Congruence Between Latent Class and K-modes Analyses in the Identification of Oncology Patients with Distinct Symptom Experiences

Nikoloas Papachristou, PhD(c)[1]
Payam Barnaghi, PhD[1]
Bruce A Cooper, PhD[2]
Xiao Hu, PhD[2]
Roma Maguire, BN, MSc, PhD[3]
Kathi Apostolidis[4]
Jo Armes, RN, PhD[5]
Yvette P. Conley, PhD[6]
Marilyn Hammer, PhD[7]
Stylianos Katsaragakis, RN, PhD[8]
Kord M. Kober, PhD[2]
Jon D. Levine, MD, PhD[9]
Lisa McCann, Bsc (Hon), MSc, PhD[3]
Elisabeth Patiraki, RN, PhD[10]
Steven M. Paul, PhD[2]
Emma Ream, RN, PhD[1]
Fay Wright, RN, PhD[11]
Christine Miaskowski, RN, PhD[2]

[1]School of Health Sciences, University of Surrey, Guilford, UK
[2]School of Nursing, University of California, San Francisco, CA
[3]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK
[4]European Cancer Patient Coalition, Brussels, Belgium
[5]Florence Nightingale Faculty of Nursing and Midwifery, King's College, London, UK
[6]School of Nursing, University of Pittsburgh, Pittsburgh, PS
[7]Department of Nursing, Mount Sinai Medical Center, New York, NY
[8]Faculty of Nursing, University of Peloponnese, Efstathiou & Stamatikis Valioti and Plateon, PC, Sparti, Greece
[9]School of Medicine, University of California, San Francisco, CA
[10]School of Health Sciences, National and Kapodistrian University of Athens, Athens, Greece
[11]School of Nursing, Yale University, New Haven, CT

Running title: Methods to Classify Distinct Symptom Phenotypes

Conflict of interest: The authors have no conflicts of interest to declare.

Number of tables: 7 and one supplemental table
Number of figures: 3
Number of references: 65
Number of words: 4633

This is a peer-reviewed accepted author manuscript of the following research output: Papachristou, N., Barnaghi, P., Cooper, B. A., Hu, X., Maguire, R., Apostolidis, K., ... Miaskowski, C. (2017). Congruence between latent class and k-modes analyses in the identification of oncology patients with distinct symptom experiences. Journal of Pain and Symptom Management. DOI: 10.1016/j.jpainsymman.2017.08.020

Acknowledgements: This study was funded by the National Cancer Institute (NCI, CA134900). Dr. Miaskowski is funded by grants from the American Cancer Society and NCI (CA168960). Dr. Wright is funded by a T32 grant from the National Institute of Nursing Research (NR008346). In addition, this project received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number 602289.

Address correspondence to:
Christine Miaskowski, RN, PhD
Professor
Department of Physiological Nursing
University of California
2 Koret Way – N631Y
San Francisco, CA 94143-0610
415-476-9407 (phone)
415-476-8899 (fax)
chris.miaskowski@ucsf.edu

ABSTRACT

Context: Risk profiling of oncology patients based on their symptom experience assists clinicians to provide more personalized symptom management interventions. Recent findings suggest that oncology patients with distinct symptom profiles can be identified using a variety of analytic methods.

Objectives: To evaluate the concordance between the number and types of subgroups of patients with distinct symptom profiles using latent class analysis (LCA) and K-modes analysis.

Methods: Using data on the occurrence of 25 symptoms from the Memorial Symptom Assessment Scale (MSAS), that 1329 patients completed prior to their next dose of chemotherapy (CTX), Cohen's kappa coefficient was used to evaluate for concordance between the two analytic methods. For both LCA and K-modes, differences among the subgroups in demographic, clinical, and symptom characteristics, as well as quality of life outcomes were determined using parametric and nonparametric statistics.

Results: Using both analytic methods, four subgroups of patients with distinct symptom profiles were identified (i.e., All Low, Moderate Physical and Lower Psychological, Moderate Physical and Higher Psychological, All High). The percent agreement between the two methods was 75.32% which suggests a moderate level of agreement. In both analyses, patients in the All High group were significantly younger and had a higher comorbidity profile, worse MSAS subscale scores, and poorer QOL outcomes.

Conclusion: Both analytic methods can be used to identify subgroups of oncology patients with distinct symptom profiles. Additional research is needed to determine which analytic methods and which dimension of the symptom experience provides the most sensitive and specific risk profiles.

Key words: symptom clusters; cancer; latent class analysis; machine learning; clustering; chemotherapy, k-modes analysis

INTRODUCTION

Both clinical experience and research findings suggest that oncology patients experience significant interindividual variability in their symptom experience.[1,2] In the era of precision medicine,[3] which focuses on the identification of patients who are at greater risk for chronic conditions like cancer, it is imperative that the optimal methods to risk profile patients based on their symptom burden is identified. In two reviews of the state of the science in symptom clusters research,[4,5] it was noted that future studies need to focus on an evaluation of the concordance between the various analytic methods that can be used to identify patients who are at greatest risk for a higher symptom burden.

Recent findings from our group[6-14] and others[15-18] have identified subgroups of patients with distinct symptom experiences using approaches like hierarchical cluster analysis and latent class analysis (LCA). In the earliest of these studies,[6,7,15,16] different clustering methods were used to create the patient subgroups. In the later studies,[9-14,18] LCA was the preferred analytic approach. While across these thirteen studies, the number of subgroups ranged from two to five, a common finding across all of these studies was the identification of a group of patients who reported low levels of symptoms and a group of patients who reported high levels of symptoms. However, none of these studies determined whether the use of two different analytic approaches produces congruent results (e.g., the percentages of patients in the “all high” groups are equal and are the same patients).

As noted in a recent review,[5] machine learning techniques may provide useful approaches to identify subgroups of patients with distinct symptom profiles. Some specific machine learning techniques that can be used for this purpose include: K-means,[19] K-modes,[20,21] spectral clustering,[22] birch,[23] or agglomerative hierarchical clustering (AHC).[24,25] For binary variables (e.g., symptom occurrence), K-means and K-modes are two centroid based algorithms that calculate the distance between each pair of data points using Euclidean distance or a simple dissimilarity measure (e.g., Hamming distance), respectively. The clusters derived

from K-means and K-modes analyses are described by the “centroid”, which is the multidimensional mean and mode, respectively, of the samples inside them.[19,21] Spectral clustering is a graph distance based algorithm that performs a dimensionality reduction before clustering the lower-dimension dataset in a similar fashion to K-means. It is used when the clusters are not linearly separated in the original space, providing better results than algorithms such as K-means (which tends to find spherical clusters).[26] Birch is a hierarchical clustering algorithm that can provide an advantage in datasets that are non-uniformly distributed and every data point is not equally important. It concentrates on densely occupied partitions and follows a hierarchical order of analysis that focuses on calculating and updating measurements that capture the natural closeness of data. Therefore, it is more robust to “noise” (i.e., data points that are not part of the underlying pattern).[23] Finally, AHC is a decision tree, bottom-up clustering method that starts with every single data point in a single cluster. In each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying a similarity criterion, until all of the data are in one cluster. A matrix tree plot visually demonstrates the hierarchy within the final cluster, where each merger is represented by a binary tree. AHC can be both informative for data display and helpful for the discovery of smaller clusters.[24]

No studies were identified that evaluated for congruence between two methods of classifying oncology patients based on their distinct experiences with common symptoms associated with cancer treatment. Based on how well the machine learning methods described above performed during our initial analyses,[27] for this paper, K-modes was selected as the method to compare with LCA. The purpose of this study, in a sample of patients (n=1329) who were undergoing chemotherapy (CTX) for breast, lung, gastrointestinal (GI), or gynecological (GYN) cancers was to evaluate the concordance between the number and types of subgroups of patients with distinct symptom experiences that were identified using LCA and K-modes analyses. We hypothesized that the number and types of subgroups would be similar using these two analytic methods.

METHODS

Patients and Settings

This study is part of a longitudinal study of the symptom experience of oncology outpatients receiving CTX. The methods for this study are described in detail elsewhere.[13,28,29] According to the study's eligibility criteria: patients were ≥18 years of age; had a diagnosis of breast, GI, GYN, or lung cancer; had received CTX within the preceding four weeks; were scheduled to receive at least two additional cycles of CTX; were able to read, write, and understand English; and gave written informed consent. Patients were recruited from two Comprehensive Cancer Centers, one Veteran's Affairs hospital, and four community-based oncology programs.

Instruments

A demographic questionnaire obtained information on age, gender, ethnicity, marital status, living arrangements, education, employment status, and income. The Karnofsky Performance Status (KPS) scale[30] was used to evaluate patients' functional status. The Self-administered Comorbidity Questionnaire (SCQ)[31] evaluated the occurrence, treatment, and functional impact of thirteen common comorbid conditions (e.g., diabetes, arthritis).

A modified version of the Memorial Symptom Assessment Scale (MSAS) was used to evaluate the occurrence, severity, frequency, and distress of 38 symptoms commonly associated with cancer and its treatment. In this study, six symptoms were added to the original list of 32 MSAS symptoms (i.e., hot flashes, chest tightness, difficulty breathing, abdominal cramps, increased appetite, weight gain). The MSAS is a self-report questionnaire designed to measure the multidimensional experience of symptoms. Patients were asked to indicate whether or not they had experienced each symptom in the past week (i.e., symptom occurrence). If they had experienced the symptom, they were asked to rate its frequency of occurrence, severity, and distress. The reliability and validity of the MSAS is well established in oncology patients.[32,33]

Three subscale scores (i.e., physical [MSAS-PHYS], psychological [MSAS-PSYCH], global distress index [MSAS-GDI]) were calculated. The MSAS-PHYS is the average of the frequency, severity, and distress ratings for twelve physical symptoms (i.e., lack of energy, feeling drowsy, pain, nausea, vomiting, change in the way food tastes, lack of appetite, dry mouth, constipation, feeling bloated, dizziness, and weight loss). The MSAS-PSYCH is the average of the frequency, severity, and distress ratings for six psychological symptoms (i.e., worrying, feeling sad, feeling nervous, feeling irritable, difficulty in sleeping, difficulty concentrating). The MSAS-GDI is the average of the distress ratings for six physical symptoms (i.e., lack of energy, feeling drowsy, pain, lack of appetite, dry mouth, constipation) and the frequency ratings for four psychological symptoms (i.e., worrying, feeling sad, feeling nervous, feeling irritable).

Quality of life (QOL) was evaluated using disease-specific (i.e., Quality of Life Scale-Patient Version (QOL-PV))[34-36] and generic (i.e., Medical Outcomes Study-Short Form-12 (SF-12))[37] measures. The QOL-PV is a 41-item instrument that measures four dimensions of QOL (i.e., physical, psychological, social, and spiritual well-being) in oncology patients, as well as a total QOL score. Each item is rated on a 0 to 10 numeric rating scale (NRS) with higher scores indicating a better QOL. The QOL-PV has established validity and reliability.[36,38-40]

The SF-12 consists of 12 questions that evaluate physical, mental, and overall health status. Individual items on the SF-12 are evaluated. In addition, the instrument is scored into physical component summary (PCS) and mental component summary (MCS) scores. These scores can range from 0 to 100. Higher PCS and MCS scores indicate a better QOL. The SF-12 has well established validity and reliability.[37]

Study Procedures

The study was approved by the Committee on Human Research at the University of California, San Francisco and by the Institutional Review Board at each of the study sites. Written informed consent was obtained from all patients. For this analysis, symptom occurrence

data from the enrollment assessment, that asked patients to report on their symptom experience for the week prior to the administration of the next cycle of CTX, were analysed (i.e., recovery from previous CTX cycle).

Data Analyses

Symptom Occurrence Data

In order to have a sufficient number of patients who endorsed each symptom, the LCA and K-modes analyses were done with the 25 symptoms that occurred in ≥30% of the patients (i.e. difficulty concentrating, pain, lack of energy, cough, feeling nervous, hot flashes, dry mouth, nausea, numbness or tingling in hands or feet, feeling drowsy, difficulty sleeping, feeling bloated, diarrhea, feeling sad, sweats, problems with sexual interest or activity, worrying, lack of appetite, dizziness, feeling irritable, hair loss, constipation, change in the way food tastes, I do not look like myself, changes in skin).

Latent Class Analysis

LCA identifies latent classes based on an observed response pattern.[41,42] It is a statistical method for finding subtypes of related cases (i.e., latent classes) from multivariate categorical data. The LCA was performed using Mplus™ Version 7.[43] Estimation was carried out with robust Maximum-Likelihood (MLR) and the Expectation-Maximization (EM) algorithm.[44] The optimal number of latent classes for this LCA was selected based on the Bayesian Information Criterion (BIC), the Vuong, Lo, Mendel, and Rubin (VLMR) likelihood ratio test, and entropy. Theoretically, the best fitting LCA model has the lowest BIC. Nevertheless, the BIC can be supplemented by an evaluation of the VLMR[45] which tests whether a model with K classes fits the data better than a model with one fewer class (the K-1 class model). When this VLMR is significant, the K-class model is considered to be a better fit for the data. When models are evaluated sequentially, with each new model having one more class than the previous model, if a model is identified for which the VLMR is not significant, then too many classes were extracted and the K-1 class model is considered to fit the data better than the current K-class

model. Furthermore, well-fitting models produce entropy values of ≥0.80.[46] In addition, the optimal fitting model should "make sense" conceptually and its classes should differ as might be expected on variables not used in the generation of the model.

K-modes analysis

K-modes is a centroid method that is optimized for use with categorical variables.[21] It defines clusters based on the number of matching categories between data points and not on their Euclidean distance (a common similarity index in agglomerative clustering methods). Although its performance is comparable to K-means,[27] the K-modes distance measurement approach is theoretically a more appropriate approach to use to cluster the categorical variable of symptom occurrence.[21,47] The K-modes analysis was implemented with PyCharm Professional Edition 4.5 and the Scikit-Learn library.[48]

The optimal number of clusters for the K-modes analysis was assessed using the Silhouette Coefficient (SC).[49] The SC represents how well each case (i.e., patient) lies within its cluster and how appropriate each case's assignment is inside a specific cluster. The average SC, called the Silhouette Index (SI), allows one to evaluate the overall quality of the separation between the clusters. The SC is calculated using its intra-cluster distance and its nearest-cluster distance.[27] The SC is bounded between -1 for inappropriate clustering and +1 for highly compact clustering. A SC around zero indicates that a case is assigned inside overlapping clusters. In general, the average SI is high when clusters are dense and well separated.

Evaluation of Congruence

In order to evaluate the congruence between the LCA and K-modes solutions (i.e., number of subgroups identified), we compared the solutions using SCI diagrams (see Figures 1A and 1B, respectively).[49] When the SC for a case is >0, its assignment to this cluster is considered appropriate. When the SC for a case is $\leq$0, this case may have equal similarities with cases in another, overlapping cluster and its assignment inside a specific cluster may not

be an appropriate fit. In addition, Cohen's kappa coefficient was used to evaluate the agreement between the two analytic approaches.

Differences in Demographic, Clinical, and Symptom Characteristics and QOL Outcomes

Descriptive statistics and frequency distributions were calculated for demographic and clinical characteristics using SPSS version 23 (IBM, Armonk, NY). For each analytic approach, differences in demographic and clinical characteristics and QOL outcomes, among the groups, were evaluated using analyses of variance, Kruskal-Wallis, and Chi Square analyses. Post hoc contrasts were calculated using the Bonferroni corrected alpha of 0.008 (0.05/6 pairwise comparisons).

RESULTS

Number of Subgroups Identified Using LCA and K-modes Approaches

For the LCA, the fit indices for the candidate models are shown in Table 1. The four class solution was selected because its BIC was lower than for the 3- and 5-class solutions. In addition, the VLMR indicated that a 4-class solution was better than a 3-class solution. However, the VLMR for the 5-class solution was not better than the 4-class solution indicating that too many classes were extracted.

Using K-modes, while the average SI for the 3-class solution was slightly larger than the average SI for the 4-class solution (Table 2), given this trivial difference and in order to compare the differences in demographic, clinical, and symptom characteristics and QOL outcomes between the two methods, we used the 4-class solution from the K-modes analysis.

As shown in Figures 2 and 3, for the LCA and K-modes analyses, respectively, the four subgroups were named based on the probability of occurrence of the 25 MSAS symptoms that occurred in $\geq$30% of the patients. The All High and All Low groups included patients who reported relatively high and low occurrence rates for most of the 25 MSAS symptoms, respectively. The Moderate Physical and Higher Psychological and Moderate Physical and Lower Psychological groups included patients who reported relatively moderate occurrence

rates for the majority of the physical symptoms and relatively higher or lower occurrence rates, respectively, for the five psychological symptoms (i.e., worrying, feeling irritable, feeling sad, feeling nervous, I don't look like myself).

The SC diagrams for all of the patient cases within each of the 4 clusters for the LCA and K-modes analyses (Figures 1A and 1B) showed that their inefficient assignments were mostly within two specific groups (i.e. Moderate Physical and Higher Psychological, Moderate Physical and Lower Psychological). Both well (SC >0) and inappropriately (SC $\leq$0) clustered cases were included within these clusters. As illustrated in the SC diagrams, K-modes assigned a larger proportion of cases to these two groups (SC >0). Of note, the two other groups (All Low, All High) were well defined and separated using both the LCA and K-modes approaches (SC >0.4).

Pairwise Agreement Between the LCA and K-modes Approaches

As shown in Table 3, the observed agreement among the four groups was 75.32% and the expected agreement was 26.08%. The two analyses separated patients into 4 distinct groups with substantial agreement beyond chance (range 0.6-0.7) as measured by the Cohen's coefficient (kappa=0.666).(50) The biggest disagreements between the LCA and K-modes approaches were between: a) the Moderate Physical and Lower Psychological (LCA) and All Low (K-modes) and b) the Moderate Physical and Higher Psychological (LCA) and All High (K-modes) groups, with 92 and 101 divergent classifications, respectively.

Group Characteristics Identified with LCA and K-modes Approaches

The All Low group consisted of 31.5% (n=419) of the sample using LCA and 40.3% (n=536) using K-modes. The probability of occurrence of the MSAS symptoms for this group ranged from 0.064 to 0.549 for LCA and 0.093 to 0.647 for K-modes.

The second largest group identified using LCA was named Moderate Physical and Higher Psychological and consisted of 31.3% (n=416) of the sample. Using K-modes, this group consisted of 21.1% (n=280) of the patients. The occurrence rates for the majority of the physical

symptoms ranged from 0.293 to 0.930 for LCA and from 0.236 to 0.939 for K-modes. For the psychological symptoms, the occurrence rates were relatively high. They ranged from 0.541 to 0.906 for LCA and from 0.582 to 0.811 for K-modes.

The third largest group identified using LCA (23.8%, n=316) was named the Moderate Physical and Lower Psychological group. Using K-modes, this group was the smallest one identified (15.4%, n=205). The probability of occurrence for the physical symptoms ranged from 0.241 to 0.987 for LCA and from 0.210 to 0.956 for K-modes. For the psychological symptoms, the range was from 0.142 to 0.282 for LCA and from 0.185 to 0.278 for K-modes.

The All High group was the smallest one for LCA (13.4%, n=178) and the second largest for the K-modes analysis (23.2%, n=308). The probability of occurrence of the MSAS symptoms for this group ranged from 0.562 to 0.994 for LCA and from 0.429 to 0.974 for K-modes.

Differences in Patient Characteristics Among the Groups Identified with LCA and K-modes Approaches

Tables 4 and 5 summarize the differences in demographic and clinical characteristics among the four groups of patients identified using LCA and K-modes, respectively. For both analyses, compared to the “All Low” group, patients in the “Moderate Physical and Higher Psychological” and the “All High” groups were significantly younger, had a lower KPS score, had a higher SCQ score, were more likely to have breast cancer, and were more likely to report depression and back pain. In addition, for both analyses, compared to the “Moderate Physical and Lower Psychological” group and the “Moderate Physical and Higher Psychological” group, patients in the “All High” group had a lower KPS score and a higher SCQ score.

Differences in Symptom Occurrence Rates Among the Groups Identified with LCA and K-modes

Supplemental Table 1 summarizes differences in symptom occurrence rates among the four groups of patients identified using LCA and K-modes. Both analyses identified two groups of oncology patients who reported moderate levels of physical symptoms but differentiated on

the occurrence of five psychological symptoms (i.e., worrying, feeling irritable, feeling sad, feeling nervous, I don't look like myself). For patients in the Moderate Physical and Higher Psychological group, worrying (LCA: 0.906, K-modes: 0.811), feeling sad (LCA: 0.813, K-modes: 0.811), and feeling irritable (LCA: 0.649, K-modes: 0.657) were among the top symptoms. In contrast, in the Moderate Physical and Lower Psychological group, worrying (LCA: 0.142, K-modes: 0.278), feeling sad (LCA: 0.161, K-modes: 0.259), and feeling irritable (LCA: 0.256, K-modes: 0.224) were among the symptoms with the lowest probability of occurrences. The remaining psychological symptoms, namely: "feeling nervous" (Moderate Physical and Higher Psychological group: LCA: 0.606, K-modes: 0.693; Moderate Physical and Lower Psychological group: LCA: 0.184, K-modes: 0.185) and "I don't look like myself" (Moderate Physical and Higher Psychological group: LCA: 0.541, K-modes: 0.582; Moderate Physical and Lower Psychological group: LCA: 0.282, K-modes: 0.259) had significant differences between the aforementioned groups for both analyses.

Across all four groups, lack of energy was the most common symptom. While the probability of its occurrence for the total sample was 0.832, values ranged from 0.549 to 0.994 for LCA and from 0.647 to 0.974 for K-modes. In addition, pain (LCA: 0.944-0.334, K-modes: 0.834-0.360), difficulty in sleeping (LCA: 0.927-0.458, K-modes: 0.896-0.537), numbness/tingling in hands/feet (LCA: 0.798-0.334, K-modes: 0.724-0.356), change in the way food tastes (LCA: 0.837-0.274, K-modes: 0.802-0.323), and feeling drowsy (LCA: 0.966-0.243, K-modes: 0.860-0.321) occurred in the top ten symptoms across all four groups for both analyses.

Differences in MSAS Summary Scores Among the Groups Identified with LCA and K-modes

Table 6 summarizes differences in the MSAS summary scores among the four groups of patients identified using LCA and K-modes. For the Physical subscale, the Psychological subscale, and the Global Distress index, the differences among the four groups followed the

same pattern for both analyses. For the MSAS total score, as well as for the total number of MSAS symptoms, the pattern observed using the LCA was in the expected direction (i.e., All Low < Moderate Physical and Lower Psychological < Moderate Physical and Higher Psychological < All High). For the MSAS total score, as well as for the total number of MSAS symptoms, the pattern observed using K-modes was as follows: All Low < Moderate Physical and Lower Psychological, Moderate Physical and Higher Psychological and All High (i.e., 0 < 1, 2, and 3), as well as Moderate Physical and Lower Psychological and Moderate Physical and Higher Psychological < All High (i.e., 1 and 2 < 3).

Differences in QOL Scores Among the Groups Identified with LCA and K-modes

Table 7 summarizes differences in MQOLS-CA subscale and total scores among the four groups of patients identified using LCA and K-modes. For the MQOLS psychological and social well-being subscales, and total QOL scores, the differences among the four groups followed the same pattern for both analyses (i.e., All Low > Moderate Physical and Lower Psychological > Moderate Physical and Higher Psychological > All High). In addition, for the physical well-being subscale scores, the differences among the four groups followed the same pattern for both analyses (i.e., All Low > Moderate Physical and Lower Psychological, Moderate Physical and Higher Psychological, and All High (i.e., 0 > 1, 2, and 3) and Moderate Physical and Lower Psychological and Moderate Physical and Higher Psychological > All High (i.e., 1 and 2 > 3)).

For the SF12, for both analyses, the MCS scores followed a similar pattern (i.e., All Low > Moderate Physical and Lower Psychological > Moderate Physical and Higher Psychological > All High). For the PCS scores, the post hoc contrasts were different depending on the method of analysis. For LCA, the pattern was All Low > Moderate Physical and Higher Psychological > Moderate Physical and Lower Psychological > All High. For the K-modes analysis, the pattern was as follows: All Low > Moderate Physical and Lower Psychological, Moderate Physical and Higher Psychological and All High (i.e., 0 > 1, 2, and 3), as well as Moderate Physical and

Higher Psychological > Moderate Physical and Lower Psychological and All High (i.e., 2 > 1 and 3).

DISCUSSION

This study is the first to evaluate for congruence between the ability of two different analytic approaches to identifiy subgroups of oncology patients with distinct symptom profiles. Using both LCA and K-modes, four groups of patients with distinct symptom profiles were identified. The Cohen's kappa coefficient of 0.666 represents a moderate level of agreement between the two approaches.[51-53] Potential reasons for only a moderate level of agreement may be related to differences in the underlying assumptions of each of the methods. LCA is a model based approach where "clusters" (i.e. classes) are defined by parametric probability distributions that can be interpreted to generate homogenous points, while the whole data set is modelled by a mixture of such distributions.[54] Its key assumption is the conditional independence of the observed variables given the latent class. Inside the same class, the presence or the absence of one symptom is viewed as unrelated to the presence or absence of all of the others. On the other hand, K-modes is a distance-based clustering method that separates clusters as data subsets that have small within-cluster distances and large separation from other clusters. K-modes tries to find clusters that bring similar observations together without making an assumption about their distribution or attempt to fit a mixture distribution. Our findings, as well as others,[54-56] suggest that further research is needed, using both approaches, to determine the most sensitive and specific method(s) to risk profile oncology patients based on symptom occurrence rates.

While the absolute percentages of patients in the four groups differed depending on the analytic approach, the specific symptom profiles within each of the four groups were very similar. In addition, previous work in heterogeneous samples of oncology patients, using a different numbers of MSAS symptoms,[9,57] found the same four phenotypic profiles identified in the current study. Across these three studies, the percentage of patients in the All Low group

ranged from 28.0%[9] to 40.3% (using K-modes in the current study) and the percentage of patients in the All High class ranged from 13.4% (using LCA in the current study) to 27.8%.[57] Across these three studies, these relatively wide ranges may be related to differences in the number and types of symptoms evaluated, the timing of the symptom assessments in relationship to cancer diagnosis and treatments, and/or the specific cancer diagnoses of the patients in each of the studies. That said, these two extreme phenotypes were identified in previous studies that used only four symptoms[6,7,10,11] or identified only two or three groups.[15-17]

Across the two previous studies[9,57] and with the two analytic methods used in the current study, the consistent phenotypic characteristics associated with membership in the All High group were younger age and poorer functional status. The association between younger age and a higher symptom burden is consistent with previous studies.[6,7] While younger patients may receive more aggressive cancer treatments,[58] equally plausible hypotheses for this association include: that older adults experience a "response shift" in their perception of symptoms;[59] that chronological age may not be an accurate representation of the biological age of oncology patients;[60] and/or that accelerated aging occurs with cancer and its treatment.[61-63]

Similar to age, the association between a higher symptom burden and poorer functional status was reported previously.[11,16,18] In the current study and in the one conducted in Norway,[57] that both used the KPS scale, compared to patients in the All Low group who had KPS scores between 85 and 95, patients in the All High group reported KPS scores in the mid-70s. This difference represents a clinically meaningful change in functional status on this scale. Given that patients typically report lower KPS scores than their clinicians,[64,65] patients should be interviewed not only about the number and severity of their symptoms but about changes in functional status during and following cancer treatment.

An equally important finding in this study and in the two previous studies[9,57] is the identification of two groups of patients who differentiated based on the occurrence of psychological symptoms. While our phenotypic data suggest that these two groups have lower

KPS scores and a higher comorbidity profile than the All Low group and better scores for both characteristics than the All High group, the demographic and clinical characteristics that distinguish between these two “Moderate” groups are not readily apparent. These findings are similar to previous reports[9,57] and warrant investigation in future studies. An evaluation of additional psychosocial characteristics (e.g., coping styles, personality, social support) may improve the phenotypic characterization of these two “Moderate” groups.

In terms of the QOL outcomes, regardless of whether a generic (i.e., SF12) or disease-specific (i.e., MQOLS-PV) measure was used, the pattern of the differences in scores were in the expected direction, namely that as the symptom phenotype worsened, QOL decreased. The one interesting finding on Table 7, relates to the PCS scores from the SF12. While none of the groups had PCS scores of $\geq$50 (i.e., the normative value for the general population in the United States), patients in the Moderate Physical and Lower Psychological group had worse scores than patients in the Moderate Physical and Higher Psychological group. This finding is consistent with the report by Astrup and colleagues.[57] Additional research is warranted to explain this finding and to determine the specific phenotypic characteristics that distinguish between these two Moderate groups.

In terms of study limitations, patients were recruited at various points in their CTX treatment. In addition, the types of CTX were not homogeneous. While we cannot rule out the potential contributions of clinical characteristics to patients' symptom experiences, the relatively similar percentages of cancer diagnoses, reasons for current treatment, time since cancer diagnosis, and evidence of metastatic disease across the four groups, suggest that the patients were relatively similar in terms of disease and treatment characteristics. Although it is possible that patients in the “All Low” group were receiving more aggressive symptom management interventions, the occurrence rates for the five most common symptoms were relatively similar across the four classes for both analyses. It is possible that using ratings of frequency, severity

or distress to create patients groups would provide additional information on inter-individual differences in the symptom experience of these patients.

Additional research is warranted using different analytic methods to optimize the identification of oncology patients with a higher symptom burden. Future studies can evaluate different machine learning approaches, as well as real time collection of different dimensions of a patient's symptom experience (i.e., occurrence, severity, distress) to determine the most sensitive and specific methods to use to risk profile patients and design and test more effective symptom management interventions.

REFERENCES

1. Esther Kim JE, Dodd MJ, Aouizerat BE, Jahan T, Miaskowski C. A review of the prevalence and impact of multiple symptoms in oncology patients. J Pain Symptom Manage 2009;37:715-736.

2. Reilly CM, Bruner DW, Mitchell SA, et al. A literature synthesis of symptom prevalence and severity in persons receiving active cancer treatment. Support Care Cancer 2013;21:1525-1550.

3. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793-795.

4. Miaskowski C. Future directions in symptom cluster research. Semin Oncol Nurs 2016;32:405-415.

5. Miaskowski C, Barsevick A, Berger A, et al. Advancing symptom science through symptom cluster research: Expert panel proceedings and recommendations. J Natl Cancer Inst 2017;109.

6. Miaskowski C, Cooper BA, Paul SM, et al. Subgroups of patients with cancer with different symptom experiences and quality-of-life outcomes: a cluster analysis. Oncol Nurs Forum 2006;33:E79-89.

7. Pud D, Ben Ami S, Cooper BA, et al. The symptom experience of oncology outpatients has a different impact on quality-of-life outcomes. J Pain Symptom Manage 2008;35:162-170.

8. Illi J, Miaskowski C, Cooper B, et al. Association between pro- and anti-inflammatory cytokine genes and a symptom cluster of pain, fatigue, sleep disturbance, and depression. Cytokine 2012;58:437-447.

9. Miaskowski C, Dunn L, Ritchie C, et al. Latent class analysis reveals distinct subgroups of patients based on symptom occurrence and demographic and clinical characteristics. J Pain Symptom Manage 2015;50:28-37.

10. Langford DJ, Paul SM, Cooper B, et al. Comparison of subgroups of breast cancer patients on pain and co-occurring symptoms following chemotherapy. Support Care Cancer 2016;24:605-614.

11. Dodd MJ, Cho MH, Cooper BA, et al. Identification of latent classes in patients who are receiving biotherapy based on symptom experience and its effect on functional status and quality of life. Oncol Nurs Forum 2011;38:33-42.

12. Doong SH, Dhruva A, Dunn LB, et al. Associations between cytokine genes and a symptom cluster of pain, fatigue, sleep disturbance, and depression in patients prior to breast cancer surgery. Biol Res Nurs 2015;17:237-247.

13. Miaskowski C, Cooper BA, Aouizerat B, et al. The symptom phenotype of oncology outpatients remains relatively stable from prior to through 1 week following chemotherapy. Eur J Cancer Care (Engl) 2016.

14. Miaskowski C, Cooper BA, Melisko M, et al. Disease and treatment characteristics do not predict symptom occurrence profiles in oncology outpatients receiving chemotherapy. Cancer 2014;120:2371-2378.

15. Ferreira KA, Kimura M, Teixeira MJ, et al. Impact of cancer-related symptom synergisms on health-related quality of life and performance status. J Pain Symptom Manage 2008;35:604-616.

16. Gwede CK, Small BJ, Munster PN, Andrykowski MA, Jacobsen PB. Exploring the differential experience of breast cancer treatment-related symptoms: a cluster analytic approach. Support Care Cancer 2008;16:925-933.

17. Reese JB, Blackford A, Sussman J, et al. Cancer patients' function, symptoms and supportive care needs: a latent class analysis across cultures. Qual Life Res 2015;24:135-146.

18. Snyder CF, Garrett-Mayer E, Blackford AL, et al. Concordance of cancer patients' function, symptoms, and supportive care needs. Qual Life Res 2009;18:991-998.

19. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics 2007:1027-1035.

20. Cao F, Liang J, Bai L. A new initialization method for categorical data clustering. Expert Systems with Applications 2009;36:10223-10228.

21. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 1998;2:283-304

22. Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 2002;2:849-856.

23. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Record 1996;25:103-114.

24. Sasirekha K, Baby P. Agglomerative Hierarchical Clustering Algorithm – A review. International Journal of Scientific and Research Publications 2013;3:1-3.

25. Zhao Y, Karypis G, Fayyad U. Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery 2005;10:141-168.

26. Dhillon IS, Guan Y, Kulis B. A unified view of kernel k-means, spectral clustering and graph cuts. In: UTCS Technical Report TR-04-25, 2005.

27. Papachristou N, Miaskowski C, Barnaghi P, et al. Comparing machine learning clustering with latent class analysis on cancer symptoms' data. Proceedings of the IEEE Healthcare Innovation Point-of-Care Technologies Conference, 2016.

28. Wright F, Hammer M, Paul SM, et al. Inflammatory pathway genes associated with inter-individual variability in the trajectories of morning and evening fatigue in patients receiving chemotherapy. Cytokine 2017;91:187-210.

29. Kober KM, Cooper BA, Paul SM, et al. Subgroups of chemotherapy patients with distinct morning and evening fatigue trajectories. Support Care Cancer 2016;24:1473-1485.

30. Karnofsky D, Abelmann WH, Craver LV, Burchenal JH. The use of nitrogen mustard in the palliative treatment of cancer. Cancer 1948:634-656.

31. Sangha O, Stucki G, Liang MH, Fossel AH, Katz JN. The Self-Administered Comorbidity Questionnaire: a new method to assess comorbidity for clinical and health services research. Arthritis Rheum 2003;49:156-163.

32. Portenoy RK, Thaler HT, Kornblith AB, et al. Symptom prevalence, characteristics and distress in a cancer population. Qual Life Res 1994;3:183-189.

33. Portenoy RK, Thaler HT, Kornblith AB, et al. The Memorial Symptom Assessment Scale - an instrument for the evaluation of symptom prevalence, characteristics and distress. Eur J Cancer 1994;30a:1326-1336.

34. Ferrell BR, Wisdom C, Wenzl C. Quality of life as an outcome variable in the management of cancer pain. Cancer 1989;63:2321-2327.

35. Padilla GV, Grant MM. Quality of life as a cancer nursing outcome variable. Adv Nurs Sci 1985;8:45-60.

36. Padilla GV, Presant C, Grant MM, et al. Quality of life index for patients with cancer. Res Nurs Health 1983;6:117-126.

37. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34:220-233.

38. Padilla GV, Ferrell B, Grant MM, Rhiner M. Defining the content domain of quality of life for cancer patients with pain. Cancer Nurs 1990;13:108-115.

39. Ferrell BR, Dow KH, Grant M. Measurement of the quality of life in cancer survivors. Qual Life Res 1995;4:523-531.

40. Ferrell BR. The impact of pain on quality of life. A decade of research. Nurs Clin North Am 1995;30:609-624.

41. Collins LM, Lanza ST. Latent class and latent transition analysis: with applications in the Social, Behavioral, and Health Science, Hoboken, NJ: John Wiley & Sons, 2010.

42. Nylund K, Bellmore A, Nishina A, Graham S. Subtypes, severity, and structural stability of peer victimization: what does latent class analysis say? Child Dev 2007;78:1706-1722.

43. Muthen LK, Muthen BO. Mplus (Version 7.4), Los Angeles, CA: Muthen & Muthen, 2015.

44. Muthen B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 1999;55:463-469.

45. Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural equation modeling 2007;14:535-569.

46. Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification 1996;13:195-212.

47. Ordonez C. Clustering binary data streams with K-means. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, 2003:12-19.

48. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825-2830.

49. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987;20:53-65.

50. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

51. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm 2013;9:330-338.

52. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276-282.

53. Steinijans VW, Diletti E, Bomches B, Greis C, Solleder P. Interobserver agreement: Cohen's kappa coefficient does not necessarily reflect the percentage of patients with congruent classifications. Int J Clin Pharmacol Ther 1997;35:93-95.

54. Anderlucci L, Hennig C. The clustering of categorical data: a comparison of a model-based and a distance-based approach. Communications in Statistics-Theory and Methods 2014;43:704-721.

55. Hennig C, Liao TF. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C 2013;62:309-369.

56. Oberski DL. Beyond the number of classes: separating substantive from non-substantive dependence in latent class analysis. Advances in Data Analysis and Classification 2016;10:171-182.

57. Astrup GL, Hofso K, Bjordal K, et al. Patient factors and quality of life outcomes differ among four subgroups of oncology patients based on symptom occurrence. Acta Oncol 2017:1-9.

58. Klepin HD, Rodin M, Hurria A. Treating older adults with cancer: geriatric perspectives. Am Soc Clin Oncol Educ Book 2015;35:e544-552.

59. Sprangers MA, Schwartz CE. The challenge of response shift for quality-of-life-based clinical oncology research. Ann Oncol 1999;10:747-749.

60. Bae CY, Kang YG, Piao MH, et al. Models for estimating the biological age of five organs using clinical biomarkers that are commonly measured in clinical practice settings. Maturitas 2013;75:253-260.

61. Henderson TO, Ness KK, Cohen HJ. Accelerated aging among cancer survivors: from pediatrics to geriatrics. Am Soc Clin Oncol Educ Book 2014:e423-430.

62. Hurria A, Jones L, Muss HB. Cancer treatment as an accelerated aging process: Assessment, biomarkers, and interventions. Am Soc Clin Oncol Educ Book 2016;35:e516-522.

63. Ness KK, Krull KR, Jones KE, et al. Physiologic frailty as a sign of accelerated aging among adult survivors of childhood cancer: a report from the St Jude Lifetime cohort study. J Clin Oncol 2013;31:4496-4503.

64. Schnadig ID, Fromme EK, Loprinzi CL, et al. Patient-physician disagreement regarding performance status is associated with worse survivorship in patients with advanced cancer. Cancer 2008;113:2205-2214.

65. Ando M, Ando Y, Hasegawa Y, et al. Prognostic value of performance status assessed by patients themselves, nurses, and oncologists in advanced non-small cell lung cancer. Br J Cancer 2001;85:1634-1639.

Figure legends

Figure 1A - Silhouette coefficient diagram for the 4-class solution using latent class analysis. The sizes of the clusters in the diagram are proportional to their size inside the total sample of patients (n=1329). The labels represent the following clslusters: 0 (All Low (n=419, 31.5%)), 1 (Moderate Physical & Lower Psychological (n=316, 23.8%)), 2 (Moderate Physical & Higher Psychological (n=416, 31.3%)) and 3 (All High (n=178, 13.4%).

Figure 1B - Silhouette coefficient diagram for the 4-cluster solution using the K-modes analysis. The sizes of the clusters in the diagram are proportional to their size inside the total sample of patients (n=1329). The labels represent the following clusters: 0 (All Low (n=536, 40.3%)), 1 (Moderate Physical & Lower Psychological (n=205, 15.4%)), 2 (Moderate Physical & Higher Psychological (n=280, 21.1%)), and 3 (All High (n=308, 23.2%)).

Figure 2 - Symptom occurrence for each of the subgroups identified using latent class analysis for the 25 symptoms on the Memorial Symptom Assessment Scale that occurred in $\geq$30% of the total sample (n=1329) at Time 1 (i.e., prior to next dose of chemotherapy).

Figure 3 - Symptom occurrence for each of the subgroups identified using K-modes analysis for the 25 symptoms on the Memorial Symptom Assessment Scale that occurred in $\geq$30% of the total sample (n=1329) at Time 1 (i.e., prior to next dose of chemotherapy).