Published in final edited form as:

Exp Aging Res. 2015 ; 41(2): 177–192. doi:10.1080/0361073X.2015.1001655.

Baseline Cluster Membership Demonstrates Positive Associations with First Occurrence of Multiple Gerontological Outcomes over 10 Years

S.J. Fodeh¹, M. Trentalange², H.G. Allore², T.M. Gill², C.A. Brandt^{1,3}, and T.E. Murphy²

¹ VA Connecticut Healthcare System, West Haven, CT, USA

² Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA

³ Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, USA

Abstract

Background: The potential of cluster analysis (CA) as a baseline predictor of multivariate gerontologic outcomes over a long period of time has not been previously demonstrated.

Methods: Restricting candidate variables to a small group of established predictors of deleterious gerontologic outcomes, various CA methods were applied to baseline values from 754 nondisabled, community-living persons, aged 70 years or older. The best cluster solution yielded at baseline was subsequently used as a fixed explanatory variable in time-to-event models of the first occurrence of the following outcomes: any disability in four activities of daily living, any disability in four mobility measures, and death. Each outcome was recorded through a maximum of 129 months or death. Associations between baseline ordinal cluster level and first occurrence of all three outcomes were modeled over a 10 year period with proportional hazards regression and compared to the associations yielded by the analogous latent class analysis (LCA) solution.

Results: The final cluster-defining variables were continuous measures of cognitive status and depressive symptoms, and dichotomous indicators of slow gait and exhaustion. The best solution yielded by baseline values of these variables was obtained with a K-means algorithm and cosine similarity, and consisted of three clusters representing increasing levels of impairment. After adjustment for age, gender, ethnic group, and number of chronic conditions, baseline ordinal cluster level demonstrated significantly positive association with all three outcomes over a 10 year period that were equivalent to those from the corresponding LCA solution.

Conclusion: These findings suggest that baseline clusters based on previously established explanatory variables have potential to predict multivariate gerontologic outcomes over a long period of time.

Keywords

gerontologic; longitudinal; cluster analysis; mortality; disability; mobility

Corresponding Author: Terrence Murphy, PhD., terrence.murphy@yale.edu, Yale Program on Aging, 300 George St, Suite 775, New Haven, CT 06511, Phone: (203) 737-2295, Fax: (203) 785-4823.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts relevant to this research.

One of the most promising techniques for identifying latent variables is cluster analysis (CA), which assigns persons similar to each other and who differ from those in other clusters (MacQueen, 1967). As an unsupervised data mining technique, CA does not rely on any outcome measure and therefore holds great potential as a predictor of longitudinal outcomes (Everitt BS, Landau S, Leese M, & Stahl D, 2011; Lloyd, 1982). In the epidemiological literature the use of CA has been limited to a few particular applications such as the characterization of dietary patterns (Reedy J et al., 2010; Siou GL, Yasui Y, Csizmadi I, McGregor SE, & Robson PJ, 2011) or risk factors for specific conditions (Ness RB et al., 2005; Pugliatti M et al., 2006). Even rarer are examples evaluating the longitudinal association between cluster assignment and a subsequent disease such as colorectal cancer (Reedy J, et al., 2010). The purpose of this study is to demonstrate whether baseline levels of cluster membership can potentially serve as useful predictors of the first occurrence of three important gerontologic outcomes over a long period of time. These three outcomes were occurrence of disability in any of four basic activities of daily living, occurrence of disability in any of four mobility activities, and death.

METHODS

Study Population

The Precipitating Events Project (PEP) is an on-going longitudinal study of 754 community-living persons, aged 70 or older, who were initially nondisabled in four basic activities of daily living (ADLs)—bathing, dressing, walking, and transferring (Katz S, 1988). The assembly of this cohort has been described in detail elsewhere (Gill TM, Desai MM, Gahbauer EA, Holford TR, & Williams CS, 2001; Hardy SE & Gill TM, 2004). Potential participants were members of a large health plan in greater New Haven, Connecticut, USA, and were excluded for significant cognitive impairment with no available proxy, life expectancy less than 12 months, plans to move out of the area, or inability to speak English (Gill TM, Hardy SE, & Williams CS, 2002). Only 4.6% of persons contacted refused screening, and 75.2% of those eligible agreed to participate and were enrolled from March 1998 to October 1999. Persons who refused to participate did not differ significantly from those who were enrolled in terms of age or sex. The study protocol was approved by the Yale Human Investigation Committee, and all participants provided verbal informed consent.

Data Collection

The fixed covariates of age at study enrollment, gender, ethnic group, and number of chronic conditions (total of 13 possible self-reported, physician-diagnosed chronic conditions: amputation of leg; arthritis; cancer (other than minor skin cancers); chronic lung disease; cirrhosis or liver disease; congestive heart failure; diabetes; fracture of wrist, arm, or spine since 50 years of age; hip fracture; hypertension; myocardial infarction; Parkinson disease; stroke) as well as all candidate variables for cluster formation, were recorded at the baseline assessment. Monthly telephone interviews that assessed the outcomes (monthly occurrence of disability in any of four basic activities of daily living, monthly occurrence of disability in

any of four mobility activities, and monthly occurrence of death) were completed through either death or up through 129 months (June 2009) with a completion rate greater than 99%.

Candidate Variables for the Formation of Clusters

A set of well established predictors of deleterious gerontologic outcomes were chosen as candidate variables for cluster formation. These candidate variables consisted of demographic variables (age in years and indicators of gender, living alone, education less than high school, non-white ethnic group), evaluation of depressive symptoms with the 11-item Center for Epidemiologic Studies – Depression (CES-D) scale (Irwin M, Artin KH, & Oxman MN, 1999), and cognitive status measured on a 0-30 point scale by the Folstein Mini-Mental State Examination (MMSE) (Folstein ME, Folstein SE, & McHugh PR, 1975). The following dichotomous criteria were also candidates for cluster generation: muscle weakness, exhaustion, low physical activity, weight loss, and slow gait Each of these dichotomous criteria were modified slightly from the original definitions by Fried (Fried LP et al., 1997; Gill TM, Gahbauer EA, Allore HG, & Han L, 2006). While this group of 13 candidate variables is fairly narrow for a data mining operation, these specific explanatory variables have strong associations for the three outcomes of interest as documented in previous analyses based on this data (Gill TM, Allore HG, Gahbauer EA, & Murphy TE, 2010; Gill TM, Gahbauer EA, Han L, & Allore HG, 2010). Because our goal was to evaluate the association between baseline cluster membership and three important gerontologic outcomes, we restricted our focus to this small group of explanatory variables.

Assessment of Longitudinal Outcomes

Telephone interviews gathered self-reported monthly observations of disability in activities of daily living (ADL) (Katz S, 1988), mobility disability, and death. ADL disability was defined as needing help with one or more of the four basic ADLs of bathing, dressing, walking, and transferring while mobility disability referred to having difficulty with at least one of the following four movement related activities: climbing a flight of stairs, walking a quarter mile, lifting and carrying 10 pounds, and driving a car. Monthly death was ascertained by review of local obituaries and/or from an informant during a subsequent telephone interview. Attrition for reasons other than death was less than 5%. Over the 129 month follow-up period, three hundred thirty (43.8%) participants died with median follow-up of 56.5 months, while 34 (4.5%) dropped out with median follow-up of 23.5 months.

Statistical Analysis

The demographic and geriatric characteristics of the cohort were described at baseline. Several combinations of clustering algorithms and similarity measures were fit to the baseline values of those variables serving as candidates for cluster formation. The subsection immediately following describes in greater detail the cluster selection process used to select the best clustering approach.

Clustering Analysis

In this study we conducted CA using the Matlab® package which flexibly allowed the fitting of different clustering algorithms with a variety of similarity measures. We applied

the following CA algorithms: agglomerative clustering, spectral clustering, and K-means. Based on the graphical metrics explained below in this section, the agglomerative and spectral algorithms did not fit this data well. Per the same criteria, the K-means algorithm fit the data well and was therefore applied. The K-means algorithm is a partitioning approach that generates K disjoint clusters of persons from the data. The K-means algorithm starts by randomly selecting a person to represent the center of each cluster. All remaining persons are then assigned to the particular cluster of highest similarity per the chosen similarity measure. This process is iteratively performed until all persons are placed in the group of clusters that maximizes overall similarity. We evaluated similarity measures based on Euclidean distance and the cosine function. While Euclidean distance between two persons is computed as the sum of the squared differences between their corresponding characteristics, cosine similarity is computed as the dot product between their vectors of normalized characteristics. When some of the cluster-defining-variables are binary, as in our case, Euclidean distance has a notable limitation, namely the tendency to yield ties in cluster assignment that are resolved through arbitrary assignment. We believe it is partially due to this tendency that the K-means algorithm, in concert with cosine similarity, emerged as the best clustering approach.

In addition to examining the clinical content of the respective clusters, selection of the final value of K was based on two graphical metrics. Both graphical metrics are output from the MATLAB software which structures data into matrices in a way that streamlines complex computations. The silhouette graph is based on the silhouette measure (Rousseeuw PJ. 1987) and can be produced using the silhouette command in MATLAB. The silhouette graph provides a compact visual summary of how well members fit into their respective clusters. Silhouette values for each participant range continuously between positive and negative one with values close to positive one indicating good cluster placement and values close to negative one suggesting better placement in a different cluster. The second graphical approach is a diagonal block plot where the relative darkness of each block, i.e. cluster, reflects its relative cohesiveness. The relative size of each diagonal block reflects the number of persons assigned to it.

Choice of Clustering Algorithm and Number of Clusters

In traditional CA an algorithm identifies the natural clusters in the data by assigning membership such that persons in one cluster are highly similar to each other and discernibly different from persons in other clusters. After the clusters are formed a mathematical criterion is used to determine how cohesive the members assigned to each cluster are, i.e., a similarity measure. Because different combinations of clustering algorithms and similarity measures have advantages and disadvantages for different data types, data exploration is needed to justify the corresponding choices. A second challenge is the estimation of the natural number of clusters in the data. Investigators usually determine the optimal number of clusters by selecting a solution with good properties that makes sense to content experts (Dhillon IS, Guan Y, & Kulis B., 2004). For instance, the K-means algorithm requires that the analyst chooses the number of clusters, i.e., K, prior to cluster assignment. This is often done by fitting a range of cluster solutions, say from K=2 through 10 clusters, which are subsequently examined by content experts for face validity. Most previous applications of

CA to gerontologic data simply apply the default clustering option provided by a statistical package and do not explicitly consider choice of algorithm or similarity measure (Fukuoka Y, Lindgren TG, Rankin SH, Cooper BA, & Carroll DL, 2007; Montero-Odasso M et al., 2009).

Latent Class Analysis

For its use of statistical models, some practitioners have advocated that clustering be performed with finite mixture distributions (Everitt BS, & Hand DJ, 1981; MacLachlan GJ & Peel D, 2001). This latter procedure has also been referred to as latent class cluster analysis (LCA) (Vermunt JK, & Magidson J, 2002) and is available in the statistical package Mplus (Muthén LK & Muthén BO, 1998-2011). LCA uses a finite mixture of distributions to calculate the probability of membership of each participant in each of the categories (classes) being considered, and consequently assigns final membership of a given person to the class exhibiting highest probability of membership (Nagin R., 2005). Because of its established presence in the biostatistical literature, and its ready accessibility in the Mplus software package, we chose LCA as a reference technique against which to evaluate the best clustering approach.

Comparing the Best Clustering Approach with Latent Class Analysis (LCA)

We compared the following operating characteristics of the solutions yielded by the best clustering approach and LCA: optimal number of categories in each preferred solution, the criteria used to determine that optimal number, and the number of persons assigned to each category. The solutions produced by the two approaches (best clustering and LCA) were also described according to the mean levels of each cluster-defining-variable at baseline. Linearity of the ordinal group levels across each of the defining variables was graphically verified for both approaches.

Time-to-Event Models of First Occurrence of Three Gerontologic Outcomes

Finally, for each of the three outcomes (first occurrence of any ADL disability, any mobility disability, and death), a Cox proportional hazards regression model of each outcome, with adjustment for age at baseline, gender, ethnic group, and number of chronic conditions (out of 13 possible) evaluated the association with baseline cluster level over a 10 year period. The model results from the best cluster approach were compared to those from the LCA solution. Statistical significance was defined by a p-value less than 0.05 and all analyses were conducted using SAS/STAT® software ® V9.3 (SAS, 2009). CA was performed using MATLAB ® Version 7.11.0.584:R2010b (MATLAB, 2007) and LCA with Mplus Version 7.11 (Muthén LK & Muthén BO, 1998-2011).

RESULTS

Description of the Cohort's Baseline Characteristics

At baseline the cohort had a mean age of 78.4 years, with 13.5% being 85 years or older. About two-thirds of the cohort was female (65%) and the vast majority were non-Hispanic white (90.5%). A minority lived alone (39.5%), and the median educational level was 12 years. With regards to cognitive status, 86 persons (11.4%) had cognitive impairment,

defined by a MMSE score < 24. As defined by a score of 20 or more on the CES-D Scale (Radloff LS, 1977), 100 persons (13.3%) exhibited depressive symptoms. Nine self-reported chronic conditions were prevalent in the following proportions: hypertension (55.2%), arthritis (30.1%), myocardial infarction (18%), cancer (16.4%), chronic lung disease (10.2%), stroke (8.6%), diabetes mellitus (8.2%), congestive heart failure (6.5%), and hip fracture (4.5%).

Based on iterative examination of the two graphical metrics described in Methods, i.e., silhouette values and diagonal block plots, removal of the following characteristics did not substantively diminish CA's ability to classify study participants: age, education less than high school, gender, living alone, MMSE in dichotomous form (indicator of score < 24), muscle weakness, nonwhite ethnic group, physical activity, and weight loss. Because the inclusion of variables that do not substantially affect the algorithm's ability to refine membership are redundant, i.e., do not help to produce more consistent clusters, they are usually dropped. This is an application of the principle of parsimony, which asserts that all other factors being equal, it is preferable to choose the simpler of two competing approaches. For this reason these nine covariates were subsequently dropped, leaving a reduced set of characteristics consisting of number of depressive symptoms, cognitive status (MMSE score in continuous form), slow gait and exhaustion. From this smaller set of variables, and based on the same two graphical criteria, we observed that the K-means algorithm, in combination with cosine similarity, formed the most cohesive and interpretable clusters from the different CA algorithms examined. Solutions for K=2 through 10 were computed, with K=3 having most clinical relevance in that the values of the four characteristics followed consistent and mostly monotonic patterns across the clusters. The Bayesian Information Criterion was also lowest for the three cluster solution.

Figures 1A and 1B respectively depict the plots of the silhouette values and the diagonal block plots for the three baseline clusters yielded by the K-means algorithm in combination with cosine similarity. The three clusters of Figure 1A are characterized by positive silhouette scores mostly between 0.4 and 0.8. In the diagonal of Figure 1B, the more intense darkness of the square blocks, relative to the non-diagonal areas, indicates high similarity among the members of each cluster level. The diagonal blocks from all three clusters are of comparable area, indicating they contain similar numbers of participants. In contrast, the diagonal block plot resulting from the combination of spectral clustering and Euclidean distance (not shown) exhibited three very light colored blocks of differing sizes that reflected poor cohesiveness within the clusters and pronounced imbalance among their proportional assignments.

Characteristics of Clustering Solutions from Cluster Analysis and Latent Class Analysis

Using the same four cluster-defining variables employed in the K-means approach (continuous measures of cognitive status and depressive symptoms and dichotomous indicators of slow gait and exhaustion), LCA was applied to the baseline values of each study participant. Table 1 compares the operating characteristics of the K-means and LCA approaches that were each fit to this data. Using different statistical, graphical, and subjective criteria, each approach independently selected three clusters/classes over their

corresponding four cluster/class solutions. To more easily discern between the CA and LCA three group solutions in Table 1 and for the duration of this article, we will refer to the final three group clustering solution as K-means.

Distribution of Cluster-Defining-Variables within Three Group Solutions from K-means and LCA

In Table 2, the groups assigned at baseline by both approaches (K-means and LCA) exhibit a generalized trend of deterioration in each defining variable as cluster level increases. For this reason cluster level was interpreted as representing ordinal level of overall impairment across the four cluster defining variables. Table 2 labels the groups in ascending order of impairment as follows: Low, Moderate, and Severe. We use these labels for the remainder of this article. We note that the clusters corresponding to Severe Impairment were characterized by average values of depressive symptoms that correspond closely to a threshold, i.e., CES-D = 20, previously cited as evidence of clinically significant depressive symptoms (Barry LC, Allore HG, Guo Z, Bruce ML, & Gill TM, 2008). Finally, the number of persons assigned to each K-means cluster was comparable across the three ordinal levels of impairment, whereas the LCA clusters exhibited decreasing population as impairment rose.

Associations between Baseline Group Membership and First Occurrence of Gerontologic Outcomes

Table 3 presents the hazard ratios and their 95% confidence intervals from both approaches (K-means and LCA) for incremental change in the ordinal level of baseline group membership and first occurrence of the three outcomes over a ten year follow-up period. Incremental rise in ordinal cluster level, as estimated by both approaches, exhibited highly significant positive associations with all three outcomes. In all cases the 95% confidence intervals strongly overlap, indicating virtual equivalence in the associations between ordinal level of impairment and first occurrence of each outcome yielded by the two approaches.

DISCUSSION

This application of CA to gerontologic outcomes features several strengths lacking in previously published examples (Eastwood et al., 2002; Michel et al., 2000; Penrod et al., 2007). Whereas those publications only considered a single clustering algorithm, the process described here explored different algorithms and similarity measures in its empirical search for the optimal solution. We have also shown that baseline cluster membership, interpreted as an ordinal level of a latent state of physical and mental impairment, demonstrated highly significant positive associations with first occurrence of three important gerontologic outcomes over a long period of time. In the gerontologic research community, there is growing belief that measures of “thinking-moving-feeling” (TMF) are manifestations of a latent entity that is a powerful predictor of gerontologic outcomes (Granholt A-C, Boger H, & Emborg ME, 2008). This latent factor is indirectly measured in terms of a person’s cognitive, physical, and emotional states and may be interpreted as the time-varying status of one’s aging process. The variables that define our clusters directly align with TMF: score from the Mini-Mental State Examination (thinking), slow gait (moving), and depressive

symptoms (feeling). The outcomes examined here, i.e., any ADL disability, any mobility disability, and mortality, are characteristic of the aging process of all older persons and therefore qualify as universal outcomes (Tinetti ME et al., 2011), i.e., those not restricted to specific diseases. In summary, these results support the potential for using baseline measures of TMF as predictors of multivariate universal outcomes.

Limitations

There are several limitations to our study that warrant comment. The two primary limitations of the data are that they are based entirely on self-reporting and are observational in nature. These limitations respectively imply the possibility of reporting-bias and preclude any claims of causality. Because previous clinical applications of CA did not mention the exploration of different algorithms and similarity measures, our empirical approach implies added complexity for the analyst. For example, the exploration of various algorithms and similarity measures implies the use a programming language such as MATLAB®, as opposed to an off-the-shelf package such as SAS or Mplus. Whereas an external validation would be preferable to our graphical internal validation, to our knowledge there is not a comparable observational dataset containing monthly assessments over a long period of time for multiple gerontologic outcomes.

CONCLUSION

In this era of inexpensive, efficient computation, epidemiology is benefitting from access to big datasets and the data mining techniques that explore them. While CA has previously been used to identify natural groupings in data and to occasionally predict disease outcomes, we have shown that the use of CA to assign baseline levels of impairment exhibits significant positive associations with first occurrence of three important gerontologic outcomes over a ten year period. Utilizing LCA as the reference method of classification in this long term analysis, CA yielded a solution providing equivalent hazard ratios. The associations demonstrated between baseline clusters representing ordinal levels of latent impairment that follow the TMF paradigm, and three outcomes that are relevant to the vast majority of older persons, represent a higher level of predictive sophistication than previously reported. These results support the feasibility of baseline cluster membership as a potential long term predictor of multivariate longitudinal outcomes among older persons living in community.

ACKNOWLEDGEMENTS

We thank Denise Shepard, BSN, MBA, Andrea Benjamin, BSN, Paula Clark, RN, Martha Oravetz, RN, Shirley Hannan, RN, Barbara Foster, Alice Van Wie, BSW, Patricia Fugal, BS, Amy Shelton, MPH, and Alice Kossack for assistance with data collection; Evelyne Gahbauer, MD, MPH for data management and programming; Wanda Carr and Geraldine Hawthorne for assistance with data entry and management; Peter Charpentier, MPH for development of the participant tracking system; Joanne McGloin, MDiv, MBA for leadership and advice as the PEP Project Director.

FUNDING

Supported by grants from VA grant HIR 08-374 HSR&D: Consortium for Healthcare Informatics (Fodeh and Brandt) and in part by CTSA Grant Number UL1 RR024139 (Brandt) from the NIH National Center for Advancing Translational Sciences (NCATS). Support was also received from the National Institute on Aging (R37AG17560 - Gill, Allore and Murphy and 1R21AG033130-01 -Murphy). The study was conducted at the Yale Claude D.

Pepper Older Americans Independence Center (P30AG21342). Dr. Gill is the recipient of an Academic Leadership Award (K07AG043587) from the National Institute on Aging. This publication was also made possible in part by CTSA Grant Number UL1 RR024139 (Murphy) from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH roadmap for Medical research.

References

- Barry LC, Allore HG, Guo Z, Bruce ML, Gill TM. Higher burden of depression among older women - the effect of onset, persistence, and mortality over time. *Archives of General Psychiatry*. 2008; 65(2):172–178. [PubMed: 18250255]
- Dhillon IS, Guan Y, Kulis B. Kernel k-means: spectral clustering and normalized cuts. Paper presented at the The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004
- Eastwood EA, Magaziner J, Wang J, Silberzweig SB, Hannan EL, Strauss E, Siu AL. Patients with hip fracture: subgroups and their outcomes. *Journal of the American Geriatrics Society*. 2002; 50(7): 1240–1249. [PubMed: 12133019]
- Everitt, BS.; Landau, S.; Leese, M.; Stahl, D. *Cluster Analysis*. 5th. John Wiley & Sons; Chichester, UK: 2011.
- Folstein ME, Folstein SE, McHugh PR. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*. 1975; 12(3):189–198. doi: doi: 10.1016/0022-3956(75)90026-6. [PubMed: 1202204]
- Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, McBurnie MA. Frailty in older adults: evidence for a phenotype. *Journals of Gerontology. Series A: Biological Sciences and Medical Sciences*. 1997; 56A(3):M146–M156.
- Fukuoka Y, Lindgren TG, Rankin SH, Cooper BA, Carroll DL. Cluster analysis: a useful technique to identify elderly cardiac patients at risk for poor quality of life. *Quality of Life Research*. 2007; 16:1655–1663. [PubMed: 17955345]
- Gill TM, Allore HG, Gahbauer EA, Murphy TE. Change in disability after hospitalization or restricted activity in older persons. *Journal of the American Medical Association*. 2010; 304(17):1919–1928. [PubMed: 21045098]
- Gill TM, Desai MM, Gahbauer EA, Holford TR, Williams CS. Restricted activity among community-living older persons: incidence, precipitants and health care utilization (with editorial). *Annals of Internal Medicine*. 2001; 135:313–321. [PubMed: 11529694]
- Gill TM, Gahbauer EA, Allore HG, Han L. Transitions between frailty states among community-living older persons. *Archives of Internal Medicine*. 2006; 166:418–423. [PubMed: 16505261]
- Gill TM, Gahbauer EA, Han L, Allore HG. Trajectories of disability in the last year of life. *N Engl J Med*. 2010; 362(13):1173–1180. doi: PMID 20357280. [PubMed: 20357280]
- Gill TM, Hardy SE, Williams CS. Underestimation of disability among community-living older persons. *Journal of the American Geriatrics Society*. 2002; 50:1492–1497. [PubMed: 12383145]
- Granhölm A-C, Boger H, Emborg ME. Mood, memory and movement: an age-related neurodegenerative complex? *Current Aging Science*. 2008; 1(2):133–139. doi: 10.2174/1874609810801020133. [PubMed: 20021382]
- Hardy SE, Gill TM. Recovery from disability among community-dwelling older persons. *Journal of the American Medical Association*. 2004; 291:1596–1602. [PubMed: 15069047]
- Irwin M, Artin KH, Oxman MN. Screening for depression in the older adult: criterion validity of the 10-item Center for Epidemiological Studies Depression Scale (CES-D). *Archives of Internal Medicine*. 1999; 159(15):1701–1704. [PubMed: 10448771]
- Katz S. Assessing self-maintenance: Activities of daily living, mobility and instrumental activities of daily living. *Journal of the American Geriatrics Society*. 1988; 31(12):721–726. [PubMed: 6418786]
- Lloyd SP. Least square quantization in PCM. Paper presented at the IEEE Transactions on Information Theory. 1982

- MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. Paper presented at the Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967
- MATLAB. Matlab User's Guide. The MathWorks; Natick, MA: USA: 2007.
- Michel JP, Hoffmeyer P, Klopfenstein C, Bruchez M, Grab B, d'Epinay CL. Prognosis of functional recovery 1 year after hip fracture: typical patient profiles through cluster analysis. [Research Support, Non-U.S. Gov't]. *Journals of Gerontology. Series A: Biological Sciences and Medical Sciences*. 2000; 55(9):M508–515.
- Montero-Odasso M, Bergman H, Beland F, Sourial N, Fletcher JD, Dallaire L. Identifying mobility heterogeneity in very frail older adults. Are frail people all the same? *Archives of Gerontology and Geriatrics*. 2009; 49:272–277. [PubMed: 18986718]
- Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*. Sixth. Muthén & Muthén; Los Angeles, CA: 1998-2011.
- Ness RB, Kip KE, Hillier SL, Soper DE, Stamm CA, Sweet RL, Richter HE. A cluster analysis of bacterial vaginosis-associated microflora and pelvic inflammatory disease. *American Journal of Epidemiology*. 2005; 162:585–590. [PubMed: 16093289]
- Penrod JD, Litke A, Hawkes WG, Magaziner J, Koval KJ, Doucette JT, Siu AL. Heterogeneity in hip fracture patients: age, functional status, and comorbidity. [Multicenter Study Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.]. *Journal of the American Geriatrics Society*. 2007; 55(3):407–413. doi: 10.1111/j.1532-5415.2007.01078.x. [PubMed: 17341244]
- Pugliatti M, Riise T, Sotgiu MA, Satta WM, Sotgiu S, Pirastru MI, Rosati G. Evidence of early childhood as the susceptibility period in multiple sclerosis: space-time cluster analysis in a Sardinian population. *American Journal of Epidemiology*. 2006; 164:326–333. [PubMed: 16754634]
- Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*. 1977; 1:385–401.
- Reedy J, Wirfaelt E, Flood A, Panagiota NM, SM Krebs-Smith, Kipnis V, Subar AF. Comparing 3 dietary pattern methods-cluster analysis, factor analysis, and index analysis-with colorectal cancer risk, The NIH-AARP Diet and Health Study. *American Journal of Epidemiology*. 2010; 171:479–487. [PubMed: 20026579]
- SAS. *SAS User's Manual (Version 9.2)*. SAS Institute; Cary, NC: 2009.
- Siou GL, Yasui Y, Csizmadia I, McGregor SE, Robson PJ. Exploring statistical approaches to diminish subjectivity of cluster analysis to derive dietary patterns, The Tomorrow Project. *American Journal of Epidemiology*. 2011; 173(8):956–967. [PubMed: 21421742]
- Tinetti ME, McAvay GJ, Chang SS, Newman AB, Fitzpatrick AL, Fried TR, Peduzzi PN. Contribution of multiple chronic conditions to universal health outcomes. *Journal of the American Geriatric Society*. 2011; 59:1686–1691.

Figure 1. Three Clusters Derived at Baseline from K-means and Cosine Similarity
Panel A - Silhouette Values
Panel B - Diagonal Block Plots
New Haven, Connecticut, United States, 1998 - 2009

Table 1

Baseline Selection of Clusters for Modeling Gerontologic Outcomes, New Haven, Connecticut, United States, 1998 - 2009

Classification Technique	Software	Candidate Solutions	Optimization Criteria	Optimal Solution
K-means Clustering Algorithm with Cosine Similarity	Matlab	3 clusters versus 4	Bayesian Information Criterion Clinical Content Graphical Metrics	3 clusters
Latent Class Analysis	Mplus	3 classes versus 4	Vuong-Lo-Mendell-Rubin Likelihood Ratio Test	3 classes

Table 2

Variables across Three Group Solutions from K-means and Latent Class Analysis (LCA), New Haven, Connecticut, United States, 1998 – 2007

Defining Variables	Classification Approach	Low Impairment Group	Moderate Impairment Group	Severe Impairment Group
		Mean	Mean	Mean
Cognitive Status (lower mean is worse)	K-means	27.05	27.31	25.87
	LCA	27.16	26.22	25.39
Depressive Symptoms (higher mean is worse)	K-means	0.13	7.72	19.11
	LCA	4.03	15.89	27.61
		Percent of Members With	Percent of Members With	Percent of Members With
Exhaustion (higher percent is worse)	K-means	1	4	34
	LCA	2	23	66
SlowGait (higher percent is worse)	K-means	32	33	63
	LCA	31	60	71
		Group Size	Group Size	Group Size
Number of Persons in Cluster	K-means	241	269	244
	LCA	479	219	56

K-means = clustering algorithm used with cosine similarity

Table 3

Time-to-Event Associations between Baseline Group and First Occurrence of Three Gerontologic Outcomes from K-means and Latent Class Analysis (LCA), New Haven, Connecticut, United States, 1998 - 2009

Hazard Ratios (95% Confidence Interval)						
Explanatory Variables	First Disability in ADLs		First Disability in Mobility		Death	
	K-means	LCA	K-means	LCA	K-means	LCA
Ordinal Level of Impairment from Baseline Classification [@]	1.18** (1.05, 1.32)	1.25** (1.09, 1.43)	1.28** (1.16, 1.41)	1.35** (1.20, 1.53)	1.20** (1.05, 1.36)	1.21* (1.04, 1.41)
Age in Years at Baseline	1.05** (1.04, 1.07)	1.05** (1.04, 1.07)	1.08** (1.06, 1.10)	1.08** (1.06, 1.10)	1.08** (1.06, 1.10)	1.08** (1.06, 1.10)
Female Sex	1.00 (0.83, 1.21)	1.00 (0.82, 1.20)	1.84** (1.56, 2.17)	1.81** (1.53, 2.13)	0.70** (0.57, 0.87)	0.70** (0.57, 0.86)
Non-White	1.00 (0.73, 1.36)	0.98 (0.72, 1.34)	1.37* (1.05, 1.78)	1.34* (1.03, 1.75)	0.81 (0.56, 1.18)	0.81 (0.56, 1.17)
Number of Comorbid Conditions (0 – 13)	1.21** (1.12, 1.30)	1.20** (1.11, 1.29)	1.23** (1.15, 1.31)	1.22** (1.14, 1.30)	1.29** (1.19, 1.40)	1.29** (1.19, 1.39)
Model Fit						
Bayesian Information Criterion ^{&} (smaller is better)	5064.83	5063.37	5115.70	5118.72	3097.64	3099.35

ADLs = activities of daily living, i.e., bathing, dressing, walking, and transferring Disability in Mobility = need help with one or more of: walking ¼ mile, flight of stairs, carrying 10 pounds, driving

[@] based on measures of depressive symptoms, exhaustion, mental status, and slow gait, updated at each 18 month comprehensive assessment

* p-value < 0.05

** p-value < 0.01

[&] differences less than five are not statistically meaningful