Acute Coronary Syndrome Symptom Clusters: Illustration of Results Using Multiple Statistical Methods

Catherine J. Ryan[1], **Karen M. Vuckovic**[1], **Lorna Finnegan**[1], **Chang G. Park**[1], **Lani Zimmerman**[2], **Bunny Pozehl**[2], **Paula Schulz**[2], **Susan Barnason**[2], and **Holli A. DeVon**[1]

[1]The University of Illinois at Chicago, IL, USA

[2]University of Nebraska Medical Center, Omaha, NE, USA

Abstract

Researchers have employed various methods to identify symptom clusters in cardiovascular conditions, without identifying rationale. Here, we test clustering techniques and outcomes using a data set from patients with acute coronary syndrome. A total of 474 patients who presented to emergency departments in five United States regions were enrolled. Symptoms were assessed within 15 min of presentation using the validated 13-item ACS Symptom Checklist. Three variable-centered approaches resulted in four-factor solutions. Two of three person-centered approaches resulted in three-cluster solutions. *K*-means cluster analysis revealed a six-cluster solution but was reduced to three clusters following cluster plot analysis. The number of symptoms and patient characteristics varied within clusters. Based on our findings, we recommend using (a) a variable-centered approach if the research is exploratory, (b) a confirmatory factor analysis if there is a hypothesis about symptom clusters, and (c) a person-centered approach if the aim is to cluster symptoms by individual groups.

Keywords

cluster analysis; symptom clusters; symptoms; latent class analysis; acute coronary syndrome

Patients with cardiac conditions (e.g., heart failure, acute coronary syndrome) or those who have received cardiac interventions (e.g., cardiac surgery, percutaneous coronary intervention) often experience multiple concurrent symptoms, known as symptom clusters (Abbott, Barnason, & Zimmerman, 2010; DeVon, Ryan, Ochs, & Shapiro, 2008; DeVon, Ryan, Rankin, & Cooper, 2010; Fukuoka, Lindgren, Rankin, Cooper, & Carroll, 2007; Heo, Doering, Widener, & Moser, 2008; Lee et al., 2010; Lindgren et al., 2008; McSweeney, Cleves, Lefler, & Yang, 2010; Ryan et al., 2007). While there is no universally accepted definition for a symptom cluster, for the purpose of this study, a symptom cluster was defined as two or more symptoms that occur concurrently with each cluster being independent of other symptom clusters (Kim, McGuire, Tulman, & Barsevick, 2005;

Corresponding Author: Holli A. DeVon, Professor and Department Head, Department of Biobehavioral Health Science, College of Nursing, The University of Illinois at Chicago, 845 S. Damen Ave., MC 802, Chicago, IL 60612, USA. hdevon1@uic.edu.

Miaskowski et al., 2017). Numerous researchers are examining a variety of analytic methods, conceptual or empirical rationale related to symptom clusters (Miaskowski et al., 2017). This may facilitate meta-analyses or comparisons between studies in the future. Clusters may be misidentified and misinterpreted if the statistical approach is not appropriate for the study's theoretical concepts, aims, hypothesis, data characteristics, researcher judgment, or the objectives of the analysis.

Variable- and Person-Centered Approaches to Symptom Cluster Analyses

Exploratory and model-based variable-centered and person-centered analytic approaches have been used in the cardiovascular literature to identify symptom clusters (Laursen & Hoff, 2006). Variable-centered approaches (e.g., factor analysis) assume that the population is homogeneous and focus on the relationships among variables, such as symptoms, across individuals. In contrast, person-centered approaches (e.g., latent class analysis [LCA]) capture population heterogeneity by grouping persons into categories or subgroups based on their similarities on a set of observed variables, such as symptoms. It is important to note that the homogeneity/heterogeneity of the symptom data is not a precondition for a cluster analysis; the cluster analysis will reveal whether the symptoms are homogeneous or heterogeneous. The objective of this secondary data analysis was to illustrate outcomes of commonly used statistical techniques in identifying symptom clusters in a cohort of patients with ACS, using variable-centered (exploratory factor analysis [EFA] with follow-up confirmatory factor analysis [CFA] and principal component analysis [PCA]) and person-centered analytic approaches (hierarchical agglomerative cluster analysis, *k*-means cluster analysis, and LCA). Specifically, we show symptom cluster results for each approach, discuss advantages and disadvantages of each approach, address considerations for choosing the best approach, and make recommendations for using these methods in symptom cluster research.

Overview of Analytic Approaches

Variable-Centered Approaches

EFA with CFA and PCA have been used to derive symptom clusters from a set of observed symptoms (Dong et al., 2016). In EFA, common variance is separated from unique variance; common variance represents the variance shared among all participants because it is related to measurement method not the constructs that they represent. Unique variance includes measurement error and the variance that is not shared among the variables (Kim, 2008). The purpose of EFA, a theory-generating approach, is to examine and understand the structure of the data by exploring correlations among the observed variables (symptoms) as a function of one or more underlying latent variables (Kim, 2008). Latent variables are factors that are not observed or measured directly but can be inferred from behavior or responses to a set of questions (Polit & Yang, 2016). When EFA is used, each factor includes a set of mutually exclusive observed variables (symptoms in this case), and the number and content of factors represents novel information related to the data set. The strength of the relationship between each observed variable (symptom) and the latent factor is measured by a factor loading. CFA also uses latent variables (factors) to account for shared variance among observed variables

(symptoms). However, CFA is based on a theoretical understanding of the relationships among the latent variables (factors) and the relationships between the observed variables and latent variables. CFA traditionally follows preliminary EFA studies that determine the number of factors in the data set. Therefore, in CFA, the number of factors is not a result; rather, the number of factors is a precondition that facilitates assessment of the quality of the data partitioning. For example, random samples can be drawn from large data sets, with half of the data used to conduct EFA. The EFA findings are then confirmed on the other half of the data using CFA. When using CFA for factor analysis, investigators must specify a model that includes how many factors will be derived from the data set, which variables will load on which factors, and whether the factors are correlated with each other (Kääriäinen et al., 2011).

PCA, primarily a data reduction technique, collapses the number of observed variables (symptoms) into smaller linear combinations (known as principal components) that still explain the same amount of variance but with fewer variables (principal components). The first identified principal component explains the largest possible variance. Subsequent components are unique because they are uncorrelated with previous components. In PCA, the common variance among variables is maximized, there is no assumed measurement error, and there is no unique variance attributed to each variable (Suhr, 2005).

Person-Centered Approaches

In contrast to variable-centered approaches, person-centered approaches (often referred to as cluster analytic methods) partition individuals into meaningful clusters in which individuals are similar to each other within a cluster but different from each other between clusters (Everitt, Landau, & Leese, 2011). Two exploratory methods (hierarchical and k-means) are heuristic approaches based on distance or optimization measures. There are two forms of hierarchical clustering: (a) agglomerative, in which the initial variables are successively merged according to the greatest increase in the classification likelihood until all pairs belong to the same cluster (bottom-up approach), and (b) divisive, in which variables are partitioned into clusters from a whole (Hansen & Jaumard, 1997). At each stage of the process, a pair of clusters is merged to maximize the likelihood that observations are similar to the predicted clusters. Prior to conducting agglomerative analysis, the investigator determines the maximum number of clusters to consider based on the clinical understanding of the data. The appropriateness of the clusters is determined using a combination of dendrograms (visual representations of possible groupings), various fit indices, and the clinical judgment of the investigators.

k-means clustering is a nonhierarchical approach that begins with a user-specified number of clusters. The final solution is the one that has the minimum total within-cluster sum of squared distances (Jain, 2010). Traditionally, the algorithm used in k-means begins with an initial partitioning of the data into similar k partitions or clusters specified by the investigator, based on knowledge of the data (Everitt et al., 2011). While any number of clusters (k) may be specified, the researcher typically selects a range (i.e., 2–10). The mean or "centroid" is computed for each partition. Each data point is assigned by the distance to its closest centroid. The new centroid for the partition is computed, and data points are

reassigned until convergence is obtained (i.e., centroid or data points no longer change). The objective is to form discrete clusters that maximize similarities within a cluster and maximize differences between clusters. Thus, each cluster is independent of others, and there is no overlap between clusters (Jain, 2010).

An important consideration with *k*-means methodology is the initialization choice or the value assigned to *k* (Aldenderfer & Blashfield, 1984). The results from the *k*-means methods are sensitive to initialization. Poor initialization can lead to improper clustering and extended time to reach convergence. Approaches to selecting *k* include partitioning using an estimate of the centroid means, randomly selecting a number, or calculating a hierarchical cluster prior to performing the analysis. Convergence implies a hard assignment of individuals to a cluster (i.e., an individual either completely belongs to a cluster or does not belong at all). In addition, because the algorithm is based on the distance from the mean, results are sensitive to outliers and best suited for large data sets with clusters that are dense and of similar size.

Symptom clusters can also be derived using latent variable mixture models, a flexible modeling approach that includes LCA and latent profile analysis (LPA). In LCA (used with symptoms that are measured categorically) and LPA (used with symptoms that are measured continuously), a latent variable captures population heterogeneity by creating categories that represent subgroups of the population, with each category representing one subgroup (Kaufman & Rousseeuw, 1990; McCutcheon, 1987; Vermunt & Magidson, 2002). LCA and LPA are probability-based clustering approaches with several advantages over hierarchical and *k*-means clustering. Latent class approaches do not require variables (symptoms) to be normally distributed (McCutcheon, 1987), and symptoms may be of varying levels of measurement (e.g., categorical, continuous, or mixed). Symptoms can be modeled as conditionally independent given the clusters and therefore may be included in more than one cluster (Rindskopf & Rindskopf, 1985). Although LCA and LPA models can be developed without covariates (demographic or other characteristics of individuals in the sample), a distinct advantage to this approach is that covariates can be included in the analysis simultaneously with latent class formation (Clark & Muthén, 2009; Vermunt & Magidson, 2002), thus improving the efficiency of predictors and the clinical value of the results.

In exploratory LCA and LPA, the numbers and sizes of classes are not known a priori (Kaufman & Rousseeuw, 1990). Instead, the researcher uses an iterative approach beginning with a two-class model and fits successive models with an increasing number of classes. The goal is to find the most parsimonious model with an acceptable fit to the data. Fit statistics including Akaike information criterion (AIC), Bayesian information criterion (BIC), and R^2 entropy are used to determine the best-fitting model. The output includes latent class probabilities and conditional probabilities for each class or cluster. Latent class probabilities describe the distribution of the latent variable with respect to the number and relative size of classes. Conditional probabilities are comparable to factor loadings in factor analysis and represent the probability of an individual in a given class of the latent variable being at a particular level or category of the observed variable (e.g., symptoms; McCutcheon, 1987).

Method

Description of the Data Set

This secondary analysis used data from a larger prospective, multicenter study that examined the influence of sex on symptom characteristics during ACS (DeVon et al., 2017). The sample included 474 patients diagnosed with ACS presenting to one of five emergency departments (EDs) in the Midwest, West, Southwest, and Pacific Northwest regions of the United States. Diagnoses were based on the clinical judgment of the attending physician and were abstracted from the medical record by trained research associates. Symptoms were assessed within 15 minutes of ED presentation using the 13-item Acute Coronary Syndrome Symptom Checklist (DeVon, Rosenfeld, Steffen, & Daya, 2014). The institutional review boards at the sponsoring institution and data collection sites granted approval prior to the start of the study. All participants gave written, informed consent.

Study Population

Individuals presenting to the ED with symptoms triggering a cardiac evaluation, who were at least 21 years old, fluent in English or Spanish, and walked in or arrived by emergency medical services, were eligible for the study. Patients were excluded if they had an exacerbation of heart failure (B-type natriuretic peptide 500 pg/ml), were transferred from a hemodialysis facility, were referred for evaluation of a dysrhythmia, or were unable to understand and provide written informed consent for the study. Demographic characteristics of the sample are described in Table 1.

Measures

The Acute Coronary Syndrome Symptom Checklist is a 13-item empirically derived instrument that has been validated in previous studies (DeVon, Hogan, Ochs, & Shapiro, 2010; DeVon & Zerwic, 2003). The 13 symptoms are chest pressure, shoulder pain, sweating, palpitations, chest discomfort, upper back pain, shortness of breath, arm pain, unusual fatigue, nausea, lightheadedness, chest pain, and indigestion. Each symptom is measured dichotomously (present/absent, resulting in nominal-level data) and analyzed individually (no summary score). Patient characteristics were collected using the ACS Patient Information Questionnaire. This demographic and clinical questionnaire was designed using the standardized reporting guidelines for studies evaluating risk stratification of ED patients with potential ACS (Hollander et al., 2004).

Statistical Analyses

Factor and cluster analyses were conducted using the three variable-centered approaches and the three person-centered approaches described previously. Please see Table 2 for a description of each approach, along with assumptions, advantages, disadvantages, examples, and statistical software programs available for each method.

EFA.—Analysis was performed with SPSS v. 24 (IBM Corp., Armonk, NY). The sample was split for EFA, with the computer program selecting a random sample of 237 (training sample) of the total sample of 474 participants for initial analysis. The oblimin technique was used for rotation because there was an a priori assumption that the factors were related.

Factors were identified based on the Kaiser rule, with Eigenvalues > 1 being considered meaningful. Each variable was assigned to the factor with the highest factor loading.

CFA.—STATA v. 14 (College Station, TX) was used for analyses of the other half of the randomly selected sample from EFA (N = 237 participants). Based on findings from the EFA, we fit a four-factor model to the data from the validation sample. Model fit statistics (standardized root mean square residual [SRMR], root mean square error of approximation [RMSEA], comparative fit index [CFI], and Tucker–Lewis Index [TLI]) were used to determine the fit of the model to the data.

PCA.—Analyses were performed with SPSS v. 24 (IBM Corp., Armonk, NY) and included all 474 participants and all variables. The oblimin technique was used for rotation. Factors were identified based on the Kaiser rule, with Eigenvalues > 1 being considered meaningful. Each variable was assigned to the factor with the highest factor loading. It is important to note that EFA and PCA are conceptually and methodologically different and may yield different results.

Hierarchical (agglomerative) cluster analyses.—SPSS v. 24, (IBM Corp., Armonk, NY) was used to generate a cluster dendrogram. We then used the R NbClust and optCluster Programs, which displayed a scree plot, to determine the number of clusters present (Charrad, Ghazzali, Boiteau, & Niknafs, 2014; Sekula, Datta, & Datta, 2017). This allowed for visualization of how individuals clustered based on similarities in their symptoms.

k-means cluster analysis.—We initially used SPSS v.24 (IBM Corp., Armonk, NY) where the number (k) of clusters is predetermined by the researcher. Analysis of variance (ANOVA) was used to maximize the differences among the clusters. To validate and visualize how individuals fell into specific clusters, we used the R NbClust and optCluster programs, which use the Calinski-Harabasz Criterion and Simple Structure Index as well as other criteria to confirm the number of clusters present and perform k-means analysis. These programs employ a slightly different algorithm than SPSS to obtain scree and cluster plots.

LCA.—Latent Gold v.3 software (Statistical Innovations, Belmont, MA) was used to perform the LCA. First, we fit two to six class models without covariates using model fit statistics (BIC, AIC, L^2) and theoretical interpretability to determine the best-fitting model. As noted previously, a major advantage of LCA is the ability to include covariates as part of the model fitting process. Therefore, a second LCA analysis was performed with age, race, and sex added as covariates to illustrate how the covariates influenced class (cluster) membership.

Results

Sample Demographic Characteristics

Of the 474 participants, the majority were White, non-Hispanic males, aged 50–60 years. The sample well-matched the population of ACS patients in the United States.

Variable-Centered Approaches

EFA.—A four-factor solution was identified with EFA using a random sample of 50% of participants. This model explained 52% of the variance after oblimin rotation and Kaiser normalization. Goodness-of-fit test (Chi-square = 26.129, df = 32, p = .758) indicated that the four-factor model fit the data well. The four factors were labeled: (a) chest symptoms, (b) exertion-like symptoms, (c) non-chest pain symptoms, and (d) gastrointestinal symptoms. See Table 3 for factor loadings.

CFA.—CFA was used to confirm the findings of EFA using the remaining half of the participants' symptom data. The four-factor model derived from EFA was confirmed with CFA (RMSEA = .30, CFI = .968, TLI = .958, and SRMR = .041). Factor loadings ranged from .43 to .75 (Table 3).

PCA.—PCA was performed using symptom data from all 474 participants and revealed four principal components. These components were similar to those identified with EFA (see Table 3 for factor loadings). EFA and PCA may yield similar results when the variables are highly correlated.

In summary, EFA (confirmed with CFA) and PCA resulted in a four-factor solution, supporting that there were four factors (clusters) and each of the factors contained similar symptoms. This is important and provides evidence that EFA and PCA accurately represented the factor structure in this sample. Depending on the objective of analysis (exploratory or theoretical), either EFA to understand data structure or PCA to reduce the number of variables could be used, but not both.

Person-Centered Approaches

Hierarchical (agglomerative) cluster analyses.—Two to six clusters were identified using an agglomerative technique. About 14 persons were included in both a six-cluster and five-cluster model. However, as the number of clusters decreased, the same 14 persons collapsed into the next cluster. This suggests that the five- and six-cluster solutions provided minimal influence over a smaller cluster model in identifying the cluster pattern.

Further analysis with the optCluster program (which uses a slightly different algorithm) provided scree cluster plots. Scree plots indicated that a three-cluster solution was appropriate and that there was minimal gain in meaning with additional cluster solutions identified in the original analysis. The three symptom clusters were labeled (a) heavy symptom burden, (b) chest symptoms, and (c) classic symptoms. The three-cluster solution also revealed noteworthy similarities and overlap between the clusters (Table 4).

k-means analysis.—_k_-means analysis initially resulted in a six-cluster solution. The clusters were labeled (a) classic symptoms, (b) chest symptoms only, (c) gastrointestinal, (d) low symptom burden, (e) heavy symptom burden, and (f) pain and discomfort. The amount of cases varied from 35 to 100 per cluster. In _k_-means analysis, the software performs multiple iterations to seek convergence and optimize outcomes. We noted that three of the clusters converged with multiple iterations, which suggested that there were really fewer

clusters. Further analysis with the optCluster program provided a scree plot that identified 2 to 3 clusters. A cluster plot confirmed a three-cluster solution with noteworthy similarities and overlap between the clusters (Table 4). These clusters were labeled, (a) low symptom burden, (b) chest symptoms, and (c) classic symptoms.

LCA.—When data were analyzed without covariates, a three-class solution had the best model fit, lower limit (LL) = −3,532.70, L^2 = 1,683.46, AIC = 7,147.38, BIC = 7,317.9, R^2 entropy =.70, and classes were labeled (a) low probability of any symptom (31% of the sample), (b) chest symptoms (32% of the sample), and (c) heavy symptom burden (37% of the sample). When individual covariates of age, race, and sex were included in the analysis, a four-class solution had the best model fit (LL = −3,458, L^2 = 5,805.122, AIC = 7,084.58, BIC = 7,310.62, R^2 entropy = .6955). The clusters were labeled (a) chest symptoms (40% of sample), (b) heavy symptom burden (25% of sample), (c) classic symptoms (19% of sample), and (d) low symptom burden (15% of sample). See Table 5 for a summary of the LCA findings and Table 6 for fit statistics to demonstrate that a three- and four-class solution were the best fit for these data.

Discussion

We identified similarities and differences in symptom clusters experienced during ACS using three variable-centered and three person-centered approaches. Like Atkas, Walsh, and Hu (2014), we found minor variations in the number of clusters in our data set. Choosing and interpreting a cluster analytic technique should be based on theoretical concepts, study aims, hypotheses, data characteristics, and the judgment of the researcher. If the research is exploratory, then EFA or CFA is recommended. If the investigators have a hypothesis about symptom clusters for a specific condition, then CFA is recommended. If the aim is to cluster symptoms by individual groups, then a person-centered approach, such as LCA, is recommended.

However, because "best practice" has yet to be identified, how to interpret and apply these results remains in question. In a recent paper on the use of cluster analysis or factor analysis to assess cooccurrence of risk behaviors, the authors differentiated the two techniques with respect to policy and intervention goals (Hofstetter, Dusseldorp, Empelen, & Paulussen, 2014). They recommended that cluster analysis should be used when the goal is to target interventions to clusters of individuals showing similar patterns of behaviors, and factor analysis should be used to target intervention strategies at behaviors that share the same underlying source. We believe that this guideline can also be applied when the study aim is to tailor interventions to clusters of individuals showing similar patterns of symptoms and factor analysis when the intervention goal is to target symptoms that share the same underlying source or mechanism.

Variable-Centered Analyses

EFA, confirmed with CFA, resulted in four factors: (a) chest symptoms, (b) exertion-like symptoms, (c) non-chest pain symptoms, and (d) gastrointestinal symptoms. Of note, the factors identified with PCA contained the same symptoms. This is important and gives us confidence that two variable-centered approaches and the confirmatory procedure accurately

represented the factor structure in this sample of the ACS population. While PCA is commonly used for symptom cluster analysis, conceptually it is not appropriate for this purpose due to measurement error and the threat to the reliability and validity of the findings. EFA and CFA could be useful when the goal is to understand and develop further an intervention to target the mechanism underlying the associations among symptoms.

Person-Centered Analyses

All three of the person-centered approaches (hierarchical agglomerative, *k*-means, and LCA) demonstrated a three-class (cluster) solution (four-class with LCA when covariates were added). However, it took a two-step statistical process to identify a three-class solution with hierarchical agglomerative and *k*-means analytic techniques, and this resulted in a potentially significant change in the symptom clusters. *k*-means with additional optimization procedures resulted in the loss of a gastrointestinal symptom cluster. Retaining this cluster is clinically relevant because patients experiencing ACS often erroneously label their symptoms as "heartburn." Triage personnel may do the same if the patient self-diagnoses "heartburn." This result illustrates the need for the researcher to understand not only the data but the clinical implications of the findings. Like other investigators, we used our clinical judgment to make a determination of the number of symptom clusters realizing that age and sex were important variables that needed to be taken into account for the analysis (DeVon et al., 2010; Ryan et al., 2007).

As previously stated, LCA has several advantages over other person-centered clustering methods (probability-based classification), including the following: variables may be of different levels of measurement, demographics or other covariates can be added to the models (Vermunt & Magidson, 2002),; and a normal distribution is not required (McCutcheon, 1987). Most noteworthy when considering potential intervention development is the inclusion of demographic variables and other covariates in the model so that clustering techniques can be used to categorize the population into subgroups that can be used to target interventions. This has the potential to improve precision health care and facilitate the design of implementation studies targeting the most high-risk patients.

The methodology selected to describe symptom clusters should be determined based on the conceptual or theoretical rationale for the study aims, the levels of measurement, research implications, and potential clinical application of the findings. The use of variable-centered approaches is recommended when the aim of the study is to cluster variables, without grouping the research participants. For example, if the researcher seeks to determine if items in a quality of life instrument cluster into domains such as health and functioning, psychological/spiritual, social and economic, and family (Ferrans, 1996). We recommend using a person-centered approach to cluster symptoms by individual groups, so that interventions can be targeted to the symptom cluster and covariate patterns of each unique group. Agglomerative and *k*-means cluster techniques are difficult to interpret and require multiple steps to identify the specific symptoms within each cluster. In addition, there is no clear definition of what the stop point would be when determining what number of clusters the researcher will accept. Therefore, we recommend the use of LCA for identifying symptom clusters in cardiac populations.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Acknowledgments

References

Abbott AA, Barnason S, & Zimmerman LS (2010). Symptom burden clusters and their impact on psychosocial functioning following coronary artery bypass surgery. Journal of Cardiovascular Nursing, 25, 301–310. doi:10.1097/JCN.0b013e3181cfbb46 [PubMed: 20539164]

Aldenderfer MS, & Blashfield RK (1984). Cluster analysis. Beverly Hills, CA: Sage.

Atkas A, Walsh D, & Hu B (2014). Cancer symptom clusters: An exploratory analysis of eight statistical techniques. Journal of Pain and Symptom Management, 48, 1254–1265. doi:10.1016/j.jpainsymman.2014.02.006 [PubMed: 24747226]

Charrad M, Ghazzali N, Boiteau V, & Niknafs A (2014). Nbclust: An r package for determining the relevant number of clusters in a data set. Journal of Statistical Software, 61(6), 1–36. doi:10.18637/jss.v061.i06

Clark S, & Muthén B (2009). Relating latent class analysis results to variables not included in the analysis. Retrieved from https://www.statmodel.com/download/relatinglca.pdf

DeVon HA, Burke LA, Vuckovic KM, Haugland T, Eckhardt AL, Patmon F, & Rosenfeld AG (2017). Symptoms suggestive of acute coronary syndrome: When is sex important? Journal of Cardiovascular Nursing, 32, 383–392. doi:10.1097/jcn.0000000000000351 [PubMed: 27281055]

DeVon HA, Hogan N, Ochs AL, & Shapiro M (2010). Time to treatment for acute coronary syndromes: The cost of indecision. Journal of Cardiovascular Nursing, 25, 106–114. doi:10.1097/JCN.0b013e3181bb14a0 [PubMed: 20142752]

DeVon HA, Rosenfeld A, Steffen AD, & Daya M (2014). Sensitivity, specificity, and sex differences in symptoms reported on the 13-item acute coronary syndrome checklist. Journal of the American Heart Association, 3, e000586. doi:10.1161/JAHA.113.000586 [PubMed: 24695650]

DeVon HA, Ryan CJ, Ochs AL, & Shapiro M (2008). Symptoms across the continuum of acute coronary syndromes: Differences between women and men. American Journal of Critical Care, 17, 14–24. [PubMed: 18158385]

DeVon HA, Ryan CJ, Rankin SH, & Cooper BA (2010). Classifying subgroups of patients with symptoms of acute coronary syndromes: A cluster analysis. Research in Nursing & Health, 33, 386–397. doi:10.1002/nur.20395 [PubMed: 20672306]

DeVon HA, & Zerwic JJ (2003). The symptoms of unstable angina: Do women and men differ? Nursing Research, 52, 108–118. [PubMed: 12657986]

Dong ST, Costa DS, Butow PN, Lovell MR, Agar M, Velikova G, . . . Fayers PM (2016). Symptom clusters in advanced cancer patients: An empirical comparison of statistical methods and the impact on quality of life. Journal of Pain and Symptom Management, 51, 88–98. doi:10.1016/j.jpainsymman.2015.07.013 [PubMed: 26300025]

Eastwood JA, Johnson BD, Rutledge T, Bittner V, Whittaker KS, Krantz DS, . . . Bairey Merz CN (2013). Anginal symptoms, coronary artery disease, and adverse outcomes in Black and White women: The NHLBI-sponsored Women's Ischemia Syndrome Evaluation (WISE) study. Journal of Women's Health, 22, 724–732. doi:10.1089/jwh.2012.4031

Everitt BS, Landau S, & Leese M (2011). Cluster analysis (5th ed.). London, England: Arnold.

Ferrans CE (1996). Development of a conceptual model of quality of life. Scholarly Inquiry for Nursing Practice, 10, 293–304. [PubMed: 9009823]

Fukuoka Y, Lindgren TG, Rankin SH, Cooper BA, & Carroll DL (2007). Cluster analysis: A useful technique to identify elderly cardiac patients at risk for poor quality of life. Quality of Life Research, 16, 1655–1663. [PubMed: 17955345]

Hansen P, & Jaumard B (1997). Cluster analysis and mathematical programming. Mathematical Programming, 79, 191–215. doi:10.1007/BF02614317

Heo S, Doering LV, Widener J, & Moser DK (2008). Predictors and effect of physical symptom status on health-related quality of life in patients with heart failure. American Journal of Critical Care, 17, 124–132. [PubMed: 18310649]

Herr JK, Salyer J, Flattery M, Goodloe L, Lyon DE, Kabban CS, & Clement DG (2015). Heart failure symptom clusters and functional status—A cross-sectional study. Journal of Advanced Nursing, 71, 1274–1287. doi:10.1111/jan.12596 [PubMed: 25533867]

Hertzog MA, Pozehl B, & Duncan K (2010). Cluster analysis of symptom occurrence to identify subgroups of heart failure patients: A pilot study. Journal of Cardiovascular Nursing, 25, 273–283. doi:10.1097/JCN.0b013e3181cfbb6c [PubMed: 20539162]

Hofstetter H, Dusseldorp E, Empelen P, & Paulussen T (2014). A primer on the use of cluster analysis or factor analysis to assess co-occurrence of risk behaviors. Preventive Medicine, 67, 141–146. doi:10.1016/j.ypmed.2014.07.007 [PubMed: 25036437]

Hollander JE, Blomkalns AL, Brogan GX, Diercks DB, Field JM, Garvey JL, . . . Zalenski R. (2004). Standardized reporting guidelines for studies evaluating risk stratification of emergency department patients with potential acute coronary syndromes. Annals of Emergency Medicine, 44, 589–598. doi:10.1016/j.annemergmed.2004.08.009 [PubMed: 15573034]

Hollander JE, Blomkalns AL, Brogan GX, Diercks DB, Field JM, . . . & Holroyd BR (2004). Standardized reporting guidelines for studies evaluating risk stratification of ED patients with potential acute coronary syndromes. Academic Emergency Medicine, 11(12), 1331–1340. [PubMed: 15576525]

Jain AK (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31, 651–666. doi:10.1016/j.patrec.2009.09.011

Jurgens CY, Moser DK, Armola R, Carlson B, Sethares K, & Riegel B (2009). Symptom clusters of heart failure. Research in Nursing & Health, 32, 551–560. doi:10.1002/nur.20343 [PubMed: 19650069]

Kääriäinen M, Kanste O, Elo S, Pölkki T, Miettunen J, & Kyngäs H (2011). Testing and verifying nursing theory by confirmatory factor analysis. Journal of Advanced Nursing, 67, 1163–1172. doi:10.1111/j.1365-2648.2010.05561.x [PubMed: 21226874]

Kaufman L, & Rousseeuw PJ (1990). Finding groups in data: An introduction to cluster analysis. New York, NY: John Wiley.

Kim HJ (2008). Common factor analysis versus principal component analysis: Choice for symptom cluster research. Asian Nursing Research, 2, 17–24. doi:10.1016/S1976-1317(08)60025-0 [PubMed: 25031108]

Kim HJ, McGuire DB, Tulman L, & Barsevick AM (2005). Symptom clusters: Concept analysis and clinical implications for cancer nursing. Cancer Nursing, 28, 270–282. [PubMed: 16046888]

Laursen B, & Hoff E (2006). Person-centered and variable-centered approaches to longitudinal data. Merrill-Palmer Quarterly, 52, 377–389. doi:10.1353/mpq.2006.0029

Lee KS, Song EK, Lennie TA, Frazier SK, Chung ML, Heo S, . . . Moser DK (2010). Symptom clusters in men and women with heart failure and their impact on cardiac event-free survival. Journal of Cardiovascular Nursing, 25, 263–272. doi:10.1097/JCN.0b013e3181cfbb88 [PubMed: 20539161]

Lindgren TG, Fukuoka Y, Rankin SH, Cooper BA, Carroll D, & Munn YL (2008). Cluster analysis of elderly cardiac patients' prehospital symptomatology. Nursing Research, 57, 14–23. [PubMed: 18091288]

McCutcheon AL (1987). Latent class analysis. Newbury Park, CA: Sage.

McSweene JC., Cleve MA., Lefle LL., & Yan S. (2010). Cluster analysis of women's prodromal and acute myocardial infarction symptoms by race and other characteristics. Journal of Cardiovascular Nursing, 25, 311-322. doi:10.1097/JCN.0b013e3181cfba15 [PubMed: 20539165]

Miaskowski C, Barsevick A, Berger A, Casagrande R, Grady P, Jacobson P, . . . Marden S (2017). Advancing symptom science through symptom cluster research: Expert panel proceedings and recommendations. Journal of the National Cancer Institute, 109, djw253. doi:10.1093/jnci/djw253

Moser DK, Lee KS, Wu J-R, Mudd-Martin G, Jaarsma T, Huang T-Y, . . . Riegel B (2014). Identification of symptom clusters among patients with heart failure: An international observational study. International Journal of Nursing Studies, 51, 1366–1372. doi:10.1016/j.ijnurstu.2014.02.004 [PubMed: 24636665]

Polit DF, & Yang FM (2016). Measurement and the measurement of change. Philadelphia, PA: Wolters Kluwer.

Riegel B, Hanlon AL, McKinley S, Moser DK, Meischke H, Doering LV, . . . Dracup K (2010). Differences in mortality in acute coronary syndrome symptom clusters. American Heart Journal, 159, 392–398. doi:10.1016/j.ahj.2010.01.003 [PubMed: 20211300]

Rindskopf D, & Rindskopf W (1985). The value of latent class analysis in medical diagnosis. Statistics in Medicine, 5, 21–27.

Rosenfeld AG, Knight EP, Steffen A, Burke L, Daya M, & DeVon HA (2015). Symptom clusters in patients presenting to the emergency department with possible acute coronary syndrome differ by sex, age, and discharge diagnosis. Heart & Lung, 44, 368–375. doi:10.1016/j.hrtlng.2015.05.008 [PubMed: 26118542]

Ryan CJ, DeVon HA, Horne R, King KB, Milner K, Moser DK, . . . Zerwic JJ (2007). Symptom clusters in acute myocardial infarction: A secondary data analysis. Nursing Research, 56, 72–81. doi:10.1097/01.NNR.0000263968.01254.d6 [PubMed: 17356437]

Sekula M, Datta S, & Datta S (2017). Optcluster: An r package for determining the optimal clustering algorithm. Bioinformation, 13, 100–103. doi:10.6026/97320630013101

Song EK, Moser DK, Rayens MK, & Lennie TA (2010). Symptom clusters predict event-free survival in patients with heart failure. Journal of Cardiovascular Nursing, 25, 284–291. doi:10.1097/JCN.0b013e3181cfbcbb [PubMed: 20539163]

Streur M, Ratcliffe SJ, Ball J, Stewart S, & Riegel B (2017). Symptom clusters in adults with chronic atrial fibrillation. Journal of Cardiovascular Nursing, 32, 296–303. doi:10.1097/jcn.0000000000000344 [PubMed: 27203271]

Suhr DD (2005). Statistics and data analysis paper 203–30 principal component analysis vs. exploratory factor analysis. In SUGI 30 proceedings (pp. 203–230). Retrieved from http://www2.sas.com/proceedings/sugi30/203-30.pdf

Vermunt JK, & Magidson J (2002). Applied latent class analysis. Cambridge, UK: Cambridge University Press.

Table 1.

Sample Demographic Characteristics.

Characteristic	*n* = 474
Age (*M, SD*)	61.7 (11.9)
Age (*n*, %)	
<50	70 (14.8)
50–59	148 (31.2)
60–69	138 (29.1)
70+	118 (24.9)
Sex (*n* [%])	
Female	131 (27.6)
Male	343 (72.4)
Race/Ethnicity (*n* [%])	
White Non-Hispanic	325 (69.0)
African American	56 (11.9)
Hispanic	45 (9.6)
Asian	13 (2.8)
Multiracial	13 (2.8)
Other	19 (4.0)
Missing	3 (0.6)

Table 2.

Summary of Symptom Cluster Analytic Approaches, Assumptions, Exemplars, Advantages, and Disadvantages.

Analytic Approach	Assumptions	Statistical Package/ Criteria for Factor Determination	Exemplar	Advantages	Disadvantages
EFA	• Variables used should be metric • Sample size should be > 200 • Homogeneous sample • At least 0.30 correlations are required between the research variables • No outliers in the data	• SPSS • SAS • STATA • R • Eigenvalue or adjusted eigen value > 1, model fit statistics	Eastwood et al. (2013) explored 67 potential anginal symptoms in women to identify the latent variables that are representative of angina symptoms in Black vs. White women. Four angina symptom clusters were found using EFA: chest/general malaise, upper body, abdominal discomfort, and typical triggers/relievers	• Explores data to provide information about the number of factors required to represent the data • All measured variables are related to every latent variable • Determines factors based on covariance between variables • Determines clusters conceptually based on cooccurrence and relatedness	• Number of variables tied to sample size • Requires EFA-generated model or theory • Requires a priori *N* of factors or factor structures
Confirmatory Factor Analysis	• Multivariate normality • Sample size > 200 • The correct a priori model specification • Data from a random sample	• SAS • STATA • Mplus • R • Model fit statistics: RMSEA, CFI, GFI, TLI, X2		• Accounts for measurement error • Is model-based • Confirms or rejects the measurement theory	• Requires EFA-generated model or theory • Requires a priori *N* of factors or factor structures
PCA	• Variables should be metric • Sample size should be > 200 • Homogeneous sample • At least 0.30 correlations are required between the research variables • No outliers in the data	• SPSS • SAS • R • STATA • Eigenvalue > 1	Jurgens et al. (2009) used PCA and and a theoretical approach (Theory of Unpleasant Symptoms) to identify the multidimensional nature of symptoms in hospitalized patients with HF based on nine symptoms from the Minnesota Living with Heart Failure Questionnaire. Three unique symptom clusters were identified: acute volume overload, emotional, and chronic volume overload. Other Example: Herr et al. (2015)	• Requires highly correlated variables • Orthogonally transforms correlated variables into a set of uncorrelated variables called principal components	• Assumes multivariate normal distribution of items • PCA minimizes distance between clusters that are not widely separated • Does not account for measurement error
Hierarchical Methods— Agglomerative	• Spherical shaped clusters • Variables are uncorrelated within clusters	• SPSS • SAS • STATA • R • Can be represented by dendrogram	Lindgren et al. (2008) sought to determine symptom clusters in elderly patients experiencing ischemic coronary heart disease in the week before hospitalization. Three clusters (groups) were identified: (a) Classic Acute Coronary Syndrome, (b) Weary, and (c) Diffuse Symptoms. The most appropriate clusters were determined using a combination of dendrograms, fit indices, and the clinical judgment of the investigators.	•Links clusters through four different methods (Single, Complete, Average, and Ward's Method) • Joins new candidate for cluster membership to existing group based on highest level of similarity to existing group • Begins with each case being own group, therefore manageable	• Requires calculation and storage of large similarity matrix • Makes only one pass through data; therefore, poor early partitioning can create problems • Can generate different solutions by reordering data in similarity matrix • Becomes unstable when cases dropped

Analytic Approach	Assumptions	Statistical Package/ Criteria for Factor Determination	Exemplar	Advantages	Disadvantages
			Other examples: Fukuoka, Lindgren, Rankin, Cooper, and Carroll (2007); Hertzog, Pozehl, and Duncan (2010); Lindgren et al. (2008); Lee et al. (2010); Moser et al. (2014); Song, Moser, Rayens, and Lennie (2010); Streur, Ratcliffe, Ball, Stewart, and Riegel (2017)	• Does not allow overlapping groups, people, or symptoms • Best with small data sets	
k-means	• Data should have roughly the same scale to squared Euclidean distance • Each group has roughly the same Error Sum of Squares (variance/ covariance matrix between objects within a group is equal) • Variance of the distribution of each attribute (variable) is spherical • All variables have the same variance • Each cluster has roughly equal number of observations	• SPSS • SAS • STATA • R • Scree plot • Cluster reduction software	McSweeney, Cleves, Lefler, and Yang (2010) used k-means for a secondary analysis of prodromal and acute symptoms of myocardial infarction in women. The authors wanted to determine the naturally occurring homogeneous subgroups based on increasing symptom frequency and severity.	• Compensates for initial poor partitioning • Reduces data • Develops taxonomies	• Requires that you define the number of clusters in advance or that you run multiple solutions to identify the best solution • Does not permit overlapping clusters • Repeatedly calculates to identify best partitioning • Based on a predetermined number of clusters • Not model-based • Multiple algorithms available to cluster the data
Latent Class Analysis (categorical variables) Latent Profile Analysis (continuous variables)	• The population is composed of different unobserved groups, or latent classes • Observations conditionally independent variables • No assumptions related to linearity, normal distribution, or homogeneity • Data level should be categorical or ordinal • Observations should be independent in each class	• Latent Gold • Mplus • LEM • MLLSA • SAS • STATA-Model-based • R • Formal criteria such as BIC, AIC, and R2 entropy to make decisions related to number of clusters	Ryan et al. (2007) used latent class analysis to identify clusters of symptoms that represented AMI based on subject demographic characteristics, with the goal of informing clinical practice. Five distinct clusters were identified that included high, medium, and low probability symptoms. They also noted that age, race, and sex were statistically significant in predicting cluster membership. Other examples: DeVon et al (2010); Riegel et al. (2010); Rosenfeld et al. (2015)	• Is model-based • Maximum likelihood estimation with fit statistics used to make decisions related to number of classes • Can incorporate covariates in the model • Used in conjunction with large data sets and to develop taxonomies • ONLY method that allows for the inclusion of covariates in the analyses	• Sparse cells and response patterns may lead to difficulties in model evaluation and identification

Note. EFA = exploratory factor analysis; RMSEA = root mean square error of approximation; SPSS = Statistical Package for the Social Sciences; CFI = comparative fit index; GFI = Goodness-of-fit index; HF = heart failure; LEM = statistical software for latent class analysis; TLI = Tucker-Lewis Index; PCA = principal component analysis; AIC = Akaike information criterion; BIC = Bayesian information criterion; MLLSA = Maximum likelihood latent structure analysis; SAS = statistical software package; STATA = statistical software package.

Table 3.

EFA, CFA, and PCA Results.

Symptoms	EFA Factor Loading (50% of Sample; *n* = 237; Highest Values)	CFA (50% of EFA Sample; *n* = 237) Factor Loading (95% CI)	PCA Factor Loading
1. Chest symptoms			
chest discomfort	.866	.721 [.628, .813]	−.744
chest pressure	.465	.582 [.491, .673]	−.650
chest pain	.391	.485 [.391, .578]	−.696
2. Exertion-like			
lightheadedness	.546	.594 [.506, .682]	.561
fatigue	.427	.586 [.506, .682]	.471
palpitations	.455	.442 [.347, .537]	.746
shortness of breath	.420	.479 [.386, .573]	.598
3. Pain non-chest			
shoulder pain	.759	.745 [.650, .842]	.752
arm pain	−.565	.555 [.464, .645]	.751
upper back pain	−.363	.443 [.346, .540]	.652
4. Gastrointestinal			
nausea	.636	.567 [.458, .676]	.730
sweating	.387	.471 [.367, .575]	.719
indigestion	.355	.432 [.327, .539]	.407

Note. EFA = exploratory factor analysis; CFA = confirmatory factor analysis; CI = confidence interval; PCA = principal component analysis.

Table 4.

Results of Analysis.

Methodology	Solution	Description of Solution
Factor Analysis (PCA, EFA, and CFA)	Four-factor solution (SPSS and SAS)	1. Chest symptoms 2. Exertion-like symptoms 3. Pain: non-chest 4. Gastrointestinal
Hierarchical Agglomerative Analysis	Three-cluster solution (optCluster)	1. Heavy symptom burden (indigestion, lightheadedness, nausea, fatigue, arm pain, upper back pain, sweating, shoulder pain) 2. Chest symptoms (chest pain, chest discomfort, chest pressure) 3. Classic symptoms (chest pain, lightheadedness, fatigue, shortness of breath, chest discomfort, chest pressure)
k-means Cluster Analysis	Six clusters (SPSS) Based on cluster centers	1. Gastrointestinal (sweating, fatigue, lightheadedness) 2. Low symptom burden (no symptoms with high probability of occurrence) 3. Heavy symptom burden (chest pressure, shoulder pain, sweating, palpitations, chest discomfort, upper back pain, shortness of breath, arm pain, fatigue, nausea, lightheadedness, chest pain, indigestion) 4. Classic symptoms (chest pressure, sweating, chest discomfort, shortness of breath, nausea, chest pain) 5. Pain & discomfort (chest pressure, shoulder pain, chest discomfort, arm pain, chest pain) 6. Chest symptoms only (chest pressure, chest discomfort, chest pain)
k-means Cluster Analysis	Three clusters (optCluster)	1. Low symptom burden (chest pain was only symptom with > 40% probability) 2. Chest symptoms (chest pressure, chest discomfort, chest pain) 3. Classic symptoms (chest pressure, shoulder pain, sweating, chest discomfort, shortness of breath, arm pain, fatigue, nausea, lightheadedness, chest pain)
Latent Class Analysis without covariates	Three classes (clusters) (Latent Gold)	1. Low probability symptoms (no symptoms have a > 50% probability of occurring) 2. Chest symptoms (chest pressure, chest discomfort, chest pain) 3. Heavy symptom burden (chest pressure, shoulder pain, sweating, chest discomfort, shortness of breath, arm pain, fatigue, nausea, lightheadedness, chest pain)
Latent Class Analysis with covariates	Four classes (clusters) (Latent Gold	1. Chest symptoms (chest pressure, chest discomfort) 2. Heavy symptom burden (chest pressure, shoulder pain, chest discomfort, shortness of breath, arm pain, fatigue, chest pain) 3. Classic symptoms (chest pressure, shortness of breath, fatigue, nausea, lightheadedness, chest pain) 4. Low symptom burden (no symptoms have a >50% probability of occurring)

Note. PCA = principal components analysis; EFA = exploratory factor analysis; CFA = confirmatory factor analysis; SAS = Statistical Analysis Software; SPSS = Statistical Package for the Social Sciences.

Table 5.

Results of Cluster Analysis by Method.

Methodology	Solution	Description of Solution
Factor Analysis (PCA, EFA, and CFA)	Four-factor solution (SPSS and SAS)	1. Chest symptoms 2. Exertion-like symptoms 3. Pain: non-chest 4. Gastrointestinal
Hierarchical Agglomerative Analysis	Three-cluster solution (optCluster)	1. Heavy symptom burden (indigestion lightheadedness, nausea, fatigue, arm pain, upper back pain, sweating, shoulder pain) 2. Chest symptoms (chest pain, chest discomfort, chest pressure) 3. Classic symptoms (chest pain, lightheadedness, fatigue, shortness of breath, chest discomfort, chest pressure)
k-means Cluster Analysis	Six clusters (SPSS) Based on cluster centers	1. Gastrointestinal (sweating, fatigue, lightheadedness) 2. Low symptom burden (no symptoms with high probability of occurrence) 3. Heavy symptom burden (chest pressure, shoulder pain, sweating, palpitations, chest discomfort, upper back pain, shortness of breath, arm pain, fatigue, nausea, lightheadedness, chest pain, indigestion) 4. Classic symptoms (chest pressure, sweating, chest discomfort, shortness of breath, nausea, chest pain) 5. Pain and discomfort (chest pressure, shoulder pain, chest discomfort, arm pain, chest pain) 6. Chest symptoms only (chest pressure, chest discomfort, chest pain)
k-means Cluster Analysis	Three clusters (optCluster)	1. Low symptom burden (chest pain was only symptom with > 40% probability) 2. Chest symptoms (chest pressure, chest discomfort, chest pain) 3. Classic symptoms (chest pressure, shoulder pain, sweating, chest discomfort, shortness of breath, arm pain, fatigue, nausea, lightheadedness, chest pain)
Latent Class Analysis without covariates	Three classes (clusters) (Latent Gold)	1. Low probability symptoms (no symptoms have a >50% probability of occurring) 2. Chest symptoms (chest pressure, chest discomfort, chest pain) 3. Heavy symptom burden (chest pressure, shoulder pain, sweating, chest discomfort, shortness of breath, arm pain, fatigue, nausea, lightheadedness, chest pain)
Latent Class Analysis with covariates	Four classes (clusters) (Latent Gold	1. Chest symptoms (chest pressure, chest discomfort) 2. Heavy symptom burden (chest pressure, shoulder pain, chest discomfort, shortness of breath, arm pain, fatigue, chest pain) 3. Classic symptoms (chest pressure, shortness of breath, fatigue, nausea, lightheadedness, chest pain) 4. Low symptom burden (no symptoms have a > 50% probability of occurring)

Note. PCA = principal components analysis; EFA = exploratory factor analysis; CFA = confirmatory factor analysis; SAS = Statistical Analysis Software; SPSS = Statistical Package for the Social Sciences.

Table 6.

Latent Class Analysis Fit Statistics.

	LL	BIC (LL)	N	L²	df
All Symptoms without Covariates					
Two-Cluster Model	−3,577.0327	7,320.303	27	1,772.142	8,164
Three-Cluster Model	*−3,532.6907*	*7,317.817*	*41*	*1,683.458*	*8,150*
Four-Cluster Model	−3,942.4829	7,323.599	55	1,603.042	8,136
Five-Cluster Model	−3,464.7162	7,354.263	69	1,547.509	8,122
Latent Class Fit Statistics—All Symptoms with Age, Race, Sex entered as Covariates					
Two-Cluster Model	−3,561.0188	7,306.683	30	6,010.456	167,912
Three-Cluster Model	*−3,511.2717*	*7,311.821*	*47*	*5,910.962*	*16,910*
Four-Cluster Model	*−3,458.3531*	*7,310.617*	*64*	*5,805.125*	*167,909*
Five-Cluster Model	−3,428.3237	7,355.190	81	5,745.066	167,909

Note. LL = Log-likelihood; BIC = Bayesian information criteria.