

Proposal

Clustering of Unhealthy Behaviors: Protocol for a Multiple Behavior Analysis of Data From the Canadian Longitudinal Study on Aging

Zack van Allen^{1,2}, MA, MBA; Simon L Bacon^{3,4}, PhD; Paquito Bernard^{5,6}, PhD; Heather Brown⁷, PhD; Sophie Desroches⁸, PhD; Monika Kastner^{9,10}, PhD; Kim Lavoie^{4,11}, PhD; Marta Marques^{12,13}, PhD; Nicola McCleary^{1,2}, PhD; Sharon Straus⁹, PhD; Monica Taljaard^{1,2}, PhD; Kednapa Thavorn^{1,2}, PhD; Jennifer R Tomasone¹⁴, PhD; Justin Presseau^{1,2}, PhD

¹School of Psychology, University of Ottawa, Ottawa, ON, Canada

²Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

³Department of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, QC, Canada

⁴Montreal Behavioural Medicine Centre, Le Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, Montreal, QC, Canada

⁵Department of Physical Activity Sciences, University of Quebec in Montreal, Montreal, QC, Canada

⁶Research Center of the Montreal Mental Health University Institute, Montreal, QC, Canada

⁷Population Health Sciences Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

⁸Department of Food and Nutrition Sciences, Laval University, Quebec City, QC, Canada

⁹Department of Medicine, University of Toronto, Toronto, ON, Canada

¹⁰North York General Hospital, Toronto, ON, Canada

¹¹Department of Psychology, University of Quebec in Montreal, Montreal, QC, Canada

¹²ADAPT Science Foundation Ireland Research Centre, Trinity College Dublin, Dublin, Ireland

¹³Comprehensive Health Research Centre, NOVA Medical School, Lisbon, Portugal

¹⁴School of Kinesiology and Health, Queen's University, Kingston, ON, Canada

Corresponding Author:

Zack van Allen, MA, MBA

School of Psychology

University of Ottawa

75 Laurier Ave E

Ottawa, ON, K1N 6N5

Canada

Phone: 1 6138660060

Email: zvana030@uottawa.ca

Abstract

Background: Health behaviors such as physical inactivity, unhealthy eating, smoking tobacco, and alcohol use are leading risk factors for noncommunicable chronic diseases and play a central role in limiting health and life satisfaction. To date, however, health behaviors tend to be considered separately from one another, resulting in guidelines and interventions for healthy aging siloed by specific behaviors and often focused only on a given health behavior without considering the co-occurrence of family, social, work, and other behaviors of everyday life.

Objective: The aim of this study is to understand how behaviors cluster and how such clusters are associated with physical and mental health, life satisfaction, and health care utilization may provide opportunities to leverage this co-occurrence to develop and evaluate interventions to promote multiple health behavior changes.

Methods: Using cross-sectional baseline data from the Canadian Longitudinal Study on Aging, we will perform a predefined set of exploratory and hypothesis-generating analyses to examine the co-occurrence of health and everyday life behaviors. We will use agglomerative hierarchical cluster analysis to cluster individuals based on their behavioral tendencies. Multinomial logistic regression will then be used to model the relationships between clusters and demographic indicators, health care utilization, and general health and life satisfaction, and assess whether sex and age moderate these relationships. In addition, we will conduct network community detection analysis using the clique percolation algorithm to detect overlapping communities of behaviors based on the strength of relationships between variables.

Results: Baseline data for the Canadian Longitudinal Study on Aging were collected from 51,338 participants aged between 45 and 85 years. Data were collected between 2010 and 2015. Secondary data analysis for this project was approved by the Ottawa Health Science Network Research Ethics Board (protocol ID #20190506-01H).

Conclusions: This study will help to inform the development of interventions tailored to subpopulations of adults (eg, physically inactive smokers) defined by the multiple behaviors that describe their everyday life experiences.

International Registered Report Identifier (IRRID): DERR1-10.2196/24887

(*JMIR Res Protoc* 2021;10(6):e24887) doi: [10.2196/24887](https://doi.org/10.2196/24887)

KEYWORDS

health behaviors; multiple behaviors; cluster analysis; network analysis; CLSA

Introduction

Two-thirds of all annual deaths in Canada are caused by 4 noncommunicable chronic diseases: cancer, cardiovascular disease, diabetes, and chronic respiratory disease [1]. Approximately 12% of Canadians 65 years or older have lived with 2 or more of these chronic diseases, also known as multimorbidity [2]. Health behaviors such as physical inactivity, unhealthy eating, smoking tobacco, and alcohol use are leading risk factors for chronic diseases and play a central role in health status and quality of life [3]. The prevalence of risky health behaviors is high, with approximately 4 in 5 Canadian adults reporting at least one of these modifiable risk factors for noncommunicable chronic diseases [2].

Although the risk factors and consequences of multimorbidity have been studied extensively [4-6], much less attention has been paid to understanding how different combinations of multiple behaviors influence individuals' life satisfaction and health. Our everyday lives are defined by multiple co-occurring health, social, family, personal, and work-related behaviors, each vying for the limited time, energy, and motivation available [7]. In spite of this, health behaviors tend to be studied and promoted largely separately from each other, resulting in guidelines and interventions for healthy aging siloed by specific behaviors. For example, Canada has distinct sets of guidelines for physical activity and sleep [8] and alcohol consumption [9]. In addition to focusing on single health behaviors, guidelines often do not consider the interconnectedness—or co-occurrence—of the various other family, social, work, hobby, and other behaviors that characterize daily life. Understanding how behaviors cluster and how such clusters are associated with physical and mental health, life satisfaction, and health care utilization may provide new opportunities to leverage this co-occurrence to promote multiple health behavior change interventions tailored to which behaviors co-occur for whom, thus better reflecting the real-world complexity of health care for aging Canadians.

International population data support the co-occurrence and clustering of health behaviors. For instance, Irish national data collected in the 2007 National Survey of Lifestyle, Attitudes, and Nutrition investigated the co-occurrence of smoking, alcohol use, physical inactivity, and unhealthy eating in adults aged 18 years and older [10]. In this cross-sectional analysis, the authors identified 6 clusters of behaviors to describe the population, which were labeled as (1) *healthy lifestyle* (characterized by

people who had never smoked, high physical activity, highest healthy eating, and moderate alcohol use), (2) *former smokers* (former smokers who reported high physical activity, moderate alcohol use, and healthy eating), (3) *temperate* (moderately active and moderate drinkers who had never smoked), (4) *physically inactive* (people with low levels of physical activity, poor eating habits, and who reporting some smoking and high alcohol use), (5) *mixed lifestyle* (those who had never smoked and reported moderate physical activity and variable alcohol consumption), and (6) *multiple risk factor* (moderate physical activity, moderate to high alcohol use, and variable healthy eating). Although the cluster labels do not directly convey the co-occurrence of the behaviors, definitions of each label describe which behaviors co-occur most in each group in terms of the extent of the performance of each behavior. Beyond defining how behaviors cluster, they also showed that membership in each cluster was associated with demographic factors and health and well-being outcomes: those in the *healthy lifestyle* cluster were more likely to be women 65 years or older, be in the highest socioeconomic status (SES) group, and exhibit low psychological distress; those in the *former smokers* cluster were more likely than the healthy lifestyle cluster to be men and be in a lower SES group; the *temperate* cluster was associated with being male and being in a lower SES group as well as being younger; the *physically inactive* cluster was associated with being male aged 18 to 29 years and being in a lower SES group and having higher psychological distress; the *multiple risk factor* cluster had lowest reported energy and vitality, highest distress, lowest self-reported health, and lowest quality of life; and the *mixed lifestyle* cluster was associated with being male and younger and being in a lower SES group, with highest distress and lower energy and vitality. These findings show how the population can be segmented by the multiple health behaviors that characterize their lives and that these segmented clusters are socially patterned and associated with different health outcomes. It remains to be seen how such clusters may hold when focusing on adults aged 45 to 85 years and extending the behaviors under consideration beyond health behaviors. Identifying these clusters and their associations with demographic factors, general health, and health care utilization can inform the tailoring of future interventions targeting multiple behavior change.

In another study, Buck and Frosini [11] used data from the Health Survey for England to investigate the clustering of smoking, excessive alcohol use, unhealthy eating, and physical inactivity among adults aged 16 to 74 years between 2003 and

2008. They showed that in 2008, only 7% of people engaged in none of the behaviors and only 5% engaged in all 4; 63% of people engaged in 1 or 2, with the remaining 25% engaging in 3 or more behaviors. The most common co-occurring behaviors included either unhealthy eating and physical inactivity (more prevalent in women) or alcohol use, unhealthy eating, and inactivity (more prevalent in men). Men were more likely than women to have 3 co-occurring unhealthy behaviors, and those 65 years and older were more likely to have 2 co-occurring unhealthy behaviors than any other age group (16-24, 25-44, 45-64, and ≥ 65 years).

Although the importance of co-occurring health behaviors is clear, the role of nonhealth behaviors should not be overlooked. Using Australian population survey data collected in 2007 and 2009 in the Household Income and Labour Dynamics of Australia survey, we showed that other behaviors co-occurred with healthy eating and to some extent differed between men and women aged 18-65 years [12]. For Australian women, some family behaviors (caring for young children) and work behaviors (being employed in a managerial position) were associated with healthy eating, whereas not working was negatively associated with healthy eating. For men, the co-occurrence of work-related behaviors with healthy eating depended on the type of job category (positive for managers and negative for laborers). The findings highlight that nonhealth behaviors also co-occur with health behaviors and underscore the connectedness of the multiple behaviors in daily life.

The examples discussed so far have examined co-occurring health behaviors in the general population. However, some studies have focused on adults aged 50 years and older. For example, one study specifically focusing on older adults involved cross-sectional data collected in Germany in 2006, involving 982 men and 1020 women aged 50 to 70 years [13]. The authors investigated how 4 health behaviors (tobacco use, alcohol use, unhealthy eating, and physical inactivity) clustered in the population of adults in one German state and reliably grouped individuals by the unhealthy behaviors that they engaged in: 25% were defined by the lack of unhealthy behaviors, 21% as inactive, 18% as low fruit and vegetable eaters, 13% as smokers with other risk behaviors, and 23% as drinkers with other risk behaviors [13]. Although some adults were only described using one unhealthy behavior, over one-third had multiple unhealthy behaviors that clustered together. Furthermore, membership in each cluster was also patterned in terms of sociodemographic factors: those with multiple co-occurring unhealthy behaviors tended to be men, living alone, and of a lower SES. This study focused on a limited set of unhealthy behaviors, and it remains unclear which other behaviors characterizing daily life may also co-occur.

A different methodological approach was taken by Shaw and Agahi [14], who used baseline data from the 1998 Health and Retirement Study [15], which collected data from adults older than 50 years, to form *health behavior profiles*. In total, 12 health behavior profiles were constructed based on combinations of smokers versus nonsmokers, physically active versus inactive, and those who reported no versus moderate versus heavy alcohol consumption. Profiles varied widely in the percentage of participants captured in each profile. The 6 most prevalent

profiles included the following: (1) *physically inactive, nondrinkers, who do not smoke* (6702/19,662, 34.1%); (2) *physically active, nondrinkers, who do not smoke* (4662/19,662, 23.7%); (3) *physically active, moderate drinkers, who do not smoke* (1986/19,662, 10.1%); (4) *physically inactive, moderate drinkers, who do not smoke* (1684/19,662, 8.6%); (5) *physically inactive, nondrinkers, who smoke* (1274/19,662, 6.5%); and (6) *physically active, nondrinkers, who smoke* (820/19,662, 4.2%).

Taken together, population survey data worldwide are converging on the idea that health behaviors cluster differently across the population and that the resulting clusters have distinct sociodemographic and health outcome patterns. However, to our knowledge, this has only been explored to a limited extent in Canadian data. Canadian National Population Survey data have shown that physical inactivity, alcohol use, and smoking co-occur [16,17]. Understanding how behaviors co-occur at a population level may provide novel ways of developing support and guidance for patients and clinicians and preventing noncommunicable chronic diseases in the general population using a multiple behavior approach.

To this end, we aim to leverage cross-sectional baseline data from the Canadian Longitudinal Study on Aging (CLSA) [18]. By 2031, 1 in 4 Canadians will be aged 65 years or older, and the CLSA aims to understand the determinants of health and wellness as people age [19]. For our purposes, we will use CLSA baseline data to address the following objectives: (1) describe how health behaviors cluster, (2) describe how other behaviors of everyday life (family, work, hobby, and community behaviors) cluster, (3) identify sociodemographic factors (sex, age group, marital status, income, country of birth, and social support availability) associated with cluster membership, (4) examine whether life satisfaction and health differ across clusters, (5) examine whether and which clusters of behavior are associated with health care utilization, and (6) examine whether clusters of health behaviors are associated with nonhealth behaviors.

Methods

Overview

The CLSA is a national longitudinal study designed to assess the biological, physical, societal, and psychosocial factors involved in healthy aging [18]. CLSA baseline data collection was conducted between 2010 and 2015 and involved 2 approaches: (1) a *tracking cohort* (n=21,241) that responded to questions administered by a 60-minute computer-assisted telephone interview and (2) a *comprehensive cohort* (n=30,097) that involved a 90-minute in-person interview and a data collection site visit. A 30-minute *maintaining contact questionnaire* was also administered by telephone to both cohorts 18 months after the initial contact to collect supplementary data from the same cohort (all of which form the baseline data collection used in our planned analysis).

Participants

Participants were recruited through random-digit dialing, provincial health registries, and the Canadian Community Health Survey on Healthy Aging [19,20]. Exclusion criteria for the

CLSA included residents living in 3 territories and First Nations reserves, full-time members of the Canadian Armed Forces, people living with cognitive impairments, and individuals living in institutions (including 24-hour nursing homes) [21]. This study included 51,338 French- and English-speaking Canadians (26,155/51,338, 50.95% female) aged between 45 and 85 years at the time of enrollment. The average participant age was 62.98 years (SD 10.43), with 26.15% (13,427/51,338) aged between 45 and 54 years, 31.98% (16,420/51,338) aged between 55 and 64 years, 23.37% (11,996/51,338) aged between 65 and 74 years, and 18.5% (9495/51,338) aged between 75 and 85 years. A full description of demographic characteristics of the sample as well as summary data across all measured variables is available in the CLSA baseline data report [19].

Variable Selection

Selection Approach

Variable selection can be performed by *objective* or *subjective* approaches. Objective methods rely on data-driven techniques (eg, forward or backward selection) and techniques such as factor analysis and principal component analysis for dimension reduction to arrive at a parsimonious set of features for inclusion in a model [22]. In contrast, subjective approaches are generally driven by expert opinions and/or theory-driven research questions. Due to the manageable number of variables in the CLSA related to our research question, objective or data-driven approaches to feature selection were not required. Rather, based on our research objectives and the data collected by CLSA, we identified by group consensus an initial set of variables assessing health behaviors, nonhealth behaviors, sociodemographic indicators, general health and well-being, and health care service utilization. Our decisions were also shaped by issues of survey design (eg, skip questions), knowledge of basic summary statistics for baseline CLSA data [19], and our own supplementary summary statistics on the baseline data. [Multimedia Appendix 1](#) provides a description of variables in each category (eg, health behaviors and nonhealth behaviors) to be used in the analyses, along with example items.

Health Behaviors

Physical activity and *sedentary behavior* were measured using the Physical Activity Scale for the Elderly [23], which assesses the frequency of sedentary behavior, walking, light physical activity, moderate physical activity, strenuous physical activity, and exercise. The items asked participants to report on their activity levels over the previous 7 days on a scale of 1 (never) to 4 (often, 5-7 days). A recent report published by Statistics Canada focusing on the relationship between physical activity and lung functioning [24] merged light and moderate physical activity together and merged strenuous physical activity and exercise together based on issues with question prompts and conceptual overlap between question items. To facilitate dimension reduction, we opted for a similar approach in which the Physical Activity Scale for the Elderly subscale items were merged to represent sitting, walking, light or moderate physical activity, and strenuous physical activity or exercise.

Fruit and vegetable consumption was assessed using one item from the Seniors in the Community Risk Evaluation for Eating

and Nutrition questionnaire [25]. The item asks respondents how many servings of fruits and vegetables they eat in a day. The original scale was reverse coded such that higher scores indicate more fruit and vegetable consumption.

Smoking behavior was measured using a skip-question framework in the CLSA. Participants who answered “no” to the question “have you smoked at least 100 cigarettes in your life” and responded “yes” to the question “have you ever smoked a whole cigarette” were subsequently asked whether they smoke occasionally, daily, or not at all in the past 30 days. Next, only participants who reported smoking occasionally or daily were asked follow-up questions pertaining to the frequency and types of tobacco products used. A descriptive analysis of the last 2 items showed that only a small minority of respondents (4845/51,338, 9.44%) engaged in occasional or daily smoking, with the majority (30,558/51,338, 59.52%) not engaging in smoking behavior in the past 30 days. Although the frequency of smoking would be a more informative metric, any cluster analysis with smoking frequency would reduce the sample size to 4845 and would only represent people who have smoked within the past 30 days. In addition, we will assign a value of 0 to each respondent who responded “no” to the question “have you ever smoked a whole cigarette,” as these individuals also did not smoke in the past 30 days. A similar approach has been applied to skip structure data when missing data represent the absence of a behavior or psychological feature [26]. Ultimately, this creates 4 levels distinguishing between people who have never smoked and people who have smoked occasionally, daily, or not at all during a 30-day window.

Alcohol use was assessed with a single item asking participants how often they drank alcohol in the past 12 months on a scale from 1 (almost every day) to 7 (less than once a week). Responses will be reverse coded so that higher values indicate greater alcohol consumption.

Finally, *sleep* was also measured with a single item. Participants were asked how many hours of sleep they get, on average, during the past month and could respond with any value between 0 and 24.

Nonhealth Behaviors

Two items representing participation in *hobbies* were selected from the general health module. One item asked how much time participants spent playing board games, crossword puzzles, cards, sudoku, or jigsaw puzzles. The second question asked how much time participants spent singing in a choir or playing a musical instrument. Both items were originally scored on a scale from 1 (every day) to 5 (once a year or less); we will reverse code these variables so that higher values represent higher frequencies.

A social participation module was included in the CLSA baseline data collection that asked respondents about their tendencies to engage in various *community activities*, including church or religious activities; attending concerts, watching plays, or visiting museums; service club or fraternal organization activities; community or professional association activities; volunteer or charity work; participation in activities with family or friends outside the household; participation in sports or

physical activities with others; participation in educational or cultural activities; and participation in other recreational activities. The CLSA contains 2 derived variables in the social participation module, one binary variable reflecting whether participants engaged in any social activities and another reflecting the frequency of participation in any activity over the past 12 months (0=no activities, 1=yearly, 2=monthly, 3=weekly, and 4=daily). We will use only the latter in our analysis.

The *caregiving* module contained questions pertaining to assisting others, how many others were assisted, the type of assistance, the people who the respondents help most often, and the personal and professional impacts of providing care to others. To reduce the number of items included in the analysis, we will use a derived variable in the CLSA data set, which indicates whether the respondent provided assistance to any person in the past 12 months (excluding aid rendered as part of a paid job or volunteer work). According to the descriptive analysis from CLSA baseline data [19], 44.4% (22,805/51,338) of participants reported aiding another person due to health conditions or other limitations. This variable will be recoded as 0 (did not provide assistance) or 1 (did provide assistance).

Finally, the CLSA participants were asked whether they used any *social networking sites* (eg, Facebook, LinkedIn, MySpace, MSNGroups, and Twitter). Our preliminary descriptive analysis showed that 44.7% (22,959/51,338) of the participants reported using social networking sites, whereas 38.0% (19,518/51,338) of the participants were not social networking site users, and 17.3% (8861/51,338) of responses were either missing or nonresponses. Although more detailed follow-up questions were subsequently posed to respondents, including these items would substantially reduce the sample size available for analysis. Given the skip-question structure of the CLSA social networking site module, we will include a binary variable representing the use (1) or nonuse (0) of social networking sites.

Sociodemographic Indicators

We will use several sociodemographic indicators in our analysis. These include age, as grouped in the CLSA data set (45-54, 55-64, 65-74, or 75-85 years); sex (male or female); country of birth (recoded as 0=Canada or 1=other); marital status (single, married or common-law, widowed, divorced, or separated); household income (five income levels); retirement status (completely retired, partly retired, or not retired); and working status (yes or no to the question “are you currently working at a job or business.” In addition, participants responded to 19 questions from the Medical Outcomes Study (MOS) Social Support Survey [27]. The MOS is scored according to 5 subscales: tangible social support, affection, positive social interaction, emotional support, and informational support. The MOS overall support index is also scored in the CLSA baseline data set. To reduce the number of constructs in our analyses, we will use the overall support index, scored from 0 (low support) to 100 (high support).

General Health and Life Satisfaction

Three single-item measures were selected from the CLSA’s general health module. These include an indicator of general

health (“in general, would you say your health is excellent, very good, good, fair, or poor?”), mental health (“in general, would you say your mental health is excellent, very good, good, fair, or poor?”), and perceptions of healthy aging (“in terms of your own healthy aging, would you say it is excellent, very good, good, fair, or poor?”). Items were originally scored from 1 (excellent) to 5 (poor) but will be reverse coded. In addition, a composite score from the Satisfaction With Life Questionnaire [28] will be used in the analysis. The Satisfaction With Life Questionnaire is scored according to Deiner [29] on a scale ranging from 1 (extremely dissatisfied) to 7 (extremely satisfied). Finally, BMI will be used as an indicator of physical health.

Health Care Utilization

Three single-item questions were selected from the CLSA’s health care utilization module. These items represent emergency department visits (“Have you been seen in an Emergency Department during the past 12 months?”), hospital admittance (“Were you a patient in a hospital overnight during the past 12 months?”), and nursing home use (“Were you a patient in a nursing home or convalescent home during the past 12 months?”). All responses will be coded as yes (1) or no (0).

Cluster Analysis Overview

We will use cluster analysis to cluster *individuals* based on their behaviors and network community detection algorithms to cluster *variables* based on the strength of conditionally independent pairwise relationships between variables.

Classifying data through the assignment of classes to objects in a data set is a common application of machine learning (ie, “set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data” [30]). Classification algorithms fall into 3 categories: supervised learning, semisupervised learning, and unsupervised learning. In supervised learning, the relationships between the input and target variables are known. An algorithm is *supervised* in that it can be trained on a data set that contains correct classifications. Data sets containing these correct classifications are referred to as *labeled data*, in contrast to *unlabeled data*, in which the correct classifications are not known. In semisupervised learning, a combination of labeled and unlabeled data is used to model the data, whereas in unsupervised learning, the model works on its own to discover patterns in unlabeled data [31].

Cluster analysis is a type of unsupervised machine learning that comprises a set of methods for identifying distinct characteristics in heterogeneous samples and clustering them into homogenous groups [32]. When the target number of clusters (*k*) is known, partitioning-based clustering arguments such as *k*-means, *k*-medoids, or model-based clustering approaches are appropriate. However, when *k* is unknown, as is the case with clusters of Canadians based on health and nonhealth behaviors, hierarchical clustering is a suitable method [32].

The hierarchical structure of the data can be obtained by clustering individual data points in a bottom-up approach (ie, agglomerative clustering) or by partitioning a single cluster into smaller clusters until each cluster is a single observation through

a top-down approach (ie, divisive clustering). Divisive methods are rarely used in practice owing to their heavy computational requirements [33]. In agglomerative hierarchical clustering, each individual data point is initially treated as its own cluster. The methodological process is as follows [34]: each data point is assigned to its own cluster, the distance (ie, the similarity or dissimilarity between each cluster) between each cluster is calculated, the pair of clusters with the shortest distance between them is selected and merged into a single cluster, the distances between the new cluster and all other clusters are recalculated, and these steps are repeated until only one cluster remains. However, a single cluster ($k=1$) is unlikely to be informative; researchers can identify the number of clusters that best describe the data (eg, $k=5$) through subjective criteria and/or with the aid of statistical tests that have been developed for this purpose.

Several measures of *distance* are widely used in practice, although the Gower distance [35] is appropriate for mixed data (binary, ordinal, and continuous). In addition to selecting a measure of distance, hierarchical agglomerative clustering also requires a linkage method to be specified to define how the distance between clusters is calculated. Different methods exist for specifying the anchor points used to calculate the distance between clusters (ie, how the distances between clusters are *linked*). For example, *single or minimum linkage* calculates the minimum distance between data points in each cluster, whereas *centroid linkage* calculates the distance between the center of each cluster [31]. No consensus exists as to which linkage method is superior, although it is recognized that final clustering solutions may differ based on the linkage method selected [33].

Network Analysis and Community Detection Overview

Another method for clustering data is through network community detection. In contrast to cluster analysis, which clusters *people* based on similarity or dissimilarity of selected features (eg, health behaviors), network community detection identifies groups of *highly connected variables* based on the strength of the connections. Before conducting the community detection analysis, a network must first be estimated. Networks consist of nodes and edges. In psychological networks, nodes represent psychological attributes (eg, emotions, behaviors, and symptoms) and edges represent relationships between nodes (eg, partial correlations) and can indicate the presence of a positive or negative relationship and the direction of the effect. Psychological networks are commonly estimated using a pairwise Markov random field (PMRF) [36,37]. In a PMRF, edges represent the conditional independence between a pair of connected variables. For cross-sectional data, the underlying models for the PMRF vary depending on the type of data. For example, a Gaussian graphical model is appropriate for multivariate normal continuous data, whereas Ising models are used for binary data, and mixed graphical models (MGMs) [38] have been developed for mixed data. Due to the large number of parameters often estimated in a PMRF, researchers often compute a regularized network by applying, for example, the least absolute shrinkage and selection operator [39] to produce a sparse or conservative network that reduces weak connections between nodes to 0, resulting in a more interpretable network. However, recent work has questioned whether regularization

is required for low-dimensional data with large sample sizes [40].

Once a network has been estimated, an additional analysis can be conducted. For example, node centrality, a family of measures meant to indicate the importance of nodes within a network [37], can be computed. Other analytical options include calculating the explained variance and performing network comparison tests. In the cross-sectional data, a network is estimated by regressing all nodes onto each other. This enables the explained variance for each node to be calculated and visualized within the network itself [38]. Network comparison tests are permutation-based hypothesis tests that allow comparisons between 2 networks (eg, health behaviors for those aged 45-54 years and 55-64 years) based on their global structure, global strength (a measure of centrality), and differences between individual edges [41].

Finally, community detection algorithms can be applied to the network. Commonly used community detection methods in psychology include the spinglass algorithm [42], the walktrap algorithm [43], leading eigenvector [44], exploratory graph analysis [45], and the clique percolation algorithm [46,47]. When compared with alternative methods, the clique percolation algorithm addresses a central challenge in identifying communities within a network, namely, the ability to assign a node to multiple communities. Overlapping communities enable the identification of *bridge nodes* (nodes that connect 2 otherwise distinct clusters) and are therefore important for hypothesis generation [48]. Cliques are fully connected networks. The number of nodes (k) that must be connected can vary, with the smallest clique being $k=3$ (a closed triangle). When more than one set of cliques are adjacent in a network, they are said to form a community. In psychological networks, where edges are weighted (eg, representing correlations), only edges that surpass a certain threshold (I) are considered when identifying cliques [46,47,49].

Proposed Analysis

Survey Weights

The CLSA baseline data set contains sample weights (to ensure representativeness of the sample), inflation weights (to improve the precision of estimates), and analytic weights (to estimate the relationships among variables) [50]. The CLSA recommends the use of inflation weights for the estimation of descriptive parameters and analytic weights for exploring the relationships among variables at a national or provincial level [50]. However, the statistical packages we have selected in our planned cluster analysis method do not have the option to include survey weights and will, therefore, not be applied. Given the relative novelty of network analysis methods in psychology, the appropriate role of survey weights in network methodology is unknown. To our knowledge, the only study that addresses this issue is the study by Lin et al [51], who opted not to include survey weights due to “a lack of established methods to do so for network models.” Therefore, weights will not be applied for the network analysis.

Preprocessing and Descriptive Statistics

All analyses will be conducted in R (R Core Team) [52]. Selected variables from the tracking (TRM) and comprehensive (COM) data sets will be merged and coalesced (eg, ALC_FREQ_TRM and ALC_FREQ_COM coalesced into ALC_FREQ) using the tidyverse package [53]. Raw data will be visualized to inspect the univariate outliers. Means and SDs for all variables will be computed for continuous data, and frequencies and percentages will be computed for categorical data, overall and by sex and age group. Missing data will be handled with listwise deletion.

Cluster Analysis

Step 1

Variables will be reverse coded, when necessary, such that higher scores indicate greater frequency. Subsequently, all continuous and ordinal variables will be mean centered using the scale function in base R [52].

Step 2

We will run cluster analyses on the overall CLSA sample (all age groups) using 5 linkage methods (eg, complete linkage, single linkage, average linkage, centroid linkage, and Ward method). The research team will decide which method produces the most interpretable clustering solution; the selected linkage method will then be applied to all subset analyses grouped by age.

Step 3

Cluster analysis will be performed on health behavior variables for 5 groups (overall sample, 4 age groups) using the base R *hclust* function supported by the package *fastcluster* [54] to optimize performance. Gower distance will be computed using the *daisy* function in the *cluster* package [55].

Step 4

Once agglomerative cluster analysis has been performed, the next step is to determine the most interpretable number of clusters to represent the data. The cluster analysis literature has produced several competing methods to accomplish this task. To navigate these analytical options, we will use the NbClust package [56]. This approach allows for a consensus approach to determine the number of clusters, which best fits the data by computing 30 different indices and reporting the level of agreement between them. The output of this analysis is a list showing the agreement between statistical tests for various clustering solutions (eg, “10 indices propose 2 as the best number of clusters” and “7 indices propose 3 as the best number of clusters”). To allow for the synthesis of data-driven and expert consensus approaches, the research team will select the most interpretable clustering solution via a majority vote from the top 3 solutions identified via NbClust. In instances of an equal number of indices recommending the same cluster solution and/or disagreement within the research team, we will follow the advice of the NbClust package authors who recommend considering indices, which have performed the best in a seminal simulation study [57].

Step 5

When the optimal clustering solution has been identified, the clusters can be characterized by the number of individuals assigned to each cluster and by mean scores on each health-related variable via one-way analysis of variance tests to determine whether the mean levels of the health behaviors vary by cluster.

Step 6

Next, multinomial logistic regression will be used to determine whether clusters are associated with sociodemographic variables, indicators of physical and mental health (eg, life satisfaction), nonhealth behaviors, and health care utilization. Multinomial logistic regression is similar to logistic regression but is appropriate when the dependent variable has more than 2 levels (which will likely be the case for the number of identified clusters). The dependent variable will be the clusters we identify in the cluster analysis. To perform the analysis, we will use the *multinom* function from the *nnet* package [58].

Network Analysis

Step 1

Networks of health behaviors will be estimated using an MGM using the MGM package [38]. To the best of our knowledge, an MGM is the only available method for estimating a psychological network with mixed data. In the mixed model, the edges between nodes represent pairwise interactions and can be interpreted as the strength of conditional dependence [38]. A total of 5 networks will be estimated (all age groups, 45-54, 55-64, 65-74, and 75-85 years). In the mixed model, we will specify *lambdaSel=EBIC* to use the Extended Bayesian Information Criteria [59] for selecting the tuning parameter controlling regularization. The hyperparameter γ in the Extended Bayesian Information Criteria will be set to the default *lambdaGam=0.25*, and only pairwise interactions will be included in the model through setting *k=2*.

Step 2

For each network from step 2, networks will be visualized with *qgraph* [60] using the *averageLayout* function to compute a joint layout across networks.

Step 3

The NetworkComparisonTest package [41] will be used to conduct permutation-based hypothesis tests to determine whether networks differ from one another based on sociodemographic variables. We will specify *it=1000* to run 1000 iterations or permutations and will plot the results from the network structure invariance test.

Step 4

Next, community detection analysis will be performed on each network estimated in step 1 using the CliquePercolation package [49]. Although multiple options are available, we will detect overlapping communities by optimizing *k* and *l*, where *k* represents the minimum clique size and *l* represents the strength of the relationships between nodes required for classification as a community. Following the study by Blanken et al [61], we will determine the optimal threshold *l* for the fixed values of

$k=3-6$ and will set the clique percolation algorithm to search through ranges of l from 0.01 to the largest edge weight in each network through increments of 0.001. We will choose the value of k that allows for the broadest community structure, and l will be selected based on the largest chi-square value for intensities with a ratio threshold over 2. The resulting networks will be visualized with colored nodes indicating community structures.

Steps 5 to 8

The processes described in steps 1 to 4 will be repeated for nonhealth behaviors and health behaviors in the same model. A summary of the proposed analysis and its connections with research questions is presented in [Table 1](#). Further graphical representations of the analytical steps are provided in [Multimedia Appendices 2 and 3](#).

Table 1. Overview of research questions and planned analysis.

Research questions	Planned analyses				
	Cluster analysis (analysis 1)	Multinomial logistic regression (analysis 1)	Network analysis (analysis 2)	Network community detection (analysis 2)	Network comparison tests (analysis 2)
Describe how health behaviors cluster in Canadians aged 45-85 years	✓	—	✓	✓	—
Describe how nonhealth behaviors cluster in Canadians aged 45-85 years	—	—	✓	✓	—
Identify sociodemographic factors associated with cluster membership; identify sociodemographic factors associated with network structures	—	✓	—	—	✓
Examine whether cluster are associated with health indicators	—	✓	—	—	—
Examine whether clusters are associated with health care utilization	—	✓	—	—	—
Examine whether clusters are associated with nonhealth behaviors	—	✓	—	—	—

Data Availability

Data are available from the CLSA for researchers who meet the criteria for access to deidentified CLSA data. We will make R scripts public so that any researcher who independently gains access to data can reproduce the results. In addition, we will publish R Markdown documents so that the code and outputs of analysis can be viewed publicly on the Open Science Framework. Deviations from the protocol plan will be noted in the final report. Any additional analysis, not specified here, will be labeled as such and published in web-based supplemental materials.

Results

Baseline data for the CLSA were collected from 51,338 participants aged between 45 and 85 years. Data were collected between 2010 and 2015. Secondary data analysis for this project was approved by the Ottawa Health Science Network Research Ethics Board (protocol ID #20190506-01H).

Discussion

The scope, size, and rigor of the CLSA data set will provide us with an unprecedented opportunity to investigate how behaviors cluster. The findings will allow us to assess alignment with findings from other countries, while extending findings in novel ways by investigating how social, family, and work behaviors cluster alongside health behaviors typically investigated. Perhaps most importantly, this study will help to inform the development of novel health behavior change interventions tailored to

subpopulations of adults defined by the behaviors that cluster within them. If behaviors co-occur, intervening on one may impact—or be impacted by—the others [62]. Interventions that target only one behavior may thus be undermined by the impact of a conflicting co-occurring behavior or miss an opportunity to leverage the enabling nature of a positively co-occurring behavior. In addition, targeting multiple co-occurring behaviors simultaneously has potential practical benefits, such as reduced expenses for intervention providers and reduced time commitments for those receiving the intervention. Understanding which behaviors co-occur and for whom is an important first step toward developing health behavior change interventions that address people's actual challenges. In the context of a behavioral intervention development and testing framework [63], the proposed analysis will inform the foundational *basic behavioral science*—regarding patterns of co-occurring risky health behaviors and their associated outcomes—which precedes early phase behavioral trials.

This study will ultimately help to develop and evaluate more targeted interventions to support healthy aging and well-being in adulthood by identifying how clusters of co-occurring health behaviors are associated with sociodemographic factors, general health, and health care utilization. However, the proposed analyses are not without limitations. For example, many of the items selected for planned analysis are self-report, which have inherent strengths and weaknesses [64]. In addition, there are several points in the proposed analysis that require subjective decision making on behalf of the research team (eg, selecting variables for inclusion in models, data preprocessing decisions, selecting missing data procedures, and interpreting and selecting

clustering solutions). We have sought to document these here to provide a clear explanation of decision processes, while acknowledging where and how researcher degrees of freedom are used to ensure transparency. Finally, we recognize that the analyses proposed are, to some extent, limited by their cross-sectional nature. Nevertheless, given the longitudinal nature of CLSA and planned future data releases, the proposed analyses have a number of novel implications and set the stage for planned future longitudinal analyses that extend the research questions to investigate changes in behavior clusters as a

function of time both between and within individuals. Thus, this proposed study will establish the foundation for future analyses. More broadly, the methodological approaches proposed for this analysis lend themselves to replication in other similar data sets internationally, and we hope that the sharing of R code will help to enable this.

Finally, we emphasize that our choices of analytic methods are hypothesis generating, not hypothesis testing. However, if replicated, such findings may ultimately find their way into clinical practice guidance and public health guidance.

Acknowledgments

This research was made possible using the data or biospecimens collected by the Canadian Longitudinal Study on Aging (CLSA). Funding for the Canadian Longitudinal Study on Aging (CLSA) is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) under grant reference: LSA 94473 and the Canada Foundation for Innovation. This research has been conducted using the CLSA data set [Baseline Tracking Dataset version 3.4 and Comprehensive Dataset version 4.0.], under Application Number [19CA012]. The CLSA is led by Drs. Parminder Raina, Christina Wolfson and Susan Kirkland. Peer review documents for the CIHR grant application are available in [Multimedia Appendix 4](#). ZVA is supported by a CIHR Doctoral Award: Frederick Banting and Charles Best Canada Graduate Scholarship. SLB is supported by a CIHR-Strategy for Patient-Oriented Research Mentoring Chair (SMC-151518) and an FRQS (Fonds de Recherche du Québec-Santé) Chair (251618). MM is funded by a Marie-Sklodowska-Curie Fellowship (grant agreement 713567) at the ADAPT (Artificial Intelligence-Driven Digital Content Technology) Science Foundation Ireland Research Centre at Trinity College. PB was supported by Université du Québec à Montréal, Institut Universitaire de Santé Mentale de Montréal, and by a salary award from FRQS. The opinions expressed in this manuscript are the authors' own and do not reflect the views of the CLSA.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Variables to be included in the analysis.

[\[DOCX File , 19 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Graphical representation of the cluster analysis and multinomial logistic regression analytical process.

[\[DOCX File , 145 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

Graphical representation of the network analysis and community detection process.

[\[DOCX File , 266 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Peer-review report by the Canadian Institutes of Health Research.

[\[PDF File \(Adobe PDF File\), 100 KB-Multimedia Appendix 4\]](#)

References

1. How healthy are Canadians? A trend analysis of the the health of Canadians from a healthy living and chronic disease perspective. Public Health Agency of Canada. 2016. URL: <https://www.canada.ca/content/dam/phac-aspc/documents/services/publications/healthy-living/how-healthy-canadians/pub1-eng.pdf> [accessed 2021-05-15]
2. CCDI Steering Committee. At-a-glance - the 2017 Canadian chronic disease indicators. Health Promot Chronic Dis Prev Can 2017 Aug;37(8):248-251 [[FREE Full text](#)] [doi: [10.24095/hpcdp.37.8.03](https://doi.org/10.24095/hpcdp.37.8.03)] [Medline: [28800294](https://pubmed.ncbi.nlm.nih.gov/28800294/)]
3. Fisher EB, Fitzgibbon ML, Glasgow RE, Haire-Joshu D, Hayman LL, Kaplan RM, et al. Behavior matters. Am J Prev Med 2011 May;40(5):15-30 [[FREE Full text](#)] [doi: [10.1016/j.amepre.2010.12.031](https://doi.org/10.1016/j.amepre.2010.12.031)] [Medline: [21496745](https://pubmed.ncbi.nlm.nih.gov/21496745/)]
4. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev 2011 Sep;10(4):430-439 [doi: [10.1016/j.arr.2011.03.003](https://doi.org/10.1016/j.arr.2011.03.003)] [Medline: [21402176](https://pubmed.ncbi.nlm.nih.gov/21402176/)]

5. Nunes BP, Flores TR, Mielke GI, Thumé E, Facchini LA. Multimorbidity and mortality in older adults: a systematic review and meta-analysis. *Arch Gerontol Geriatr* 2016 Nov;67:130-138 [doi: [10.1016/j.archger.2016.07.008](https://doi.org/10.1016/j.archger.2016.07.008)] [Medline: [27500661](https://pubmed.ncbi.nlm.nih.gov/27500661/)]
6. Prados-Torres A, Calderón-Larrañaga A, Hanco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. *J Clin Epidemiol* 2014 Mar;67(3):254-266 [doi: [10.1016/j.jclinepi.2013.09.021](https://doi.org/10.1016/j.jclinepi.2013.09.021)] [Medline: [24472295](https://pubmed.ncbi.nlm.nih.gov/24472295/)]
7. Sniehotta FF, Presseau J, Allan J, Araújo-Soares V. "You Can't Always Get What You Want": A novel research paradigm to explore the relationship between multiple intentions and behaviours. *Appl Psychol Health Well Being* 2016 Jul 27;8(2):258-275 [FREE Full text] [doi: [10.1111/aphw.12071](https://doi.org/10.1111/aphw.12071)] [Medline: [27230935](https://pubmed.ncbi.nlm.nih.gov/27230935/)]
8. Tremblay MS, Warburton DE, Janssen I, Paterson DH, Latimer AE, Rhodes RE, et al. New Canadian physical activity guidelines. *Appl Physiol Nutr Metab* 2011 Feb;36(1):36-47 [FREE Full text] [doi: [10.1139/H11-009](https://doi.org/10.1139/H11-009)] [Medline: [21326376](https://pubmed.ncbi.nlm.nih.gov/21326376/)]
9. Butt P, Gliksman L, Beirness D, Paradis C, Stockwell T. Alcohol and health in Canada: a summary of evidence and guidelines for low-risk drinking. Canadian Centre on Substance Abuse. 2011. URL: <https://www.uvic.ca/research/centres/cisur/assets/docs/report-alcohol-and-health-in-canada.pdf> [accessed 2021-05-22]
10. Conry MC, Morgan K, Curry P, McGee H, Harrington J, Ward M, et al. The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life. *BMC Public Health* 2011 Sep 06;11(1):692 [FREE Full text] [doi: [10.1186/1471-2458-11-692](https://doi.org/10.1186/1471-2458-11-692)] [Medline: [21896196](https://pubmed.ncbi.nlm.nih.gov/21896196/)]
11. Buck D, Frosini F. Clustering of unhealthy behaviours over time : implications for policy and practice. London: The Kings Fund. 2012. URL: https://www.makeeverycontactcount.org/media/1045/018_-clustering-of-unhealthy-behaviours-over-time-aug-2012.pdf [accessed 2021-05-15]
12. Brown H, Presseau J. Work me not into temptation: exploring the relationship between work and healthy eating in dieters using data from the HILDA survey. *Aust Econ Rev* 2018 Jul 30;51(3):368-381 [doi: [10.1111/1467-8462.12269](https://doi.org/10.1111/1467-8462.12269)]
13. Schneider S, Huy C, Schuessler M, Diehl K, Schwarz S. Optimising lifestyle interventions: identification of health behaviour patterns by cluster analysis in a German 50+ survey. *Eur J Public Health* 2009 Jun 22;19(3):271-277 [doi: [10.1093/eurpub/ckn144](https://doi.org/10.1093/eurpub/ckn144)] [Medline: [19164433](https://pubmed.ncbi.nlm.nih.gov/19164433/)]
14. Shaw BA, Agahi N. A prospective cohort study of health behavior profiles after age 50 and mortality risk. *BMC Public Health* 2012 Sep 18;12(1):803 [FREE Full text] [doi: [10.1186/1471-2458-12-803](https://doi.org/10.1186/1471-2458-12-803)] [Medline: [22989155](https://pubmed.ncbi.nlm.nih.gov/22989155/)]
15. Heeringa S, Connor J. Technical description of the health and retirement study sample design. Sampling Section, Institute for Social Research, University of Michigan, Ann Arbor, MI. 1995. URL: <https://hrspubs.sites.uofmhosting.net/sites/default/files/biblio/HRSSAMP.pdf> [accessed 2021-05-15]
16. deRuiter WK, Cairney J, Leatherdale S, Faulkner G. The period prevalence of risk behavior co-occurrence among Canadians. *Prev Med* 2016 Apr;85:11-16 [doi: [10.1016/j.ypmed.2015.11.026](https://doi.org/10.1016/j.ypmed.2015.11.026)] [Medline: [26658026](https://pubmed.ncbi.nlm.nih.gov/26658026/)]
17. deRuiter WK, Cairney J, Leatherdale ST, Faulkner GE. A longitudinal examination of the interrelationship of multiple health behaviors. *Am J Prev Med* 2014 Sep;47(3):283-289 [doi: [10.1016/j.amepre.2014.04.019](https://doi.org/10.1016/j.amepre.2014.04.019)] [Medline: [25145617](https://pubmed.ncbi.nlm.nih.gov/25145617/)]
18. Raina PS, Wolfson C, Kirkland SA, Griffith LE, Oremus M, Patterson C, et al. The Canadian longitudinal study on aging (CLSA). *Can J Aging* 2009 Sep;28(3):221-229 [doi: [10.1017/S0714980809990055](https://doi.org/10.1017/S0714980809990055)] [Medline: [19860977](https://pubmed.ncbi.nlm.nih.gov/19860977/)]
19. Raina P, Wolfson C, Kirkland S, Griffith L. The Canadian Longitudinal Study on Aging (CLSA) Report on Health and Aging in Canada : findings from baseline data collection 2010-2015. Canadian Longitudinal Study on Aging. 2018. URL: https://ifa.ngo/wp-content/uploads/2018/12/clsa_report_en_final_web.pdf [accessed 2021-05-15]
20. Wolfson C, Raina PS, Kirkland SA, Pelletier A, Uniat J, Furlini L, et al. The Canadian community health survey as a potential recruitment vehicle for the Canadian longitudinal study on aging. *Can J Aging* 2009 Sep 01;28(3):243-249 [doi: [10.1017/s0714980809990031](https://doi.org/10.1017/s0714980809990031)]
21. Raina P, Wolfson C, Kirkland S. Canadian Longitudinal Study on Aging (CLSA) : Protocol. Canadian Longitudinal Study on Aging. URL: <https://www.clsa-elcv.ca/doc/511> [accessed 2021-05-15]
22. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. *Pulm Ther* 2016 Jun 22;2(1):19-41 [FREE Full text] [doi: [10.1007/s41030-016-0017-z](https://doi.org/10.1007/s41030-016-0017-z)] [Medline: [27512723](https://pubmed.ncbi.nlm.nih.gov/27512723/)]
23. Washburn RA, Smith KW, Jette AM, Janney CA. The physical activity scale for the elderly (PASE): development and evaluation. *J Clin Epidemiol* 1993 Feb;46(2):153-162 [doi: [10.1016/0895-4356\(93\)90053-4](https://doi.org/10.1016/0895-4356(93)90053-4)]
24. Dogra S, Good J, Gardiner P, Copeland J, Stickland M, Rudoler D, et al. Effects of replacing sitting time with physical activity on lung function: an analysis of the Canadian Longitudinal Study on Aging. *Health Rep* 2019 Mar 20;30(3):12-23 [FREE Full text] [doi: [10.25318/82-003-x201900300002-eng](https://doi.org/10.25318/82-003-x201900300002-eng)] [Medline: [30892662](https://pubmed.ncbi.nlm.nih.gov/30892662/)]
25. Keller HH, Goy R, Kane S. Validity and reliability of SCREEN II (Seniors in the community: risk evaluation for eating and nutrition, Version II). *Eur J Clin Nutr* 2005 Oct 13;59(10):1149-1157 [doi: [10.1038/sj.ejcn.1602225](https://doi.org/10.1038/sj.ejcn.1602225)] [Medline: [16015256](https://pubmed.ncbi.nlm.nih.gov/16015256/)]
26. Borsboom D, Cramer AO. Network analysis: an integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol* 2013 Mar 28;9(1):91-121 [doi: [10.1146/annurev-clinpsy-050212-185608](https://doi.org/10.1146/annurev-clinpsy-050212-185608)] [Medline: [23537483](https://pubmed.ncbi.nlm.nih.gov/23537483/)]
27. Sherbourne CD, Stewart AL. The MOS social support survey. *Soc Sci Med* 1991 Jan;32(6):705-714 [doi: [10.1016/0277-9536\(91\)90150-B](https://doi.org/10.1016/0277-9536(91)90150-B)]
28. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. *J Pers Assess* 1985 Feb;49(1):71-75 [doi: [10.1207/s15327752jpa4901_13](https://doi.org/10.1207/s15327752jpa4901_13)] [Medline: [16367493](https://pubmed.ncbi.nlm.nih.gov/16367493/)]

29. Diener E. Understanding scores on the satisfaction with life scale. URL: <http://labs.psychology.illinois.edu/~ediener/Documents/Understanding%20SWLS%20Scores.pdf> [accessed 2006-02-13]
30. Murphy KP. Machine Learning: A Probabilistic Perspective. Massachusetts, United States: The MIT Press; 2012:1-1104
31. Fung G. A comprehensive overview of basic clustering algorithms. 2001. URL: https://sites.cs.ucsb.edu/~veronika/MAE/clustering_overview_2001.pdf [accessed 2021-05-15]
32. Rapkin B, Luke D. Cluster analysis in community research: epidemiology and practice. *Am J Community Psychol* 1993;21(2):247-277 [doi: [10.1007/BF00941623](https://doi.org/10.1007/BF00941623)]
33. Xu R, Wunsch II D. Survey of clustering algorithms. *IEEE Trans Neural Netw* 2005 May;16(3):645-678 [doi: [10.1109/tnn.2005.845141](https://doi.org/10.1109/tnn.2005.845141)]
34. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The application of unsupervised clustering methods to Alzheimer's disease. *Front Comput Neurosci* 2019 May 24;13:31 [FREE Full text] [doi: [10.3389/fncom.2019.00031](https://doi.org/10.3389/fncom.2019.00031)] [Medline: [31178711](https://pubmed.ncbi.nlm.nih.gov/31178711/)]
35. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics* 1971 Dec;27(4):857-871 [doi: [10.2307/2528823](https://doi.org/10.2307/2528823)]
36. van Borkulo CD, Borsboom D, Epskamp S, Blanken TF, Boschloo L, Schoevers RA, et al. A new method for constructing networks from binary data. *Sci Rep* 2014 Aug 01;4(1):5918 [FREE Full text] [doi: [10.1038/srep05918](https://doi.org/10.1038/srep05918)] [Medline: [25082149](https://pubmed.ncbi.nlm.nih.gov/25082149/)]
37. Epskamp S, Borsboom D, Fried EI. Estimating psychological networks and their accuracy: a tutorial paper. *Behav Res Methods* 2018 Feb 24;50(1):195-212 [FREE Full text] [doi: [10.3758/s13428-017-0862-1](https://doi.org/10.3758/s13428-017-0862-1)] [Medline: [28342071](https://pubmed.ncbi.nlm.nih.gov/28342071/)]
38. Haslbeck JM, Waldorp LJ. mgm: Estimating time-varying mixed graphical models in high-dimensional data. *J Stat Soft* 2020 Dec;93(8):1-8 [FREE Full text] [doi: [10.18637/jss.v093.i08](https://doi.org/10.18637/jss.v093.i08)]
39. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc: Series B (Methodological)* 2018 Dec 05;58(1):267-288 [doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)]
40. Williams DR, Rast P. Back to the basics: rethinking partial correlation network methodology. *Br J Math Stat Psychol* 2020 May 17;73(2):187-212 [doi: [10.1111/bmsp.12173](https://doi.org/10.1111/bmsp.12173)] [Medline: [31206621](https://pubmed.ncbi.nlm.nih.gov/31206621/)]
41. van Borkulo C, Boschloo L, Borsboom D, Penninx BW, Waldorp LJ, Schoevers RA. Comparing network structures on three aspects: a permutation test. *ResearchGate*. 2017. URL: https://www.researchgate.net/profile/Claudia_Van_Borkulo/publication/314750838_Comparing_network_structures_on_three_aspects_A_permutation_test/links/58c55ef145851538eb8af8a9/Comparing-network-structures-on-three-aspects-A-permutation-test.pdf [accessed 2021-05-22]
42. Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Phys Rev E* 2006 Jul 18;74(1):016110 [doi: [10.1103/physreve.74.016110](https://doi.org/10.1103/physreve.74.016110)]
43. Pons P, Latapy M. Computing communities in large networks using random walks. In: *Computer and Information Sciences - ISICIS 2005*. Berlin, Heidelberg: Springer; 2005:284-293
44. Newman ME. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 2006 Sep 11;74(3):036104 [doi: [10.1103/physreve.74.036104](https://doi.org/10.1103/physreve.74.036104)]
45. Golino HF, Epskamp S. Exploratory graph analysis: a new approach for estimating the number of dimensions in psychological research. *PLoS One* 2017 Jun 8;12(6):e0174035 [FREE Full text] [doi: [10.1371/journal.pone.0174035](https://doi.org/10.1371/journal.pone.0174035)] [Medline: [28594839](https://pubmed.ncbi.nlm.nih.gov/28594839/)]
46. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005 Jun 09;435(7043):814-818 [doi: [10.1038/nature03607](https://doi.org/10.1038/nature03607)] [Medline: [15944704](https://pubmed.ncbi.nlm.nih.gov/15944704/)]
47. Farkas I, Ábel D, Palla G, Vicsek T. Weighted network modules. *New J Phys* 2007 Jun 28;9(6):180 [doi: [10.1088/1367-2630/9/6/180](https://doi.org/10.1088/1367-2630/9/6/180)]
48. Jones PJ, Ma R, McNally RJ. Bridge centrality: a network approach to understanding comorbidity. *Multivariate Behav Res* 2019 Jun 10:1-15 [doi: [10.1080/00273171.2019.1614898](https://doi.org/10.1080/00273171.2019.1614898)] [Medline: [31179765](https://pubmed.ncbi.nlm.nih.gov/31179765/)]
49. Lange J. An introduction to the clique percolation community detection algorithm. 2019. URL: <https://cran.r-project.org/web/packages/CliquePercolation/vignettes/CliquePercolation.html> [accessed 2021-05-15]
50. Sampling and computation of response rates and sample weights for the tracking (Telephone Interview) participants and comprehensive participants. *Canadian Longitudinal Study on Aging (CLSA) Technical Document*. 2011. URL: <https://www.clsa-elcv.ca/doc/1041> [accessed 2021-05-15]
51. Lin S, Fried EI, Eaton NR. The association of life stress with substance use symptoms: a network analysis and replication. *J Abnorm Psychol* 2020 Feb;129(2):204-214 [doi: [10.1037/abn0000485](https://doi.org/10.1037/abn0000485)] [Medline: [31670531](https://pubmed.ncbi.nlm.nih.gov/31670531/)]
52. R: A language and environment for statistical computing. The R Foundation. URL: <http://www.R-project.org/> [accessed 2021-05-15]
53. Wickham H. Easily install and load the 'Tidyverse'. R package version. 2021. URL: <https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf> [accessed 2021-05-15]
54. Müllner D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *J Stat Soft* 2013;53(9):1-8 [doi: [10.18637/JSS.v053.i09](https://doi.org/10.18637/JSS.v053.i09)]
55. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster: cluster analysis basics and extensions. R Package Version. 2021. URL: <https://cran.r-project.org/web/packages/cluster/cluster.pdf> [accessed 2021-05-15]
56. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R package for determining the relevant number of clusters in a data set. *J Stat Soft* 2014;61(6):1-36 [doi: [10.18637/jss.v061.i06](https://doi.org/10.18637/jss.v061.i06)]

57. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 1985 Jun;50(2):159-179 [doi: [10.1007/bf02294245](https://doi.org/10.1007/bf02294245)]
58. Ripley B, Venables W, Ripley M. Package ‘nnet’. R Package Version. 2021. URL: <https://cran.r-project.org/web/packages/nnet/nnet.pdf> [accessed 2021-05-15]
59. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 2008 Sep 01;95(3):759-771 [doi: [10.1093/biomet/asn034](https://doi.org/10.1093/biomet/asn034)]
60. Epskamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network visualizations of relationships in psychometric data. *J Stat Soft* 2012;48(4):1-8 [doi: [10.18637/jss.v048.i04](https://doi.org/10.18637/jss.v048.i04)]
61. Blanken T, Deserno M, Dalege J, Borsboom D, Blanken P, Kerkhof G, et al. The role of stabilizing and communicating symptoms given overlapping communities in psychopathology networks. *Sci Rep* 2018 Apr 11;8(1):5854 [FREE Full text] [doi: [10.1038/s41598-018-24224-2](https://doi.org/10.1038/s41598-018-24224-2)] [Medline: [29643399](https://pubmed.ncbi.nlm.nih.gov/29643399/)]
62. Presseau J, Tait RI, Johnston DW, Francis JJ, Sniehotta FF. Goal conflict and goal facilitation as predictors of daily accelerometer-assessed physical activity. *Health Psychol* 2013 Dec;32(12):1179-1187 [doi: [10.1037/a0029430](https://doi.org/10.1037/a0029430)] [Medline: [22888818](https://pubmed.ncbi.nlm.nih.gov/22888818/)]
63. Bacon SL, Campbell TS, Lavoie KL. Rethinking How to Expand the Evidence Base for Health Behavior Change in Cardiovascular Disease Prevention. *J Am Coll Cardiol* 2020 May 26;75(20):2619-2622 [FREE Full text] [doi: [10.1016/j.jacc.2020.03.055](https://doi.org/10.1016/j.jacc.2020.03.055)] [Medline: [32439011](https://pubmed.ncbi.nlm.nih.gov/32439011/)]
64. Del Boca FK, Noll JA. Truth or consequences: the validity of self-report data in health services research on addictions. *Addiction* 2000 Nov;95 Suppl 3:S347-S360 [doi: [10.1046/j.1360-0443.95.11s3.5.x](https://doi.org/10.1046/j.1360-0443.95.11s3.5.x)] [Medline: [11132362](https://pubmed.ncbi.nlm.nih.gov/11132362/)]

Abbreviations

ADAPT: Artificial Intelligence–Driven Digital Content Technology

CIHR: Canadian Institutes of Health Research

CLSA: Canadian Longitudinal Study on Aging

FRQS: Fonds de Recherche du Québec–Santé

MGM: mixed graphical model

MOS: Medical Outcomes Study

PMRF: pairwise Markov random field

SES: socioeconomic status

Edited by G Eysenbach; peer-reviewed by T Webb, CA Almenara, E Brindal; externally peer-reviewed by the Canadian Institutes of Health Research. See the Multimedia Appendices for the peer-review report; Submitted 09.10.20; accepted 19.04.21; published 11.06.21.

Please cite as:

van Allen Z, Bacon SL, Bernard P, Brown H, Desroches S, Kastner M, Lavoie K, Marques M, McCleary N, Straus S, Taljaard M, Thavorn K, Tomasone JR, Presseau J

Clustering of Unhealthy Behaviors: Protocol for a Multiple Behavior Analysis of Data From the Canadian Longitudinal Study on Aging

JMIR Res Protoc 2021;10(6):e24887

URL: <https://www.researchprotocols.org/2021/6/e24887>

doi: [10.2196/24887](https://doi.org/10.2196/24887)

PMID: [34114962](https://pubmed.ncbi.nlm.nih.gov/34114962/)

©Zack van Allen, Simon L Bacon, Paquito Bernard, Heather Brown, Sophie Desroches, Monika Kastner, Kim Lavoie, Marta Marques, Nicola McCleary, Sharon Straus, Monica Taljaard, Kednapa Thavorn, Jennifer R Tomasone, Justin Presseau. Originally published in *JMIR Research Protocols* (<https://www.researchprotocols.org>), 11.06.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR Research Protocols*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.researchprotocols.org>, as well as this copyright and license information must be included.