ORIGINAL ARTICLE

K-Means Cluster Analysis of Rehabilitation Service Users in the Home Health Care System of Ontario: Examining the Heterogeneity of a Complex Geriatric Population

Joshua J. Armstrong, MSc, Mu Zhu, PhD, John P. Hirdes, PhD, Paul Stolee, PhD

ABSTRACT. Armstrong JJ, Zhu M, Hirdes JP, Stolee P. K-means cluster analysis of rehabilitation service users in the home health care system of Ontario: examining the heterogeneity of a complex geriatric population. Arch Phys Med Rehabil 2012;93:2198-205.

Objective: To examine the heterogeneity of home care clients who use rehabilitation services by using the K-means algorithm to identify previously unknown patterns of clinical characteristics.

Design: Observational study of secondary data.

Setting: Home care system.

Participants: Assessment information was collected on 150,253 home care clients using the provincially mandated Resident Assessment Instrument–Home Care (RAI-HC) data system.

Interventions: Not applicable.

Main Outcome Measures: Assessment information from every long-stay (>60d) home care client that entered the home care system between 2005 and 2008 and used rehabilitation services within 3 months of their initial assessment was analyzed. The K-means clustering algorithm was applied using 37 variables from the RAI-HC assessment.

Results: The K-means cluster analysis resulted in the identification of 7 relatively homogeneous subgroups that differed on characteristics such as age, sex, cognition, and functional impairment. Client profiles were created to illustrate the diversity of this geriatric population.

Conclusions: The K-means algorithm provided a useful way to segment a heterogeneous rehabilitation client population into more homogeneous subgroups. This analysis provides an enhanced understanding of client characteristics and needs, and could enable more appropriate targeting of rehabilitation services for home care clients.

Key Words: Cluster analysis; Geriatrics; Home care services; Population characteristics; Rehabilitation.

WITH THE AGING of the population and pressures to limit the use of inpatient/hospital services, home-based services are an increasingly important component of the health care system. In the Canadian province of Ontario, as well as many other jurisdictions, many older adults rely on services from the home care sector to continue living in their own homes. Included in the array of services offered by home care in Ontario are rehabilitation therapies (physical therapy [PT] and occupational therapy [OT]), which have been demonstrated to be effective for supporting the independence of older persons in home-based settings.[1-7]

As the number of older home care clients needing PT and OT services is expected to increase in the years to come,[7] a better understanding of the population characteristics and the differing needs of these clients is needed in order to effectively plan and efficiently deliver rehabilitation services. Currently in Ontario, over 1 million OT and PT visits to home care clients are provided per year through Community Care Access Centres, the organizations that coordinate home care services and long-term care placement in Ontario.[8] Despite the importance of these therapeutic services and the large volume of service provision, little is currently known about home care rehabilitation clients and whether there are distinct subgroups that exist within this population that could benefit from specialized services. This study aims to enhance the understanding of this care population by using census level data and the K-means clustering algorithm.

Efforts to understand the health and service needs of an aging population have commonly focused on identifying predictors and risk factors related to outcomes such as service use,[9,10] institutionalization,[11,12] functional improvement/decline,[13,14] and mortality.[15,16] Less emphasis has been placed on understanding the individual differences and naturally occurring groupings within study populations.[17] Medical research studies tend to treat all patients as the same, controlling for individual patient differences as covariates in statistical models.[18] Although important findings can arise from using a traditional approach that statistically controls for individual differences, disregarding the heterogeneity found in patient populations can conceal meaningful patterns in patient characteristics.

Typically in medical and epidemiologic research, the focus is on population averages, and heterogeneity is actively sup-

List of Abbreviations

ADLs	activities of daily living
CHESS	Changes in Health, End-Stage Disease, and Signs and Symptoms Scale
IADLs	instrumental activities of daily living
OT	occupational therapy
PT	physical therapy
RAI-HC	Resident Assessment Instrument–Home Care

From the School of Public Health and Health Systems, University of Waterloo (Armstrong, Hirdes, Stolee); Department of Statistics and Actuarial Science, University of Waterloo (Zhu); and the Schlegel – University of Waterloo Research Institute for Aging (Stolee), Waterloo, ON, Canada.

Presented to the Gerontological Society of America, November 21, 2011, Boston, MA; and the Canadian Association on Gerontology, December 3, 2010, Montreal, QC, Canada.

Supported by the Canadian Institutes of Health Research (grant no. ETG-92249).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

Reprint requests to Paul Stolee, PhD, School of Public Health and Health Systems, University of Waterloo, 200 University Ave W, Waterloo, ON, N2L 1P3, Canada, e-mail: *stolee@uwaterloo.ca*.

In-press corrected proof published online on Aug 1, 2012, at www.archives-pmr.org.

0003-9993/12/9312-01295$36.00/0

http://dx.doi.org/10.1016/j.apmr.2012.05.026

pressed and treated as noise[19,20]; however, ignoring individual differences by focusing on averages can lead to misleading research results and can potentially harm patients. For example, population heterogeneity has been shown to cause errors in calculating the cost-effectiveness of interventions, as well as in the results of regression models and decision analytic models.[21-24] Carefully considering patient heterogeneity may be necessary to obtain unbiased model outcomes when making uniform recommendations for a heterogeneous population,[25] such as policy changes within health services. The impact of heterogeneity is potentially even stronger in geriatric populations because of the complexities in clinical status that are associated with aging.[26,27] Older patients are much more likely to have differing patterns of numerous medical problems, multiple chronic diseases, atypical presentations of medical conditions, multiple prescription medications (polypharmacy), cognitive impairments, and sensory impairments.[28-36] These increased amounts of clinical complexities found in older adults lead to geriatric patient populations that are challenging to manage.[37] Although many research articles often mention the heterogeneity found in geriatric patient populations, only a handful of more recent studies have directly focused on uncovering the heterogeneity found among older patients, including patients with hip fracture,[38,39] cardiac problems,[40,41] and frailty.[17,42]

METHODS

In this article, we aimed to explore the heterogeneity of home care clients who use rehabilitation services, discover previously unidentified patterns of clinical characteristics, and create client profiles to illustrate the different subgroups found within this complex client population. This study used data collected based on the Resident Assessment Instrument–Home Care (RAI-HC).[43,44] The RAI-HC assessment system has been mandated for use for all clients expected to use home care services for more than 60 days, which provides Ontario census-level data on long-stay home care clients. For this exploratory cluster analysis, we used the RAI-HC data of 150,253 clients who received rehabilitation services (OT or PT) within the first 3 months of their initial home care assessment. Data were collected between April 2005 and August 2008. Ethics approval was granted by the University of Waterloo's Office of Research.

The RAI-HC is one of a suite of standardized assessment tools developed by the international interRAI consortium.[45] The instrument contains a wide variety of assessment items including demographic information, cognition, physical functioning, disease diagnoses, nutrition/hydration status, environmental assessment, and service use. Collected by frontline workers using data entry software, the assessment system has checks during data input that constrain item entries as nonmissing, within correct ranges, and with logical checks.

Because of the large number of variables available within this assessment data (≥300), variables to be used in the K-means cluster analysis were selected through quantitative variable selection techniques (Proc Varclus in SAS 9.1)[a] and consultation with fellow researchers from the infoRehab team (http://www.inforehab.uwaterloo.ca), who have knowledge of rehabilitation services for home care clients. In total, 37 variables were used for the K-means cluster analysis; these included activities of daily living (ADLs), instrumental activities of daily living (IADLs), disease diagnoses, sex, Changes in Health, End-Stage Disease, Symptoms and Signs (CHESS), a health instability outcome measure,[46] and age. A complete list of all the measures used and their brief descriptions can be found in table 1. This table also includes the proportion of the entire rehabilitation service user population.

Thirteen ADLs and IADLs items were included, which covered a variety of physical functioning domains. These items were reduced to a dichotomous form (independent vs dependent). Fourteen disease diagnosis categories were also included, and these were also coded dichotomously as present/not present. Additional variables included in the cluster analysis were sex, age (standardized using Euclidean distance), presence of daily pain, multiple recent falls (≥2 vs <2), unsteady gait, problems with decision-making, presence of pressure ulcers, CHESS score (≥2 vs <2), and home living status (alone vs living with caretakers).

K-Means Cluster Analysis

In this study, we sought to discover patterns of clinical features by using the K-means algorithm,[47] a popular data partition method widely used in many fields including data mining, pattern recognition, decision support, machine learning, and image segmentation. This algorithm is considered by the data mining and machine learning community to be an unsupervised learning technique, because it searches for patterns among input variables without using an outcome variable to dictate how the pattern is formed. In other words, K-means clustering is a way to use data to uncover natural groupings within a heterogeneous population. To uncover patterns, the algorithm starts by first assigning data points into random groups. The group centers are then calculated and the group memberships are reassigned based on the distances between each data point and the group centers. This process is repeated until there are no changes in the group memberships from the previous iteration. With the exception of age, all variables used in the K-means cluster analysis were dichotomized. This allows for easy interpretation as a mean score of a dichotomous variable directly relates to the proportion of clients with a score of 1.

To perform this analysis, we used the procedure FASTCLUS in SAS.[a] We chose the K-means algorithm and the SAS implementation in particular because of its suitability for analyzing relatively large data sets, as well as its use of a spacing heuristic for initial group assignments in order to avoid suboptimal solutions.[47] All analyses were performed in SAS version 9.1.[a]

Because of the fact that K (the number of clusters) needs to be specified prior to analysis, we used an iterative process to explore a range (2–20) of possible cluster solutions. For each possible cluster solution, we examined 3 statistical criteria: (1) the cubic clustering criterion,[48] (2) the pseudo F,[49] and (3) the squared multiple correlation. To compare these 3 criteria over the range of solutions, each of 3 statistics were graphed by the number of potential clusters.

1-Year Service Outcomes

In addition to the clinical variables collected with the RAI-HC, service outcome data were available for each of these clients for the time period of 1 year after the clients' initial home care assessment. These service outcomes indicate if the clients left the home care system and include: successful completion of care plan (released to live at home), hospitalization, long-term care placement, and mortality. After the cluster analysis, service outcomes were used to examine outcome differences between the clusters.

RESULTS

The first column in table 1 presents the variables used in the cluster analyses, as well as the full sample baseline demographic, functional, and health characteristics. For the entire sample, the average age ± SD was 76.8±13.2 years, 12.6% were