Brief Methodological Report

Cancer Symptom Clusters: An Exploratory Analysis of Eight Statistical Techniques

Aynur Aktas, MD, Declan Walsh, MSc, FACP, FRCP(Edin), and Bo Hu, PhD
Department of Solid Tumor Oncology, Section of Palliative Medicine and Supportive Oncology, (A.A., D.W.), Taussig Cancer Institute and Department of Quantitative Health Sciences (B.H.), Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA

Abstract

Context. Statistical methods to identify symptom clusters (SC) have varied between studies. The optimal statistical method to identify SC is unknown.

Objectives. Our primary objective was to explore whether eight different statistical techniques applied to a single data set produced different SC. A secondary objective was to investigate whether SC identified by these techniques resembled those from our original study.

Methods. We reanalyzed a symptom data set of 1000 patients with advanced cancer. Eight separate cluster analyses were conducted on both prevalence and severity of 38 symptoms. Hierarchical cluster analysis identified clusters at *r*-values of 0.6, 0.5, and 0.4. For prevalence and severity, the Spearman correlation and Kendall tau-b correlation, respectively, measured the similarity (distance) between symptom pairs. Sensitivity analysis of the prevalence data was done with Cohen kappa coefficient as a similarity measure. The *K*-means clustering method validated clusters.

Results. Hierarchical cluster analysis identified similar cluster configurations from the 38 symptoms using an *r*-value of 0.6, 0.5, or 0.4. A cutoff point of 0.6 yielded seven clusters. Five of them were identical at all three *r*-values used: 1) fatigue/anorexia-cachexia: anorexia, dry mouth, early satiety, fatigue, lack of energy, taste changes, weakness, and weight loss (>10%); 2) gastrointestinal: belching, bloating, dyspepsia, and hiccough; 3) nausea/vomiting: nausea and vomiting; 4) aerodigestive: cough, dysphagia, dyspnea, hoarseness, and wheeze; 5) neurologic: confusion, hallucinations, and memory problems. Regardless of the threshold, there were always some symptoms (e.g., pain) that did not cluster with any others. Seven clusters were validated by *K*-means analysis.

Conclusion. Seven SC identified from both prevalence and severity data were consistently present irrespective of the statistical analysis used. There were only minor variations in the number of clusters and their symptom composition between analytical techniques. All seven clusters originally identified were confirmed. Four consistent SC were found in all analyses: aerodigestive, fatigue/anorexia-cachexia, nausea/vomiting, and upper GI. Our results support

Address correspondence to: Declan Walsh, MSc, FACP, FRCP(Edin), Cleveland Clinic, The Harry R. Horvitz Center for Palliative Medicine, 9500 Euclid Avenue, Desk M76, Cleveland, OH 44195, USA. E-mail: walshtd@tcd.ie

Accepted for publication: March 10, 2014.

Published by Elsevier Inc. on behalf of American Academy of Hospice and Palliative Medicine.

0885-3924/$ - see front matter
http://dx.doi.org/10.1016/j.jpainsymman.2014.02.006

the clinical importance of the SC concept. J Pain Symptom Manage 2014;48:1254–1266. *Published by Elsevier Inc. on behalf of American Academy of Hospice and Palliative Medicine.*

Key Words
Cancer, cluster analysis, statistics, symptoms

Introduction

Previous research and clinical experience suggest that multiple cancer symptoms are inter-related and frequently occur together in groups or clusters.[1] A symptom cluster (SC) can be defined as a stable group of two or more symptoms that predictably co-occur and are independent of other clusters.[1] This implies a common or interactive mechanism among clustered symptoms. Thus, identification of symptoms in clusters may facilitate understanding of their pathophysiology and might lead to new symptom management strategies.[2]

Cluster analysis is a classification tool used to group variables (e.g., symptoms) or units (e.g., patients). The latter approach identifies homogeneous patient subgroups based on symptom prevalence or severity.[3,4] Knowledge of patient clusters may help identify high-risk populations for early intervention.[4] This approach differs from clustering symptoms because individuals can experience multiple SC.[5] SC can be determined based on a priori clinical assumptions about relationships among symptoms (nausea and vomiting) or by statistical analysis.[6] The latter are obtained from large symptom data sets.[7] These data-driven methods, for example, cluster or factor analysis, may reveal hidden data patterns. They may identify complex relationships among multiple symptoms that might otherwise remain undiscovered by routine clinical observation.

Various statistical methods have identified SC; there is no accepted best practice analytical approach.[7] Application of different methodologies to identify SC might contribute to some of the variations in SC identified.[8] Methodological differences make study comparisons and interpretation of cluster validity and reliability difficult. If inconsistencies existed, the clinical relevance of those SC identified would be questionable. Therefore, reanalysis of a single data set that used multiple (rather than one) cluster techniques would be of interest. We previously derived seven SC from 25 symptoms with a prevalence of more than 15% in advanced cancer by hierarchical cluster analysis (HCA).[9] In this report, we have extended this previous work on SC analysis. This was an exploratory study that applied eight analytical techniques to different symptom numbers, domains, or statistical methods. We also evaluated 13 additional symptoms not analyzed in the original report.[9] Aside from HCA (hierarchical method), the *K*-means (nonhierarchical analysis) was applied separately to both our prevalence and severity data to confirm the original observations.

Hierarchical cluster analysis is an exploratory technique that identifies groups of objects (e.g., individuals) or variables (e.g., symptoms) based on similarity between them. Symptoms within the same cluster resemble each other—but differ from those in other SC. Measures of similarity group the symptoms by HCA.[10] For this analysis, an agglomerative hierarchical method considered each symptom as an individual cluster; it then merged the closest pair of SC at each step until all symptoms were in a single cluster.[11] Finally, the optimum number of SC was selected of all cluster solutions. This agglomerative hierarchical structure is represented graphically as a tree-like diagram (or dendrogram) with the correlation coefficient at each joint. In this report, cluster analysis extracted SC by average linkage between groups. The distance between two SC was defined as the average distance between symptoms.

The *K*-means cluster method is preferred with large data sets. It classified a given data set through a certain number (K) of clusters determined a priori. It then assigned each symptom to a SC based on proximity to the cluster mean. Each cluster center was recomputed as the mean of the points in that SC. The process of symptom assignment and recomputation of the means was repeated until the process converged.[12] It generated a