Network Analysis of Cancer Patients and Symptoms: Implications for Symptom Management and Treatment

Suresh K. Bhavnani[1,2] PhD, Gowtham Bellala[3], Arunkumaar Ganesan[3], Rajeev Krishna[4], MD Phd, Paul Saxman[2], Clayton Scott[1,3] PhD, Maria Silveira[5] MD MA MPH, Charles Given[6], PhD

[1]Center for Computational Medicine & Bioinformatics, [2]MICHR, [3]EECS, [4]Dept. of Psychiatry, [5]HSRD, Ann Arbor VAMC, Univ. of Michigan, Ann Arbor, MI; [6]Dept. of Family Medicine, Michigan State University

Abstract

Although many cancer patients experience multiple concurrent symptoms, most studies have focused on the analysis of single symptoms. Furthermore, the few studies that have analyzed how symptoms co-occur across patients have used methods such as factor analysis that have *a priori* assumptions of how the data is structured. To address these limitations, we used networks to visualize how 665 cancer patients reported 18 symptoms, and verified the results using appropriate quantitative methods. The results suggest that symptoms co-occur in a nested structure, where there is a small set of symptoms that co-occur in many patients, and larger inclusive sets of symptoms that co-occur among a few patients. These results (1) demonstrate how network analyses can reveal complex relationships between patients and symptoms while avoiding *a priori* assumptions, and (2) provide implications for cancer treatment and management that address complexities in symptom co-occurrence.

Introduction

Although cancer patients experience on average between 11-13 symptoms [1], most research has focused on the etiology, progression, and treatment of single symptoms. Furthermore, because of the additive impact of multiple symptoms, patients with many co-occurring symptoms generally fare worse than others. Understanding how symptoms co-occur in patients can therefore lead to more efficient assessment and management of symptoms, with the goal of significantly improving the overall function and quality of life for cancer patients.

To address this need, recent research has used data reduction methods such as factor analysis and hierarchical clustering to identify symptom clusters in different granularities of data [1]. For example, hierarchical cluster analysis was used to identify a cluster of five symptoms (e.g., hot flashes and weight gain) in menopausal women with breast cancer [2], and factor analysis was used to identify three clusters of symptoms across patients of all types of cancer [3]. While these early studies have made important inroads into identifying symptom clusters, several researchers have admitted that such methods produce results that are inherently unverifiable [2]. For example, there is no objective method to select cut-off points in a dendrogram (generated by hierarchical clustering [4]) to identify disjoint clusters. More importantly, these methods are based on *a priori* assumptions about the existence of disjoint symptom clusters, potentially masking more complex relationships in the data.

Inspired by the importance of symptom cluster research, but concerned about the *a priori* assumptions of the current methods, we used networks to first visualize the complex relationship between cancer patients and symptoms. This approach enabled us to visually inspect the data, with minimal assumptions about the underlying relationships. These visual observations then enabled us to select the appropriate methods to quantitatively verify the nature of the co-occurrence. Such a multi-method approach helped us to arrive at a new understanding of how symptoms co-occur across cancer patients, with insights about the treatment and management of co-occurring symptoms.

Method

Our research began with the question: *How do symptoms co-occur across cancer patients?* To address this research question, we made critical decisions regarding *data selection*, *data representation* and *data analysis* as discussed below:

Data Selection. We conducted a secondary analysis on data collected in a published study on cancer symptom management [5]. The data consisted of 671 cancer patients undergoing chemotherapy. The patients reported 18 symptoms using the M.D. Anderson Symptom Inventory which measures symptom severity ranging from 0 (not present) to 10 (worst imaginable). Six patients did not report any symptoms and therefore were dropped from the analysis, resulting in a total of 665 patients. The patients varied on a number of different disease and demographic variables including type and stage of cancer, gender and age. Similar to several previous studies [3], the focus of our analysis was to analyze

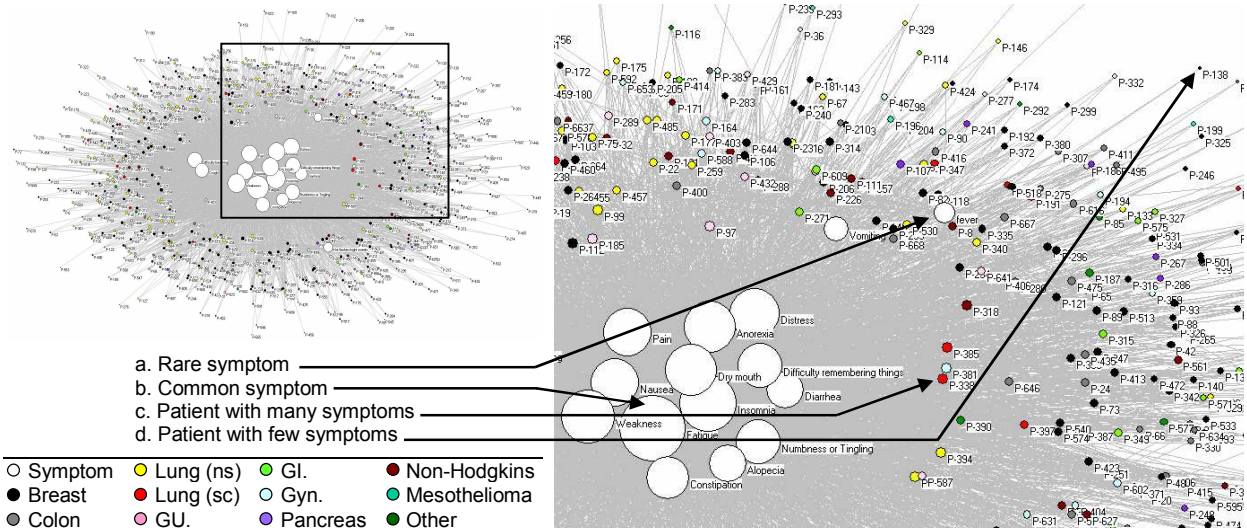

Figure 1. A bipartite network in the top left (automatically generated by the *Fruchterman Rheingold* algorithm [6]) shows the high overlap of 18 symptoms (white nodes) across 665 patients (solid colored nodes). The size of the nodes is proportional to the edges that connect to them. Therefore common symptoms have large nodes, whereas rare symptoms have smaller nodes. The inset shows the common symptoms in the center of the network, and rare symptoms that are off center. The patients that have many symptoms are closer to the center and closer to the symptoms they have; the colors represent each patient's type of cancer.

how symptoms co-occurred across all 665 patients in a single observation, and use insights from that analysis for more focused partitioning of the data (e.g., controlling for disease stage) in the future.

Data Representation. Networks are increasingly being used to analyze a wide range of phenomena, such as how diseases relate to genes [6]. A network is a graph consisting of nodes and edges; nodes represent one or more types of entities (e.g., patients or symptoms), and edges between the nodes represent a specific relationship between the entities (e.g., a patient has reported a symptom). Figure 1 shows a bipartite network (where edges exist only between two different types of entities) of patients and their symptoms. Our analysis first considers an unweighted network, with edges indicating symptom prevalence at any severity. We then analyze a network with edges weighted by severity (1-10). Finally, nodes are colored to represent disease type (e.g, black nodes represent breast cancer patients).

Networks have two advantages for analyzing complex relationships. (1) They do not require *a priori* assumptions about the data, such as whether the data are hierarchically clustered or contain fuzzy clusters. Instead, by using a simple pair-wise representation of nodes and edges, networks enable identification of complex relationships using the above representation. (2) They can be rapidly visualized and analyzed using a set of network algorithms to reveal global regularities in the data. For example, Figure 1 shows how a *force-directed* layout algorithm [5] helps to visualize the relationship between diseases and genes. The algorithm pulls together nodes that are tightly connected, and pushes apart nodes that are not. As shown, the result is that patients that have similar symptoms (e.g., *P-338* and *P-381* on the right hand side of the symptoms in Figure 1) are placed close to each other, and close to their symptoms (e.g., *Difficulty Remembering Things*). The networks were created using *Pajek* (version 1.24).

Data Analysis. We used 6 visual and statistical analysis methods: (1) To understand the overall relationship of patients and symptoms we visually analyzed the bipartite network. (2) To provide a quantitative verification of the observed overlap of symptoms across patients, we plotted the mean number of patients sharing symptom sets of different sizes. (3) To assess the degree of clustering in the network, we used the *RGraph* algorithm [7] which measures modularity (existence of clusters) in bipartite networks. (4) To understand the structure of symptom co-occurrence, we transformed the bipartite network using a method called a *one-mode projection* [6]. Here, all patient nodes were removed, and an edge was placed between two symptoms if they shared one or more patients as shown in Figure 3. The resulting network represented how pairs of symptoms co-occurred across patients. (5) To verify our observations of the structure underlying symptom co-occurrence, we used agglomerative hierarchical clustering [3] using the Ward2 clustering method. To identify the sets of symptoms that co-occur *together*

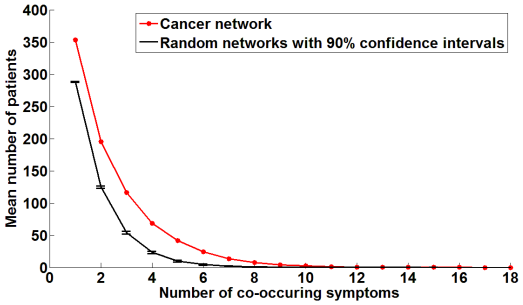

Figure 2. The mean number of symptoms shared by different numbers of patients suggests that many patients share 1-4 symptoms, and a decreasing number of patients share more than 4 symptoms. The area under the curve is significantly different from the curve of random networks of the same size.

across patients, we developed an algorithm to identify the most frequent co-occurring symptoms of different sizes. (6) To understand the role of symptom severity, we redid the above analyses with severity scores. To test the significance of all our findings, we compared them to random bipartite networks of the same size.

Results

We first analyzed a bipartite network where the edges had a weight of 1 representing severity at any level. The analysis revealed six patterns related to cancer patients and symptoms:

1. Many common symptoms, few rare symptoms. As shown in Figure 1, the bipartite network visually represents the explicit relationships between the 665 patients and 18 symptoms. The size of a node is proportional to its *degree* (number of edges that connect to that node), and the color of the nodes represent the cancer types. There are 15 commonly-occurring symptoms in the center of the network, and 3 less common symptoms off center. For example, *Fatigue* (a) is the most commonly occurring symptom with 602 edges each connected to a patient. In contrast, *Fever* (b) is off center with only 64 edges. This pattern of connections results in a high mean and standard deviation in symptom degree (Mean=287.61, SD=132.68).

2. High overlap of symptoms across patients. The patients form a ring around the 18 symptoms in the center. Patients close to the inner set of symptoms have many symptoms compared to patients in the outer ring. For example, the patient *P-338* (c) has 16 symptoms, whereas the patient *P-138* (d) has 1 symptom. This pattern of connections also results in a high mean and standard deviation in the degree of patients (Mean=7.78, SD=3.20). This network topology where there are many high degree patients (in the ring) connecting to a smaller number of high degree symptoms (in the center), suggests a high overlap in the number of symptoms for most patients (resulting in a gray mass of indistinguishable edges). This is verified through an analysis of curves shown in Figure 2, which plots the mean number patients sharing symptom sets of different sizes. The symptom overlap in the cancer network (as measured by the area under this curve) was significantly ($p < .01$) more than the overlap in 1000 random networks of the same size.

3. Absence of patient or symptom clusters. Figure 1 shows the absence of patient, symptom, or patient-symptom clusters. Most of the symptoms are in an indistinguishable mass in the center, and the patients and cancer types are evenly distributed around the symptoms. Modularity, as measured by the *RGraph* algorithm [7], was extremely low at 0.067 (cooling factor [c]=0.999, iteration factor [f]=1) for symptoms, indicating that the symptoms exhibit no significant clustering beyond what would be expected by chance.

4. Hierarchy of symptom occurrence. The range in symptom degree, with a high overlap across patients, and absence of disjoint clusters suggested that the symptoms were hierarchically structured. To verify this observation we analyzed the one-mode projection that showed how symptoms co-occurred. Figure 3 shows the pair-wise relationship between symptoms, where the edge weights between two nodes represent how many times the connected symptoms co-occurred. As shown, there are highly co-occurring symptoms at the core the network (*Fatigue* and *Insomnia* co-occur as a pair most frequently with 442 patients), with a systematic decrease in the edge weights toward the periphery (*Fever* and *Vomiting* co-occur the most infrequently in 12 patients). Besides suggesting a hierarchical structure of symptom co-occurrence, this core-periphery topology also suggests a nested structure.

5. Nested structure of symptoms. Because the one-mode projection is designed to show only the pair-wise association between symptoms, it cannot reveal the boundaries of sets of symptoms, which is required to reveal the nested nature of the symptoms. We therefore generated a dendrogram by using the agglomerative hierarchical clustering method. As shown in Figure 3b, the depth of the resulting dendrogram is 9. In addition, it takes only 8 steps (edits) to transform this tree to a tree that is perfectly nested (maximally lopsided). This suggests that symptoms are nested in their co-occurrence pattern. Furthermore, although we could select an arbitrary cut-off point to identify disjoint clusters, there is

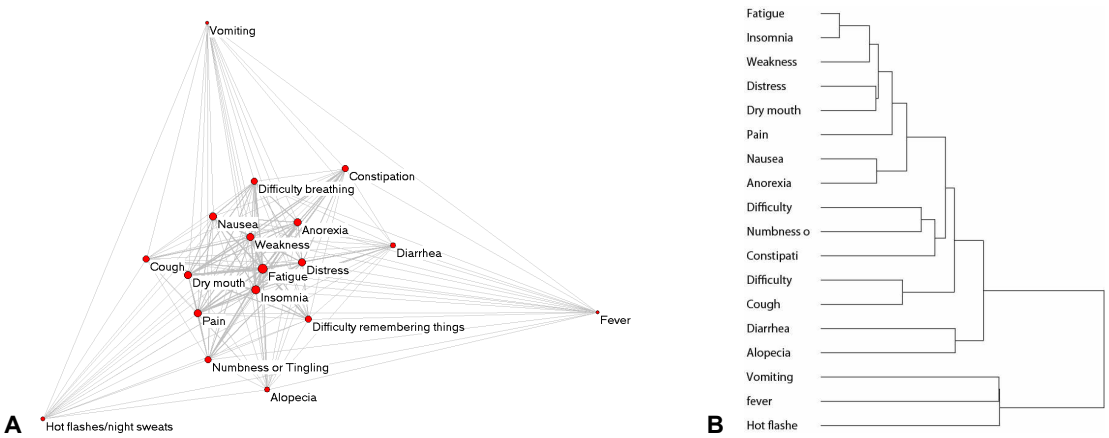

Figure 3. A. The one-mode projection of the bipartite network (shown in Figure 1), reveals how pairs of symptoms co-occur across patients. The edge thickness is proportional to the number of times two symptoms co-occur in a patient. Highly co-occurring symptom are pulled together and because of the hierarchical structure, have also been pulled to the center. **B**. The dendrogram suggests the nested structure of symptom co-occurrence.

actually no natural break in the dendrogram to reliably determine such clusters.

The above tree depth and number of edits for the network were compared against dendrograms generated from 1000 random networks of the same size. The results revealed that the probability of the nested structure of cancer symptoms occurring by chance was less than 0.1 percent ($p < 0.001$).

Because of the agglomerative nature of the dendrogram, it conceals specific co-occurrence frequencies. For example, although *Weakness* (third from the top in Figure 3B) is closest to both *Fatigue* and *Insomnia*, the method conceals how frequently *Weakness* co-occurs with either of them. We therefore used an exhaustive search algorithm to identify the most frequently co-occurring symptoms for different set sizes. Figure 4 shows the resulting block diagram that lists the most frequently co-occurring symptoms, ranging from 1 to the maximum set size of 16 co-occurring symptoms. With the exception of set sizes 7 and 11, the most frequently occurring symptom sets are a proper subset of the next larger set size. The analysis therefore verified the strongly nested nature of symptom co-occurrence.

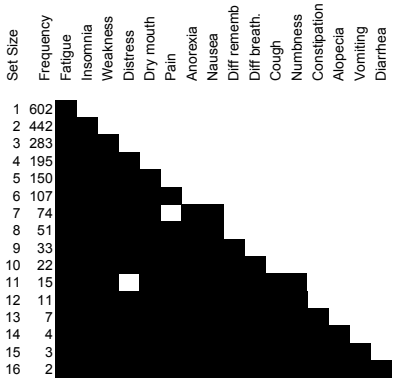

Figure 4. The most frequently co-occurring symptoms for each size of symptom set. With the exception of set sizes 7 and 11, the symptoms follow a strongly nested pattern.

6. Effect of symptom severity. Because the data contained symptom severity, we redid the entire analysis by adding symptom severity to the edge weights in the bipartite network. The main results did not change. Modularity was low at 0.1, and symptoms were strongly nested with the dendrogram exhibiting high depth and few edits from perfect nesting. Finally, we analyzed only patients with breast cancer which was the most common type of cancer in the data set. Although a detailed report is beyond the scope of the current analysis, the co-occurrence pattern of symptoms was also strongly nested.

Discussion

Based on the clinical literature, we hypothesized that our analysis would identify disjoint symptom clusters. However, no matter how we partitioned the data (by cancer type, age, etc. details of which are not reported in this paper), we repeatedly found the absence of such clusters. Disjoint clusters occur infrequently in random networks, and therefore if they occur are highly suggestive of a meaningful underlying process. Fortunately, our seemingly null results led us to probe deeper into the structure of co-occurring symptoms using multiple methods, starting with visualizations and verifying observations through existing and new quantitative methods. This exploratory process led us to the conclusion that symptom co-occur in a nested pattern rather in disjoint clusters. Furthermore, the comparison of the results with equivalent random networks has led us to conclude that cancer symptom co-occurrence is more complex than we originally

expected, but not random as we subsequently feared. These results could explain why our predecessors have found it difficult to agree on a single definition for clusters, or on their content [1].

Implications for Clinical Practice and Research

Our findings have implications for both clinical practice and future research. Currently between 15 to 27 symptoms are assessed upon every clinic visit during chemotherapy, at a time when patients are already burdened with the stress of therapy. Efficient means for assessing symptoms are therefore needed not only during office visits, but also at home where there is increasing interest in using telephonic or web-based symptom monitoring.

The absence of disjoint symptom clusters precludes a simple approach of asking a few questions to eliminate candidate symptom clusters. We therefore foresee our results applied to developing a simple computational system for symptom-assessment. The system will initially present a list of common symptoms ranked by frequency or severity. Each time a symptom is selected, the remaining symptoms are re-ranked based on their co-occurrence in the data with the already selected symptoms. For example, a patient presenting *Fatigue* **with** *Insomnia* may next be asked about *Weakness*, while a patient presenting *Fatigue* **without** *Insomnia* many next be asked about *Dry Mouth* as the latter most frequently co-occurs in patients with *Fatigue* but not *Insomnia*. Such a process should save time and reduce excess burden on the patient by obtaining a complete picture of the patient's symptoms through a small set of targeted questions.

The nested structure of cancer symptoms also suggests that the underlying biochemical mechanism in chemotherapy may involve a single mediator which causes additional symptoms as its concentration increases. Alternatively, it may involve a chain reaction where each intermediate state causes another symptom. Future research will need to confirm our results, and test such emergent hypotheses. Finally, the results imply that symptom cluster researchers should be wary of methods that have *a priori* assumptions by (1) visualizing their data to develop hypotheses about the underlying structure of symptom co-occurrence, (2) selecting appropriate multiple methods to verify observations realizing the limitations of single methods, and (3) developing new methods if current methods do not suffice.

Conclusions and Future Research

Inspired by the research on symptom clusters, but concerned by the limitations of using methods with *a priori* assumptions, we used networks to visually analyze how symptoms co-occurred across cancer patients. These observations were then verified through a series of carefully selected existing and new quantitative methods. Although the results consistently showed the absence of symptom clusters, the multi-method approach revealed a strongly nested structure, where a small set of symptoms co-occurred in many patients, and a progressively larger set of symptoms co-occurred with a decreasing number of patients. This result reveals a more complex co-occurrence organization of symptoms across patients than previously reported. The result also suggested that a computational approach, if designed carefully to fit into current work practice, could guide clinicians to ask patients a small number of questions about symptoms based on their co-occurrence in a subset of data that best matches the patient.

Because symptoms could be caused by a number of factors that change over time including the disease itself, co-morbid conditions, treatment, and other symptoms, our future research aims to probe deeper into the large number of variables that could be related to symptoms. Our aim is to help clinicians accurately identify, predict, and treat co-occurring symptoms, with the ultimate goal of improving compliance with therapy, and the overall quality of life for cancer patients.

Acknowledgements

This study is funded by NIH grant # UL1RR024986. We thank B. Given and C. Given (PIs for CA 79280 and CA 30724 respectively) for the data, and Y. Cui for assistance in processing the data.

References

1. Fan G, Filipczak L, Chow E. Symptom Clusters in Cancer Patients: A Review of the Literature. *Current Oncology* 2007; 14(5):173-179
2. Glaus A, Boehme CH, Thurlimann B, et al. Fatigue and menopausal symptoms in women with breast cancer undergoing hormonal cancer treatment. *Ann Oncol* 2006;17:801–6.
3. Chen ML, Tseng HC. Symptom clusters in cancer patients. *Support Care Cancer* 2006;14:825–30.
4. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis, 1998, NJ: Prentice-Hall.
5. Sikorskii A, Given CW, Given B, et al. Symptom management for cancer patients: a trial comparing two multimodal interventions J Pain Symptom Manage 2007; 34(3):253-64.
6. Newman M. The structure and function of complex networks. *SIAM Review* 2003; 45(2):167-256.
7. Guimera R, Sales-Pardo M, & Amaral LAN. Module identification in bipartite and directed networks, *Phys. Rev. E* 2007;76, 036102.