes, and gain more robust research data on health systems. In EHR, data are stored in different functional modules. Structured data is often stored in medication and diagnosis modules of the EHR and ready to be directly used by data mining applications. The encounter notes and clinical reports are unstructured data. These unstructured data are text written by clinicians, and often contain descriptions of the patients' symptoms. The unstructured data usually need to be pre-processed, and the information needs to be extracted for analysis. Previous research [12] [19] [24] demonstrates the need for extracting clinical signs and symptoms from patient medical records and further analyzing the associations between them and other attributes, such as diseases.

Cancer patients commonly experience symptoms such as pain, depression, and fatigue as a consequence of undergoing chemotherapy treatment, and these symptoms may persist, or develop, even

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342164

after the chemotherapy ends. The literature shows individual differences has associations with the symptoms and patients experience [15] [16]. Most of the current research is done by analyzing symptoms in isolation [14], Few research focuses on multiple symptoms and symptom clusters [3]. In contrast, most of the research gathers the patient-report symptoms through patient surveys [6][8] [11], instead of using EHR data. In this research, we make use of EHR data and focus on two types of cancer: breast cancer and colorectal cancer. We mainly analyze three symptom clusters and their associations with other clinical attributes, such as the pathologic stage of cancer, number of comorbidities, diabetes, etc.

To extract the symptoms from the clinical notes and reports, first, UMLS MetaMap is used to identify the phrases and terms from semantic categories 'Sign or Symptom' or 'Mental and Behavioral Dysfunction'. The clinician defines some seed terms for the three symptom clusters – 'Fatigue', 'Depression' and 'Peripheral Neuropathy'. Then, we develop a modified k-means algorithm to identify all other related symptoms from EHR data. Patients are represented as vectors of document frequency of the symptoms for clustering. The patient clusters are formed based on the severity and the combination of the symptoms. Conditional entropy based on different clinical attributes is used to interpret the clusters and investigate whether the cluster distribution has any association with some of the clinical attributes. Finally, association rule mining is used to identify the informative association rule between the symptom clusters and the clinical attributes. The results show that patients with certain attributes have less chance to develop symptoms in the symptom clusters, such as depression. Our methods can be generalized to analyze symptom clusters of other diseases.

The rest of paper is organized as follows, section 2 provides related work in the literature, section 3 details the methodologies for symptom cluster generation, patient clustering, and association rule mining. Results are given in section 4. Section 5 concludes this research.

2 RELATED WORK

Natural Language Processing (NLP) and machine learning techniques have been used to extract information from the clinical reports and notes to create predictive models for clinical decision support. Most of the previous research relies on existing ontology such as MeSH or WordNet, to identify the associations between different medical terms, such as various diseases [9] [22] [23]. UMLS MetaMap which is a natural language processing tool that uses various sources such as MeSH ontology to identify terms. The limitation of ontology is that it does not include the different representations of the same term or concept. Different descriptions for the same symptom happen quite often in the clinical notes within the EHR, because physicians have their preferences of recording notes. For example, 'fatigue' might be described as 'no energy' etc.

In recent years, the distributed representation of words which is called embedding gained interest in the research areas of text mining, natural language processing, and health informatics [17] [18] [20]. Word-based embedding has been studied for biomedical text classification and clustering [20] [25], where a word is a basic unit for the text documents and the word embedding is learned through neural networks. However, in the biomedical domain, clinical or medical concepts often contain more than one word. Many symptoms are often described as more than one word. The most recent text embedding, such as universal sentence encoder [5] analyzes the co-occurrences of the text segments based on the content of a given text document collection. It is not limited to generate word embedding, but capable of generating phrase or sentence embeddings. Research [5] shows that it is better than the word embedding, such as Word2Vec [17] at transfer learning. To the best of the authors' knowledge, it has not been investigated for clinical symptom analysis.

Research has been done on investigating symptom clusters in cancer and other diseases, although most of them use survey data. Cheung et al. [7] applied Principle Component Analysis (PCA) for the 1366 cancer patient cohort to determine the inter-relationships of the nine symptoms. They identified two major symptom clusters from their study cohort. One included fatigue, drowsiness, nausea, decreased appetite, and dyspnea; the other included anxiety and depression. Marshall et al. [10] investigated symptom clusters in women with breast cancer using social media data. The k-medoid clustering method was used to cluster the symptoms. The similarity measure was developed based on the frequency of the co-occurrences of the symptoms. Although symptom and patient clustering can be done in a unified framework proposed by Wang et al. [21], semantic similarity between the symptoms can not be well captured through the framework. In summary, no semantic analysis or NLP algorithm was involved in the research mentioned above, and most of the studies were based on surveys.

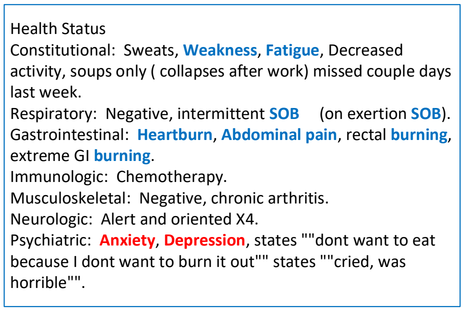
The research presented in this paper is different from previous research as it includes: (1) uses EHR clinical notes for symptom analysis; (2) investigates universal sentence coder and a modified seeds based k-means algorithm for symptom and patient clustering; (3) explores the symptom clusters and their association with clinical attributes in the EHR.

3 STUDY COHORT

The study cohort consists of patients with a primary diagnosis of breast cancer (BC) or colorectal cancer (CRC) who have electronic medical records in a large academic medical center's EHR system. BC and CRC patients are identified using the International Classification of Diseases (ICD). The ICD codes for BC are 174 (ICD-9) and C50 (ICD-10); the ICD codes for CRC are 153-154 (ICD-9) and C18-C20 (ICD-10). Through these ICD codes, we identify 4567 cases of BC and 2166 cases of CRC having received chemotherapy within the ten years of 2007-2017. For each case, we extract clinical notes, demographics data including gender, race and smoking status, cancer characteristic data including the pathologic stage of cancer, age at diagnosis, and the number of comorbidities.

4 METHODOLOGIES

In this study, we make use of both structured data and unstructured data of the EHR for analysis. The patient self-report symptoms are extracted from the unstructured clinical notes or reports, the other clinical attributes are extracted from the structured data. The following sections provide details on the methods used in this study.



Health Status
Constitutional: Sweats, **Weakness**, **Fatigue**, Decreased activity, soups only (collapses after work) missed couple days last week.
Respiratory: Negative, intermittent **SOB** (on exertion **SOB**).
Gastrointestinal: **Heartburn**, **Abdominal pain**, rectal **burning**, extreme GI **burning**.
Immunologic: Chemotherapy.
Musculoskeletal: Negative, chronic arthritis.
Neurologic: Alert and oriented X4.
Psychiatric: **Anxiety**, **Depression**, states ""dont want to eat because I dont want to burn it out"" states ""cried, was horrible"".

Figure 1: Symptom Extraction using UMLS MetaMap

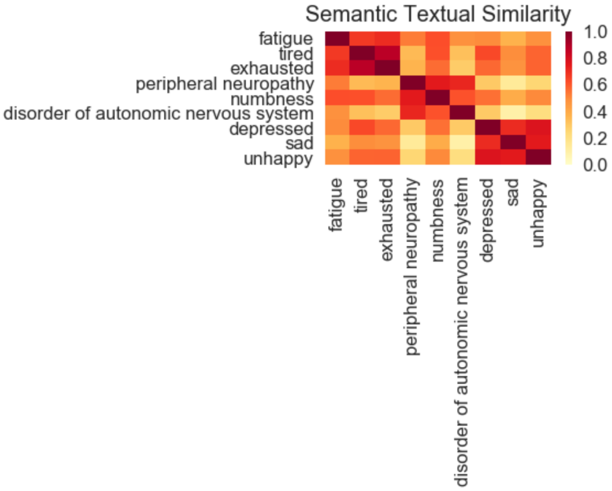


Figure 2: Symptom similarity scores using embeddings from the USE

4.1 NLP for Symptom Extraction and Clustering

4.1.1 ***Symptom Extraction.*** To extract symptoms, we utilize UMLS MetaMap [1] which is a natural language processing tool that uses various sources to categorize the phrases or terms in the text to different semantic types. In this research, we focus on three groups of symptoms – ‘fatigue’, ‘depression’ and ‘peripheral neuropathy’. So, the terms that are mapped to ‘Sign or Symptom’ or ‘Mental and Behavioral Dysfunction’ are extracted. Figure 1 provides an example of clinical notes and some of the terms mapped into the two semantic types using UMLS MetaMap. In this example, ‘weakness’ and ‘fatigue’, ‘heartburn’, ‘SOB’, ‘burning’ and ‘abdominal pain’ are mapped as ‘Sign or Symptom’, and ‘anxiety’ and ‘depression’ are mapped as ‘Mental and Behavior Dysfunction’. They are extracted as symptoms to be further analyzed in this study. The negation detection functionality is turned on for symptom extraction. Therefore, negating symptoms are not included.

4.1.2 ***Embedding Generation .*** Since the focus of this research is on three symptoms groups, some extracted symptoms, such ‘nausea’ etc., are not considered. Our clinician defines some seed symptoms to represent the symptom clusters. Table 1 lists all the seed symptoms for each symptom cluster. These seed symptoms cannot cover all terms that have the same semantic meaning and occur in the clinical notes. Embeddings are used to evaluate the semantic similarities between all extracted symptoms and the seed symptoms to create symptom clusters. Since 2018, various neuron networks have been investigated for generating sentence or even paragraph embeddings. Universal Sentence Encoder (USE) is introduced by Cer et al. [5] in 2018. The USE transformer-based encoding model uses the encoding sub-graph of the transformer architecture to generate embeddings. It computes context-aware representations of words by considering the ordering and identity of all the other words in a sentence. Given a symptom term consisted of more than one word, it computes the symptom embedding by computing the element-wise sum of the representations of each word embedding. The semantic similarities between the symptoms can be then calculated by measuring cosine distance between the embeddings. Figure 2 is the heatmap that shows the symptom similarity using embeddings generated from the USE. It includes some of the seed symptoms from each cluster. The darker the color is, the higher semantic similarity score is. Although the figure shows three clusters, some symptoms from different clusters also show the semantic relationship. For example, the semantic similarity between ‘tired’ and ‘depressed’ is higher than ‘sad’ and ‘peripheral neuropathy’.

4.1.3 ***Symptom Clustering.*** Based on the embeddings of all extracted symptoms and the seed symptoms in the clusters, we develop the modified k-means algorithm to identify other symptoms in the clinical notes which are different from the seed symptoms but have similar semantic meaning. The modified k-means is different from the seeded k-means algorithm in the literature [4]; it assumes that each seed symptom is within one defined cluster. Given a non-seed symptom (S_m) which is a symptom not listed in Table 1 but occur in a clinical note, if the average cosine similarity to all of the seed symptoms ($< S_1, \ldots, S_n >$) in the cluster (C) is bigger than the lowest average cosine similarities between the seed symptoms within that cluster, the non-seed symptom is added to the cluster. It is described as Equation 1, where $S_i \in C$ and $S_j \in C$. After applying this modified k-means clustering process, some symptoms extracted from the clinical notes are added to the symptom clusters. For example, ‘nerve pain’ is added to the cluster ‘Peripheral Neuropathy’; ‘extreme exhaustion’ is added to the ‘fatigue’ cluster. Some symptoms, such as ‘finger numbness’ which has a body part to one of the seed symptom, are also identified through this process.

$$\begin{cases} S_m \in C, & \text{if } \dfrac{\sum_{j=1}^{n} cosine(S_m, S_j)}{n} > \min\limits_{i=1}^{n} \dfrac{\sum_{j=1}^{n} cosine(S_i, S_j)}{n} \\ S_m \notin C, & \text{otherwise} \end{cases} \tag{1}$$

Cluster name	Seed Symptoms
Fatigue	fatigue, listless, weary, weariness, lethargic, lethargy, no energy, tired, sleepy, drowsy, exhausted, exhaustion, worn out, drained
Peripheral Neuropathy	peripheral neuropathy, numbness, tingling, burning, crawling, disorder of autonomic nervous system, other idiopathic peripheral autonomic neuropathy
Depression	depression, depressed, sad, unhappy, no appetite, failure to thrive, despair, misery, melancholy, hopeless, down-hearted, despondent, discouraged, hopeless

Table 1: Seed Symptoms for Symptom Clusters

4.2 Patient Clustering and Result Interpretation Method

One objective of this research is to understand the associations between the symptom clusters and other clinical attributes of the study cohort. Hence, we investigate patient clustering based on the symptom clusters and then interpret the patient clusters by using other clinical attributes. We construct patient representation using symptoms clusters. Each patient is presented as a three-dimensional vector. Each dimension presents a symptom cluster. The document frequency of the symptoms in each cluster is calculated for each dimension. In our study, each document is a clinical report. If a clinical report contains one or more than one symptom in one cluster, the document frequency (DF) for that cluster is counted as 1. We don't consider the term frequency (TF) in this study. The hypothesis is that term frequency in one report does not reflect the prominence of the symptom to the patient. Whereas, the more frequently the same symptoms are reported to the physician (at different visits) indicates greater bother or severity caused by the symptoms. Equation 2 provides the patient representation (*pt*) in this study, where df_{C_f} is the document frequency of the symptoms in the cluster 'Fatigue' – C_f; df_{C_p} is the document frequency of the symptoms in the cluster 'Peripheral Neuropathy' – C_p; df_{C_d} is the document frequency of the symptoms in the cluster 'Depression' – C_d;

$$pt = \left(df_{C_f}, df_{C_p}, df_{C_d}\right) \tag{2}$$

The k-means clustering is used here to generate patient clusters. K-means clustering is rather easy to implement and has been successfully used in the biomedical domain. K-means clustering algorithms try to assign the data in the data set to one of the predefined numbers of clusters. The aim is to minimize the sum of the distance of each point within the cluster to the cluster center. In this study, the k value is determined by the sum of the squared distances (*SSD*) from the data instances to the cluster centers. The k is selected when there is no significant change in *SSD* between two consecutive *k*.

To interpret the patient clusters, we use the external clustering evaluation method – conditional entropy to the other clinical attributes, such as the comorbidity of diabetes. The hypothesis is that the cluster distribution might associate with some clinical attributes. For example, it is possible that some clusters have a high frequency of 'depression', and the smoking status of the patients in those clusters are 'currently smoker'. In this research, we consider the following clinical attributes and patient demographics: gender (for CRC only), race, smoking status, stage of cancer, age at diagnosis, comorbidity of diabetes, and the number of comorbidities. The conditional entropy is calculated as Equation 3, where C_i is cluster i, k is the total number of categories of a clinical attribute or demographic attribute T, n_i is the total number of patient in cluster i, n_{ij} is the number of patient in cluster i have attribute value of j. Conceptually, the more of a cluster's membership has split into different categories of an attribute, the higher the conditional entropy. The lower the conditional entropy is, the better the clustering with respect to the attribute. For a perfect clustering, the conditional entropy value should be 0.

$$H\left(T|C_i\right) = -\sum_{j=1}^{k}\left(\frac{n_{ij}}{n_i}\right)log\left(\frac{n_{ij}}{n_i}\right) \tag{3}$$

4.3 Deriving Symptoms Association Rules

Other than patient clustering, we also apply the association rule mining (ARM) to the data to see whether there are associations between the symptom clusters and multiple clinical or demographic attributes. ARM algorithm is originally designed by [2] to find items that occur simultaneously and frequently in database transactions. In this study, we use it to find co-occurred clinical attribute values or symptoms in patients. The clinical attributes are extracted from the structured field of the EHR and many of them are categorical type. For example, the gender of female or male, level of depression, and so on. Some attributes are continuous type, such as age at diagnosis, are converted into categorical type based on the distribution of the data. Given a set of attribute values $A = \left\{a_1, a_2, \ldots, a_m\right\}$ and a set of patients $PT = \left\{pt_1, pt_2, ..., pt_n\right\}$ where $pt_i = \left\{a_{i1}, a_{i2}, ..., a_{im}\right\}$ and $a_{ij} \in A$, ARM finds a set of rules in the form of equation 4:

$$X \longrightarrow Y \tag{4}$$

where $X, Y \subseteq$ A are sets of attribute values called attribute sets, and $X \cap Y = \emptyset$.

The importance of a rule is defined by the support rate (S) and the confidence rate (C), as shown in equations 5 and 6. The support rate threshold indicates the frequency of the rule. The confidence rate threshold suggests the frequency of the attribute values Y appearing in transactions that contain attribute values X. In other words, the confidence rate reflects the level of reliability of a rule. In this research, we set the support rate threshold to be 0.1, and confidence rate threshold to be 0.9.

$$S\left((X \longrightarrow Y\right) = \frac{|X \cup Y|}{|PT|} \tag{5}$$

Age at Diagnosis	0: <40, 1: 40-50, 2: 50-65, 3: >65
Smoking Status	0: never smoker, 1: former smoker, 2: current light smoker, 3: current heavy smoker
Comorbidity	0: <=2, 1: 3-5, 2: >5
Fatigue	0: 0 (No fatigue), 1: 1-2 (Some fatigue), 2: >2 (Much fatigue)
Depression	0:0 (No depression), 1: >0 (Depression)
Peripheral Neuropathy (PN)	0: 0 (No PN), 1: 1-2 (Some PN), 2: >2 (Much PN)

Table 2: Converting Continuous Clinical and Demographic Attributes to Categorical values

$$C\big((X \rightarrow Y\big) = \frac{|X \cup Y|}{|X|} \qquad (6)$$

5 EXPERIMENTAL RESULTS

In this research, we employ the proposed methods on two patient cohorts – breast cancer and colorectal cancer. Specifically, we investigate symptoms and their associations with other attributes for two different time frames: (A) From the start of the chemotherapy till 12 months after the chemotherapy based on the clinician's suggestion that the symptoms of patients might show some difference between patients with and without specific clinical attributes, such as comorbidity of diabetes. (B) From 48 months to 54 months after the chemotherapy based on the survival analysis of the study cohort. The survival analysis shows that the survival rates of these two types of cancer show noticeable changes starting from 48 months after the chemotherapy.

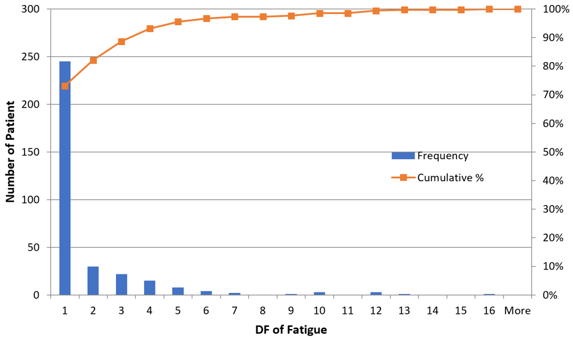


Figure 3: Document Frequency Distribution of the Symptom 'Fatigue' based on BC Patient Cohort of Time Frame A

Attribute	Tf A	Tf B
Ave. Age at Diagnosis	54.4	54.2
Smoker Ratio	0.76	0.77
Ave. Number of Comorbidity	1.55	1.59
Ave. DF of Fatigue	1.33	0.99
Ave. DF of Depression	0.62	0.32
Ave. DF of Peripheral Neuropathy	0.10	0.01
Diabetes Pt Ratio	0.15	0.13

Table 3: Attribute Values of the Breast Cancer Patients of the Two Time Frames (Tf)

As mentioned in the previous section, other than the clinical reports for symptom extraction and analysis, we extract other clinical and demographic attributes. Some of the attributes are converted into the categorical values based on the data distribution of the study cohort to interpret the patient clustering and derive association rules. Figure 3 provides the document frequency distribution of the symptom 'fatigue' based on the BC patient cohort of the time frame A. Majority of the patient has fatigue less than two times in the time frame. We find similar data distribution for BC and CRC patients for both time frames. Hence, we convert this symptom attribute into three categories: 0 – 'no fatigue'; 1 – 'some fatigue' and 2 – 'much fatigue'. In this study, we convert the age at diagnosis based on the distribution of the data and preventive care guidelines of the BC and CRC. Table 2 lists all the attributes and the conversion to the categorical values. The other attributes, such as race, gender, and TNM stage group are already categorical types. No converting is needed.

Before patient clustering and analysis, we remove the data instances that have missing data. For example, patients with no clinical reports in the selected two time frames are not included in this study.

5.1 Breast Cancer Cohort Clustering and Analysis

After data cleaning, there are 335 and 274 breast cancer patients for time frame A and B respectively. Table 3 provides the statistics of each attribute and document frequencies of the symptoms for these two groups of patients. From time frame A to time frame B, the statistic values do not change much for most of the attributes. However, the DF of the symptoms reduced. The DF of peripheral neuropathy significantly reduced. The smoker ratios for both time frames are over 0.75, which means many of them are current smokers.

K-means algorithm is applied to these two patient groups respectively to identify the patient clusters based on the three symptom clusters. Then, conditional entropy is used to interpret whether the clustering results associate with the selected clinical attributes. Table 4 and 5 shows the clusters, the number of patient in each cluster along with the calculated conditional entropy of the clinical and demographic attributes. We find that the patient clusters are generated based on the amount of the symptoms. So, we provide some description of each cluster based on the symptoms. The results show that for both time frames, the largest cluster is the one has no fatigue, depression, and peripheral neuropathy. From time frame A to time frame B, the percentage of the patient without any of these three clusters of symptoms increases by about 2.6%. Whereas, the portion of the patient only have depression increases from 5.4%

Cluster ID: Description	# of Pt(%)	Comorbidity	Diabetes	Smoking	Age at Diagnosis	TNM Stage	Race
0: Much fatigue, No depression, Much peripheral neuropathy	3(0.9%)	0.91	0	0.91	0.91	1.58	0
1: No fatigue, No Depression, No peripheral neuropathy	149(44.5%)	1.02	0.58	1.02	1.77	2.62	0.48
2: Much fatigue, Depression, No peripheral neuropathy	31(9.2%)	0.88	0.45	0.82	1.92	2.58	0.55
3: Some fatigue, Depression, No peripheral neuropathy	29(8.7%)	0.94	0.47	0.94	1.55	2.24	0.21
4: Much fatigue, Depression, No peripheral neuropathy	19(5.7%)	0.91	0.62	0.59	1.45	2.14	0.77
5: Some fatigue, No depression, No peripheral neuropathy	75(22.4%)	0.99	0.66	0.93	1.81	2.60	0.76
6: No fatigue, Depression, No peripheral neuropathy	18(5.4%)	1.12	0.85	0.80	1.95	2.33	1.05
7: Much fatigue, Depression, Some peripheral neuropathy	5(1.5%)	0.72	0.72	0.97	0.72	1.52	0.72
8: Much fatigue, No depression, Some peripheral neuropathy	2(0.6%)	1	0	0	1	1	0
9: No fatigue, No depression, Some peripheral neuropathy	1(0.3%)	0	0	0	0	0	0
10: Some fatigue, No depression, Some peripheral neuropathy	3(0.9%)	1.58	0	0	1.58	1.58	0.91

Table 4: Clusters and Conditional Entropies of Time Frame A – Breast Cancer

Cluster ID: Description	# of Pt(%)	Comorbidity	Diabetes	Smoking	Age at Diagnosis	TNM Stage	Race
0: Some fatigue, No depression, No peripheral neuropathy	70(25.5%)	1.12	0.66	0.92	1.75	2.56	0.73
1: No fatigue, Depression, No peripheral neuropathy	20(7.3%)	0.92	0.60	0.74	1.76	2.00	0.72
2: No fatigue, No depression, No peripheral neuropathy	129(47.1%)	0.91	0.49	0.86	1.76	2.79	0.65
3: Much fatigue, No depression, No peripheral neuropathy	20(7.3%)	1.05	0.46	0.81	1.64	2.28	0.88
4: Some fatigue, Depression, No peripheral neuropathy	26(9.5%)	1.19	0.61	0.84	1.94	2.48	0.39
5: Much fatigue, Depression, Some peripheral neuropathy	9(3.3%)	0	0.50	0.76	0.50	1.75	0.76

Table 5: Clusters and Conditional Entropies of Time Frame B – Breast Cancer

Rules	C	S
Age at diagnosis=40-50 Comorbid=0-2 –> Diabetes=No	1	0.18
TNM Path Stage Group=3A –> Peripheral neuropathy=No	1	0.11
Age at diagnosis<40 –> Peripheral neuropathy=No	1	0.12
TNM Path Stage Group=1A Fatigue=No Peripheral neuropathy=No Diabetes=No –> Depression=No	0.95	0.11
Smoking Status=Former Smoker Fatigue=No –> Depression=No Peripheral neuropathy=No	0.95	0.11

Table 6: Derived Association Rules from Time Frame A - Breast Cancer

to 7.3%, the percentage of the patient have all three clusters of symptoms increases from 1.5% to 3.3%.

As described in the methods section, the lower the conditional entropy is, the better the clustering is concerning the attribute. For time frame A, cluster 0, 8, 9, and 10 are tiny clusters which have less than five patients. Hence, the results show more conditional entropy values 0. The patients in these small clusters all do not have diabetes. The patients in cluster 8 and 10 are all current light

Rules	C	S
Fatigue=No 149 –> Peripheral neuropathy=No	1	0.54
TNM Path Stage Group=3A –> Peripheral neuropathy=No	1	0.30
Age at diagnosis=40-50 Fatigue=No –> Peripheral neuropathy=No	1	0.25
TNM Path Stage Group=1 Comorbid=0-2 –> Depression=No	0.97	0.11
Age at diagnosis<40 –> Diabetes=No	0.94	0.11

Table 7: Derived Association Rules from Time Frame B - Breast Cancer

smoker. For time frame B, cluster 5 has 0 conditional entropy on comorbidity, which means the comorbidity values of this cluster are the same. The details show that all patients in cluster 5 have 0 to 2 comorbidity. The age at diagnosis and diabetes status both have low entropy of 0.5. The details show that all patients except one have 'age at diagnosis' between 50 and 65, and all patient except one have no diabetes.

Attribute	Tf A	Tf B
Ave. Age at Diagnosis	59.1	59.0
Smoker Ratio	0.22	0.15
Ave. Number of Comorbidity	3.55	3.95
Ave. DF of Fatigue	1.13	0.91
Ave. DF of Depression	0.53	0.70
Ave. DF of Peripheral Neuropathy	0.06	0.03
Diabetes Pt Ratio	0.16	0.13

Table 8: Attribute Values of the Breast Cancer Patients of the Two Time Frames (Tf)

5.2 Derived Rules from the Breast Cancer Cohorts

ARM has been applied to the BC patient cohorts of the two selected time frames to investigate associations between clinical attribute values and symptoms. There are 304 rules and 336 rules identified from time frame A and B, respectively. These rules satisfy the support rate threshold 0.1 and confidence rate threshold 0.9. Table 6 and 7lists some of the informative rules for time frame A and B.

In time frame A, we find that the patients diagnosed younger than 40 or patients at a higher pathologic stage of cancer less likely to have peripheral symptom neuropathy. The patients between 40 and 50 with less comorbidity do not have diabetes. If the patient at an earlier pathologic stage of cancer without fatigue or peripheral neuropathy, most likely she also has no depression. If the patient is a former smoker, and no fatigue and peripheral neuropathy symptoms, most likely she also has no depression.

In time frame B, we find that over 50% of the chance that if a patient has no fatigue, she also has no peripheral neuropathy symptoms. If a patient is at the earlier pathologic stage of cancer and less comorbid, she less likely to have depression. In this study cohort, if a patient is diagnosed with breast cancer younger than 40, likely she has no comorbidity of diabetes. Only one rule shows up in both time frames. That is if a patient has a higher pathologic cancer stage of 3A, less likely she has peripheral neuropathy symptom.

5.3 Colorectal Cancer Cohort Clustering and Analysis

After data cleaning, there are 179 and 118 colorectal cancer patients for the time frame A and B, respectively. Table 8 provides statistics of each attribute and document frequencies of the symptoms for these two groups of patients. From time frame A to time frame B, the smoker ratio reduces. The DF of the symptom fatigue and peripheral neuropathy reduces. However, the DF of the symptom depression increases a little. Comparing to the breast cancer cohort, in general, the smoker ratio in colorectal cancer cohort is much lower, whereas, the number of comorbidity in colorectal cancer cohort is much higher. From the symptom clusters point of view, for time frame A, breast cancer patient cohort have slightly more symptoms of these three clusters comparing to the colorectal cancer cohort. For time frame B, colorectal cancer cohort has slightly more depression on average.

K-means algorithm is applied to these two patient groups respectively to identify the patient clusters based on the symptom clusters. Then, conditional entropy is used to interpret whether the clustering results associate with the selected clinical attributes. Table 9 and 10 show the clusters, the number of patient in each cluster along with the calculated conditional entropy of the clinical and demographic attributes. Similarly, the clusters are generated based on the amount of the symptoms. Similar to breast cancer cohort, for both time frames, the biggest cluster is the one that has no fatigue, depression, and peripheral neuropathy. From time frame A to time frame B, the percentage of the patient without any of these three clusters of symptoms increases by about 3.2%. Whereas, the percentage of the patient only have much fatigue increases from 8.4% to 11.0%. Different from the breast cancer cohort, for both time frames, no cluster have all three types of symptoms.

For the entropy calculation, we add gender to the colorectal cancer cohort analysis. For time frame A, cluster 5 and 6 are tiny clusters which have three patients in each. These two clusters contain patients with much peripheral neuropathy. The difference is that the patients in cluster 6 have much fatigue symptoms. The cluster 5 are all never smoker, whereas, the patients in cluster 6 are either current smokers or former smokers. Because the race of all patients in some clusters is white, conditional entropy for those clusters is 0 for attribute race. Cluster 1 and 3 have conditional entropy for diabetes below 0.5. The details show that except one patient in each cluster has comorbidity diabetes, and the rest patients do not have diabetes. For time frame B, cluster 7 has only one patient, so all conditional entropy values are 0. Cluster 4 contains patients with much fatigue and no other two symptoms, and the conditional entropy is 0 on diabetes. The data shows all patients in cluster 4 are non-diabetic. Cluster 6 has two patients with much

Cluster ID: Description	# of Pt (%)	Comorbidity	Diabetes	Smoking	Age at Diagnosis	TNM Stage	Race	Gender
0: No fatigue, No depression, No peripheral neuropathy	72(40%)	1.49	0.71	1.49	1.80	2.31	0.82	0.99
1: No fatigue, Depression, No peripheral neuropathy	16(8.9%)	1.29	0.33	1.92	1.67	2.38	0	0.95
2: Some fatigue, No depression, No peripheral neuropathy	33(18.4%)	1.46	0.53	1.45	1.60	2.48	0.44	0.95
3: Much fatigue, Depression, No peripheral neuropathy	15(8.4%)	1.53	0.35	1.37	1.74	2.23	0.35	0.83
4: Much fatigue, No depression, No peripheral neuropathy	27(15.1%)	1.27	0.69	1.63	1.56	2.18	0	0.97
5: Some or no fatigue, No depression, Much peripheral neuropathy	3(1.7%)	0.91	0.91	0	0.91	0.91	0	0.91
6: Much fatigue, No depression, Much peripheral neuropathy	3(1.7%)	1.58	0.91	0.91	0.91	0.91	0	0.91
7: Some fatigue, Depression, No peripheral neuropathy	10(5.6%)	0.92	0.72	1.76	1.48	2.05	0	0.71

Table 9: Clusters and Conditional Entropies of Time Frame A – Colorectal Cancer

Cluster ID: Description	# of Pt (%)	Comorbidity	Diabetes	Smoking	Age at Diagnosis	TNM Stage	Race	Gender
0: Some fatigue, Depression, No peripheral neuropathy	6(5.1%)	1.46	0.91	1.45	1	1.79	0	1
1: Some fatigue, No Depression, No peripheral neuropathy	24(20.3%)	1.56	0.41	1.24	1.98	2.20	0.49	0.98
2: No fatigue, No depression, No peripheral neuropathy	51(43.2%)	1.49	0.52	1.42	1.59	2.25	0.38	0.99
3: Much fatigue, Depression, No peripheral neuropathy	13(11.0%)	1.58	0.89	1.88	1.23	2.13	0.78	0.78
4: Much fatigue, No depression, No peripheral neuropathy	10(8.4%)	1.48	0	1.57	1.76	1.77	0.46	0.88
5: No fatigue, Depression, No peripheral neuropathy	11(9.3%)	1.43	0.68	1.32	1.49	1.44	0.43	0.99
6: Much fatigue, No depression, Much peripheral neuropathy	2(0.8%)	1	1	0	1	1	0	0
7: No fatigue, Depression, Much peripheral neuropathy	1(0.8%)	0	0	0	0	0	0	0

Table 10: Clusters and Conditional Entropies of Time Frame B – Colorectal Cancer

Rules	C	S
Smoking=Never smoker Gender=Male 48 –> Depression=No	0.94	0.25
TNM Path Stage Group=3B Fatigue=No 36 –> Depression=No Peripheral neuropathy=No	0.92	0.18

Table 11: Derived Association Rules from Time Frame A - Colorectal Cancer

Rules	C	S
Fatigue=Much Fatigue Gender=Male –> Peripheral neuropathy=No	1	0.14
Age at diagnosis=40-50 –> Peripheral neuropathy=No Depression=No Diabetes=No	0.9	0.15

Table 12: Derived Association Rules from Time Frame B - Colorectal Cancer

fatigue and peripheral neuropathy, but no depression. Both these patients are never smoker and female patients, which is reflected by the value 0 for conditional entropy of smoking and gender.

5.4 Drived Rules from the Colorectal Cancer Cohorts

ARM has been applied to the colorectal cancer patient cohorts of the two selected time frames to explore the associations between clinical attribute values and symptoms. There are 375 rules and 566 rules identified from time frame A and B, respectively. These rules satisfy the support rate threshold 0.1 and confidence rate threshold 0.9. Table 11 and 12 lists some of the informative rules for time frame A and B.

In time frame A, we find that if the patient is a former smoker and male, most likely, he has no depression. If a patient is at a higher cancer stage 3B, but no fatigue symptom, he or she likely doesn't have depression and peripheral neuropathy. In time frame B, we find if a patient is male and has much fatigue symptoms, he does not necessarily have the peripheral neuropathy symptom. If the patient's age at diagnosis is between 40 and 50, based on the data of this cohort, there is a high chance he/she has no depression, peripheral neuropathy, and diabetes.

6 CONCLUSIONS AND FUTURE WORK

The EHR data has not been extensively used to understand the patient-reported symptoms for patients who have chronic diseases, such as cancer. In this study, we develop a framework to make use of both structured and unstructured data to identify three symptom clusters and understand their association with other clinical attributes. The framework includes components of NLP, clustering, and ARM. It has been applied two both breast cancer and colorectal cancer patient cohorts. Through these methods, we find differences between symptoms associations for breast cancer and colorectal cancer and between different time frames after chemotherapy. These finding can help clinicians better manage patients' symptoms based on their variability.

There are limitations to the methods and results. The document frequency is used to measure the severity level of the symptoms, which means the patients need to have more physician visit to demonstrates the severity. Additionally, the number of patients for the clustering analysis is not very big to generalize the findings for breast cancer and colorectal cancer symptom management. In the future, we will include more patients and try different methods to capture the severity level of the symptoms. We also plan to expand this framework to other symptom clusters and chronic diseases.

ACKNOWLEDGMENTS

This study was supported by the Research Investment Funding grant from Indiana University School of Nursing. The authors acknowledge Regenstrief Institute, Inc.'s participation in this project.

REFERENCES

[1] [n.d.]. *MetaMap – A Tool For Recognizing UMLS Concepts in Text.* https://metamap.nlm.nih.gov/.

[2] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Acm sigmod record*, Vol. 22. ACM, 207–216.

[3] Nancy E Avis, Beverly Levine, Sarah A Marshall, and Edward H Ip. 2017. Longitudinal examination of symptom profiles among breast cancer survivors. *Journal of pain and symptom management* 53, 4 (2017), 703–710.

[4] Sugato Basu, Arindam Banerjee, and Raymond Mooney. 2002. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002.* Citeseer.

[5] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[6] Chen X Chen, Susan Ofner, Giorgos Bakoyannis, Kristine L Kwekkeboom, and Janet S Carpenter. 2018. Symptoms-Based Phenotypes Among Women With Dysmenorrhea: A Latent Class Analysis. *Western journal of nursing research* 40, 10 (2018), 1452–1468.

[7] Winson Y Cheung, Lisa W Le, and Camilla Zimmermann. 2009. Symptom clusters in patients with advanced cancers. *Supportive care in cancer* 17, 9 (2009), 1223–1230.

[8] Claire J Han, Kerryn Reding, Bruce A Cooper, Steven M Paul, Yvette P Conley, Marilyn Hammer, Fay Wright, Frances Cartwright, Jon D Levine, and Christine Miaskowski. 2019. Symptom Clusters in Patients with Gastrointestinal Cancers Using Different Dimensions of the Symptom Experience. *Journal of pain and symptom management* (2019).

[9] S. Logeswari and K. Premalatha. 2013. Biomedical document clustering using ontology based concept weight. In *Proceedings of the International Conference on Computer Communication and Informatics.* 1–4. https://doi.org/10.1109/ICCCI.2013.6466273

[10] Sarah A Marshall, Christopher C Yang, Qing Ping, Mengnan Zhao, Nancy E Avis, and Edward H Ip. 2016. Symptom clusters in women with breast cancer: an analysis of data from social media and a research study. *Quality of Life Research* 25, 3 (2016), 547–557.

[11] Melissa Mazor, Janine K Cataldo, Kathryn Lee, Anand Dhruva, Bruce Cooper, Steven M Paul, Kimberly Topp, Betty J Smoot, Laura B Dunn, Jon D Levine, et al. 2018. Differences in symptom clusters before and twelve months after breast cancer surgery. *European Journal of Oncology Nursing* 32 (2018), 63–72.

[12] Patrick A McKee, William P Castelli, Patricia M McNamara, and William B Kannel. 1971. The natural history of congestive heart failure: the Framingham study. *New England Journal of Medicine* 285, 26 (1971), 1441–1446.

[13] Stephen L. Meigs and Michael Solomon. 2016. Electronic Health Record Use a Bitter Pill for Many Physicians. *Perspect Health Information Management* 13, 1d (2016).

[14] Christine Miaskowski, Bradley E Aouizerat, Marylin Dodd, and Bruce Cooper. 2007. Conceptual issues in symptom clusters research and their implications for quality-of-life assessment in patients with cancer. *Journal of the National Cancer Institute Monographs* 2007, 37 (2007), 39–46.

[15] Christine Miaskowski, Bruce A Cooper, Steven M Paul, Marylin Dodd, Kathryn Lee, Bradley E Aouizerat, Claudia West, Maria Cho, and Alice Bank. 2006. Subgroups of patients with cancer with different symptom experiences and quality-of-life outcomes: a cluster analysis.. In *Oncology nursing forum*, Vol. 33.

[16] Christine Miaskowski, Laura Dunn, Christine Ritchie, Steven M Paul, Bruce Cooper, Bradley E Aouizerat, Kimberly Alexander, Helen Skerman, and Patsy Yates. 2015. Latent class analysis reveals distinct subgroups of patients based on symptom occurrence and demographic and clinical characteristics. *Journal of pain and symptom management* 50, 1 (2015), 28–37.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems.* 3111–3119.

[18] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan.* 39–43.

[19] Parikshit Sondhi, Jimeng Sun, Hanghang Tong, and ChengXiang Zhai. 2012. SympGraph: a framework for mining clinical notes through symptom relation graphs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 1167–1175.

[20] Stephan Tulkens, Simon Suster, and Walter Daelemans. 2016. Using Distributed Representations to Disambiguate Biomedical and Clinical Concepts. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing.*

[21] Fei Wang, Jiayu Zhou, and Jianying Hu. 2014. Densitytransfer: A data driven approach for imputing electronic health records. In *2014 22nd International Conference on Pattern Recognition.* IEEE, 2763–2768.

[22] I. Yoo, X. Hu, and I.-Y. Song. 2006. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. In *Proceedings of the First International Workshop on Text Mining in Bioinformatics.* 84–89.

[23] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou. 2007. A comparative study of ontology based term similarity measure on pubmed document clustering. In *Proceedings of the International Conference on Database Systems for Advanced Applications.* 115–126.

[24] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. Human symptoms–disease network. *Nature communications* 5 (2014), 4212.

[25] Yongjun Zhu, Erjia Yan, and Fei Wang. 2017. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making* 17 (2017), 95–103.