Identifying Symptom Clusters Through Association Rule Mining

Mikayla Biggs[1](✉), Carla Floricel[3], Lisanne Van Dijk[2],
Abdallah S. R. Mohamed[2], C. David Fuller[2], G. Elisabeta Marai[3],
Xinhua Zhang[3], and Guadalupe Canahuate[1](✉)

[1] University of Iowa, Iowa City, IA, USA
{mikayla-biggs,guadalupe-canahuate}@uiowa.edu
[2] University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3] University of Illinois in Chicago, Chicago, IL, USA

Abstract. Cancer patients experience many symptoms throughout their cancer treatment and sometimes suffer from lasting effects post-treatment. Patient-Reported Outcome (PRO) surveys provide a means for monitoring the patient's symptoms during and after treatment. Symptom cluster (SC) research seeks to understand these symptoms and their relationships to define new treatment and disease management methods to improve patient's quality of life. This paper introduces association rule mining (ARM) as a novel alternative for identifying symptom clusters. We compare the results to prior research and find that while some of the SCs are similar, ARM uncovers more nuanced relationships between symptoms such as anchor symptoms that serve as connections between interference and cancer-specific symptoms.

Keywords: Association rule mining · Symptom clusters · PRO

1 Introduction

Cancer patients experience a range of symptoms during and after treatment [1–3]. Research on these symptoms, their prevalence, relationships, and progression can improve disease prognosis and inform the appropriate treatment [4,5]. Symptom cluster (SC) research aims to identify co-occurring symptoms (e.g., pain, fatigue, dry mouth) and to understand the underlying mechanisms that drive these clusters [6]. This research is facilitated by increasingly available Patient-Reported Outcome (PRO) data, collected via questionnaires, that allows patients to rate the occurrence and severity of their symptoms.

The M.D. Anderson Symptom Inventory (MDASI) [7], and its head-and-neck (HN) cancer module [8], are short, validated questionnaires that patients record each visit. Three key groups comprise the 28 MDASI-HN survey questions: 13 core items for common symptoms to all cancers, nine items specific to HN, and six items regarding symptom interference with daily activity. Patients rate their symptoms using a 0–10 scale, from "not present" to "as bad as you can imagine"

A. Tucker et al. (Eds.): AIME 2021, LNAI 12721, pp. 491–496, 2021.
https://doi.org/10.1007/978-3-030-77211-6_58

(core and HN), respectively from "did not interfere" to "interfered completely" (interference). Preliminary SCs in the MDASI-HN data have been identified using factor and cluster analysis [9,10].

This paper introduces association rule mining (ARM) [11] as an alternative for identifying symptom clusters. To the best of our knowledge, this is the first ARM application in the SC domain. This work's main contribution is to offer an alternative methodology for defining new and interesting relationships for SC research using PRO data. We model each PRO response as a patient transaction and process PROs during and after treatment to identify acute and late symptom clusters, respectively. We furthermore model the severity of the symptoms. We present a graph-based visualization for the most significant association rules to identify symptom clusters for both acute and late stages. Finally, we evaluate this methodology on a real HN cancer patient dataset.

2 Modeling Symptom Clusters with ARM

The ARM approach can use any PRO; in this work, we focus on the MDASI-HN questionnaire. The M.D. Anderson Symptom Inventory (MDASI) is a multi-symptom patient-reported outcome measure to assess both the severity of cancer symptoms and symptom interference with daily life. Table 1 shows a sample of the symptoms described in the MDASI-HN survey and the short symptom labels used to refer to the MDASI-HN symptoms to improve readability.

ARM has two steps: the first one is to identify frequent item-sets (FIS) from the data, and the second is to generate the association rules from the FIS. The Apriori algorithm identifies the frequent items in the data set using a set of core metrics. Support is a measure of absolute frequency, i.e., the fraction of sets that contain items A and B. Confidence ($A \rightarrow B$) is a measure of correlative frequency. It tells us how often the items A and B occur together, given the number times A occurs. Lift indicates the strength of a rule over the random occurrence of A and B. The higher the lift, the more significant the association. A lift greater than 1.0 implies that the relationship between the antecedent and the consequent is more significant than expected if the two were independent. With a lift of 1.0, we can say that the relationships appear as expected and are not significantly associated. For example, with the rule $\{fatigue\} \rightarrow \{drowsy\}$ with 50% support, and 80% confidence we could say that these two symptoms

Table 1. The 28 MDASI-HN symptoms organized into 3 symptom categories

Category	Symptom labels
Common cancer	Pain, fatigue, nausea, sleep, distress, SOB, memory, appetite, drowsy, drymouth, sad, vomit
Head & Neck	Numb, mucus, swallow, choke, voice, skin, constipation, taste, mucositis, teeth
Interference	General_activity, mood, work, enjoy, relations, walking, enjoy

are experienced together by 50% of the patients, and "if a patient experiences fatigue, they are 80% likely to experience drowsiness'.'

Since symptom severity is non-binary data, we generate two categories for each symptom and use the labels low and severe to distinguish them. For one questionnaire, symptoms with a rating greater than 0 are considered occurring symptoms. A symptom is *low* if the patient rated its severity less than five and *severe* otherwise. The data models the transactions with one unique PRO for each patient, and the two items being "bought" together, indicating low or severe, are concurrent symptoms. We consider symptom clusters at two different time points. Acute symptoms refer to symptoms experienced during treatment (about six weeks from the start of treatment). For late symptoms, patients survey the PROs up to 18-months post-treatment. Symptoms with missing scores (NaN) were replaced with 0 s. Patients with no PRO recorded during the acute or late phases were not included in the time frame analysis.

3 Experimental Results

The dataset used for these experiments consists of MDASI-HN responses for a cohort of 823 patients. The patient surveys were broken into acute and late time points with two items per symptom (low and severe) used to capture the severity of the symptoms. A total of 643 patients had at least one acute PRO, and 745 patients had at least one late PRO. Figure 1 shows the symptom's overall support for low and severe symptoms during the acute and late time frames. As shown, in the acute stage, many patients experienced both low and severe symptoms during treatment. In contrast, symptoms experienced in the late stage have a lower severity than during the acute phase. We used minimum support of 20% for both the acute and late as it is the minimum cutoff between both stages for consistency in our analysis of each.

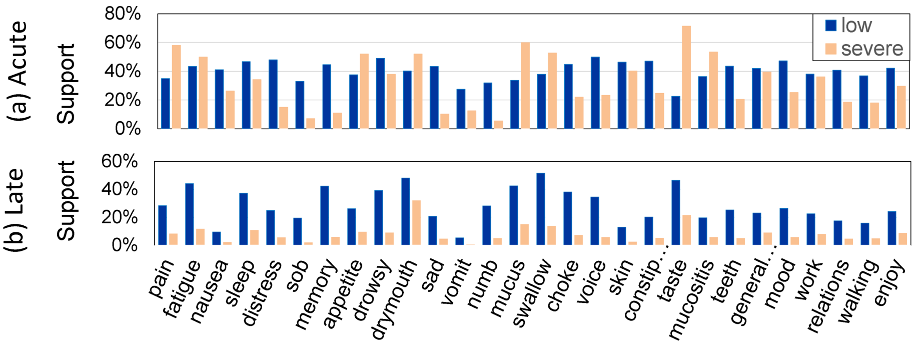
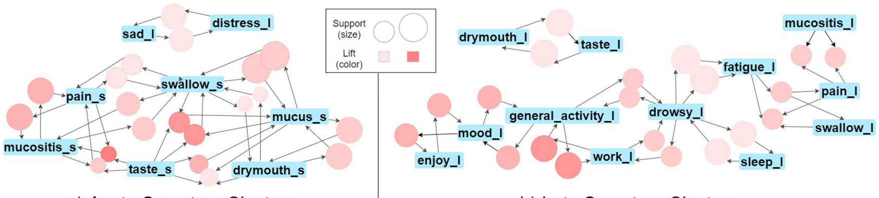


Fig. 1. Symptom Severity in the (a) acute and (b) late stages. Acute: > half of patients experience low severity symptoms, while a sizable 20% experience severe symptoms. Late: patients experience mostly low rated symptoms with highest prevalence in fatigue, drymouth, swallow, and taste.

Table 2. Top association rules for acute and late symptoms. Top five rules for each stage with the highest lift. The symptom's subscripts l and s stand for low and severe ratings, respectively.

Acute Stage				Late Stage			
antecendent	consequent	confidence	lift	antecendent	consequent	confidence	lift
$\{\text{pain}_s, taste_s\}$	$\{\text{mucositis}_s\}$	.85	2.82	$\{\text{general\_activity}_l\}$	$\{\text{work}_l\}$	.79	2.96
$\{\text{mucus}_s, taste_s\}$	$\{\text{swallow}_s\}$	.77	2.71	$\{\text{enjoy}_l\}$	$\{\text{mood}_l\}$	.75	2.84
$\{\text{swallow}_s, taste_s\}$	$\{\text{mucus}_s\}$	.89	2.70	$\{\text{fatigue}_l, swallow_l\}$	$\{\text{pain}_l\}$	.77	2.35
$\{\text{mucus}_s, taste_s\}$	$\{\text{drymouth}_s\}$	.75	2.64	$\{\text{pain}_l, fatigue_l\}$	$\{\text{swallow}_l\}$	.80	2.28
$\{\text{drowsy}_l\}$	$\{\text{fatigue}_l\}$	.76	2.19	$\{\text{drowsy}_l\}$	$\{\text{fatigue}_l\}$	.83	2.19



a) Acute Symptom Clusters

b) Late Symptom Clusters

Fig. 2. Symptoms Association Rule Graph. The graph encoding shows the top 20 association rules for (a) acute and (b) late symptoms. In the acute state there is a large cluster of severe symptoms. In the late stage, drowsy and fatigue appear to be anchor symptoms connecting a cluster of interference symptoms with a cluster of cancer symptoms.

Table 2 shows the top 5 association rules with the highest lift for the acute and late stages. The top rule for the acute stage involves pain, taste, and mucositis. While this association is clinically valid, since mucositis presents as small painful oral ulcers in patients, it notably could interfere with oral functions like taste. Other studies have shown pain to cluster more closely to fatigue than mucositis [10,12]. For late symptoms, the top three rules include interference symptoms rated with low severity. The acute symptoms showed that HN-related and common cancer symptoms were more prevalent than in late-stage analysis. Notably, rules involving drowsy and fatigue with low severity are among the top rules for both the acute and late stages. Previous studies have also supported the association between these two symptoms, drowsy and fatigue, as a symptom cluster [9,10]. Caution is advised when interpreting ARM relationships, as rules are not indicating causality but rather the probability of co-occurrence. To help visualize the symptom clusters, we adopt a graph representation for association rules [13]. Figure 2 shows the top 20 association rules sorted by lift for acute and late symptoms. The circles encode rules with size and color representing the support and lift metrics. The blue rectangles encode symptoms. An arrow pointing towards a circle means that the associated symptom is an antecedent for the association rule. If the arrow points towards a symptom, that symptom is the consequent for the association rule.

For acute symptoms, two clusters are consistent with previously reported clusters for HN cancer [10]. For late symptoms, there are four identifiable clusters. Interestingly, drowsy and fatigue seem to be anchor symptoms between interference and HN-related symptoms, a relationship that more traditional approaches for symptom cluster research cannot capture. Furthermore, we found that pain is associated with both mucositis and fatigue. These findings highlight that symptoms could appear in different clusters with the ARM algorithm, providing a more accurate model for the complex relationships between symptoms. In contrast, highly occurring symptoms would cluster together earlier when symptoms are *partitioned* into clusters, as in hierarchical clustering.

4 Conclusion

We introduce association rule mining as a powerful approach to identify patient symptom clusters and uncover interesting relationships between symptoms. Our approach models PRO data as transactions, visualizes the most significant association rules in symptom clusters, and captures the severity of symptoms in both acute and late stages. When applied to PRO data from head and neck cancer patients, this approach correctly identified higher symptom prevalence and severity during treatment and a gradual decrease after treatment. The new acute symptom clusters found include severely rated HN-related and common cancer symptoms. The late symptom clusters found include more interference symptoms and low severity symptoms. Our analysis identifies new anchor symptom clusters that connect interference and HN-related symptoms, offering new opportunities for targeted interventions that could positively affect cancer patients' quality of life while supporting previously identified SCs. In the future, we plan to include clinical variables such as staging, dose, and organs at risk [14,15] into the ARM analysis to determine whether patient characteristics are related to individual symptoms or symptoms clusters.

References

1. Christopherson, K.M., et al.: Chronic radiation-associated dysphagia in oropharyngeal cancer survivors. Clinic. Transl. Rad. Oncology **18**, 16–22 (2019)
2. Wentzel, A., et al.: Precision toxicity correlates of tumor spatial proximity to organs at risk in cancer patients. Radiother. Oncol. **148**, 245–251 (2020)
3. Wentzel, A., et al.: Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration. IEEE Trans. Vis. Comp. Graph. **26**(1), 949–959 (2019)
4. Marai, G.E., et al.: Precision risk analysis of cancer therapy with interactive nomograms and survival plots. IEEE Trans. Vis. Comp. Graph. **25**(4), 1732–1745 (2018)
5. Sheu, T., et al.: Conditional survival analysis of patients with locally advanced laryngeal cancer. Sci. Rep. **7**, 43928 (2017)
6. Miaskowski, C., et al.: Advancing symptom science through symptom cluster research. J. Nat. Cancer Instit. **109**(4) (2017)
7. Cleeland, C., et al.: Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. Cancer **89**, 1634–46 (2000)

8. Rosenthal, D.I., et al.: Measuring head and neck cancer symptom burden. Head Neck J. Sci. Specialt. **29**(10), 923–931 (2007)
9. Skerman, H.M., et al.: Multivariate methods to identify cancer-related symptom clusters. Res. Nurs. Health **32**(3), 345–360 (2009)
10. Rosenthal, D.I., et al.: Patterns of symptom burden during radiotherapy or concurrent chemoradiotherapy for H&N cancer. Cancer **120**(13), 1975–1984 (2014)
11. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proceedings 20th International Conference on Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
12. Kirkova, J., Aktas, A., Walsh, D., Davis, M.P.: Cancer symptom clusters: clinical and research methodology. J. Palliat. Med. **14**(10), 1149–1166 (2011)
13. Hahsler, M.: arulesviz: interactive visualization of association rules with r. R J. **9**(2), 163 (2017)
14. Tosado, J., et al.: Clustering of largely right-censored oropharyngeal HNC patients to improve outcome prediction. Sci. Rep. **10**(1), 1–14 (2020)
15. Luciani, T., et al.: A spatial neighborhood methodology for computing & analyzing lymph node carcinoma similarity in precision medicine. J. Biomed. Info. **5** (2020)