CUSTOMER SEGMENTATION APPROACHES: A COMPARISON OF METHODS WITH DATA FROM THE MEDICARE HEALTH OUTCOMES SURVEY

by

Gina Pugliano McKernan

B.A., Washington & Jefferson College, 2006

M.A., University of Pittsburgh, 2009

Submitted to the Graduate Faculty of

the School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Gina Pugliano McKernan

It was defended on

November 20, 2017

and approved by

Suzanne Lane, Professor/Chair, Research Methodology

Suzanne Lane, Associate Professor, Research Methodology

Lauren Terhorst, Associate Professor, Department of Occupational Therapy

Dissertation Advisor: Clement Stone, Professor, Research Methodology

CUSTOMER SEGMENTATION APPROACHES: A COMPARISON OF METHODS WITH DATA FROM THE MEDICARE HEALTH OUTCOMES SURVEY

Gina Pugliano McKernan, Ph.D.

University of Pittsburgh, 2017

Model-based segmentation approaches are particularly useful in healthcare consumer research, where the primary goal is to identify groups of individuals who share similar attitudinal and behavioral characteristics, in order to develop engagement strategies, create products, and allocates resources tailored to the specific needs of each segment group. Despite the growing research and literature on segmentation models, many healthcare researchers continue to use demographic variables only to classify consumers into groups; while failing to uncover unique patterns, relationships, and latent traits and relationships. The primary aim of this study was to 1) examine the differences in outcomes when classification methods (K-Means and LCA) for segmentation was used in conjunction with continuous and dichotomous scales; and 2) examine the differences in outcomes when prediction methods (CHAID and Neural Networks) for segmentation was used in conjunction with binary and continuous dependent variables and a variation of the classification algorithm. For the purpose of comparison across methods, data from the Medicare Health Outcome Survey was used in all conditions. Results indicated that the best segment class solution was dependent upon both the method and treatment of the inputs and dependent variable for both classification and prediction problems. When the input depression scale was dichotomized, the K-Means model yielded a 6 segment best-class-solution, whereas the LCA model yielded 9 distinct segment classes. On the other hand, LCA models yielded the same segment solution (9 classes), irrespective of the treatment of the depression scale. Similarly, differences in outcomes were identified when the dependent variable was continuous vs. binary when prediction models were

used to segment survey respondents. When the outcome was dichotomous, CHAID models resulted in a 5-segment solution, compared to a 6-segment solution for Neural Networks. On the other hand, the binary dependent variable produced a 4-segment solution for both CHAID and Neural Network models. In addition, the interpretation of the segment class profiles is dependent upon both method and condition (input and treatment of dependent variable).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my sincere appreciation to my committee members, Dr. Clem Stone, Dr. Suzanne Lane, Dr. Feifei Ye, and Dr. Lauren Terhorst for their time, review, and feedback throughout the entire dissertation process. To Dr. Stone, Dr. Lane, Dr. Ye, and the late Dr. Kim- thank you for being incredible teachers. I've learned so much from all of you. Dr. Stone, I will be forever grateful to you for your support, patience, and dedication to my academic and professional success over the (many!) years I've been at Pitt. It's been an honor and a pleasure learning from you.

Thank you to my family- my parents, Laure, Pat, and Joe; sister, Krista; husband, Shawn; and the many in-laws, aunts, uncles, and cousins for your encouragement and understanding over the many years I've been doing the school thing while trying to balance everything else. To my dear friend Jan- you have always reminded me to "put one foot in front of the other." And to Teagan, you have been my inspiration and motivation since your Day 1. You are already so much smarter than me, and I hope you never let anyone discourage you or hinder you from achieving your dreams. I'm so proud to be your mommy.

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

1.1.1 Background

A deep understanding of its customer base is a crucial component to the success of any consumer-facing company, organization, or industry. Customer segmentation, or dividing customers into groups based on similarities, is a powerful tool used to identify unique customer wants, needs, and future behavior. At its core, segmentation methods are leveraged to reduce the number of individuals being dealt with into meaningful subgroups that share well-defined characteristics (Teichert, Shehu, & Wartburg, 2008). Customer segmentation has the ability to identify underserved customer groups, and allocate resources (marketing, specific media, spending) accordingly. By adjusting efforts to meet specific customer needs, companies may see benefits such as increased customer engagement, awareness, sales, increased website and in-store traffic, and many more. Insights gained from a segmentation analysis can be used to prioritize new products, develop tailored marketing campaigns to target individual segments of healthcare consumers, and determine the potential profitability of each segment. In addition, segmentation analysis can be very valuable when designing marketing experiments, such as testing the effectiveness of a campaign in specific customer groups. Segmentation models have been used to price airline tickets, guide targeted messaging techniques, identify health provider communication preferences, and predict bankruptcy and credit card evasion, (Teichert et al., 2008, Bhatnagar & Ghose, 2004, Kuo, Ho, & Hu et al., 2002).

Customer segmentation is often employed by first gathering all possible information about the current customer market base. This often involves utilizing actual purchase data, surveying the customers about their attitudes, beliefs, past and future behavior, and collecting demographic data from customers themselves and/or outside vendors or data brokers. Once all information is gathered, researchers employ a variety of statistical techniques to place customers into "buckets" or segments, in which they score similarly on the aforementioned demographic, attitudinal, and behavioral variables. In most cases, the computational procedure assigns a number to an individual, representing their most likely segment class. It is up to the researcher to determine the characteristics unique to that segment, which often involves examining the descriptive statistics (means, medians) of the segment characteristics.

Despite the widespread proliferation of segmentation studies in market research and in practice, marketing managers and executives have historically relied upon intuition and/or an analysis of socio-demographic variable only to segment their customers (Teichert, Shehu, & Wartburg, 2008). Segmentation researchers have also been faced with the substantial challenge of determining the appropriate number of segments. It's common for researchers to choose too few segments, resulting in a lack of insightful information or distinguishing characteristics. On the other hand, a larger number of segments may be inherently more difficult to deal with, but may offer valuable information to the user. Few studies have examined convergent validity across methods and datasets. Therefore, a comparison of methods utilizing the same dataset would greatly contribute to the literature surrounding customer segmentation.

The data source for this study, the Medicare Health Outcomes Survey (HOS), is an assessment of the physical functioning and mental well-being of Medicare beneficiaries over time. The survey was implemented in 1998, nationally in Medicare managed care organizations

(MCOs), as part of Medicare Healthcare Effectiveness Data and Information Set (HEDIS®). The goal of the HOS program has been to gather valid, reliable, and clinically meaningful data that are used by: MCOs, providers, and quality improvement organizations to monitor and improve health care quality; CMS to assess the performance of MCOs and reward high performers; Medicare beneficiaries, their families, and advocates when making health care purchasing decisions, and health researchers to advance the state-of-the-science in functional health outcomes measurement, and quality improvement interventions and strategies.

Insights from segmentation analyses have been widely used in financial services, retail and other sectors to influence consumer purchasing behavior. Yet, the design of most health-related products, services, and interventions remains remarkably unaffected by segmentation-related analytics. Health care organizations need to move beyond a "one size fits all" approach, and strive to customize services for individual needs. There are many ways to segment customers, the obvious being demographic dimensions such as gender, age, race, etc. Although these dimensions are important, they usually provide little understanding about how and why consumers go about shopping for, selecting, using, and evaluating health care services. A more useful construct is to look at "psychographic" dimensions — people's priorities, attitudes, and values.

Segmentation offers insights into health care consumers' behaviors and attitudes – critical information in an environment where health care is moving rapidly towards patient-centered care which is premised upon individuals becoming more active participants in managing their health care. Awareness of consumers' preferences and styles needs to be taken into consideration and strategies to encourage and support consumer engagement in health care are important for providers, health plans, and bio-pharma companies. Increased access to health information can help consumers make better and more informed decisions leading to better quality of care, health

outcomes, and satisfaction with care. Providing consumers with more information may change their behavior in a way that reduces health costs. Segments give valuable ‘clues’ as to how health care organizations may more specifically target and personalize products and services for health care consumers.

- Financial preparedness/economic profile
- Healthcare utilization/satisfaction
- Test/treatment use and compliance
- Care preferences (traditional approach, want to make healthcare decisions on own)
- Health status (diet, exercise, wellness/preventative activities)
- Views of healthcare reform
    - Healthcare spending
    - How healthcare is meeting their needs

Results of an effective healthcare segmentation can change the healthcare delivery landscape, by engaging with individuals more effectively by taking advantage of deeper insights (going beyond disease-based classification; integrating behavior change as a core component of new care delivery models; and adopting a multi-stakeholder approach to support primordial and prevention initiatives.

The purpose of this study is to bridge the gap in the literature between several of the most commonly used methods for segmentation, while utilizing a common data set. Potential differences in outcomes by segmentation method will be examined with respect to responses from Cohort 15 (initially surveyed in 2012; re-surveyed in 2014) on the Medicare Health Outcomes Survey. Segmentation can essentially answer two types of problems: classification and prediction. Therefore, methods for classification (latent class analysis and k-means clustering) will be

evaluated against each other. Similarly, outcomes of prediction methods (neural networks and chi-square automatic interaction detector or CHAID) will be compared. A secondary goal of this research is to assess the performance of the segmentation models across different types of independent (for prediction)/input (for classification) and outcome variables. Input variables will vary according to classifying algorithm (boosting vs. bagging) and scale; whereas the dependent variable for prediction will exist in both a continuous and binary state for comparison purposes. The four segmentation models will be evaluated according to model fit/accuracy, the number of segment classes yielded, and interpretation of segment classes by variables included in the model.

1.1.2 Statistical approaches

In order to identify customer groupings and gain a deeper understanding of their similarities and differences, statistical methods beyond simple descriptive analyses should be employed. The more rigorous segmentation methods, and those that will be investigated in this research include: latent class analysis, k-means clustering, neural networks, and decision trees or CHAID.

Latent class analysis (LCA) is a frequently used method to classify individuals based on some underlying or unobservable (latent) variables and relationships (Patterson, Dayton, & Graubard, 2002), by means of assigning a probability of belonging to each latent class or segment to every individual in the dataset. The K-means algorithm uses a partitioning (non-hierarchical) solution to identify similar groups of individuals or customers on selected characteristics by calculating distances from a cluster center; for variables in the dataset those with similar distances are clustered together. Neural network models are used in segmentation studies when the relationships between variables are complex or obscure (Kumar, Rao, Soni, 1995). These models have the distinct advantage of approximating nonlinear functions and work well by identifying

hidden trends or relationships in large datasets with many independent variables that are non-linear and may appear unrelated (Rygielski, Wang, & Yen, 2002). The Chi Square Automatic Interaction Detector (CHAID) procedure recursively partitions a population into separate and distinct groups, by searching for the predictor variables that best differentiate among the individuals with respect to dichotomous or continuous classifications for variables (MacCarty & Hastak, 2007). When the CHAID objective is met: the variance of the dependent (target) variable is minimized within the groups, and maximized across the groups (Borden, 1995). An advantage to the CHAID model is that the output is easy to interpret and highly visual, resembling an organizational chart.

With the availability of large datasets and advanced data mining techniques and software, issues surrounding predictor variable selection cannot be ignored (Dash & Liu, 1997, Guyon & Elisseeff, 2003). Often, the researcher is faced with the problem of too many variables (columns) and too few cases (rows), thus causing problems with overfitting and statistical testing of hypotheses. The goal of any analysis is to explain the data in the simplest way, but by completely ignoring predictors, bias is introduced. Therefore, the researcher is tasked with selecting variables that produce the best, parsimonious model; one that performs well in terms of accuracy and statistical tests, while ensuring the integrity of the original dataset.

Fortunately, procedures such as neural networks, CHAID, k-means clustering, and latent class analysis within many software packages have built-in procedures for reducing the overall number of variables in a model. These procedures commonly use the R-Square method with a sequential forward selection process to select the input variable that has the highest correlation coefficient with the target.

The benefits to variable reduction involve both statistical and practical outcomes. Models containing a reasonable number of predictors can be much easier to interpret and visualize. In

terms of logistic and operational benefits, fewer variables take less computational processing and storage. Also, if a model is used for prediction needs, then the researcher can cut costs by not measuring redundant predictors. Statistically, reducing the number of predictors has been shown to increase model performance and accuracy while reducing dimensionality.

1.2 PURPOSE

The primary aim of this study is to examine the differences in outcomes between models for two types of segmentation analysis problems. For classification purposes, latent class analysis and k-means clustering will be compared. For segmentation studies in which there is a defined research question and/or prediction need, outcomes of neural networks and CHAID models will be compared. Response data from Medicare Health Outcomes Survey (HOS) will be used for all segmentation models. A secondary aim will be the evaluation of the treatment of independent and dependent variables within the segmentation model. The varieties of dependent variables studied in this analysis include dichotomous (binary) and continuous outcomes for prediction models. The outcome for classification models is the probability of segment class membership. The types of input variables include: dichotomous and categorical scales for classification models, and continuous and categorical responses for prediction models. Criteria used to determine the best segment class solution will also be examined across models, which include: the number of segment classes yielded and segment size, model fit and accuracy, and variable importance.

1.3 RESEARCH QUESTIONS

Research Question #1: Do distinct segmentation methods for classification questions result in different outcomes, such as the number of segment classes, segment class size, and important variables, when the scales of the inputs (continuous vs. binary) are varied? How can/should a researcher interpret potential differences?

Specifically,

1.1. How do the segments differ by number of classes, size, and differentiating variables by method?

1.2. How do the segments vary by number of classes, size, and important variables by method, given the dichotomization of the depression scale?

1.3. What are the potential implications resulting from different outcomes across methods and inputs??

Research Question #2: Do distinct segmentation methods for prediction questions result in different outcomes, such as the number of segment classes, segment class size, and important variables, when the dependent variable is continuous vs. binary? How can/should a researcher interpret potential differences?

Specifically,

2.1. How do the segments differ by number of classes, size, and important predictor variables by method when the outcome is binary?

2.2. How do the segments differ by number of classes, size, and important predictor variables by method when the outcome is continuous?

2.3. What are the potential implications resulting from different outcomes across methods and treatment of dependent variable?

Research Question #3: How does the treatment of the dependent and independent variable affect prediction model results?

Specifically,

3.1. What are the differences in model fit/accuracy, number of segment classes, and important predictor variables when the dependent variable is dichotomous, as compared to a continuous outcome for prediction models?

3.2. What are the differences in model fit/accuracy, number of segment classes, and important input variables when a dichotomous vs. a continuous scale is used for classification models?

3.3. What are the differences in model fit and accuracy when boosting vs. bagging algorithms are used in prediction models?

1.4 SIGNIFICANCE OF THE STUDY

Since few studies have been able to compare differences in outcomes across methods using the same dataset, this study has the unique ability to identify the advantages/disadvantages of a particular method, and provide a recommendation for model usage based on defined conditions. Accordingly, results of this study will enrich the literature and practice by addressing issues of evaluating model fit, segment class comparisons, and determining the appropriate segment size(s).

While this study is designed to inform segmentation model choice decisions, results from this research can be used in conjunction with additional data to inform marketing executives, government, and health care providers about Medicare customers unique health needs, potential care gaps, preferences for healthcare, communication and engagement needs and limitations,

health management and others. Integration of these results with future research may include the following: evaluating outcomes of care and targeting quality improvement; examination of the relationship between BMI and quality of life in community-living older adults; a study of health-related quality of life in older adult survivors of selected cancers.

The majority of research surrounding the Medicare Health Outcomes Survey has been focused primarily on psychometrics and validating the instrument, with studies on non-response bias, weight adjustments, imputing summary scores, and applications of the survey to other instruments (Gandek, Sinclair, Kosinski and Ware, 2004; McCall, Khatutsky, Smith and Pope, 2004). In addition, most of the "analysis" within these studies has been descriptive or correlational. Specifically, there have been studies which examined relationships between Medicare Advantage contract characteristics and quality of care ratings; those that assessed multiple chronic medical conditions and health-related quality of life; and research that highlighted multiple risk factors and the likelihood of patient-physician communication and health maintenance services in Medicare health plans (Xu, Burgess, Cabral, Soria-Saucedo, and Kazis, 2015; Grace, Elliott, Giordano, Burroughs and Malinoff). As a result, the current study adds a level of depth to understanding the unique needs, preferences, attitudes, and behaviors surrounding the Medicare population.

2.0 LITERATURE REVIEW

A comprehensive review of the literature on customer/group segmentation methods and considerations, with a focus on the most relevant modern methods, will be presented in this chapter. As indicated by the information presented below, there is a lack of convergent validity between methods, as no single data set is used across studies. A brief history of segmentation in the healthcare literature, as well as the types of dependent variables examined in this study will also be presented.

2.1 OVERVIEW OF SEGMENTATION METHODS

The practice of segmenting individuals into homogenous groups has been utilized in consumer research since the 1950's, due to the technique's ability to inform managers of specific customer needs and benefit orientation (Bonoma & Shapiro, 1984). From both a research and marketing perspective, segmentation can be defined as identifying meaningful sub-groups of individuals, by reducing the number of individuals being dealt with into a manageable number of clusters that are mutually-exclusive and well-defined (Teichert et al., 2008). Statistical segmentation methods can fill the gaps of the intuition and demographic methods which were previously employed. Statistical methods can inform customer preferences and classes of individuals who would likely be receptive to targeted offerings, strategies, and engagements (Teichert et al., 2008, Bhatnagar & Ghose, 2004, Kuo, Ho, & Hu et al., 2002, Bonoma & Shapiro, 1984). The methods have evolved in last several

decades to include more complex, model-based/optimizing approaches; as compared to simple regression or clustering. (Kim, Fong, & Desarbo, 2012).

There are two primary reasons one would want to use segmentation methods in research: 1) to understand the classification or groupings of individuals with no predefined outcome or target variable, or 2) to uncover the characteristics that differentiate groups of individuals with respect to a predicted outcome. Classification models divide the data into segments, or clusters, of records that have similar patterns of input fields. These models could be particularly useful for identifying patterns or groups of interest in a customer base. The value of the classification model is determined by its ability to capture interesting segments in the data and provide useful descriptions of those segments. On the other hand, prediction models use the values of the input fields to predict the value of the output or target field. Prediction models can help an organization to predict a known result, such as whether a patient is likely to remain loyal to their primary care physician or be compliant with treatment.

2.2 SEGMENTATION IN HEALTHCARE

In health communication research, audience segmentation is often based on demographic characteristics (e.g., gender and/or ethnicity) or health status (e.g., disease or diagnosis). Studies in consumer segmentation research have suggested, however, that segmenting people by their demographic and behavioral characteristics alone may not be as effective as applying psychosocial or behavioral segmentation schemes, which use relevant attitudes, perceptions, and behaviors to classify and segment consumers (Maibach, Maxfield, Ladin, & Slater, 1996; Morris, Grossman, Barkdoll, & Gordon, 1987; Slater, 1996). The latter approach may offer additional insight into

consumers' values, motivations, and prominent behavioral cues. For example, in their 2005 study to understand physical activity behaviors, Boslaugh and colleagues found that a multivariable segmentation strategy (which included psychosocial and health status factors) uncovered more behaviorally homogeneous subgroups, while only using demographic characteristics was virtually analogous to not segmenting at all in their study.

A more refined healthcare segmentation strategy involves examining audiences by their health information-seeking (HIS) behaviors and attitudes. While many researchers have established psychobehavioral segmentation schemes grounded on consumers' knowledge, attitudes, and self-efficacy for certain behaviors (Arora, Ayanian, & Guadagnoli, 2005; Hibbard, Mahoney, Stockard, & Tusler, 2005; Williams & Heller, 2007), none have incorporated HIS styles in their research. HIS behaviors and attitudes have been found to be highly interwoven with the way consumers consider their health care options and engage in health activities (Lambert & Loiselle, 2007). Though physicians historically have been the "gatekeepers" of health information (Meissner,Potosky, & Convissor, 1992), patients and consumers now have access unlimited information via the internet, advertising, social media, and others (Fox & Fallows, 2003; Hartley & Coleman, 2008).

Additionally, researchers have suggested that grouping all older healthcare consumers into one age-based category is extremely short-sighted. Moschis (1996) emphasizes that the Medicare-aged healthcare consumer market warrants segmentation above all other markets, due to the fact that with aging comes great variance with respect to lifestyles, wants and needs, and healthcare consumption practices. Initially, Leventhal (1991) suggested segmenting older consumers on the basis of chronological age only, and then took into consideration features such as financial standing, marital status, and health. Lumpkin (1985) presented a segmentation model constructed

on retirement and age to establish different segments, which was subsequently refined to include different shopping orientations and lifestyle factors (Lumpkin et al., 1985).

Continuously, research has shown that segmentation models that contain attitudinal, lifestyle, and preference characteristics are much more accurate with respect to predicting responses to consumer behavior as compared to models based on chronological age groupings (Moschis and Mathur, 1993). The one caution to these models is that often demographic differences between segments are slight, and the most beneficial models of older consumer segments incorporate a wide selection of variables, including demographic data (Ostroff, 1989).

2.3 MODERN METHODS

2.3.1 Latent class analysis

Latent class analysis (LCA) is a frequently used method to classify individuals based on some underlying or unobservable (latent) variables and relationships (Patterson, Dayton, & Graubard, 2002). Latent class analysis (LCA) is a statistical method used to identify a set of discrete, mutually exclusive latent classes of individuals based on their responses to a set of observed categorical variables. LCA also assigns a probability of belonging to each class or segment. In multiple-group LCA, both the measurement part and structural part of the model can vary across groups, and measurement invariance across groups can be empirically tested. Most statisticians credit Lazarsfeld and Henry (1968) with the origins of latent class analysis and Goodman (1974) with the computational breakthroughs that made it practical. Goodman's maximum likelihood approach remains the standard way to estimating parameters in the latent class model (Thompson, 2007).

In the social and behavioral sciences, many constructs are often referred to as latent variables. Latent variables cannot be directly identified and instead must be implied from multiple observed variables (Lanza, Collins, Lemmon, & Schafer, 2007). Latent class models divide a population into mutually exclusive and exhaustive subgroups (Lanza et al., 2007). Latent class analysis comes from a group of general mixture models, containing two parts: a measurement model and a structural model. The measurement model for LCA is a multivariate regression model that describes the relationships between a set of observed variables and a set of categorical latent variables (Muthén, 2007). In a traditional latent class model, the following parameters are estimated:

- γ (gamma) parameters, which represent latent class membership probabilities
- ρ (rho) parameters, which are variable-response probabilities conditional on latent class membership. The ρ parameters represent the relationship between the observed variables and the latent classes. Since these parameters are used to describes the relationship between a set of observed variables and the underlying latent classes, this relationship can be referred to as the "measurement piece" of the model
- β (beta) parameters are logistic regression coefficients for covariates predicting class membership. When additional covariates are introduced to predict latent class membership, this adds the "structural piece" to the measurement model

When multiple subgroups are reflected in the model, a grouping variable is included and both sets of parameters (γ, ρ) are be conditioned on group. If one or more covariates are included, the β parameters must also be estimated. Of note, when covariates are in the model, only ρ and β

parameters are actually estimated, and the γ parameters are calculated as functions of β and the covariates.

Latent class analysis relies on a contingency table created by cross-tabulating all indicators of the latent class variable. First, a latent class model with n_c classes (subgroups) is estimated from a set of *M* dichotomous items. Next, included in the model is a covariate denoted *X* which may be either continuous or dichotomous (0 or 1 coded). Let the vector $Y_i = (Y_{i1},...,Y_{iM})$ represent individual *i*'s responses to the *M* items, where the possible values of Y_{im} are 1, …, r_m. Let $Li = 1,2,...,n_c$ be the latent class membership of individual *i*, and let *I(y = k)* be the indicator function; that is, a function that equals 1 if *y* equals *k*, and 0 otherwise. Let X_i represent the value of the covariate for individual *i*; the covariate may be related to the probability, γ, of membership in each latent class, but is assumed to be otherwise unrelated to Y_i. Then the contribution by individual *i* to the likelihood is:

$$P(\mathbf{Y}_i = \mathbf{y} \mid X_i = x) = \sum_{l=1}^{n_c} \gamma_l(x) \prod_{m=1}^{M} \prod_{k=1}^{r_m} \rho_{mk|l}^{I(y_m=k)} \text{ , (1)}$$

The β parameters are the coefficients in logistic regressions using the covariate *X* to predict latent class membership. The γ parameters can be expressed as functions of the β parameters as follows:

$$\gamma_l(x) = P(L_i = l|X_i = x) = \frac{\exp(\beta_{0l}+x\beta_{1l})}{\sum_{j=1}^{n_c}\exp(\beta_{0j}+x\beta_{1j})} = \frac{\exp(\beta_{0l}+x\beta_{1l})}{1+\sum_{j=1}^{n_c-1}\exp(\beta_{0j}+x\beta_{1j})} \text{ , (2)}$$

for = 1,... ,n_c. Note that the last two expressions on the right are equal because it is assumed that the last (i.e., the n_c^{th}) class is used as the reference class. The reference class has its βs constrained to zero, because the relative probabilities of being in the other classes are being compared to the probability of this reference class. It is necessary to choose one class and set its βs to zero for model identification, under the further constraint that the probabilities for all classes must sum to one for each individual. The choice of reference class does not affect the final fitted probability estimates for any individual or class.

This model allows for the estimation of log odds that individual *i* falls in latent class relative to the baseline class. For example, if class 2 is the reference class, then the log odds of membership in class 1 relative to class 2 for an individual with value on the covariate is:

$$log\left(\frac{\gamma_1(x)}{\gamma_2(x)}\right) = \beta_{01} + \beta_{11}x, \quad (3)$$

The foundation of LCA is the notion that individuals within a particular segment share a common joint probability distribution based on observed variables (Teichert, Shehu, & von Wartburg (2008). LCA was first introduced as a way to analyze survey data; specifically with the goal detecting a latent attitude measured by multiple response items (Nylund, Asparouhov, & Muthén, 2004). LCA is best used when the data has been dichotomized (assigning values of "0" and "1" to variable responses), as in the probability of a customer making a purchase decision (purchase or not purchase) (Bhatnagar & Ghose, 2004). In addition, LCA can be used to identify characteristics of segment classes with or without the presence of a certain disease or those who are likely to exhibit compliance with a treatment decision. LCA can identify patterns in multiple outcome variables and has the benefit of creating segments that are highly associated with actual behavior and preferences (Teichert et al., 2008). In addition, LCA allows for the inclusion of individuals with missing data on some of the dependent variables, reducing potential misclassification.

In the context of segmenting consumers with respect to their perceptions and behavior, one can assume that there are *k* latent segments in the markets, and all members of a particular segment share the same latent inclination (e.g., to make a purchase) within a specified category (Bhatnagar & Ghose, 2004). Let this latent or intrinsic inclination of all members in a particular segment *k* for making a purchase in category *j* be β_j^k, and since it is unknown (a priori) which segment a

particular consumer belongs, let the probability of a given consumer in segment *k* be π^k. The unconditional probability of an inclination in consumer *i* in category *j* is:

$$\text{Prob(Inclination)}_j^i = \sum_k \frac{\exp(\beta_j^k)}{1+\exp(\beta_j^k)}\pi^k, \quad (4)$$

The segment probability has a logit distribution,

$$\pi^k = \frac{\exp(\lambda^k)}{1+\sum_{k=1}^{K-1}\exp(\lambda^k)}, \quad (5)$$

where *K* is the total number of segments. Maximizing the sample likelihood, $\Lambda = \prod_i \Lambda^i$ allows for the measurement of preference for specific categories, and the parameters are conditional on a pre-specified number of latent segments (Bhatnagar & Ghose, 2004). The process begins by first estimating and comparing a single-segment model, a two segment solution, and a three-segment model, etc.

Currently, a commonly accepted criterion for determining the final number of segment classes in latent class analysis does not exist (Nylund, Asparouhov, & Muthén, 2007). Consequently, in a latent class analysis, the researcher must decide on the number of classes, which isn't always an exact science (Nylund et al., 2004). Running several iterations of the LCA with increasing number of classes, and identifying the solution that splits that data the best (creates uniqueness in the classes) is one, non-scientific and somewhat time-consuming way of determining the desirable number of segments. Although LCA models are considered "nested models" due to their varying numbers of classes, chi-square difference tests, in the form of log-likelihood is not appropriate given the inability of the model to meet regularity assumptions (Nylund et al, 2007). On the other hand, several studies have reported an evaluation of model fit by examination of information criteria, such as Bayesian information criteria (BIC) and Constrained Akaike Information Criteria (CAIC) (Teichert, Shehu, & von Wartburg, 2008, Bhatnagar & Ghose, 2004).

The "winning" model would be the case in which the BIC is minimized, given by BIC = $-2\Lambda + k\ln(N)$, where Λ is the log-likelihood, k is the number of parameters and N is the sample size (Bhatnagar & Ghose, 2004).

BIC has been found useful as a model fit criterion by some researchers. For example, Bhatnagar & Ghose (2004) demonstrated that BIC was minimized for a 3 segment solution (8218.486), as compared to a 2 segment (8480.096) and 4 segment (8228.013) solution of an analysis of online shoppers. The researchers were then able to classify online shoppers by their preferences for online shopping (amount of money spent on computer software/hardware, home electronics and other services), and demographic variables, such as age, gender, income, and education.

2.3.2 K-means clustering

K-means clustering is one of the most widespread methods for market-based segmentation analysis (Hruschka & Natter, 1999). It is appealing due to its simplicity for use and application to multiple scenarios (Kim & Ahn, 2008). The K-means algorithm uses a partitioning (non-hierarchical) solution to identify similar groups of individuals (clusters) or customers on selected characteristics by calculating distances from a cluster center, and those with similar distances are clustered together. As demonstrated in Figure 1, the colored dots represent individuals or data points with similar characteristics, resulting in three distinct clusters. For large amounts of data, K-means algorithms have been found to perform very well (Kuo, Ho, & Hu, 2002).

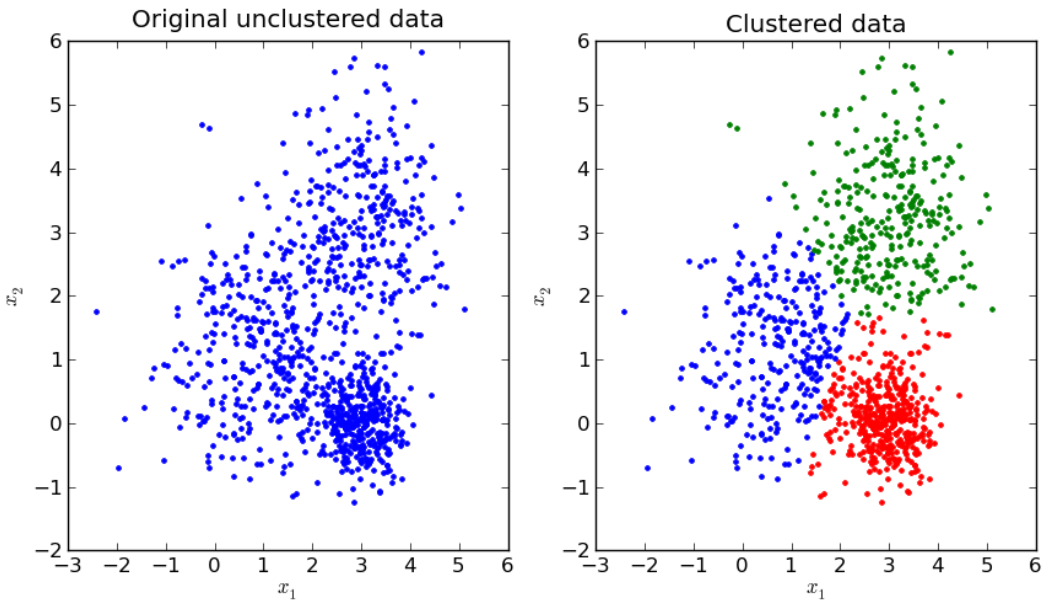


Figure 1. Visualization of K-Means Clustering

K-means analysis begins with the researcher randomly selecting the starting point and number of clusters (Kuo et al., 2002). The K-means procedure is iterative in nature and repeats the process of calculating new centroids of clusters until the clusters stop changing or the a priori conditions have been met, referred to as a "hill-climbing" strategy (Kim & Ahn, 2008). Researchers have demonstrated in simulation studies,that K-means clustering algorithms produce very low misclassification rates, which only diminish in error as the number of clusters increase (Kim & Ahn, 2008, Kuo et al., 2002). For example, the mean misclassification rate for a zero-error simulated data set for a conventional (two-stage) K-means algorithm was 0.583%; compared to 1.778% mean misclassification using a self-organized feature map (neural network) (Kuo et al., 2002).

K-means is one of the simplest algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) fixed a priori (Hruschka & Natter, 1999). The process is iterative in nature; where each observation is assigned to exactly one cluster whose mean yields the smallest

within-cluster sum of squares. Then, new means are calculated to become the centroids of the observations in the clusters. The process continues until convergence, or when the assignments of observations to clusters do not change. Finally, this algorithm aims at minimizing an objective function known as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2 , (6)$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Let $X = \{x_1, x_2, x_3, \ldots . x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots . v_c\}$ be the set of centers (centroids).

Algorithmic steps for k-means clustering

1) Randomly select 'c' cluster centers. Start with a large value (e.g. 15) and keep removing centroids

2) Calculate the distance between each data point and centroids.

3) Assign the data point x_i to the cluster center whose distance from the cluster center is the minimum of all the cluster centers, $J(V)$.

4) Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i} \right) \sum_{j=1}^{c_i} x_i , (7)$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3.

As a caution, researchers have demonstrated that clustering centroids can be very sensitive to outliers (Chaturvedi, Carroll, Green, & Rotondo, 1997). Since one person can only be assigned to a single cluster, outliers tend to distort the cluster centroids and membership. Overlapping cluster centers (as opposed to partitions) have been shown to increase interpretability and parsimony of the models by reducing the total number of segments and removing outliers. Another potential drawback of K-means clustering is that it is highly dependent on the initial seed, and there is no specific method for enhancing the initial seed (Kim & Ahn, 2008).

2.3.3 Comparison of LCA and K-means

Several studies have attempted to describe the benefits (or similarities) in using a latent class versus a k-means approach to segmentation problems (Chatirvedi, Green, & Carroll, 2001, , Sarstedt & Ringle, 2010, Schreiber & Perkarik, 2014). In their 2014 paper, Schreiber & Perkarik identified several reasons why latent class analysis was preferred over k-means for classifying museum visitors. The K-means model could not provide a clear distinction for the segment class solution (i.e., it could vary between 2-8 clusters), while the latent class model was able to accurate identify five distinct segments of museum visitors. Similarly, researchers found that finite mixture models (i.e., latent class models) were superior to k-means, as indicated by unobserved heterogeneity and measurement error within the k-means models (Sarstedt & Ringle, 2010). On the other hand, a large review of finite mixture models in market segmentation was not able to provide a recommendation for a preferred method (Tuma & Decker, 2013). These authors suggest that it is

not simply choosing a method, but ensuring that estimation methods, algorithm convergence, and model selection criteria are the critical factors in model evaluation and selection.

2.3.4 **Neural networks**

Neural network models are becoming increasingly popular in healthcare segmentation research due to their ability to explain relationships between variables that are complex or obscure (Kumar, Rao, Soni, 1995). These models have the distinct advantage of approximating nonlinear functions and work well with large datasets where the structure may be unknown. They have the advantage of identifying hidden trends or relationships in large datasets with many independent variables that are non-linear and may appear unrelated (Rygielski, Wang, & Yen, 2002). Neural network models are

Neural networks consist of several layers: input, output, and hidden layers, with weights attached to each connection (Yao, Tang, Po, Tan, 1998). Patterns are offered to the network via the input layer, which communicates to one or more hidden layers where the actual processing is completed by a system of weighted connections. The hidden layers then link to an output layer, where the final solution can be found. The nodes are connected from the input layer outward, and the error between the predicted output value and the actual value is fed back through the network to update the initial values assigned to the weights. The goal of this process is to minimize the error between anticipated and expected outputs (Yao et al., 1998). The attractiveness of a neural network is its ability to map an extract pattern from the input layers to the output; hence the reasons they are used to answer classification and segmentation questions. The ability to recognize patterns are trained into the hidden layers, and when new relationships are presented, the model is able to identify them (Yao et al., 1998).

As shown in Figure 2, a node can have several inputs and outputs. The input layer consists of one node for each independent variable. For this example, let the inputs (x_i's) represent survey factors: feeling depressed, difficulty breathing, physical activity, history of cancer, etc. The external inputs received by the input layer nodes are directly fed into the nodes in the hidden layer (y_i's), which lie in between the input and output layers (z_i's). The hidden layers, have no a priori meaning, but represent a "block" of summed weighted inputs that pass as a response to a non-linear function to create the output layer. Each block of response contains different coefficient weights that are learned by the neural network via training. In simplest terms, the hidden layer's job is to transform the inputs into something that the output layer can use. In this example, the output layer corresponds to the dependent variable(s), which represent general health status, number of unhealthy days, and mortality. The number associated with each link (represented by an arrow) is referred to as a weight (w) or coefficient. The weights in the network can be updated from the errors calculated during training.

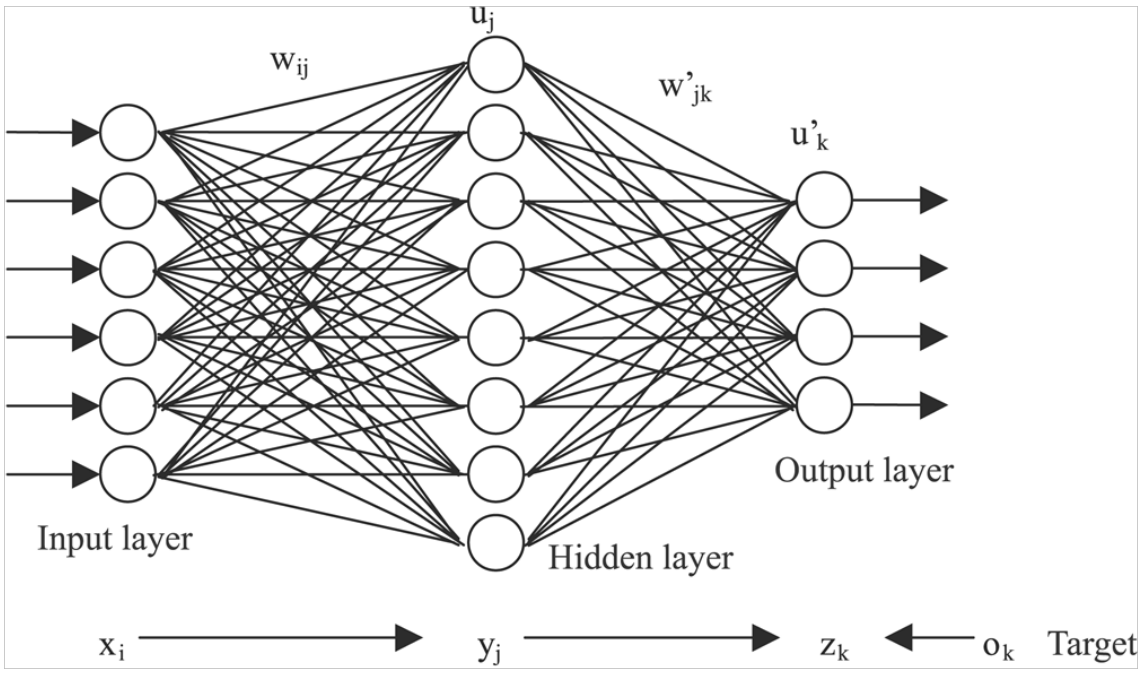


Figure 2. Visualization of Neural Network

Given such a network and set of external input values, it is possible to compute the final output of a node *i*. First, compute the net input into node *i*, net_i; and then, convert net_i by applying a transformation function (discussed below). The net_i received by node *i* is equal to the weighted sum of all the inputs fed into it by all nodes *j*, whose output is connected to *i*- that is $\mathrm{net}_i = \sum_j w_{ji} o_j$, where o_j is the output of node *j*, and w_{ji} is the weight on link *ji*. In the next step, the output of node i is computed by applying a sigmoid transformation function as: $o_i = 1/[1 + \exp(-net_i)]$. Thus, each hidden-layer node and output node produces a value between 0 and 1 because of the sigmoid transformation function.

The choice of activation function in the output layer is strongly constrained by the type of problem that is being modeled, where the role of the activation function is to transform a set of inputs into an output. Basic neural networks opperate by adding up the inputs and feeding the sum into a function -- the activation function -- to determine the neuron's output. The activation function is typically sigmoid (a special case of the logistic function) , while the main constraint is

that it can't just be linear. Neural networks with a linear activation function can only accommodate a single hidden layer, regardless of the network architecture (Leshno, Lin, Pinkus, & Schocken, 1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. For example:

- A regression problem may have a single output neuron and the neuron may have no activation function.
- A binary classification problem may have a single output neuron and use a sigmoid activation function to output a value between 0 and 1 to represent the probability of predicting a value for the class 1. This can be turned into a crisp class value by using a threshold of 0.5 and snap values less than the threshold to 0 otherwise to 1.
- A multi-class classification problem may have multiple neurons in the output layer, one for each class. In this case a softmax activation function may be used to output a probability of the network predicting each of the class values. Selecting the output with the highest probability can be used to produce a crisp class classification value.

For the purposes of this study, the single-hidden-layer Multi-Layer Perceptron (MLP) transformation function will be used. An MLP can be viewed as a logistic regression classifier where the input is first transformed using a learnt non-linear transformation Φ. This transformation projects the input data into a space where it becomes linearly separable. A single hidden layer is sufficient to make MLPs a universal approximator.

Formally, a one-hidden-layer MLP is a function $f: R^D \rightarrow R^L$, where D is the size of input vector x and L and is the size of the output vector $f(x)$, such that, in matrix notation:

$$f(x) = G\left(b^{(2)} + W^{(2)}\left(s\left(b^{(1)} + W^{(1)}x\right)\right)\right),$$

with bias vectors $b^{(1)}. b^{(2)}$; weight matrices $W^{(1)}, W^{(2)}$, and activation functions *G* and *s*.

The vector, $h(x) = \Phi(x) = s\left(b^{(1)} + W^{(1)}x\right)$ constitutes the hidden layer. $W^{(1)} \epsilon\, R^{DxD_h}$ is the weight matrix connecting the input vector to the hidden layer. Each column $W_{-I}^{(1)}$ represents the weights from the input units to the i-th hidden unit. Typical choices for *s* include *tanh*, with $\tanh(a) = (e^a - e^{-a})/(e^a + e^{-a})$, or the logistic *sigmoid* function, with $sigmoid(a) = 1/(1 + e^{-a})$. The output vector is then obtained as: $o(x) = G\left(b^{(2)} + W^{(2)}h(x)\right)$.

To train an MLP, all of the parameters must be obtained. The set of parameters to learn is the set: $\theta = \left\{W^{(2)}, b^{(2)}, W^{(1)}, b^{(1)}\right\}$. Obtaining the gradients $\partial l/\partial\theta$ can be achieved through the backpropagation algorithm (a special case of the chain-rule of derivation).

Neural networks require that input variables be transformed into binary variables (Rygielski et al., 2002), although the outcome can be either categorical (≥ 2 categories) or continuous. It is also common practice that variables are grouped into categories, with the number of hidden layers corresponding to the number of categories (Kumar et al., 1995). All input nodes are connected to the matching hidden node for that category, and all hidden nodes are connected to a single output node. The final output of the node is the predicted value of the decision (0 or 1), since the data are binary, which can be interpreted as acceptance/rejection, purchase/not purchase, etc.

Because of the complexity of the models, neural networks tend to be more accurate than other models, such as CHAID or regression-based models (Rygielski et al., 2002). In an early study of supermarket buyer decisions (carry vs. not carry new product), researchers demonstrated the neural network's ability to consistently outperform logistic regression models in terms of classification accuracy (Kumar et al., 1995). When the buyer's decision was to accept a new

product, logistic regression models correctly identified the “accept” decision across all product categories 41.4% of the time; whereas neural networks were able to correctly classify over fifty percent of cases. Consequently, the models also have a tendency of producing complicated or non-optimal solutions.

2.3.5 Chi Square Automatic Interaction Detector (CHAID)

Chi Square Automatic Interaction Detector (CHAID) has been extremely popular since the 1980’s, when it became available in SPSS. CHAID models are valuable in the sense that they remove the judgmental and subjective deliberations intrinsic to the segmentation process, by mapping data (customers) onto pre-defined classes (Levin & Zahavi, 2001). The CHAID procedure creates subsets of the categories on each classification variable that maximize between-group and minimizes within-group differences (Borden, 1995). CHAID begins by creating a node that contains every individual in the dataset. It then searches for the predictor variable that best differentiates among the individuals with respect to a dichotomous variable (MacCarty & Hastak, 2007). Next, it splits the original node on the winning predictor variable into as many subgroups that are significantly different from each other with respect to the dichotomous variable. The procedure repeats itself with the newly created nodes until there are no additional significant splits, as demonstrated in Figure 3, in which income level, age, and number of credit cards are used to predict credit rating (good vs. bad). Income level is the best at differentiating good credit vs. bad; followed by age at the low/medium income level, and number of credit cards when income is medium or higher.

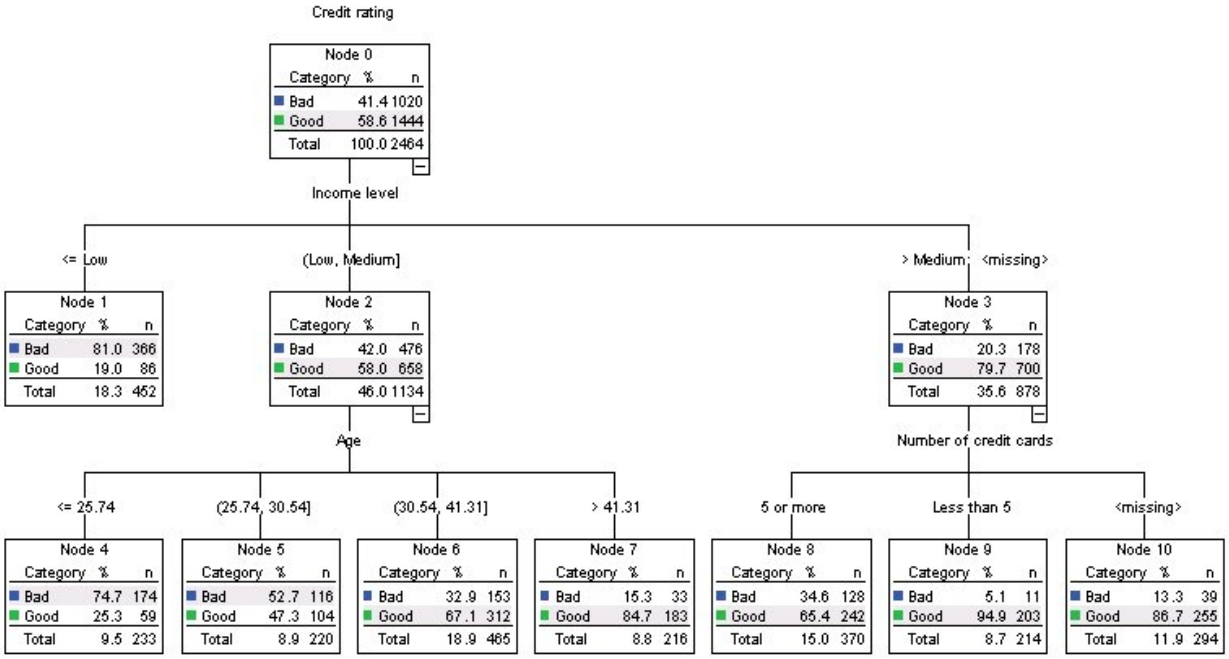


Figure 3. Example of CHAID Decision Tree

CHAID models can handle many independent variables (MacCarty & Hastak, 2007, Levin & Zahavi, 2001), but tend to perform better with small samples. When the independent variables are categorical, CHAID models tend to perform better than other methods (e.g. neural networks). The results of a CHAID model are a set of rules (which have undergone a rigorous experimental and calibration process) that can easily be explained to cosumers (Rygielski, Wang, & Yen, 2002, Levin & Zahavi, 2001). Decision trees, such as CHAID models, have the distinct advantage over logistic regression models, as that they contain built-in processes to pick the "best" predictors affecting the terminal nodes, by sorting through all possible combinations of variables to grow a tree and selecting the greatest spilt (Levin & Zahavi, 2001).

Most computer packages (e.g. SAS, SPSS) automatically finds the best split on an input unless too many candidate splits would have to be examined, whereas the user specifies this limit. The limited search on an input begins with a split into many branches and then proceeds in steps of merging pairs of branches. This heuristic search also considers switching input values between

branches after every merge, thereby examining many more candidates than just merging. The best split examined is then adopted in the final model (MacCarty & Hastak, 2007).

Several studies have compared the results of CHAID methods to "judgment-based" logistic regression and RFM (recency, frequency, and monetary value) models, using a common data set (per study) and evaluation criteria (MacCarty & Hastak, 2007, Levin & Zahavi, 2001). Results are not consistent across the studies and suggest that the methods may be data/use case specific. For decades, RFM analysis was the preferred method of classifying customers in direct marketing, despite its lack of statistical rigor (McCarty & Hastak, 2007). RFM analyses are typically employed by assigning a weight to each of the variables: recency (when a purchase was made), frequency (how often purchases were made), and monetary value (how much money was spent) and creating a weighted score for each person in the database. In their 2007 study, MacCarty & Hastak examined each model's ability to identify customers who actually to responded to a direct mail offer, when the procedure is used to select a pre-specified (10%, 20%, 50%) percentage of customers who are likely to respond. Similarly, Levin & Zahavi (2001) studied the differences in classification of buyers vs. non-buyers between RFM, logistic regression, and CHAID, using actual response data from a direct-mail offer marketing campaign. Both studies use a gain percentage (percentage of respondents actually reached – percentage of respondents likely to respond). MacCarty & Hastak (2007) reported that CHAID and logistic regression perform similarly to each other, and outperform RFM in the identification of likely responders to a mail-order offer when the response-rate for an offer is very low (<5%), representing a classic setting for database marketers. Specifically, the RFM approach was only able to identify 34.6% of respondents in the holdout sample, as compared to 39.2% (yielding a 4.6% statistically significant difference) in the test sample; whereas the differences in proportions of test and holdout out

samples for CHAID and logistic regression were 0.8% and 0.7%, respectively (MacCarty & Hastak, 2007). On the other hand, Levin & Zahavi indicate that CHAID models outperform RFM, but under-perform, as compared to traditional logistic regression methods (2001) CHAID models were able to accurately classify between forty and sixty percent of buyers, as compared to 35% (RFM) and 63% (logistic regression), for the top 10% of the audience. The differences in classification accuracy do diminish slightly, as the precision decreases; RFM is able to correctly identify 79.7% of 50% of the audience, while CHAID ranges between 79.8%- 87.7% and logistic regression correctly classifies 89.7% of the top half of the database (Levin & Zahavi, 2001).

2.3.6 Comparison of CHAID and neural networks

Accordingly, there are a limited number of several studies that compare and contrast neural networks and CHAID models on their ability to predict meaningful segment classes (Chen, 2016, Olson, Cao, Gu, & Lee, 2009, Tso & Yao, 2005). In a recent study designed to create a model for detecting companies likely to produce fraudulent financial statements, researchers employed various data mining techniques, including CHAID and neural networks (Chen, 2016). Results indicated that CHAID models performed better than neural networks, in terms of prediction accuracy (92.69%) and lowest Type I error rate (7.31%). Similarly, in a study of modeling techniques for the prediction of energy consumption, a comparison of regression analysis, CHAID, and neural networks, resulted in the CHAID model containing the fewest number of significant factors (Tso & Yau, 2006). The CHAID model was preferred due to its ability to produce the most accurate model with the simplest structure. On the other hand, several studies have produced findings in which neural network and CHAID model performance is highly comparable. Olsen and colleagues (2009) used both models to segment premier meat purchasing customers according

to their purchase behavior. Neural network and CHAID models produced identical results in terms of model prediction accuracy; CHAID 83.7% and neural network 84.9%. The others did note that while the neural network model was superior in its ability to fit non-linear data, it can be difficulty to apply; whereas CHAID models are easily interpretable.

2.3.7 Summary of methods

Across many studies and many methods of customer segmentation analysis, one of the key features of segmentation, as compared to a descriptive analysis of customer demographics, is a segmentation's ability to uncover patterns and relationships that may provide deep, meaningful insights that were previously undiscovered. For example, in Bhathagar et al.'s 2004 latent class segmentation of e-shoppers study, results indicated that there were less demographic differences in the classes than current research suggested. Similarly, in Dennis et al.'s 2001 K-means analysis of retail shopping behavior, the model uncovered variables that were not originally hypothesized to contribute to the classification of customers, such as "other shoppers are nice people" or the environment is "lively and exciting." As expected, the model also identified variables related to the physical store (selection of merchandise, cleanliness) as most significant (Dennis et al., 2001).

Segmentation analyses using advanced data mining and statistical techniques have resulted in the development of target marketing, in which customer segments or groups are purposefully targeting according to the group's needs (Kuo et al., 2002). It is no longer the case that companies spend millions of dollars on a single campaign, but attempt to deliver content to target audiences who will likely benefit from the message. Results of an online shopping behavior study provided interesting insights as a product of a latent class segmentation analysis. Researchers found that consumers were more concerned with potential "losses," such as identity theft, reliability of the

online site to deliver the product advertised, than perceived "gains," such as the convenience of shopping online, and that getting the lowest price was a lower-priority item (Bhathagar et al., 2004). A CHAID model was used in an analysis of responders to a direct mail marketing campaign, and was able to differentiate those who are likely to respond to a certain mailing from those who aren't (McCarty & Hastak, 2007). The CHAID model was superior to a more traditional approach (examining recency, frequency, and amount of customer purchases) because the analysis could occur after a test mailing was sent. Researchers were then able to use the rules that described the terminal node to identify which customers should receive the mailing (McCarty & Hastak, 2007).

Prediction methods, including neural networks and CHAID, have helped to elevate marketing from a product-centered to a customer-centered view (Rygielski et al., 2002). Since it is extremely costly to acquire new customers, marketers are focusing on engaging with their current customer base. Retailers are keeping better databases of customers and by identifying priorities based on particular behavior (i.e., preference for generic or designer brands), they are able to develop more cost-effective promotional campaigns. In Rygielski and colleagues' work (2002), a CHAID model was used to identify current mortgage customers who may be interested in home equity loans. The model identified 16 segments of customers, in which the results of the segmentation became an independent variable in a logistic regression to predict the probability of a customer responding to a promotion. The model demonstrated a high correlation between the actual and predicted response rate, which in turn allowed the company to increase the actual response rate by 30%, and in turn, increase sales by $36 million (Rygielski et al., 2002).

Recent literature has demonstrated that latent class analysis and k-means clustering are two very successful segmentation methods for uncovering previously unknown or complex relationships in survey and response data (Patterson et al, 2002, Lanza et al., 2007). These methods

are particularly attractive for researchers interested in classifying groups of individuals, customers, or patients for decision-making, risk stratification, and marketing purposes. Likewise, segmentation prediction models such as CHAID and neural networks have shown to be increasingly superior to regression-type models for identifying best customers, or those most likely to respond in a desirable way (Kumar et al., 1995, Rygielski et al., 2002). Utilization of these models has the potential to increase customer or patient retention, tailor messaging or marketing strategies, and save money or funding by focusing efforts on a narrower group of individuals.

2.4 DETERMINATION OF BEST SEGMENT CLASS SOLUTION

Perhaps the greatest challenge of segmentation analysis, is determining the "correct" number of classes. Some models, such as latent class, provide information criteria (BIC/CAIC). Bhathagar et al. chose a three-segment model to their e-shoppers classification problem, using the smallest BIC and CAIC values (2004). The choice of a cluster solution is often a "trade-off" between the marketing plan and data quality (Vellido et al., 2009). Choosing a segment solution often involves a degree of "story telling" or choosing a solution that makes sense, and provides meaningful, interesting information to a marketing manager. In addition to reviewing information criteria, researchers have used visualization techniques to determine the appropriate number of clusters. A "quality score" plot, which is similar to a scree plot in factor analysis can be used to identify the optimal number of clusters (the slope of the line goes flat as clusters become irrelevant).

Methods for identifying the number of segment classes are dependent on the analysis paradigm. Kuo and colleagues (2002) used the K-means algorithm to classify customers who shop at a "3C" store or stores that focus on computers, communication, and consumer electronics.

Attitudinal data was collected via interviews and surveys, and the results were analyzed using three variations of the K-means algorithm. In order to compare models, Wilk's lambda and cross-validation using the confusion matrix were examined (Kuo et al., 2002). Wilk's lambda is equal to the proportion of the total variance in the discriminant scores not explained by differences among the groups. Smaller values of Wilks' lambda indicate greater discriminatory ability of the function. Cross-validation is executed by partitioning sample data into a training (or model-building) set, which is used to develop the model, and a validation (or prediction) set, which is used to evaluate the predictive ability of the model. By comparing the model fit to the training set and the model refit to the validation set can provide a measure of consistency. In the Kuo example, the solution that had the smallest Wilk's lambda was considered to be the best. In latent class analysis, BIC (Bayesian information criteria) is used to evaluate model fit. Typically, the preferred model is chosen in the case in which the BIC is minimized (Bhatnagar et al., 2004). Neural networks also have model fit information criteria associated with their performance. For example, in Kumar's analysis of supermarket decisions (whether or not to carry particular products), neural networks and logistic regression methods were employed to identify segments. In this case, the groups represented factors affecting the supermarket decisions, such as financial, marketing, competition, and potential growth (Kumar et al., 1995). Results indicated that the neural network performed better than logistic regression, as it had a smaller residual mean square and a larger C-Index (computing concordant pairs; values closest to 1 indicate perfect fit) (Kumar et al., 1995).

With respect to choosing the best segment solution, there is no single criterion value (i.e., BIC) and some level of "visual inspection" occurs. For example, Bhatnagar et al. examined meaningful differences on demographic variables in their latent class analysis of e-shoppers (2004). The researchers began with a 1-segment model, and then proceeded to a 2, 3, and 4 segment

solution, stopping at the point in which the additional segment parameters did not result in a significant increase in model fit (Bhatnagar et al., 2004). Bucklin and Gupta (1992) estimate a nested multinomial logit-mixture model to identify response segments in a market research study of household goods. To relate segment membership to observable demographic characteristics, they regress the posterior segment membership probabilities on those demographic variables. Their exploratory analysis appears to be intended primarily to characterize the segments on the basis of demographics rather than for purposes of classifying households into response segments. In contrast, the methodology proposed here in which a priori segment membership probabilities are made functions of demographic variables, can be used for both characterizing segments and assigning households to preference and response segments on the basis of their demographic characteristics

Examining the model's classification accuracy is another way of validating its performance. Many researchers suggest splitting the data into training and test sets and examining the percentage of accurately classified segments (Kumar et al., 1995, Kuo et al., 2002, McCarty & Hastak, 2007). Classification tables have a unique advantage of comparing classification rates across methods, as no particular statistic unique to a specific method is required. Often, classification tables are employed by using a cutoff value of 0.5 on multiple criteria: percent of accepts (or "yes" responders) correctly classified, percent of rejects (or "no" responders) correctly classified, and overall percent of correct classifications (Kumar et al., 1995). For example, Kumar and colleagues (1995) used a classification table to compare results across a neural net and logistic regression model in their supermarket analysis. The data showed that the neural net was superior at classification, as it improved classification rates over logistic regression by more than 17%.

2.5 BOOSTING AND BAGGING ALGORITHMS

Prediction accuracy of neural networks and decision tree models can be improved upon by utilizing various classification algorithms. Boosting and bagging are two of the most popular methods for creating an ensemble of decision tree classifiers by manipulating the training data given to a "base" learning algorithm (Dietterich, 1999). Both of these methods function by taking a base learning algorithm and invoking it over and over with different training sets. The main difference is that bagging algorithms do no actively change the distribution of the training set, whereas boosting methods do, based on the performance of previous classifiers (Bauer & Kohavi, 1999). While both methods have been shown to be superior to random forest (decision trees are grown by random sampling with replacement from training data), there are cases in which one method is preferred over the other (Zhou & Tang, 2002, West, Dellana, & Qian, 2004, Banfield, Hall, Bowyer, & Kegelmeyer, 2007, Bauer & Kohavi, 1999).

2.5.1 Boosting

Boosting is a re-sampling classification strategy, with a probability distribution that is dependent on the misclassification rate for each observation (Zhou & Tang, 2002). Boosting employs an iterative algorithm that constructs an ensemble by sequentially training each ensemble member with unique training sets that increase the prominence of certain hard-to-learn examples misclassified by earlier ensemble members. Boosting maintains a probability distribution, $D_t(t)$, over the original data available for training. In each iteration, a classifier(t) is trained by sampling with replacement from this distribution. After training and testing, the probability of incorrectly classified training examples is increased and the probability of correctly classified examples is

decreased. The ensemble decision is obtained by a weighted vote of all ensemble members. Schwenk and Bengio applied adaptive boosting methods to neural network ensembles and report that boosting can significantly improve neural network classifiers (West et al., 2004). They conclude that boosting is always superior to bagging, although the differences are not always significant

In the mid-90's, machine-learning researchers (namely, Freund and Schapire) formally introduced a boosting algorithm for incremental refinement of an ensemble by emphasizing hard-to-classify data examples. This algorithm, referred to as AdaBoost, creates classifiers using a training set with weights assigned to every example (Banfield, et al, 2007). Examples that are incorrectly classified by a classifier are given an increased weight for the next iteration. Freund and Schapire showed that boosting was often more accurate than bagging when using a nearest neighbor algorithm as the base classifier, though this margin was significantly diminished when using CHAID form of decision trees. Results were reported for 27 datasets, comparing the performance of boosting with that of bagging using CHAID as the base classifier. The same ensemble size of 100 was used for boosting and bagging. In general, 10-fold cross-validation was done, repeated for 10 trials, and average error rate reported. For datasets with a defined test set, an average of 20 trials was used with this test set. Boosting resulted in higher accuracy than bagging on 13 of the 27 datasets, bagging resulted in higher accuracy than boosting on 10 datasets, and there were 4 ties (Banfiled et al., 2007). The differences in accuracy were not evaluated for statistical significance.

2.5.2 **Bagging**

Bagging (or bootstrap aggregating) was first made popular in machine learning research by Breiman in 1996 and Zhang in 1999 (Bauer and Kohavi, 1999, West et al., 2005). According to Breiman's methods, a bagging ensemble is formed by perturbing the training data, creating a unique training set for each ensemble member by sampling with replacement over a uniform probability distribution on the original data (West et al., 2005). Bagging creates an ensemble of classifiers by sampling with replacement from the set of training data to create new training sets called "bags," where the number of items in each bag is the same as the number of items in the set of training data and a separate classifier is trained from each bag. Each classifier's training set is generated by randomly drawing (with replacement) N samples, where N is the size of the original training set (Maclin & Opitz, 1997). Unlike boosting, bagging generates diverse classifiers only if the base learning algorithm is unstable, where small changes to the training data set create large changes to the learned classifier (Dietterich, 1999).

Bagging operates using bootstrap sampling. Given a training data set D containing m examples, bootstrap sampling draws a sample of training examples, D_i, by selecting m examples uniformly at random with replacement from D. The replacement means that examples may be repeated in D_i. Bagging then creates k bootstrap samples and trains k classifier on each bootstrap sample. A new instance is classified by taking a weighted majority of the k learned classifiers (using equal weights). The result is an ensemble of classifiers.

In his early work, Breiman investigated bootstrap replicates to create diverse learning sets for classification trees and tested them on both real and simulated data sets (West et al., 2005). He reported a reduction in test set misclassifications (comparing the "bagging estimator" to the "single best" estimator) ranging from 6% to 77% and concluded that a vital element for the success of

bagging is the instability of the estimators. If disturbing the learning set can cause a significant change in the predictor constructed, then bagging can improve the generalization accuracy. For classification problems, Breiman demonstrates that if a model's prediction is "order-correct" for most inputs, then an aggregated predictor or bagging model can be transformed into a nearly optimal predictor (West et al., 2005). Using an approach similar to Breiman's work, Zhang (1999) implemented bagging with an ensemble of thirty multiplayer perceptron neural network models. He used learning sets created by bootstrap replicates and found that the bagging estimator is more accurate and more robust than the "single best" neural network.

One advantage of bagging is the ability to test the accuracy of the ensemble without removing data from the training set, as is done with a validation set. Breiman referred to the error observed when testing each classifier on examples not in its bag as the "out-of-bag" error, and suggested that it might be possible to stop building classifiers once this error no longer decreases as more classifiers are added to the ensemble. The effectiveness of this technique has not yet been fully explored in the literature.

2.5.3 Summary

Since much of the literature and research papers on ensemble classifiers compare "newer" methods (such as boosting) with bagging and report improvement, it might be expected that one boosting could be considered the "winning" method. This was not the case when the results are looked at in terms of statistically significant increases in accuracy on individual datasets. Banfield and colleagues (2007) experimentally evaluated bagging and other randomization-based approaches to creating ensembles for decision trees. Of the 57 data sets considered, 37 showed no statistically significant improvement over bagging for any of the other techniques, using either the 10-fold or

5x2 cross-validation. Results indicated that while the gain over bagging is often small, there is a consistent pattern of gain. In addition, boosting-by-resampling resulted in better accuracy with a much larger ensemble size than has generally been used, and that at this larger ensemble size it does offer some performance advantage over bagging. Of note, the increase in accuracy is statistically significant in only a fraction of the data sets used (Banfield et al., 2007).

West and colleagues (2005) examined several ensemble strategies in an effort to improve credit scoring and bankruptcy prediction models. Their aggregate analysis did not find any significant difference in accuracy between the perturbation strategies, yet each method investigated achieved a statistically significant reduction in error in at least one application. The bagging strategy was most effective for the Australian credit and the bankruptcy data set, both characterized by smaller training samples, fewer feature variables, and less noise. The boosting strategy was e1ective only for the bankruptcy data, the smallest data set with the fewest number of feature variables and the least amount of noise (West et al., 2005). Further investigation regarding a preferred ensemble method would only enhance the ambivalent literature.

2.6 OUTCOME VARIABLES

Outcome variables for segmentation models are dependent upon the underlying method; classification or prediction. Classification models do not support the inclusion of a dependent variable, thus the “outcome” for comparison and assessment of this class of models includes the probability of segment class membership. Segmentation methods for prediction can accommodate both continuous and categorical outcomes. Since research examining differences in the use of categorical vs. continuous variables is inconsitent, dependent variables of both levels will be

considered in this study (Beauducel & Herzberg, 2006; Dolan, Muthe´n et al., 1997; Nussbeck, Eid, & Lischetzke, 2006).

2.6.1 Continuous vs. categorization

In many areas of research, continuous variables are often converted into categorical variables by grouping response values into ≥ 2 categories. A common case for categorization is when the continuous variable may have only a few values and should be regarded as ordinal rather than continuous. For example, the continuous variable of household income, which can have an extremely large range due to outliers, may be categorized into 3 groups: low, medium, high. Category cut-offs are defined by using the continuous value to determine the appropriate category for each measurement, where the proportion of observations in the categories must add up to 1. Treating the variable as continuous allows for an estimation of the linear component of the relationship, while the categorical version is able to explain much more complicated relationships (Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Several studies that have used continuous maximum likelihood estimation methods have resulted in underestimated parameter estimates when the number of categories is very small (e.g., two to three). Accordingly, this bias is often minimized as the number of categories increases, and ML parameter estimates approach higher accuaracy with a minimum of 4 or 5 categories (Beauducel & Herzberg, 2006; Dolan, Muthe´n et al., 1997; Nussbeck, Eid, & Lischetzke, 2006). Beauducel and Herzberg's 2006 study explicitly compared compared continuous and categorical methodologies, whereas results indicated higher accuarcy and performace with categorical methodology over and above continuous methods for parameter estimation.

On the other hand, categorization of continuously distributed variables has often been associated with various problems, including, but not limited to: multiple hypothesis testing with pairwise comparisons of categories, assumptions of homogeneity within groups, and issues of generalization and comparison of results due to the specific cut points used to define categories (Bennette & Vickers, 2012). Several researchers have cautioned that dichotomization leads to a substantial loss of power and inadequate correction for confounding factors (Bennette & Vickers, 2012; Naggara, Raymond, Guilbert, Roy, Weill, & Altman, 2011; Fedorov, Mannino, & Zhang, 2009). Fedorov et al. (2009) suggested that 100 continuous observations are statistically equivalent to at least 157 dichotomized observations. In addition, Becher and colleagues (1999) found that models with a categorized exposure variable removed only 67% of the confounding controlled when the continuous version of the variable was used. It has been reported in biomedical research where a focus is on risk of disease, dichotomization can lead to pooling groups with patients of different risk probabilites, leading to a high propensity for missclassification (Naggara et al., 2011).

While there has been much discussion over the potential pitfalls of categorizing continuous data, the ease of interpretation and decision rules have made the procedure attractive to some researchers (Irwin & McClelland, 2003). In particular, when communicating results to a non-statistically oriented audience (i.e., marketing campaign managers), the use of categorized variables can be very effective at presenting information visually, and in an easily digestible way. In addition, there are several situations in which the categorization of continuous data is advantageous (Cohen, 1983). If there are privacy or anonymity concerns about the continuous data in which a unique individual might be identified given the original data, categorization of age and/or level of education can protect an individual's anonymity. In addition, categorization of a

continuous dependent variable has shown to be considerably successful with high-dimensional (many attributes) data of small sample sizes (Leon, Soo, & Williamson, 2011). Finally, if the dependent variable is not linear, categories can be used to separate the data into linear segments for modeling rather than as one continuous model.

2.6.2 Probability of segment class membership

For segmentation methods involving classification problems, there is no observable dependent variable. In these cases, the outcome is the probability of segment class membership, which will be assessed using latent class analysis and k-means clustering. As previously discussed, the latent class estimation process works as follows:

1. Initially, select random estimates of each group's utility or part-worth, which represent the probability of belonging to each segment class
2. Use each group's estimated utilities to fit each respondent's data, and estimate the relative probability of each respondent belonging to each group.
3. Using those probabilities as weights, re-estimate the logit weights for each group.
4. Accumulate the log-likelihood over all groups.
5. Continue repeating steps 2 - 4 until the log-likelihood fails to improve by more than some small amount (the convergence limit).

Latent class reports the part worth/utility for each and every subgroup or segment, for a given individual. Latent class analysis does not assume that each respondent is "in" one group or another. Rather, each respondent is considered to have some non-zero probability of belonging to

each group. If the solution fits the data very well, then those probabilities approach zero or one (Gupta, S., & Chintagunta, 1994).

In the context of a multinomial logit-mixture model being used to explain the probability of a household belonging to a particular segment by that household's demographic characteristics, Dayton and Macready (1988) provided a type of latent class model in which the probability of latent class membership is functionally related to concomitant variables. These concomitant variables include demographic and other consumer-specific characteristics that assume known, fixed values. The relationship between the concomitant variable latent class models (Dayton and Macready 1988) and the multinomial logit-mixture model (Kamakura and Russell 1989) is the following: if the logit probability of segment membership in the multinomial logit-mixture model is made an explicit function of concomitant (demographic) variables, and if the latent class probabilities and the membership probabilities of the Dayton and Macready (1988) model are logit, then the resulting models are equivalent.

3.0 METHODOLOGY

This study is designed to compare and contrast segment class groupings and model performance by segmentation purpose and methods using the Medicare Health Outcomes Survey (HOS) as the consistent data set. The study will occur in two phases, as demonstrated in Figure 4. First, the distributions and correlations of the response variables within the Medicare HOS will be explored. Missing data will be examined and appropriate imputation procedures will be used when necessary. Next, outcomes of the four segmentation models will be assessed and compared within the classification or prediction schema. The segmentation models will be evaluated according to model fit/accuracy, the number of segment classes yielded, differentiating segment characteristics, level of dependent variable (continuous vs. categorical) and classifying algorithm for prediction model, and the variation of a continuous vs. categorical scale for classification models. In addition, variable importance for each segmentation model will be examined. The segment classes will be interpreted according to profiling variables (not used in the model) and variables included in the final models.

The primary goal of this study is to compare the most commonly used segmentation prediction (neural networks and CHAID) and classification (LCA and k-means) methods in healthcare marketing applications, while utilizing a single dataset. This chapter is divided into 4 sections; shared variables and procedures, classification methods, prediction methods, and summarization of results. The shared variables and procedures section presents a description of the Medicare HOS instrument and response variables, an overview of the common independent variables, a description of data handlings (such as the treatment of missing data) and an overview of common procedures such as segment class interpretations. The classification and prediction

models sections outline the preparation and construction of data unique to each model and present the specific model building steps and assessment of output. The final section provides a framework for summarizing results.

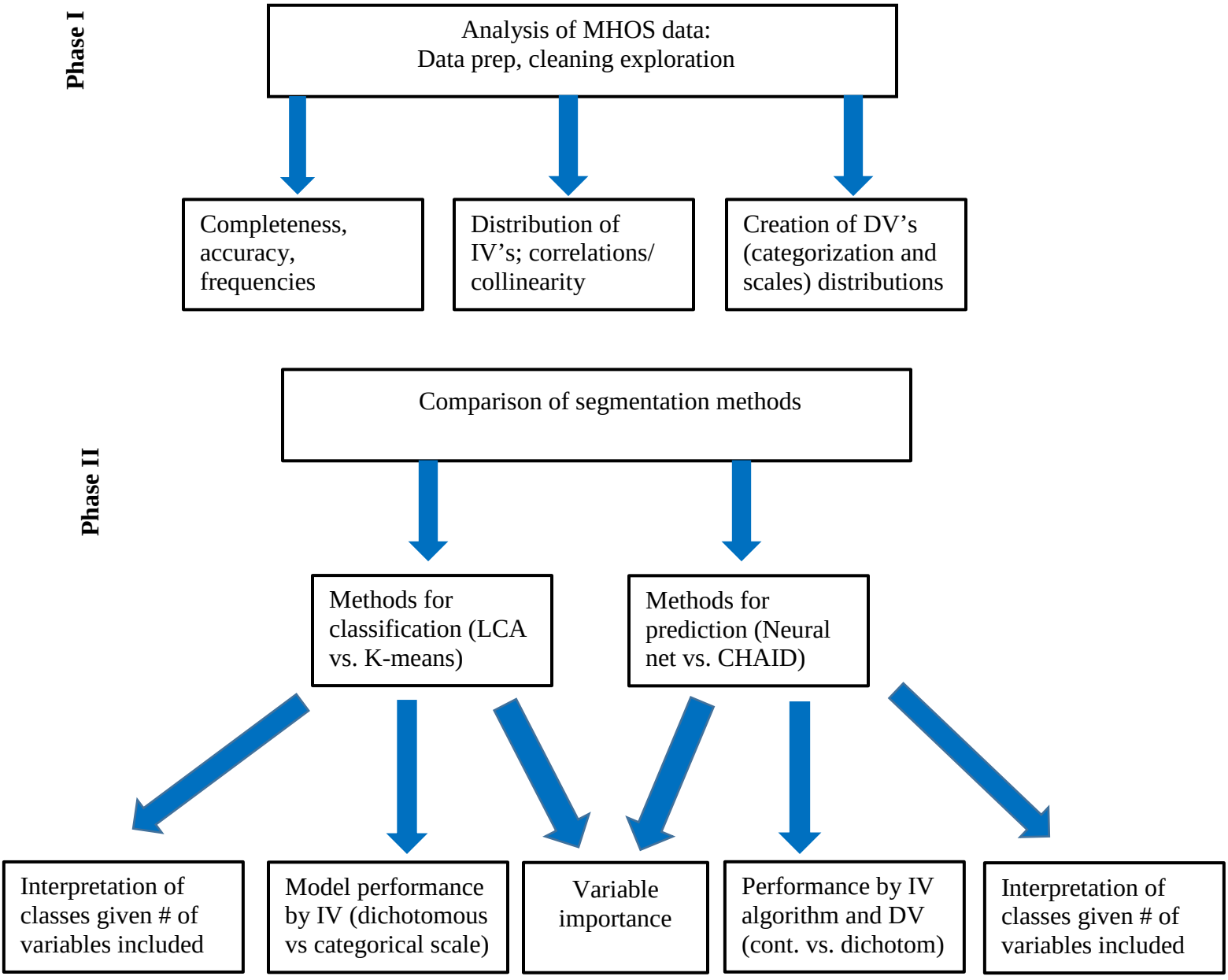


Figure 4. Phases of Study Execution

3.1 SHARED VARIABLES AND PROCEDURES

The following section will provide a detailed description of the data and procedures common to both classification and prediction methods.

3.1.1 Data

3.1.1.1 Medicare Health Outcomes Survey instrument version 3.0

The Medicare Health Outcomes Survey (HOS) is a 68-item instrument designed to measure health outcomes of Medicare recipients, across managed care plans with Medicare Advantage contracts (www.hosonline.org). The survey is comprised of several components: the Veterans RAND 12-Item Health Survey, demographics and disability status, questions designed to address HEDIS® Effectiveness of Care measures, and additional questions to measure physical and mental health functioning and case-mix and risk-adjustment. Case-mix adjustments were used so that all Medicare Advantage Organizations (MAO's) were as comparable as possible in terms of socio-demographic characteristics (age, gender, race, etc.), chronic conditions, baseline health status, and other design variables (Haffer & Bowen, 2004). All beneficiaries who completed the HOS at baseline and had a baseline physical-component (PCS) or mental component (MCS) score were included in the analysis of death outcomes. Beneficiaries age 65 or older who completed the HOS at baseline and follow up, and for whom PCS and MCS scores could be computed at both time points, were included in the analysis of PCS and MCS outcomes. Case-mix variables of demographics and health as well as selected survey design variables are risk adjusted to make equitable health outcome comparisons across MAOs. Risk-adjustment is a statistical technique that adjusts for variations in patient outcomes that stem from differences in existing patient

characteristics rather than differences in performance between MAOs. The risk-adjusted outcomes are aggregated for the respondents in each MAO, and yield the MAO level performance measurement results (Haffer & Bowen, 2004). The survey has been deployed annually since 1998, with a random sample (1000-1200) of Medicare beneficiaries from Medicare Advantage plans with 500+ members selected to participate. Follow-up measures are completed every two years on baseline respondents (i.e. Cohort 1 was surveyed in 1998 and resurveyed in 2000; Cohort 19 was surveyed in 2016 and will be resurveyed in 2018). For the purpose of this study, data from Cohort 15 (initially surveyed in 2012; re-surveyed in 2014) will be analyzed, and only baseline responses will be considered, due to the decrease in sample size at follow-up (Table 1). Beneficiaries were eligible for re-measurement if they had sufficient data to derive physical health or mental health scores at baseline and remained in the same plan at follow-up. Surveys were administered by mail and via phone in those instances when beneficiaries failed to respond to the second mailing or returned an incomplete survey. A copy of the instrument can be found in Appendix A, and a description of the available respondent demographics can be found in Table 2.

Table 1. Survey Results

Cohort	Date Fielded	Reporting Units (MAOs)	Sample Size	Ineligible Surveys	Completed Surveys	Response Rate	% 79.5-100% Complete Cases
Baseline 15	April 2012	506	591,823	13,602	297,974	51.5%	87.2%
Follow-up 15	April 2014	421	147,235	1,295	105,432	72.2%	78.4%

Final data file: 296,320 eligible surveys

Table 2. Demographics of Respondents

Variable	*Frequency*
Age: Less than 65	14.94%
Age: 65-74	50.61%
Age: 75 and older	34.45%
Race: White	80.56%
Race: AA	11.6%
Race: Other	7.84%
Female	57.55%
Married	51.12%
Less than high school education	24.08%
High school education or GED	34.71%
Greater than high school education	42.21%
Obese	32.61%
Overall Health: Excellent	6.29%
Overall Health: Very good	23.13%
Overall Health: Good	35.85%
Overall Health: Fair	26.03%
Overall Health: Poor	8.7%

3.1.1.2 Input variables

Medicare HOS original variables

A description of the variables included in the Medicare HOS can be found in Table 3. Input variables are indicated by an "I;" dependent variables by a "DV;" profile variables by a "P." Certain variables, such as demographic variables, have been chosen to be profile variables (opposed to inputs) due to their inherit nature to differentiate. For example, it could be expected that respondents' health status differs by age, and age would automatically create classes of respondents in the data. Profile variables aren't as useful as attitudinal and behavioral inputs in uncovering interesting differences in the segment classes, and are therefore used only to describe and type the segment groups. Also included in Table 3 are the proportions of missing responses by each variable. For prediction models (neural networks and CHAID), all available inputs will be used as independent variables in the analysis, provided the level of missingness is not greater than

15%. Variables with more than 15% missing response data will be excluded from the analysis. For classification models (LCA and k-means), the full set of input variables will be used to identify classes of similar response patterns in the data.

Table 3. Description of Medicare HOS Variables

Field #	*Variable*	*Description*	*Role*	*%Missing*
1	CASE_ID	ID number		
2	AGE	Age category	P	0
3	RACE	Race category	P	13.9
4	GENDER	Gender	P	3.8
5	MRSTAT	Marriage status	P	10.8
6	EDUC	Education level	P	11.3
7	BMICAT	BMI category	P	13.2
8	C15VRGENHTH	General health status	P	6.2
9	C15VRMACT	Moderate activities limited	I	7.4
10	C15VRSTAIR	Climbing stairs limited	I	9.8
11	C15VRPACCL	Accomplishing less due to physical health	I	7.9
12	C15VRPWORK	Work or activities limited by physical health	I	9.9
13	C15VRMACCL	Accomplishing less due to emotional health	I	7.9
14	C15VRMWORK	Work or activities limited by emotional health	I	10.3
15	C15VRPAIN	Degree pain interfered with normal work	I	8.1
16	C15VRCALM	Felt calm and peaceful	I	8.3
17	C15VRENERGY	Energy level	I	8.8
18	C15VRDOWN	Felt downhearted and blue	I	9.3
19	C15VRSACT	Amount of time health interfering with social activities	I	8.3
20	C15VRPHCMP	Physical Health compared to 1 Year Ago	I	8.2
21	C15VRMHCMP	Emotional Health compared to 1 Year Ago	I	9.4
22	C15ADLBTH	Difficulty Bathing	I	8.9
23	C15ADLDRS	Difficulty Dressing	I	8.9
24	C15ADLEAT	Difficulty Eating	I	9.1
25	C15ADLCHR	Difficulty Getting in or out of Chairs	I	9.1
26	C15ADLWLK	Difficulty Walking	I	9.1
27	C15ADLTLT	Difficulty Using Toilet	I	9.1
28	C15HDPHY	Number of Days Physical Health Not Good	DV	11.9
29	C15HDMEN	Number of Days Mental Health Not Good	I	11.5
30	C15HDACT	Number of Days Poor Health interfered w/activities	I	18.8
31	C15CHSTEX	Chest Pain-Exercise	I	10.1
32	C15CHSTRST	Chest Pain-Resting	I	9.9
33	C15SOBFLT	Short of Breath lying flat	I	10.1
34	C15SOBSIT	Short of Breath sitting or resting	I	10.4

Table 3 continued

35	C15SOBWLK	Short of Breath walking less than 1 block	I	10.7
36	C15SOBSTR	Short of Breath climbing 1 flight stairs	I	11.3
37	C15FTNUMB	Numbness or Loss of feeling in feet	I	9.8
38	C15FTSENS	Tingling burning in feet	I	9.9
39	C15FTHC	Decreased feeling of hot or cold in feet	I	10.4
40	C15FTSRS	Sores that do not heal on feet	I	10.1
41	C15PNART	Arthritis pain	I	10.6
42	C15READ	Can see to read newspaper	I	9.9
43	C15HEAR	Can hear most things	I	12.0
44	C15CCHBP	Hypertension or High Blood Pressure	I	9.7
45	C15CC_CAD	Angina Pectoris or Coronary Artery Disease	I	11.0
46	C15CC_CHF	Congestive Heart Failure	I	10.6
47	C15CCMI	Myocardial Infarction or Heart Attack	I	10.4
48	C15CCHRTOTH	Other Heart Conditions	I	10.6
49	C15CCSTROKE	Stroke	I	10.1
50	C15CC_COPD	Emphysema	I	10.2
51	C15CCGI	Inflammatory Bowel Diseases	I	10.6
52	C15CCARTHIP	Arthritis of hip or knee	I	10.4
53	C15CCARTHND	Arthritis of hand or wrist	I	10.5
54	C15CCOSTEO	Osteoporosis	I	10.9
55	C15CCSCIATI	Sciatica	I	10.7
56	C15CCDIABET	Diabetes	I	10.0
57	C15CCANYCA	Any Cancer (other than skin cancer)	I	9.8
58	C15CACOLON	Under Treatment for Colon Cancer	I	60.4
59	C15CALUNG	Under Treatment for Lung Cancer	I	60.8
60	C15CABRST	Under Treatment for Breast Cancer	I	60.9
61	C15CAPROS	Under Treatment for Prostate Cancer	I	62.2
62	C15PNBACK	Back Pain Interfered w/Activities in Past 4 Weeks	I	10.1
63	C15DEP2WK	Sad/Blue for Two + Weeks in Past Year	I	10.2
64	C15DEPYR	Depressed for Much of Past Year	I	10.3
65	C15DEP2YR	Depressed for Two + Years in Life	I	10.8
66	C15DEPWEEK	Depressed for How Much of the time in Past Week	I	10.4
67	C15CMPHTH	General Health compared to peers	I	10.1
68	C15SMOKE	Smoke every day	I	9.9
69	C15MUILKG	Urine Leakage in Past 6 Months	I	11.3
70	C15MUIMAG	Magnitude of Urine Leakage Problem	I	54.4
71	C15MUITLK	Talked with Doctor About Urine Leakage	I	59.5
72	C15MUITRT	Received Treatment for Urine Leakage	I	60.0
73	C15PAOTLK	Talked with Doctor About Physical Activities	I	12.3
74	C15PAOADV	Advised to Increase or Maintain Activities	I	13.7
75	C15FRMTLK	Talked with Doctor about Falling or Balance Problem	I	11.1
76	C15FRMFALL	Fell in Past 12 Months	I	10.1
77	C15FRMBAL	Problem with Walking or Balance in Past 12 Months	I	10.5

Table 3 continued

78	C15FRMPREV	Talked with Doctor about How to Prevent Falls	I	12.5
79	C15OTOTEST	Had a Bone Density Test for Osteoporosis	I	11.5
80	C15CMPWHO	Who completed Survey	P	3.8
81	C15SRVDISP	Baseline Survey Disposition	P	0
82	C15SRVMODE	Baseline Survey Round	P	0
83	C15PCTCMP	Baseline Percent of Survey Completed	P	0
84	C15SRVLANG	Baseline Survey Language	P	0
85	R15VRGENHTH	General health status FOLLOW UP		31.0
86	R15VRMACT	Moderate activities limited FOLLOW UP		31.6
87	R15VRSTAIR	Climbing stairs limited FOLLOW UP		33.4
88	R15VRPACCL	Accomplishing less due to physical health FOLLOW UP		31.9
89	R15VRPWORK	Work or activities limited by physical health FOLLOW UP		33.7
90	R15VRMACCL	Accomplishing less due to emotional health FOLLOW UP		31.8
91	R15VRMWORK	Work or activities limited by emotional health FOLLOW UP		33.7
92	R15VRPAIN	Degree pain interfered with normal work FOLLOW UP		31.8
93	R15VRCALM	Felt calm and peaceful FOLLOW UP		31.8
94	R15VRENERGY	Energy level FOLLOW UP		32.1
95	R15VRDOWN	Felt downhearted and blue FOLLOW UP		32.4
96	R15VRSACT	Amount of time health interfering with social activities FOLLOW UP		31.8
97	R15VRPHCMP	Physical Health compared to 1 Year Ago FOLLOW UP		31.7
98	R15VRMHCMP	Emotional Health compared to 1 Year Ago FOLLOW UP		32.4
99	R15ADLBTH	Difficulty Bathing FOLLOW UP		31.9
100	R15ADLDRS	Difficulty Dressing FOLLOW UP		32.0
101	R15ADLEAT	Difficulty Eating FOLLOW UP		32.1
102	R15ADLCHR	Difficulty Getting in or out of Chairs FOLLOW UP		32.1
103	R15ADLWLK	Difficulty Walking FOLLOW UP		32.3
104	R15ADLTLT	Difficulty Using Toilet FOLLOW UP		32.1
105	R15DIFMEALS	Difficulty Preparing Meals FOLLOW UP		32.1
106	R15DIFMONEY	Difficulty Managing Money FOLLOW UP		32.1
107	R15DIFMEDS	Difficulty Taking Medication as Prescribed FOLLOW UP		32.2
108	R15HDPHY	Number of Days Physical Health Not Good FOLLOW UP		34.7
109	R15HDMEN	Number of Days Mental Health Not Good FOLLOW UP		34.6
110	R15HDACT	Number of Days Poor Health interfered w/activities FOLLOW UP		38.7

Table 3 continued

111	R15DIFSEE	Blind or Serious Difficulty Seeing FOLLOW UP	32.2
112	R15DIFHEAR	Deaf or Serious Difficulty Hearing FOLLOW UP	32.3
113	R15DIFREMEM	Difficulty concentrating, remembering, or making decisions FOLLOW UP	32.5
114	R15DIFERRND	Difficulty doing errands FOLLOW UP	32.5
115	R15DIFMPROB	Memory problems interfered with activities in past month FOLLOW UP	33.4
116	R15CCHBP	Hypertension or High Blood Pressure FOLLOW UP	32.3
117	R15CC_CAD	Angina Pectoris or Coronary Artery Disease FOLLOW UP	33.1
118	R15CC_CHF	Congestive Heart Failure FOLLOW UP	32.9
119	R15CCMI	Myocardial Infarction or Heart Attack FOLLOW UP	32.9
120	R15CCHRTOTH	Other Heart Conditions FOLLOW UP	32.9
121	R15CCSTROKE	Stroke FOLLOW UP	32.6
122	R15CC_COPD	Emphysema FOLLOW UP	32.6
123	R15CCGI	Inflammatory Bowel Diseases FOLLOW UP	32.8
124	R15CCARTHIP	Arthritis of hip or knee FOLLOW UP	32.7
125	R15CCARTHND	Arthritis of hand or wrist FOLLOW UP	32.9
126	R15CCOSTEO	Osteoporosis FOLLOW UP	33.0
127	R15CCSCIATI	Sciatica FOLLOW UP	33.0
128	R15CCDIABET	Diabetes FOLLOW UP	32.5
129	R15CCDEP	Depression FOLLOW UP	32.9
130	R15CCANYCA	Any Cancer (other than skin cancer) FOLLOW UP	32.9
131	R15CACOLON	Under Treatment for Colon Cancer FOLLOW UP	67.1
132	R15CALUNG	Under Treatment for Lung Cancer FOLLOW UP	67.5
133	R15CABRST	Under Treatment for Breast Cancer FOLLOW UP	67.5
134	R15CAPROS	Under Treatment for Prostate Cancer FOLLOW UP	68.5
135	R15CAOTHER	Under Treatment for Other Cancer FOLLOW UP	67.5
136	R15PAINDACT	Pain interfered with activities in past 7 days FOLLOW UP	32.8
137	R15PAINSACT	Pain kept you from socializing in past 7 days FOLLOW UP	32.8
138	R15PAINRATE	Average pain rating in past 7 days FOLLOW UP	33.6
139	R15DEPNOPLS	Little interest or pleasure in doing things in past 2 weeks FOLLOW UP	33.5
140	R15DEPDOWN	Feeling down, depressed, or hopeless in past 2 weeks FOLLOW UP	33.7
141	R15CMPHTH	General Health compared to peers FOLLOW UP	32.4
142	R15SMOKE	Smoke every day FOLLOW UP	32.3
143	R15MUILKG	Urine Leakage in Past 6 Months FOLLOW UP	33.2
144	R15MUIMAG	Magnitude of Urine Leakage Problem FOLLOW UP	65.3
145	R15MUITLK	Talked with Doctor About Urine Leakage FOLLOW UP	69.2

Table 3 continued

146	R15MUITRT	Received Treatment for Urine Leakage FOLLOW UP		69.2
147	R15PAOTLK	Talked with Doctor About Physical Activities FOLLOW UP		33.8
148	R15PAOADV	Advised to Increase or Maintain Activities FOLLOW UP		34.9
149	R15FRMTLK	Talked with Doctor about Falling or Balance Problem FOLLOW UP		33.1
150	R15FRMFALL	Fell in Past 12 Months FOLLOW UP		32.5
151	R15FRMBAL	Problem with Walking or Balance in Past 12 Months FOLLOW UP		32.6
152	R15FRMPREV	Talked with Doctor about How to Prevent Falls FOLLOW UP		33.8
153	R15OTOTEST	Had a Bone Density Test for Osteoporosis FOLLOW UP		33.3
154	R15CMPWHO	Who completed Survey FOLLOW UP		30.5
155	R15SRVDISP	Follow Up Survey Disposition		0
156	R15SRVMODE	Follow Up Survey Round		0
157	R15PCTCMP	Follow Up Percent of Survey Completed		0
158	R15SRVLANG	Follow Up Survey Language		0
159	COHORT	COHORT ID		
160	P15PLREGCDE	Reported Plan CMS Region Code	P	0
161	SAMPLED	Follow Up Sample Indicator		
162	SFLAG	Dead, Disenroll, Inval, Resp, Nonresp (Analytic Sample)	DV	0

A table containing the frequencies of all response variables can be found in Appendix B.

3.1.2 Procedures

3.1.2.1 Preparation of Medicare HOS data

Completeness and accuracy

Response data will be examined according to completeness and accuracy. The percent of survey completed by respondent will be examined, in addition to the proportion of missing values by survey question. The variance in demographics by segment and region, and responses by item will

be calculated and evaluated for tolerance. In addition, the total sampling error will be considered. Of note, the survey vendors attest to the accuracy of their data collection process and its conformance with these Medicare HOS Quality Assurance Guidelines and Technical Specifications V2.0. These technical specifications include decision rules and quality control processes to validate the accuracy of key entry and electronic scanning procedures.

Missing data

Real time processing applications that are highly dependent on the data often suffer from the problem of missing input variables. Various heuristics of missing data imputation such as mean substitution and hot deck imputation also depend on the knowledge of how data points become missing. There are several reasons why the data may be missing, and as a result, missing data may follow an observable pattern. Exploring the pattern is important and may lead to the possibility of identifying cases and variables that affect the missing data. Having identified the variables that predict the pattern, a proper estimation method can be selected.

According to Little and Rubin, and Burk, there are three types of missing data mechanisms. Considering X and Y as random variables, the three categories of missing data are:

1. *Missing Completely at Random (MCAR)* which occurs if the missing value for the input vector has no dependence on any other variable in the database such that the inputs with missing entries are the same as the complete inputs. That is, the probability of data Y, being missing is not dependent on either X or Y , i.e. is not dependent on either missing or complete values in the same record or any other record in the database.
2. *Missing at Random (MAR)* occurs if the missing value for the input vector has dependence on other variables in the data set, such that the pattern in which the data becomes missing

is traceable. That is, the probability of data Y being missing is dependent only on X the existing values in the database and not on any missing data.

3. *Missing Not at Random (MNAR)* occurs when the missing value for the input vector depends on the other missing values such that the existing data in the database cannot be used to approximate the missing values. This is also known as the non-ignorable case. The probability that Y is missing is dependent on the missing data.

Missing data procedures specific to method

LCA

LCA handles missing data on the indicators in order to make use of all available data and it is not necessary to delete cases that have partial data. Missing data are handled with a full-information maximum likelihood (FIML) technique, which assumes that data are missing at random (MAR). However, even when the MAR assumption is not met, this procedure performs better than casewise deletion. Note, however, that cases missing values on one or more covariates or on the grouping variable are removed from the analysis.

Neural Networks

Neural network models within Enterprise Miner uses the expectation-maximization (EM) algorithm to handle missing data. The EM algorithm is a general technique for fitting models to incomplete data. Expectation maximization has been proven to work better than methods such as listwise, pairwise data deletion, and mean substitution because it assumes incomplete cases have data missing at random rather than missing completely at random. Expectation maximization capitalizes on the relationship between missing data and the unknown parameters of a data model. The interdependence between model parameters and missing values suggests an iterative method where missing values based on assumed values for the parameters are predicted first, then the

predictions are used to update the parameter estimates, and the process is repeated. The sequence of parameters converges to maximum-likelihood estimates that implicitly average over the distribution of the missing values.

K-Means

By default, observations with all missing values are excluded from the analysis, and observations with any missing values are excluded by specifying the NOMISS option. The distance between an observation with missing values and a cluster seed is obtained by computing the squared distance based on the nonmissing values, multiplying by the ratio of the number of variables, *n*, to the number of variables having nonmissing values, *m*, and taking the square root. The IMPUTE option fills in missing values in the OUT= output data set.

When there are missing values in columns with simple data types (not nested), k-Means interprets them as missing at random. The algorithm replaces missing categorical values with the mode and missing numerical values with the mean. When there are missing values in nested columns, k-Means interprets them as sparse. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. PROC HP CLUS deals with observations with missing values by scaling the distance obtained from all non-missing variables and requests imputation of missing values after the final assignment of observations to clusters. If an observation that is assigned (or would have been assigned) to a cluster has a missing value for variables used in the cluster analysis, the missing value is replaced by the corresponding value in the cluster seed to which the observation is assigned (or would have been assigned).

CHAID

One of the advantages of decision trees is that they are more efficient at dealing with high dimension data than parametric regression techniques (Westreich 2010). Additionally, they are

able to flexibly deal with missing data. There are two ways that they can deal with missing data. The first way is to treat missing observations as a new category. This will allow the difference between missingness and non-missingness of the variables to be seen. The second way is to construct surrogate variables. For a given split, if the original variable is missing, a surrogate variable that mimics the behavior of the original variable will be used for that split. In Enterprise Miner, there is the option to define more than one surrogate variable. As such they are not limited and can efficiently manage missing data issues.

Variable and computational limits

There are several limitations within the segmentation software methods that need to be considered when performing the analyses. Variable limitations and computational resources will briefly be addressed within each segmentation method, in addition to the number of measurement occasions each method can accommodate.

LCA

The limit on the number of indicators poLCA can handle is 999. However, as indicators are added the size of the contingency table (and often model complexity) increases substantially.

K-Means Clustering (PROC HP CLUS- Computational Resources)

Let n = number of observations, v = number of variables, c = number of clusters, p = number of passes over the data set. The memory required is approximately:

$4(19v + 12cv + 10c + 2\max(c + 1))$ bytes. The overall time required by PROC HP CLUS is roughly proportional to $nvcp$ if c is small with respect to n. Initial seed selection requires one pass over the data set. If the observations are in random order, the time required is roughly proportional to $nvc + vc^2$ unless you specify REPLACE=NONE. In that case, a complete pass might not be necessary,

and the time is roughly proportional to *mvc*, where $c \leq m \leq n$. For greatest efficiency, the variables in the VAR statement should be listed in order of decreasing variance.

CHAID

In order to eliminate irrelevant variables, an initial variable selection is made. The initial variable selection results in a reduction of the size of the input set to a manageable number. The tree algorithm is then applied to the selected inputs one at a time, and for each input the resulting SAS code is saved. The SAS code shows the ranges into which the input is split and also gives the mean of the target for each interval of the input. In the terminology of a tree, these input ranges are also called leaf nodes or simply leaves of the tree. Inputs that do not produce meaningful trees are eliminated. The decision tree application within enterprise miner uses a sample of at most 20,000 observations. The sample is used to prevent the excessive time and memory consumption that can occur with large data sets. However, it is possible to override the default sample size.

Neural Networks

Using a smaller number of inputs can greatly reduce the time required to train the network, as well as improving the prediction results. If there is prior knowledge that an input (predictor) is not useful in predicting the target, then it should be excluded. The explore window and the multiplot node in Enterprise Miner can be used to create exploratory plots to identify important inputs. When many inputs are available, the choice of network architecture is especially important. For example, multi-layer perceptrons (MLPs) tend to be better at ignoring irrelevant inputs than are some radial basis function (RBF) networks. Having many inputs also reduces the number of hidden inputs that can be used, since the number of weights connecting an input layer and a hidden layer is equal to the product of the number of units in each.

3.1.2.2 Model comparisons and interpretation of segment classes

The following measures will be used to examine model-by-method differences: number of segment classes yielded, model fit/accuracy, and segment differentiating characteristics. Model fit criteria specific to each method is presented in Table 6.

Number of segment classes yielded

As previously mentioned, and perhaps the greatest challenge of segmentation analysis, is determining the “correct” number of classes, and the method by which the final number of classes has been determined. Several categories of models provide specific information criteria, which can be used in accordance with additional measures to identify the best segment number solution (i.e., LCA provides BIC). The choice of a cluster solution is often a “trade-off” between the marketing plan and data quality (Vellido et al., 2009). Choosing a segment solution often involves a degree of “story telling” or choosing a solution that makes sense, and provides meaningful, interesting information to a marketing manager. In addition to reviewing information criteria, researchers have used visualization techniques to determine the appropriate number of clusters. A “quality score” plot, which is similar to a scree plot in factor analysis can be used to identify the optimal number of clusters (the slope of the line goes flat as clusters become irrelevant). Size constraints are extremely important in segmentation analysis. To minimize error, segments need to be "not-too-small" and "not-too-big." If a segment is too small, the probability of making an error increases due to the lack of statistical significance. If the segment is too big, then if a "good" segment is somehow eliminated from the treatment condition (Type I error) which have the potential for large foregone profits; and if a "bad" segment makes it to the treatment (Type II error), often large out-of-pocket costs are incurred.

Model fit/accuracy

All models will be validated by utilizing a cross-validation procedure. Specifically, the 10-fold cross-validation will be used, according to the following procedure:

1. Randomly partition the original sample into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data.
2. Repeat this process 10 times with each of the 10 subsamples used exactly once as the validation data.
3. All 10 of the results from the folds are averaged to produce a single estimation (mean squared error)

In addition to cross validation methods, each segmentation method has specific model fit criteria. Latent class models will also be evaluated according to the BIC and a preferred model is chosen when the BIC is minimized. K-means models utilize Wilk's lambda as a fit statistic, and the "best" model is the solution that yields the smallest Wilk's lambda. Neural network models will be assessed according to the residual mean square error (smaller is better) and C-index (larger is better).

For the condition of CHAID and network models when the outcome is binary (death at follow-up), classification tables will be used. Classification tables provide a description of the percentage of correctly classified "events," as illustrated in Figure 5. For binary response data, the response is either an event or a nonevent, and the prediction models model the probability of the event. From the fitted model, a predicted event probability can be computed for each observation. If the predicted event probability exceeds or equals some cut point value $z \in [0, 1]$, the observation is predicted to be an event observation; otherwise, it is predicted as a nonevent.

Partition		Predicted		
		0	1	Percent Correct
Training	0	112	0	100%
	1	9	31	77.50%
	Overall Percent	79.61%	20.40%	94.08%
Holdout	0	0	0	0%
	1	0	0	0%
	Missing	1	1	
	Overall Percent	50.00%	50.00%	0.00%

Figure 5. Classification Table Example

Since a "true" class outcome for classification models cannot be observed, misclassification rates for classification models will be derived by comparing several sources of error. For latent class models, the misclassification rate will be presented as the cumulative difference between the estimated class population shares and predicted class membership. The misclassification rate for K-means models is the cumulative difference between the population proportion in the original segment class solution and the proportion of members in each segment class of a 10-fold cross validation.

Evaluation of differentiating characteristics and profiling of segments

While there is not a specific statistical procedure or method for evaluation differentiating characteristics of segments by varying segment classes, this step is crucial in the appraisal of overall segmentation model performance. The differences in attitudes, behaviors, preferences, health status and demographic characteristics of the segment classes by method become the basis for any future engagement as a result of this research.

To further evaluate model outcomes, each segment class within each model condition will be profiled according to the following procedure:

- First, assign respondents to segments according to their highest probability of segment class membership
- Generate segment-based demographic and attitudinal profiles, using demographic variables not included in the model to profile the segments, and attitudinal variables included in the model to enhance the profiles. The profiles will highlight the differentiating characteristics to create meaningful personifications of each segment class

3.2 CLASSIFICATION MODELS

3.2.1 Latent class analysis

Latent class analysis will be performed in R, utilizing the poLCA program developed by Drew Linzer and Jeffrey Lewis (https://cran.rproject.org/web/packages/poLCA/poLCA.pdf). In its simplest form, poLCA fits a latent class model given a specific data set, with the minimum specifications: number of latent classes, the items measuring the latent variable, and the number of response categories per item. Additionally, other parameter restrictions can be specified, along with multiple-groups analysis and measurement invariance across groups. Continuous and categorical covariates can be included in the COVARIATES statement in order to explore the relationship between individual covariates and the probability of latent class membership. Starting values can be random or user-provided. Since Goodman's maximum likelihood approach has

remained the standard way of estimating parameters in the latent class model, no other estimation methods will be considered in this study.

3.2.1.1 Data preparation

Within poLCA, response categories for items measuring the latent class variable must be coded with sequential integer values from 1 to *R*, where *R* is the number of response categories for that item. All covariates are treated as numeric in the statistical model. Categorical covariates must be dummy coded and continuous covariates should be transformed to z scores, allowing easier interpretation by producing standardized logistic regression coefficients (as demonstrated in Table 4.

Only one grouping variable can be included, although two or more grouping variables can be crossed to create a single grouping variable (e.g., gender and education level can be crossed to create a four-level grouping variable: female high school education, male high school education, female college education, male college education). The grouping variable is then coded using consecutive integers (1,2,3,4).

3.2.1.2 Convergence and additional specifications

poLCA employs the EM (expectation-maximization) algorithm to produce maximum likelihood estimates of all model parameters; meaning the algorithm iterates between the Expectation (E) step and the Maximization (M) step until either the convergence criterion is achieved or the maximum number of iterations is reached. The convergence index used in poLCA is TOL (tolerance); which is a value for judging when convergence has been reached (1e-10). When the one-iteration change in the estimated log-likelihood is less than TOL, the estimation algorithm stops iterating, as a result

of finding the maximum log-likelihood. Of note, a larger value for the convergence criterion results in convergence in fewer iterations, but noticeable additional improvement in parameter estimates is possible. A smaller value for the convergence criterion requires more iterations to converge, but once convergence is reached, little improvement in estimation is possible. The researcher must set the number of repetitions (using different values of probs.start) and the maximum number of iterations through which the EM algorithm will cycle. The default number of repetitions for this study is 10, with the maximum iterations = 50000. In addition, the researcher must specify the number of latent classes for the model to assume, using the NCLASS statement.

3.2.1.3 Model output

The model outputs data files containing all final parameter estimates (gamma, beta, or rho); allowing for easy comparison across model classes. The number of lines in each parameter set varies by the number of groups, covariates, indicators, and the number of response categories for each indicator. The researcher must then merge the segment membership probabilities back to the original data set containing all of the original variables. The final segment is assigned according to the highest probability. For example, if an individual's probability for belonging to segment class 3 is 0.45 and his or her probability for belonging to Segment Class 4 is 0.89, the individual would be assigned to Segment Class 5.

3.2.1.4 Model assessment

In order to identify the optimal model, a sequence of models with varying number of classes (two, three, four…*n*) are produced. Each class will be evaluated according to their information criteria (AIC/BIC), difference in fit (G^2 likelihood) and model interpretability. For example, each class

should be distinguishable from the others on the basis of the item-response probabilities, no class should be trivial in size (i.e., with a near-zero probability of membership), and it should be possible to assign a meaningful label to each class. Desirable models would include goodness of fit indices that exceed .9 (Byrne, 1994) and BIC near 0 (Rafferty, 1995). As described in Section 3.1.2.2, models will be evaluated using the criteria listed above, as well as the cross-validation procedure and classification table procedure. The most parsimonious model will be the one whose error is no more than one standard error above the error of the best model (cross-validation) and the one with the highest proportion of correctly classified cases (classification tables).

3.2.2 **K-means clustering**

The HP CLUS procedure will be used to perform K-means clustering in SAS Enterprise Miner®. PROC HP CLUS performs a cluster analysis on the basis of distances computed from one or more quantitative variables. The observations are divided into clusters such that every observation belongs to a single cluster. By default, the HP CLUS procedure uses Euclidean distances, so the cluster centers are based on least squares estimation. Each iteration reduces the least squares criterion until convergence is achieved. The initialization method used by the HP CLUS procedure makes it robust against the detection of outliers, since outliers often appear as clusters with only one member. Since the majority of the response variable categories are ordinal/nominal, the inputs will need to be transformed using effect coding with binary indicators. This is due to the fact that k-means uses a distance measure to organize the clusters. Each level of the categorical variable will be compared to a fixed reference level. For example, let race = 1 as the reference group and compare the mean of “Number of Days Physical Health not Good” for each level of race = 2 and 3, as demonstrated below:

Table 4. Effect Coding of Response Variables

	Transformed Variable		
Level of race	RACE_CAT 1	RACE_CAT 2	RACE_CAT 3
1 (White)	1	-1	-1
2 (Black)	-1	1	-1
3 (Other)	-1	-1	1

The HP CLUS procedure then uses the data set containing the transformed variables an as input and creates an output data set. This output data set contains all original variables and two created variables, *CLUSTER* and *DISTANCE*. The variable *CLUSTER* contains the cluster number to which each observation has been assigned. The variable *DISTANCE* gives the distance from the observation to its cluster seed.

In simplest terms, the process for clustering based on Euclidean distance will be as follows:

1. Specify *k*, the number of clusters to be generated
2. Choose *k* points at random as cluster centers
3. Assign each observation to its closet cluster using Euclidean distance
4. Calculate the centroid (mean) for each cluster, use it as a new cluster center*
5. Reassign all instances to the closest cluster center
6. Iterate until the cluster centers don't change anymore

*Of note, the calculation of new cluster means uses the least-squares estimator, so no additional estimation methods will be considered.

3.2.2.1 Model output

PROC HP CLUS produces a table of summary statistics for the clusters, which include: the number of observations in the cluster (frequency) and the root mean squared standard deviation, the largest Euclidean distance from the cluster seed to any observation within the cluster, the number of the nearest cluster, and the distance between the centroid of the nearest cluster and the centroid of the current cluster. In addition, the pseudo F statistic, approximate expected overall R square, and cubic clustering criterion (CCC) are available for comparison of different values of the clustering assignments. Of note, values of the cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters, but they should be taken with caution; large negative values can indicate outliers.

3.2.2.2 Model assessment

Similar to LCA, models with increasing numbers of classes (*k=1, 2, …n*) will be compared. Models with varying segment class number solutions will be compared by examining the change in within-cluster dispersion under the null distribution. Specifically, a gap statistic will be computed for each set of model comparisons:

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(w_k)$$

where *k* is the number of clusters employed and W_k is the pooled within-cluster sum of squares around the cluster means under sample size *n.* When the gap statistic is maximized, the model has exhibited the highest level of discriminatory power between the clusters. In addition, models will be compared according to their interpretability, or profiling of the segment classes by variables not included in the model, as well as the distributions of variables included, as described in Section

The most desirable models will have Wilks' lambda approaching zero, as smaller values of Wilks' lambda indicate greater discriminatory ability of the model to separate cases into groups. As documented with latent class models, a well-fitting k-means model will have the largest proportion of correctly classified cases as well as the smallest cross-validation error.

3.2.3 Varying the inputs: Categorization of a response scale

For segmentation methods designed to address classification questions, one additional condition in the design is the variation of a response scale (specifically, the depression screening scale). There are four questions within the depression screening scale, which will be combined into a single binary categorization of depression:

1. "In the past year, have you had 2 weeks or more during which you felt sad, blue, or depressed; or when you lost interest or pleasure in things that you usually cared about or enjoyed?" The response categories are: Yes (1) and No (2). **(Q36, Field#63)**
2. "In the past year, have you felt depressed or sad much of the time?" The response categories are: Yes (1) and No (2). **(Q37, Field#64)**
3. "Have you ever had 2 years or more in your life when you felt depressed or sad most days, even if you felt okay sometimes?" The response categories are: Yes (1) and No (2). **(Q38, Field#65)**
4. "How much of the time in the past week did you feel depressed?" The response categories are: Rarely or none of the time (1), Some or a little of the time (2), occasionally or a moderate amount of the time (3), Most or all of the time (4). **(Q39, Field#66)**

Values were assigned to the responses according to the following:

Table 5. Transformation of Depression Scale

Question 36 Responses	***Transformed Value***
Yes	1
No	0
Question 37 Responses	***Transformed Value***
Yes	1
No	0
Question 38 Responses	***Transformed Value***
Yes	1
No	0
Question 39 Responses	***Transformed Value***
Most or all of the time	1
Occasionally or a moderate amount of time	1
Some or a little bit of the time	0
Rarely or none of the time	0

Model performances using the scale containing its 4 questions (Cronbach's α =0.76) will be compared against model performances with a dichotomization of the scale: sum of items ≥ 1 (depressed= 1) vs. sum of items < 1 (depressed= 0). For the baseline depression measure, the Medicare beneficiary is considered to have a positive depression screen when he or she answers "Yes" to any of the four depression questions (Questions 36, 37, 38 or 39) in the 2012 HOS 2.0. (*2012-2014 Cohort 15 Analytic* Public Use File Data User's Guide, 2015).

In addition, individuals will also be classified by their responses to the Veteran's RAND 12 Item Health Survey (VR-12), an additional instrument included within the Medicare HOS. The VR-12 is a generic, health-related quality of life survey, designed to estimate disease burden. The survey is comprised of fields 8-19 as indicated in Table 3.

3.3 PREDICTION MODELS

3.3.1 Neural networks

Neural networks for segmentation model comparisons will be analyzed utilizing SAS Enterprise Miner version 14.1. A neural network model will be built to predict the following dependent variables: the number of days physical health not good and a categorization of days physical health not good. Within Enterprise Miner, the neural networks (and all other models) are generated using a point-and-click interface, in which a “diagram” or model space is created. Each step within the model building process is referred to as a “node” that can be connected to a previous and subsequent step. Initially, the data source node is placed onto the diagram, and the “roles” (targets, inputs, variable types, etc.) are defined. A data partition code is then connected to the input data source node, in which the percentages for data partitioning (test and training) sets are established. For the purpose of this study, the training data set will contain 70% of the data, where the test set is 30%. Transformations will not be used in this analysis, but if they were desired, a transform variables node could be added at this point. (If this analysis were a regression problem rather than a classification problem, transformation of the target variable might be necessary for a good model fit if the error variance depends on the target value, especially when using the usual squared-error loss.)

Neural network models, like linear/logistic regression models (and unlike tree models), generally require imputation of missing predictor variables, which can be performed by connecting a replacement node after the data partition node, and changing the default imputation method to tree imputation. After the data has been prepared, the neural network node is placed in the diagram after the replacement node. The number of hidden units is automatically configured based on the

number of inputs and the sample size (or can be directly specified with prior knowledge regarding the data structure and modeling needs, which is beyond the scope of this study). The “usual type” of a neural network’s structure will be used when the dependent variable is binary (multilayer perceptron). With this option, there will be a single hidden layer with logistic activation functions, by default. For the ordinal dependent variable, a generalized linear model (no hidden layer) will be used, and a radial basis function is selected for continuous outcomes. The training of a neural network involves the optimization of some type of objective function (often the likelihood function) starting from a given set of initial weights (the parameters of a neural network model). Because training can take excessive amounts of time for large data sets with many variables, a maximum training time can be set. Upon completion, Enterprise Miner produces an output table which includes the fitted parameter values for the neural network (one for each connection, plus appropriate bias terms). These parameters are identified by the connections with which they are associated, as defined by the node labels in the “from” and “to” columns. Input nodes are identified by the corresponding variable name (e.g. RACE), hidden nodes are labeled with a layer number and an index (e.g. H12 for the second node in the first hidden layer), and the output node is labeled with the output variable name. A graphical representation will also be requested. The graphs here show the performance of the model on the training and validation data sets as a function of training (optimization) iteration number, where the criterion being graphed is by default average error.

Specifically, the steps to performing the neural network model are as follows:

1. Receive inputs
2. Weigh inputs. Each input that is sent into the neuron must first be weighted, i.e. multiplied by some value (often a number between -1 and 1). When creating a perceptron, random weights are assigned.

3. Take each input and multiply it by its weight.
4. Sum inputs.
5. Generate output. The output of a perceptron is generated by passing that sum through an activation function. In the case of a simple binary output, the activation function is what tells the perceptron whether to “fire” or not.

3.3.1.1 Model output

Enterprise Miner displays a variety of outputs for neural network models, including the number of input and output units, the number of hidden layers and units and activation functions. The synaptic weights are available in the form of coefficient estimates which demonstrate the relationship between the units in a given layer to the units in the following layer. The synaptic weights are based on the training sample even if the active dataset is partitioned into training, testing, and holdout data. A model summary displays a summary of the neural network results by partition and overall, including the error, the relative error or percentage of incorrect predictions, the stopping rule used to stop training, and the training time. The error is the sum-of-squares error when the identity, sigmoid, or hyperbolic tangent activation function is applied to the output layer. Relative errors or percentages of incorrect predictions are displayed depending on the dependent variable measurement levels; dependent variables of scale contain the average overall relative error and the average percent of incorrect predictions are displayed for the categorical or binary DV. In addition, a classification table giving the number of cases classified correctly and incorrectly for each dependent variable category is reported. Independent variables included in the model are evaluated according to a sensitivity analysis, which computes the importance of each predictor in determining the neural network. The assigned segment categories are available in a new data set, by using the Save Data node.

3.3.1.2 Model assessment

As with the previous models, iterations (variations of the DVs and classifying algorithm) of neural networks will be assessed according to the number of segment classes produced, the proportion of individuals within each segment class, interpretations of the variables within each model and the model's prediction accuracy. A cumulative gains chart, as presented in Figure 6, will be used to compare the model performance in predicting the respondents in the highest decile of number of unhealthy days; the greater the area between the lift curve and the baseline, the better the model. By default, the Assessment node displays a cumulative % Response lift chart. The individual cases are sorted from left to right by individuals who are most likely to have a lot of unhealthy days as predicted by each model. The sorted group is then divided into ten deciles along the X axis. The left-most decile represents the 10% of the individuals who are most likely to have a high proportion of unhealthy days. The vertical axis represents the actual cumulative response rate in each decile. For the binary target, the lift chart does not adjust for the expected loss, it considers only the event posterior probabilities. In addition, desirable models will have a C-Index approaching 1 and a smaller residual mean square and cross-validation error.

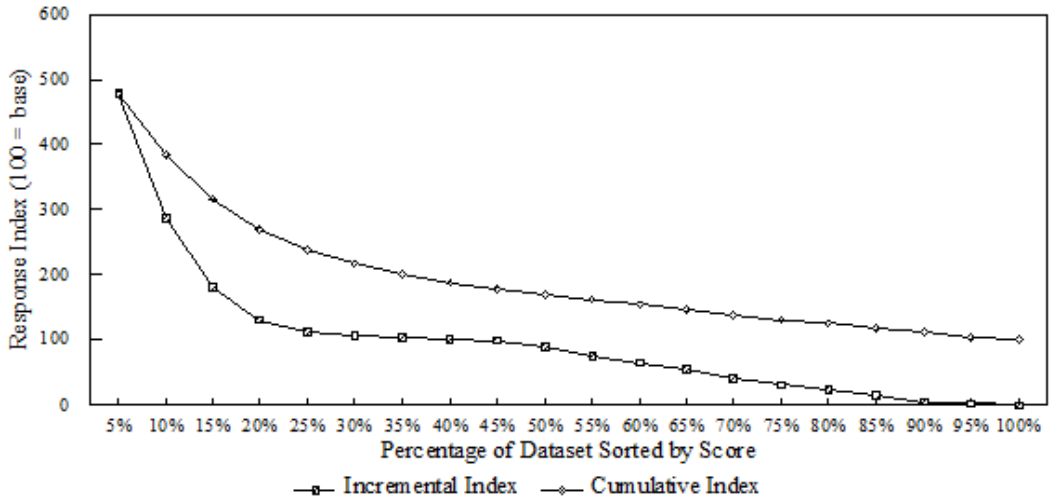


Figure 6. Neural Network Gain Chart

3.3.2 Chi Square Automatic Interaction Detector (CHAID)

SAS Enterprise Miner version 14.1 will also be used to assess the performance of CHAID segmentation models. The first several steps are identical to the Neural Network procedure as described above. The SAS implementation of trees allows binary and multi-way splits based on nominal, ordinal, and interval inputs. Since CHAID can accommodate both categorical and continuous variables, both the number of days physical health not good and it's categorical transformation will be analyzed. If root node is continuous, the branches will represent classes of the node; if the root node is categorical, the branches will represent specific ranges along the scale of the node. At each split, a question is asked, which has an answer in terms of the classes or range of the variable being split. The questions are defined in terms of some impurity measure, reflecting how uniform the resulting cases must be in the splits. Each branch is split further using the classes or ranges of other variables. At each split, the node that is split is named parent node, and the nodes which split into are called the child nodes. This process continues until some stopping rule is satisfied or splitting is impossible. There is no need to transform the values because tree node will group continuous variables into bins automatically.

3.3.2.1 Model output

In addition to producing a visual representation of the importance of individual predictors and their ability to split the data, a variety of outputs for CHAID models is available in Enterprise Miner. The assessment table and assessment plot display the training and validation assessment values for each of the decision trees (one decision tree is produced for each model condition). These views reveal how large a tree is needed for a sufficient fit, and whether the problem of overfitting is present in large trees. If the classification for the training and validation data are similar across all

subtrees, no overfitting is present in the training data. For the binary outcome, the misclassification rate is used as the model assessment measure. In addition, the classification matrix summarizes the prediction for each level of the binary target variable. Included in the visualization of the trees are node statistics (means or frequencies), splitting variables, and splitting rules. The segment assignment values are also included by saving the new data set as a SAS file.

3.3.2.2 Model assessment

The performance of the variety of CHAID models will be examined using a lift chart, as described in Figure 7. Within Enterprise Miner, the "Assess" node will be employed to create lift charts for each of the model conditions. The lift chart plots the values in the *Index (%)* column in the table. This chart compares the percentage of records in each increment that are hits with the overall percentage of hits in the training dataset, using the equation:

(hits in increment / records in increment) / (total number of hits / total number of records). As with neural networks, a desirable model will have the greatest area between the lift curve and baseline, along with minimal cross-validation error.

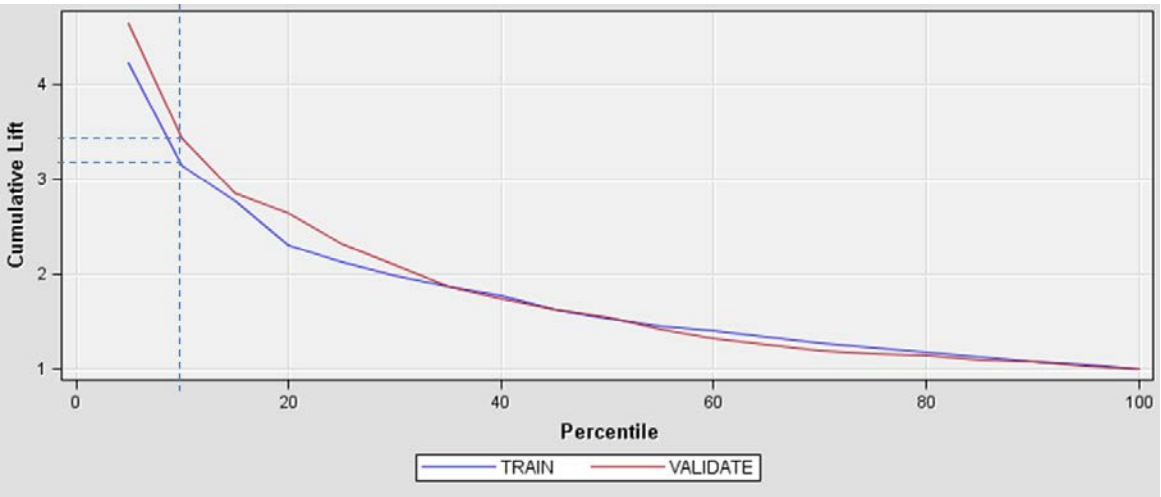


Figure 7. Lift Chart for CHAID Models

3.3.3 **Varying the DV: Continuous vs. binary outcome**

For segmentation prediction models, the dependent variable will take on both continuous and binary distributions for comparative purposes:

1. Continuous: Number of Days Physical Health Not Good

   Of note, in the modeling/analysis exercise, a log transformation of the continuous variable was used to approximate a more normal distribution.

2. Binary: Categorization of Number of Days Physical Health Not Good. An additional condition of this design will be the examination of the continuous split:

   (i) Even split 1= (≤15 days physical health not good) and 0 (>15 days physical health not good)

   (ii) Uneven split 1 (≤5 days physical health not good) and 0 (>6 days physical health not good)

The rationale behind the split was derived from research conducted by the CDC regarding the measurement of healthy days. An in-depth review of the National Health and Nutrition Examination Survey (NHANES) results over a span of five years (1993-1998) indicated that the mean unhealthy days by self-rated general health for those in "good health" is 5.3 years (Centers for Disease Control and Prevention, 2000)

The distribution of each condition of the dependent variable can be found in Appendix C.

3.3.4 Varying the IV: Boosting and bagging

For segmentation methods designed to address prediction questions, variation of the ensemble classifying algorithm will serve as an additional IV condition. Since multiple studies were unable to identify differences in accuracy by ensemble methods, this study will attempt to uncover any differences by classifying strategy (West et al., 2005) Model performances in which the boosting algorithm was used will be compared to model outcomes where bagging was applied.

Boosting and bagging will be implemented alongside neural networks and CHAID in Enterprise Miner, by means of the Gradient Boosting Node is on the model tab of the Enterprise of the tools bar. This node uses a partitioning algorithm, which searches for an optimal partition of the data defined in terms of the values of a single variable. The optimality criterion depends on how another variable, the target, is distributed into the partition segments. The more similar the target values are within the segments, the greater the worth of the partition. The partitions are then combined to create a predictive model. The model is evaluated by goodness-of-fit statistics defined in terms of the target variable. These statistics are different than the measure of worth of an individual partition.

Gradient boosting is a boosting approach that resamples the analysis data set several times to generate results that form a weighted average of the re-sampled data set. Tree boosting creates a series of decision trees which together form a single predictive model. A tree in the series is fit to the residual of the prediction from the earlier trees in the series. The residual is defined in terms of the derivative of a loss function. For squared error loss with an interval target the residual is simply the target value minus the predicted value. Each time the data is used to grow a tree and the accuracy of the tree is computed. The successive samples are adjusted to accommodate previously computed inaccuracies. Because each successive sample is weighted according to the

classification accuracy of previous models, this approach is sometimes called stochastic gradient boosting. Boosting is defined for binary, nominal, and interval targets.

Like decision trees, boosting and bagging make no assumptions about the distribution of the data. For an interval input, the model only depends on the ranks of the values. For an interval target, the influence of an extreme value theory depends on the loss function. The Gradient Boosting node offers a Huber M-estimate loss which reduces the influence of extreme target values.

In the Group Processing node in SAS Enterprise Miner, bagging uses sampling with replacement to train models in parallel and combining the predicted probabilities. All observations have the same weight applied. As illustrated in Figure 8, multiple decision trees can be trained by using different samples of the training data. First, a different random seed is specified for each of the sample nodes. Next a Decision Tree or Neural Network node is connected and the default Ensemble node is used to average the predicted probabilities of all connected models.

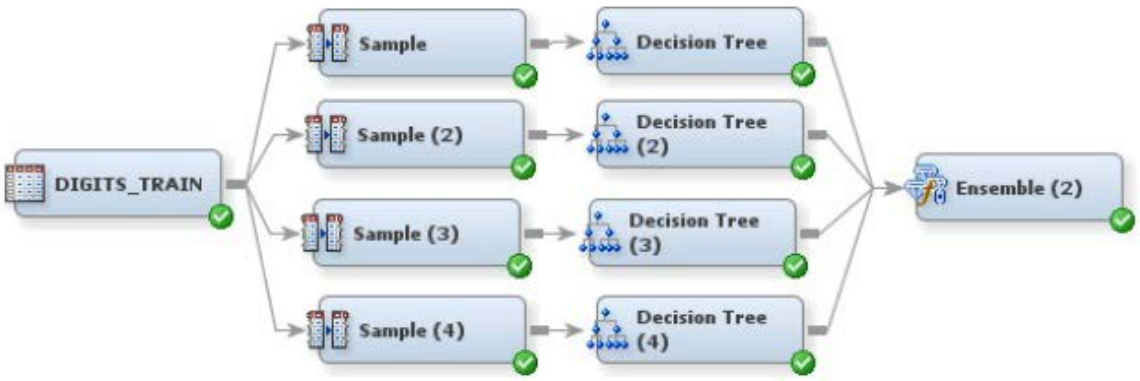


Figure 8. Illustration of Bagging in Enterprise Miner

3.4 SAMPLE SIZES AND MEAN RESPONSE FOR CLASSIFICATION AND PREDICTION MODELS

3.4.1 Classification models

Due to the way missing data are handled by the different methods, the sample sizes varied slightly for classification models, with LCA utilizing 94,472 out of 260,984 respondents (36%), and K-Means utilizing 104,564 (40% of respondents). While there was approximately a 10% reduction in the LCA sample size, the response proportions for all levels of the input variables (and profiling variables) were nearly identical between LCA and K-Means, as presented in Table 6 below. Thus, any comparisons between the methods should not be compromised by the different samples that were analyzed.

Table 6. Response Proportions for Scale and Profile Variables

Variable	*LCA Sample (n=94,472)*	*k-Means Sample (n=104,564)*
Demographic (Profile Variables)		
Age: Less than 65	14.83%	14.99%
Age: 65-74	50.48%	50.54%
Age: 75 and older	34.69%	34.47%
Female	57.66%	57.30%
Less than high school education	24.78%	24.07%
High school education or GED	34.74%	34.76%
Greater than high school education	40.49%	41.17%
Obese	32.83%	32.64%
Married	51.10%	51.28%
Depression Variables (Dichotomized and Original Scale)		
Depress1 (Depressed)	65.26%	65.00%
Depressed 2 weeks or more past year	29.43%	30.26%
Depressed in the past year	20.47%	21.24%
Depressed in the past 2 years	22.53%	23.4%
Depressed most of the time (past week)	6.94%	7.03%

Table 6 continued

Depressed some of the time (past week)	12.22%	12.55%
Depressed rarely (past week)	19.38%	19.95%
Depressed none of the time (past week)	61.46%	60.47%
Veteran's RAND Scale		
General Health: Excellent	6.37%	6.17%
General Health: Very good	23.13%	22.8%
General Heath: Good	35.8%	35.93%
General Health: Fair	25.95%	25.89%
General Health: Poor	8.75%	9.21%
Moderate activities: Limited a lot	25.92%	25.72%
Moderate activities: Limited a little	33.35%	34.11%
Moderate activities: Not limited at all	40.53%	40.18%
Climbing stairs: Limited a lot	32.56%	33.05%
Climbing stairs: Limited a little	34.02%	33.59%
Climbing stairs: Not limited at all	33.42%	33.35%
Accomplish less (physical): None of the time	32.48%	32.07%
Accomplish less (physical): A little of the time	19.18%	18.81%
Accomplish less (physical): Some of the time	22.3%	22.73%
Accomplish less (physical): Most of the time	16.64%	16.79%
Accomplish less (physical): All of the time	9.4%	9.6%
Limited in work (physical): None of the time	58.78%	58.15%
Limited in work (physical): A little of the time	15.05%	14.98%
Limited in work (physical): Some of the time	13.59%	13.64%
Limited in work (physical): Most of the time	8.02%	8.27%
Limited in work (physical): All of the time	4.57%	4.97%
Accomplish less (emotional): None of the time	55.52%	54.61%
Accomplish less (emotional): A little of the time	15.18%	15.41%
Accomplish less (emotional): Some of the time	15.09%	15.38%
Accomplish less (emotional): Most of the time	9.34%	9.61%
Accomplish less (emotional): All of the time	4.87%	4.99%
Did not work as carefully: None of the time	34.19%	34.19%
Did not work as carefully: A little of the time	18.06%	17.62%
Did not work as carefully: Some of the time	20.99%	21.35%
Did not work as carefully: Most of the time	16.02%	15.73%
Did not work as carefully: All of the time	10.74%	11.1%
Pain interfered with work: Not at all	28.56%	27.9%
Pain interfered with work: A little bit	25.7%	25.38%
Pain interfered with work: Moderately	18.26%	18.95%
Pain interfered with work: Quite a bit	19.15%	19.3%
Pain interfered with work: Extremely	8.32%	8.48%
Felt calm and peaceful: All of the time	15.57%	15.21%
Felt calm and peaceful: Most of the time	40.17%	40.07%
Felt calm and peaceful: A good bit of the time	13.2%	13.54%
Felt calm and peaceful: Some of the time	18.17%	18.6%
Felt calm and peaceful: A little of the time	9.44%	9.17%
Felt calm and peaceful: None of the time	3.45%	3.42%

Table 6 continued

Have a lot of energy: All of the time	7.16%	7.11%
Have a lot of energy: Most of the time	26.36%	25.86%
Have a lot of energy: A good bit of the time	15.35%	15.13%
Have a lot of energy: Some of the time	23.45%	23.3%
Have a lot of energy: A little of the time	17.59%	18.06%
Have a lot of energy: None of the time	10.09%	10.54%
Felt downhearted/blue: All of the time	3.64%	3.82%
Felt downhearted/blue: Most of the time	5.28%	5.4%
Felt downhearted/blue: A good bit of the time	6.00%	6.08%
Felt downhearted/blue: Some of the time	19.8%	20.45%
Felt downhearted/blue: A little of the time	26.22%	26.15%
Felt downhearted/blue: None of the time	39.06%	38.1%
Health interfered with social activities: All of the time	5.94%	6.21%
Health interfered with social activities: Most of the time	11.01%	11.22%
Health interfered with social activities: Some of the time	19.44%	19.64%
Health interfered with social activities: A little of the time	15.9%	15.95%
Health interfered with social activities: None of the time	47.7%	46.98%
Number of days mental health not good	5.32	5.39
Number of days physical health not good	7.71	7.79

3.4.2 **Prediction models**

Both CHAID and Neural Network models utilized the entire sample (*n=260,984)* for classifying respondents into segment classes. See appendix B for the frequency of response for each variable.

3.5 EVALUATION AND SUMMARIZATION OF RESULTS

3.5.1 Model performance

In order to answer the first three research questions, model outcomes were compared according to the evaluation plan below. Each model was evaluated according to: the number of segments yielded (best segment class comparisons); the descriptions of the segments and the differentiating characteristics between segment classes; and model fit statistics, including information criteria and classification accuracy. For latent class classification models, the misclassification rate was presented as the difference between the estimated class population shares and predicted class membership. The misclassification rate for K-means models is the cumulative difference between the proportion of membership in the original segment class solution and the 10-fold cross validation solution. Misclassification rates for prediction models will be assessed using the mean proportion of misclassified observations, *predicted – actual*. For prediction models with continuous outcomes, average squared error will be reported, which is the sum of squared error (SSE) divided by the number of cases, N. For neural networks and decision trees, there is no known unbiased estimator, thus why SSE is divided by N and not the degrees of freedom.

Table 7. Evaluation Plan

Method	*Variables*	*# of Segments*	*Segment Descriptions*	*Differentiating Characteristics*	*Model Fit Statistics*	*Segment Profile*
Segmentation Method: Prediction						
Dependent Variable as a Continuous Outcome (# Days Physical Health not Good)						
Neural Networks	K_a	N_a			Average Squared Error	
CHAID	K_b	N_b			Average Squared Error	
Dependent Variable as a Categorical Outcome (Categorization of Days Physical Health not Good)						
Neural Networks	K_a	N_a			Classification Tables/ % Gain	

Table 7 continued

CHAID	K_b	N_b	Classification Tables/ % Lift
Independent Variable: Boosting vs. Bagging			
Neural Networks	K_a	N_a	Classification Tables/ % Gain
CHAID	K_b	N_b	Classification Tables/ % Lift
Segmentation Method: Classification			
Independent Variable as a Categorical Scale			
Latent Class	K_a	N_a	Classification Tables/ BIC
K-Means Clustering	K_b	N_b	Classification Tables/ Gap
Independent Variable as a Dichotomous Scale			
Latent Class	K_a	N_a	Classification Tables/ BIC
K-Means Clustering	K_b	N_b	Classification Tables/ Gap

3.5.2 **Variable importance**

In addition, variable importance across models were compared. Variable importance within PROC LCA was evaluated according to the ρ parameters generated by the output. Variables are ranked according to their ρ coefficient; with values ranging from 0 to 1; 1 indicating highest importance. Since K-Means models do not produce the same parameters, the process of examining intra-cluster variable similarity was used. This process works by calculating the average similarity of each variable to its centroid. A variable that has *high* similarity between a centroid and its objects is likely more important to the clustering process than a variable that has *low* similarity. Although similarity magnitude is relative, variables can be ranked by the degree to which they help to cluster the objects in each cluster. By default, CHAID models produce diagram containing an illustration of splits; variables that are most important in distinguishing categories are listed at the top of the diagram and lessen in importance when moving down the tree. To determine variable importance

for neural networks, the Metadata node is selected and the role of the target variable is changed to "rejected" and the role of the predicted targets (clusters) are changed to "target." Then, a decision tree can be used to visual inspect and determine variable importance, according to the procedure above.

4.0 RESULTS

The purpose of this study was to examine the differences in segmentation outcomes between models for two types of analysis problems in marketing research: classification models and prediction models. For classification models that attempt to identify a set latent classes of individuals based on their responses to a set of observed variables, latent class analysis and k-means clustering were compared. For studies in which there is a defined prediction question, the identification of segment classes using neural networks and CHAID models were compared. Response data from Medicare Health Outcomes Survey (HOS) were used for all segmentation model comparisons. The use of continuous variables versus dichotomized continuous variables within the segmentation models was also evaluated, in addition to classification algorithms for prediction models.

4.1 COMPARISON OF CLASSIFICATION MODELS

The variables from the HOS dataset used for this analysis included C15DEP2WK (2 weeks of depression), C15DEPYR (depression most of the year), C15DEP2YR (depressed 2 years or more), C15DEPWEEK (depressed during past week), C15VRGENHTH (general health), C15VRMACT (moderate activities), C15VRSTAIR (climbing several flights of stairs), C15VRPACCL (physical health limiting amount accomplished), C15VRPWORK (physical health interfere with work), C15VRMACCL (emotional health limiting amount accomplished), C15VRMWORK (emotional health interfere with work), C15VRPAIN (pain interfering with work), C15VRCALM (feel calm

and peaceful), C15VRENERGY (lots of energy), C15VRDOWN (feel downhearted and blue), C15VRSACT (health interferes with social activities, and DEPRESS1 (dichotomization of depression scale); with the demographic variables of age, race, gender, marital status, education, BMI, and smoking history used as profiling variables (after the segmentation models were completed).

For the different classification models, determination of the best segment class solution involved the following considerations:

1. Estimation of Clusters. For the K-Means method, the best segment class solution would be indicated by a maximization of the gap statistic, which represents the change in cluster dispersion. When the gap statistic is maximized, the model has exhibited the highest level of discriminatory power between the clusters. For the LCA method, a winning solution would be reflected in the model that demonstrates the minimization of the BIC criteria. BIC is a measure of information loss, therefore the best solution would be one that minimizes the loss and provides the best approximation of the data.
2. Misclassification. In addition to information criteria, the misclassification rate of each segment class solution will be considered. For latent class classification models, the misclassification rate is presented as the difference between the estimated class population shares and predicted class membership. The misclassification rate for K-means models is the cumulative difference between the proportion of membership in the original segment class solution and the 10-fold cross validation solution. However, misclassification alone cannot be the rationale behind determination for the best segment class solution, as often models diminish in error and misclassification rates as

the number of clusters increase. The goal is to find the most parsimonious solution, in which total intra-cluster variance or error function is minimized.

3. Segment class profiles. The distributions of demographic variables were examined for each segment class across conditions. Heat maps were used to identify similarities and differences across segment classes. A heat map is an Excel table where the data are visualized using color and conditional formatting. For the purpose of demographic variable comparisons, values that are close to the mean value (of each variable, individually) are coded in shades of YELLOW. Values that are higher than the mean value are displayed in GREEN; whereas values much lower than the mean are presented in RED.

As previously discussed, the best segment class solution for Latent Class models is the one in which the BIC criteria is minimized. As demonstrated in Table 8 below, the BIC criteria decreases as the number of segment classes increases, through Segment 9, but then increases for the 10-segment solution for both methods under LCA. Similarly, a K-Means model identifies a best segment class solution when the Gap statistic is maximized. When the K-Means model utilizes the original depression scale, the Gap statistic increases as the number of segment classes increases, until a 9-segment solution is reached. The Gap statistic then decreases at 10 segments. Examining the dichotomized depression scale, the Gap statistic is maximized at 6 segments.

Table 8. Comparison of Information Criteria by Solution for Classification Methods

Method	*5 Segments*	*6 Segments*	*7 Segments*	*8 Segments*	*9 Segments*	*10 Segments*
Original Depression + Veteran's RAND 12-item Scales						
Latent Class Analysis	8145053	8060049	7977870	7920759	**(7880830)**	7886125
K-Means Clustering	.1227	.1636	.1642	.1655	**.2003**	.1865
Dichotomization of Depression Scale + Veteran's RAND 12-item Scales						
Latent Class Analysis	7449813	7364669	7303344	7258120	**(7225709)**	7244217
K-Means Clustering	.1101	**.2273**	.1563	.1873	.1967	.2094

In addition to consideration of the information criteria described above, misclassification rates for each method/condition were examined. In both treatments of the scale (full-scale and dichotomization), K-Means models resulted in higher misclassification rates; 43% misclassified using the original scale, and 29% misclassified in the dichotomous condition. The misclassification rate for Latent Class models was considerably lower than K-Means, with only 4% misclassified when the original scale was included, and 8% for the dichotomized scale. Misclassification rates did not change dramatically for the LCA models as the number of classes increased, and were essentially the same in the dichotomous condition. In the case of the K-Means models, the misclassification rates were considerably higher as the number of segment classes increased, from the preferred solution to the same solution + one additional segment class. As indicated in Table 9, for the K-Means model utilizing the original depression scale, the misclassification rate doubled from a 9-segment to 10-segment solution.

Table 9. Comparison of Misclassification Rates by Solution for Classification Methods

Method	*5 Segments*	*6 Segments*	*7 Segments*	*8 Segments*	*9 Segments*	*10 Segments*
Original Depression + Veteran's RAND 12-item Scales						
Latent Class Analysis	6%	7%	7%	8%	**4%**	5%
K-Means Clustering	56%	52%	54%	54%	**43%**	90%
Dichotomization of Depression Scale						
Latent Class Analysis	10%	9%	9%	9%	**8%**	8%
K-Means Clustering	31%	**29%**	46%	44%	41%	41%

Segment class profiles for each of the winning solutions + the solution with an additional segment class were compared using heat maps for each of the classification methods and conditions. The heat map is a table in Excel where each row of responses (within variable, across segment classes) is shaded according to its relationship to the 50th percentile, as compared to the sample mean (see Figure 9 for an example). As can be seen in Figure 9, the lowest value in the row is shaded red, values around the 50th percentile are shaded yellow, and the highest value in the

row is shaded green. The values in between the maximum, 50th percentile, and minimum are shaded according to a gradient color scale. As demonstrated in Figure 9, the mean response for Age=1 is 14.94, which is shaded in yellow. The highest value is represented by Segment 4 (Age 1= 38.69%), which is shaded in green. The lowest value is contained within Segment 1 (Age 1=4.51%), which is shaded in red. This approach allows for a quick, visual inspection of similarities and differences within segment classes and across segment class solutions.

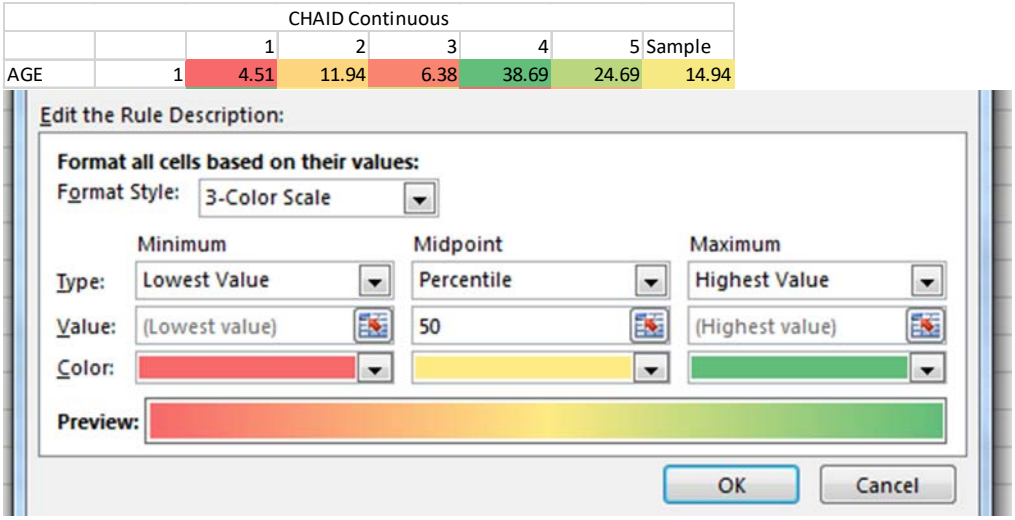


Figure 9. Illustration of Conditional Formatting for Heat Maps

4.1.2 K-Means

As demonstrated in Figure 10 below, the 6-class segment solution for the K-Means Dichotomous model provided adequate variability in responses across each segment class. For example, Segments 3 and 4 had essentially opposite response patterns, where Segment 3 was younger, more likely to be un-married, not obese and male. The 7-class solution for the K-Means Dichotomous

model resulted in an additional segment, Segment 4b, which was had little variation in responses and provided no discriminatory information.

		1	2	3	4	5	6	Sample
AGE	1	13.14	15.79	36.65	4.15	30.04	5.64	14.99
	2	44.19	45.11	28.67	67.85	40.36	55.44	50.54
	3	42.67	39.1	34.69	28	29.6	38.93	34.47
MRSTAT	1	48.69	40.05	39.74	60.94	41.7	56.6	51.28
	2	51.31	59.95	60.26	39.06	58.3	43.4	48.72
RACE	1	81.34	68.94	77.25	85.11	73.08	83.29	80.58
	2	11.51	19.66	13.96	8.37	16.95	9.47	11.62
	3	7.15	11.4	8.79	6.52	9.98	7.24	7.8
BMICAT	1	60.59	67.64	56.18	78.07	60.28	69.33	67.36
	2	39.41	32.36	43.82	21.93	39.72	30.67	32.64
EDUC	1	24.1	38.24	33.26	15.92	33.4	19.76	24.07
	2	36.14	31.8	34.58	33.8	34.96	35.9	34.76
	3	39.76	29.97	32.16	50.27	31.64	44.33	41.17
GENDER	1	39.85	42.48	38.2	47.21	40.23	43.09	42.7
	2	60.15	57.52	61.8	52.79	59.77	56.91	57.3

		1	2	3	4b	4	5	6	Sample
AGE	1	12.18	13.98	35.99	10.45	3.13	28.05	4.86	14.99
	2	44.67	45.02	29	58.09	69.85	41.16	56.23	50.54
	3	43.15	41	35	31.45	27.02	30.79	38.91	34.47
MRSTAT	1	50.32	40.39	39.81	48.31	63.35	41.77	58.06	51.28
	2	49.68	59.61	60.19	51.69	36.65	58.23	41.94	48.72
RACE	1	82.51	66.27	77.12	77.38	86.13	72.7	84.78	80.58
	2	10.82	21.67	14.08	13.4	7.66	17.18	8.47	11.62
	3	6.68	12.05	8.8	9.21	6.21	10.12	6.74	7.8
BMICAT	1	61.18	67.7	56.2	69.92	79.69	61.27	70.2	67.36
	2	38.82	32.3	43.8	30.08	20.31	38.73	29.8	32.64
EDUC	1	22.63	41.29	33.14	26.74	14.19	33.71	18.17	24.07
	2	36.73	29.78	34.54	34.82	33.33	34.69	35.79	34.76
	3	40.64	28.93	32.32	38.44	52.48	31.6	46.04	41.17
GENDER	1	40.6	43.28	38.19	40.45	48.56	40.26	44.03	42.7
	2	59.4	56.72	61.81	59.55	51.44	59.74	55.97	57.3

Figure 10. Heat Maps for Comparison of 6- and 7-Segment Class Solutions: K-Means Dichotomous

The K-Means full scale model with an additional segment class (10-Segment solution) produced two very similar segments- Segments 7 and 9, which provides no additional discriminatory information. The segments had almost identical response percentages for age, marital status, race, BMI, and gender. On the other hand, the 9-Segment class solution resulted in distinct segment classes with adequate variability in responses. While Segments 7 and 9 exhibited similar responses for the Race categories, they strongly differed in age (40% of Segment 9 was 75 or older, vs. 27% in Segment 7) and BMI (30% of Segment 7 is obese, vs. 20% in Segment 9).

		1	2	3	4	5	6	7	8	9	10	Sample
AGE	1	13.83	27.05	25.8	17.25	49.92	12.99	3.64	25.23	4.6	39.34	14.99
	2	42.04	45.17	50.68	38.03	29.29	44.96	68.98	36.81	55.17	25.51	50.54
	3	44.13	27.78	23.52	44.72	20.79	42.05	27.38	37.97	40.23	35.14	34.47
MRSTAT	1	48.6	39.99	36.25	44.33	35.83	38	61.72	46.91	58.4	38.68	51.28
	2	51.4	60.01	63.75	55.67	64.17	62	38.28	53.09	41.6	61.32	48.72
RACE	1	81.39	73.55	68.69	75.36	73.09	67.59	85.29	78.39	84.4	77.83	80.58
	2	11.32	16.63	19.37	15.28	17.05	22.36	8.19	13.21	8.9	13.75	11.62
	3	7.28	9.82	11.93	9.36	9.86	10.05	6.52	8.4	6.7	8.42	7.8
BMICAT	1	59.45	63.46	65.26	62.75	52.89	66.76	78.46	57.57	69.85	57.65	67.36
	2	40.55	36.54	34.74	37.25	47.11	33.24	21.54	42.43	30.15	42.35	32.64
EDUC	1	24.71	32.28	35.48	33.01	32.72	42.13	15.43	31.06	18.93	36	24.07
	2	36.11	34.74	32.9	35.74	34.21	31.22	33.73	34.95	36.04	33.44	34.76
	3	39.18	32.98	31.61	31.25	33.07	26.65	50.84	33.98	45.04	30.56	41.17
GENDER	1	40.39	38.68	37.96	39.27	36.93	42.64	47.43	40.94	44.7	38.95	42.7
	2	59.61	61.32	62.04	60.73	63.07	57.36	52.57	59.06	55.3	61.05	57.3

Figure 11. Demographic Distribution of K-Means Full Scale 10-Segment Solution

		1	2	3	4	5	6	7	8	9	Sample
AGE	1	13.64	27.19	29.65	19.15	47.85	13.03	3.65	26.47	4.75	14.99
	2	42.35	45.19	47.77	34.84	26.74	44.72	68.97	36.81	55.09	50.54
	3	44.01	27.62	22.58	46.01	25.41	42.25	27.38	36.72	40.15	34.47
MRSTAT	1	48.59	39.8	34.96	43.7	36.79	38.73	61.68	45.84	58.12	51.28
	2	51.41	60.2	65.04	56.3	63.21	61.27	38.32	54.16	41.88	48.72
RACE	1	81.06	73.43	67.42	77.06	75.06	68.21	85.29	77.75	84.21	80.58
	2	11.52	16.74	20.09	14.38	15.5	22.04	8.2	13.94	9	11.62
	3	7.43	9.82	12.49	8.56	9.44	9.74	6.5	8.31	6.79	7.8
BMICAT	1	59.79	63.52	63.9	61.65	54.3	67.12	78.46	57.29	69.76	67.36
	2	40.21	36.48	36.1	38.35	45.7	32.88	21.54	42.71	30.24	32.64
EDUC	1	24.82	32.13	37.26	33.58	34.25	40.98	15.46	31.1	19.11	24.07
	2	36.27	34.87	32.47	35.44	33.52	31.62	33.74	35.29	35.96	34.76
	3	38.91	33.01	30.27	30.98	32.23	27.4	50.8	33.62	44.93	41.17
GENDER	1	40.35	38.59	38.99	39.38	37.53	42.43	47.39	40.05	44.5	42.7
	2	59.65	61.41	61.01	60.62	62.47	57.57	52.61	59.95	55.5	57.3

Figure 12. Demographic Distribution of K-Means Full Scale 9-Segment Solution

4.1.3 Latent Class Analysis

The 10-segment LCA full-depression-scale solution resulted in several near-identical segments, as indicated in Figure 13 below. Segments 2 and 4 were very similar to each other; with exhibiting near-identical response patterns for age, marital status, race, BMI, education, and gender. Similarly, Segments 6, 7, and 9 exhibited many parallel responses with respect to age, marital status, race, and gender.

		1	2	3	4	5	6	7	8	9	10	Sample
AGE	1	24.59	2.38	7.34	4.78	11.1	31.51	43.89	9.07	51.4	19.17	14.83
	2	36.95	62.07	47.88	66.18	57.76	43.8	25.19	46.99	28.42	33.05	50.48
	3	38.46	35.56	44.78	29.04	31.15	24.7	30.91	43.94	20.19	47.78	34.69
MRSTAT	1	47.46	61.8	53.96	62.7	48.38	35.54	38.27	48.6	35.89	44.29	51.45
	2	52.54	38.2	46.04	37.3	51.62	64.46	61.73	51.4	64.11	55.71	48.55
RACE	1	77.12	88.56	84.47	84.86	77.47	69.5	76.05	77.52	72.15	80.6	80.34
	2	13.56	5.96	10.83	8.02	12.05	19.84	16.09	13.43	18.51	10.97	11.92
	3	9.32	5.47	4.7	7.12	10.47	10.66	7.86	9.05	9.34	8.43	7.73
BMICAT	1	55.1	73.49	61.59	77.53	74.01	65.14	54.37	67.32	56.93	59.75	67.17
	2	44.9	26.51	38.41	22.47	25.99	34.86	45.63	32.68	43.07	40.25	32.83
EDUC	1	28.79	14.53	24.68	15.13	26.57	35.25	34.84	28.27	33.49	30.97	24.78
	2	35.19	34.62	37.66	32.8	36.1	35.76	34.32	37.18	32.37	32.86	34.74
	3	36.02	50.85	37.66	52.07	37.33	28.99	30.85	34.55	34.14	36.16	40.49
GENDER	1	39.22	45.57	43.55	47.09	39.48	37.36	40.9	39.64	36.42	40.32	42.34
	2	60.78	54.43	56.45	52.91	60.52	62.64	59.1	60.36	63.58	59.68	57.66

Figure 13. Demographic Distribution of LCA Full Scale 10-Segment Solution

On the other hand, the 9-segment LCA full-depression-scale solution offers more variability in responses between segment classes. For example, respondents who are younger than 65 are differently distributed across each of the 9 segment classes (Figure 14). And while Segments 7 and 8 exhibit similar response patterns for marital status, race, and BMI, they differ in age, education, and gender.

		1	2	3	4	5	6	7	8	9	Sample
AGE	1	12.66	21.03	44.41	8.23	30.1	36.12	5.86	1.39	11.14	14.83
	2	43.11	32.12	24.58	46.29	47.38	35.9	65.42	61.67	59.95	50.48
	3	44.23	46.85	31.01	45.48	22.51	27.97	28.72	36.94	28.91	34.69
MRSTAT	1	47.08	46.5	37.42	52.17	37.09	41.13	63.23	60.86	49.11	51.45
	2	52.92	53.5	62.58	47.83	62.91	58.87	36.77	39.14	50.89	48.55
RACE	1	75.81	80.51	75.8	83.27	71.09	71.6	85.51	87.5	76.92	80.34
	2	15.61	12.08	16.24	11.69	18.88	19.51	7.46	5.93	11.81	11.92
	3	8.58	7.42	7.96	5.04	10.03	8.89	7.02	6.57	11.26	7.73
BMICAT	1	67.01	58.66	53.44	63.21	60	57.49	77.08	74.15	71.92	67.17
	2	32.99	41.34	46.56	36.79	40	42.51	22.92	25.85	28.08	32.83
EDUC	1	28.79	29.65	34.67	28.47	34.25	34.04	14.2	14.92	27.11	24.78
	2	38.22	35.49	32.51	37.43	35.91	33.1	32.93	33.97	33.76	34.74
	3	33	34.86	32.82	34.09	29.83	32.86	52.87	51.11	39.13	40.49
GENDER	1	39.24	40.12	41.4	41.17	39.2	39.37	47.51	42.99	39.22	42.34
	2	60.76	59.88	58.6	58.83	60.8	60.63	52.49	57.01	60.78	57.66

Figure 14. Demographic Distribution of LCA Full Scale 9-Segment Solution

Similar results were found using the dichotomous depression scale as an input. The 10-segment solution resulted in 2 pairs of segments with similar response patterns. Segments 4 & 9 and 5 & 7 are very similar, reporting almost identical response percentages for age, marital status, race, BMI, education, and gender. For example, 52% of respondents were female in both segments 4 and 9, along with 63% married, and 52% college educated in both groups.

		1	2	3	4	5	6	7	8	9	10	Sample
AGE	1	7.77	8.82	24.02	5.43	20.09	39.52	20.6	4.6	2.41	40.7	15
	2	53.46	47.29	45.53	65.13	36.05	31.49	43.64	66.07	60.91	26.23	50.61
	3	38.78	43.89	30.45	29.44	43.86	28.98	35.76	29.33	36.69	33.08	34.39
MRSTAT	1	50.5	51.73	37.97	63.71	44.66	41.36	42.82	57.04	63.5	39.58	51.1
	2	49.5	48.27	62.03	36.29	55.34	58.64	57.18	42.96	36.5	60.42	48.9
RACE	1	79.64	83.8	67.44	85.96	80.48	74.23	74.48	83.9	88.85	76.88	80.5
	2	11.11	10.95	19.23	7.99	12.44	16.37	15.7	8.33	6	15.25	11.66
	3	9.25	5.25	13.33	6.05	7.08	9.4	9.82	7.77	5.15	7.88	7.84
BMICAT	1	71.9	62.79	64.58	80.36	56.52	55.6	63.85	76.35	71.7	54.95	67.23
	2	28.1	37.21	35.42	19.64	43.48	44.4	36.15	23.65	28.3	45.05	32.77
EDUC	1	22.83	22.48	41.12	15.05	26.65	32.03	32.8	16.52	12.61	36.35	24.19
	2	37.36	37.69	30.58	32.74	34.82	34.28	36.18	33.69	34.97	33.02	34.57
	3	39.81	39.83	28.3	52.22	38.54	33.69	31.02	49.79	52.42	30.64	41.24
GENDER	1	41.94	41.41	41.27	47.53	39.5	38.28	39.11	42.05	47.18	41.04	42.51
	2	58.06	58.59	58.73	52.47	60.5	61.72	60.89	57.95	52.82	58.96	57.49

Figure 15. Demographic Distribution of LCA Dichotomized Scale 10-Segment Solution

And while the 9-segment solution (the "best" solution) resulted in segments that were similar in marital status, race, BMI, and education, they did exhibit differences in age and gender (Segments 3 & 6; 5 & 7). Of note, these differences may not be meaningful enough for healthcare and marketing researchers, and they could ultimately decide to combine segment classes for targeted efforts.

		1	2	3	4	5	6	7	8	9	Sample
AGE	1	13.75	21.23	39.4	20.61	2.96	38.93	4.72	7.71	8.18	15
	2	55.5	44.67	26.66	35.79	60.4	32.1	66.31	53.53	48.77	50.61
	3	30.75	34.1	33.94	43.59	36.64	28.97	28.97	38.76	43.05	34.39
MRSTAT	1	43.4	43.7	38.35	44.26	61.98	40.23	62.5	51.28	54.01	51.1
	2	56.6	56.3	61.65	55.74	38.02	59.77	37.5	48.72	45.99	48.9
RACE	1	72.63	73.24	76.73	80.31	88.21	74.8	85.51	80.03	83.01	80.5
	2	16.11	16.33	14.25	12.44	6.45	16.46	7.91	10.18	11.18	11.66
	3	11.26	10.42	9.02	7.25	5.35	8.74	6.58	9.79	5.81	7.84
BMICAT	1	71.15	63.29	56.97	57.31	71.25	55.7	79.91	72.25	62.45	67.23
	2	28.85	36.71	43.03	42.69	28.75	44.3	20.09	27.75	37.55	32.77
EDUC	1	31.06	33.06	37.14	26.86	13.85	33.1	14.45	22.83	22.26	24.19
	2	34.01	35.23	32.69	35.48	34.89	33.55	33.4	35.28	37.24	34.57
	3	34.92	31.7	30.17	37.66	51.26	33.35	52.14	41.89	40.5	41.24
GENDER	1	40.61	39.6	40.23	40.04	45.06	38.05	46.76	42.55	42.25	42.51
	2	59.39	60.4	59.77	59.96	54.94	61.95	53.24	57.45	57.75	57.49

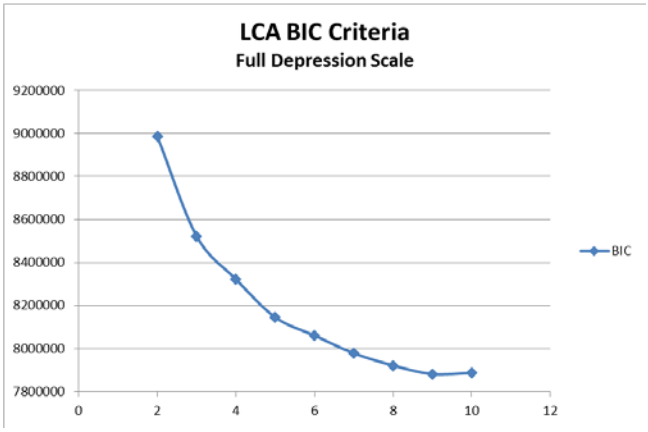
Figure 16. Demographic Distribution of LCA Dichotomized Scale 9-Segment Solution

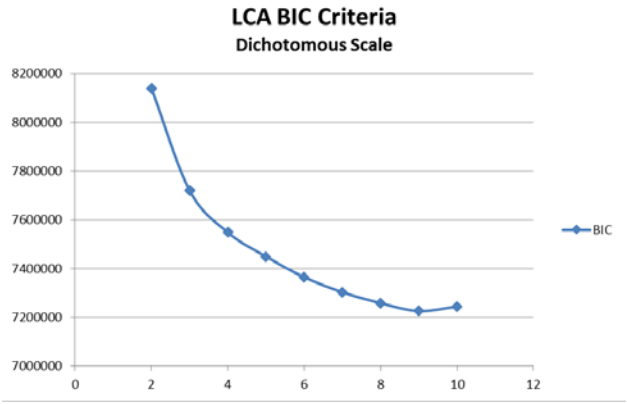
4.1.4. Selection of the preferred model solution

When the original depression scale plus the Veteran's RAND 12-item scale were used to segment the Medicare HOS respondents, both LCA and K-Means clustering resulted in a 9-segment best class solution. LCA also produced a 9-class solution when the depression scale was dichotomized; while K-Means resulted in only 6 segments (Table 10). The best segment class solutions were selected when the BIC criteria were minimized (LCA) and gap statistics were maximized (K-Means). In addition, the best segment class solutions resulted in segments in which the distance between clusters were maximized (Figure 17). The size of the segment classes within each solution were similar by method (K-Means/LCA) and scale (full/dichotomous), as illustrated in Figure 18. The segment class memberships ranged from five to twenty-nine percent.

Table 10. Best Segment Class Solution for Classification Methods

Method	*# of Segments*	*Gap Statistic (BIC)*
Original Depression + Veteran's RAND 12-item Scales		
Latent Class Analysis	9	(7880830)
K-Means Clustering	9	.16
Dichotomization of Depression Scale		
Latent Class Analysis	9	(7225709)
K-Means Clustering	6	.23





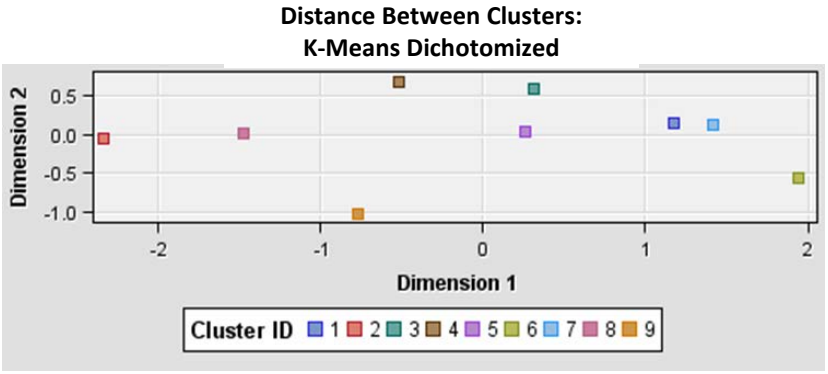
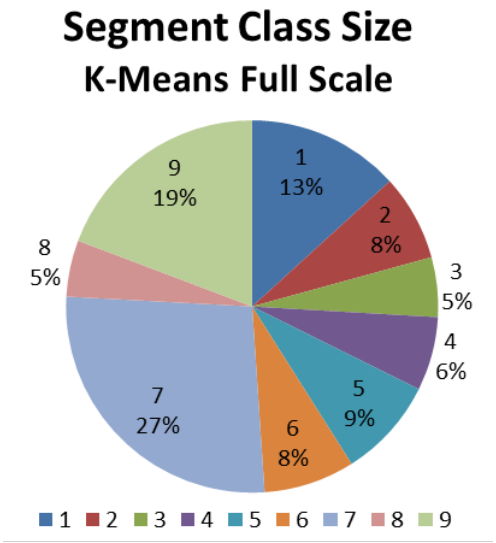
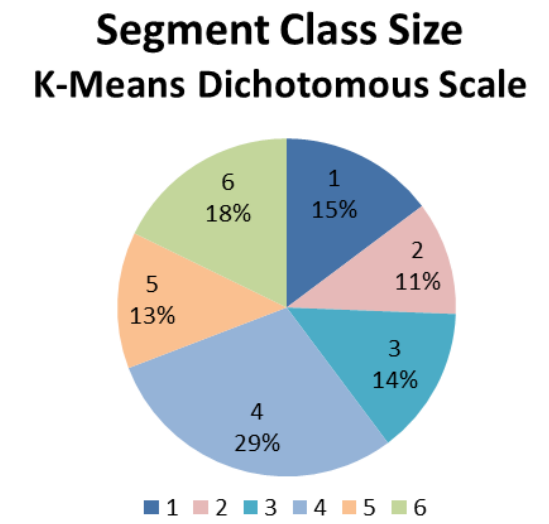
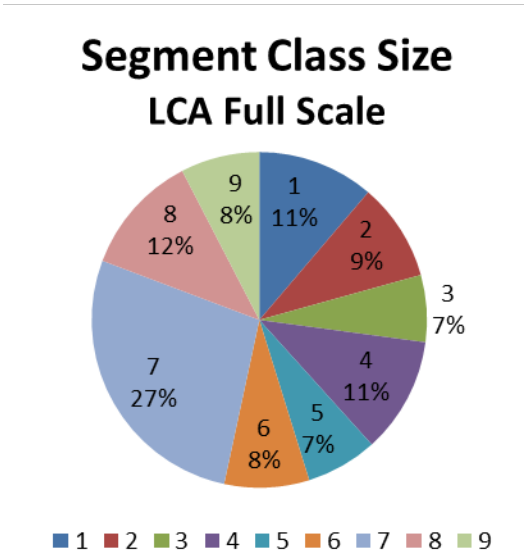


Figure 17. BIC Criteria and Cluster Distances







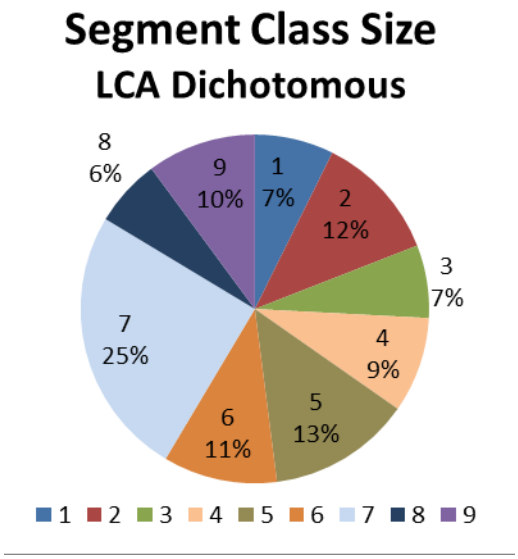


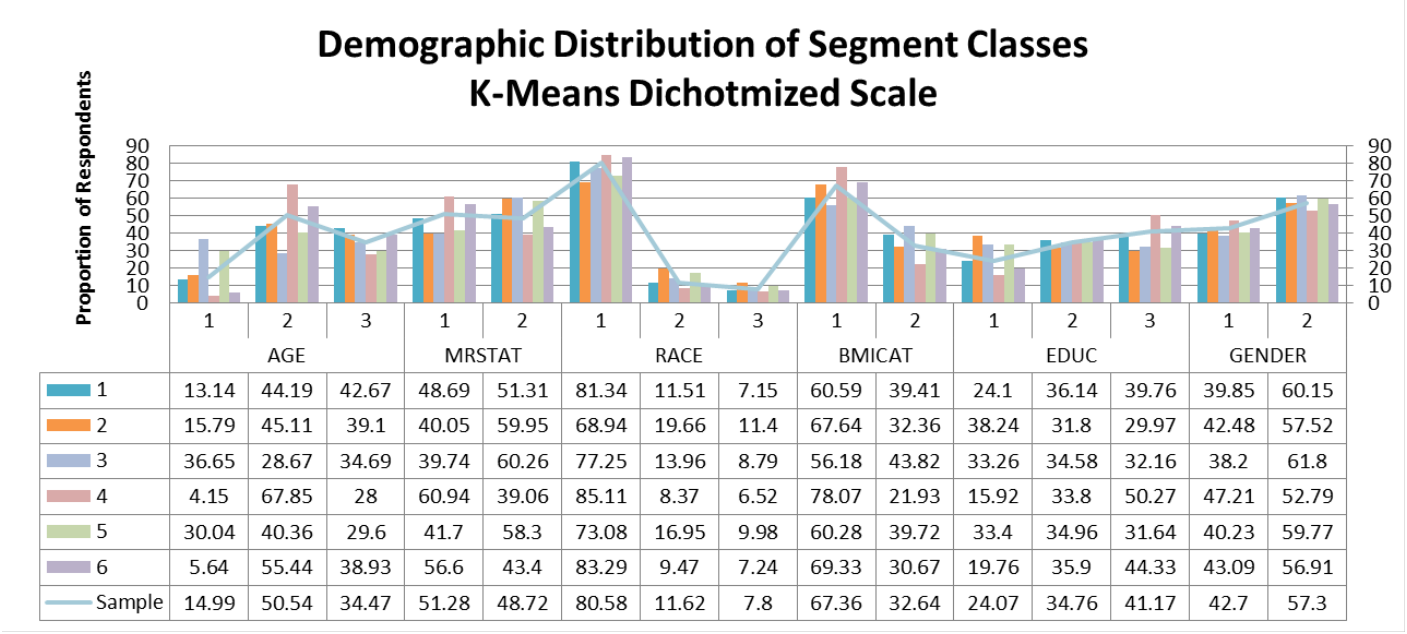
Figure 18. Segment Class: Member Distribution for Classification Models

4.1.5 **Segment class profiles**

Each of the methods (K-Means and LCA) resulted in segment classes with appropriate demographic variability (with the exception of Education =2; High school education or GED). Demographic variability was assessed by comparing the mean proportion of respondents in each class using heat maps and additional graphical visualizations, as discussed in Section 4.1 and demonstrated in Figure 19 below, which displays the same information as heat maps, but graphically.

For example, the K-Means dichotomous model yielded 6 very different segment classes. Segment 1 was the oldest segment, mildly unhealthy but non-smokers, and skewed female. Segment 2 was the least educated, but most ethnically diverse, and the majority of the responses were evenly distributed across categories. Segment 3 was the youngest segment, most likely to be obese, highest proportion of female respondents, higher proportion of depression and more likely to have pain that interferes with their lives. They are also heavy smokers and had the highest reported mean number of unhealthy days (~25/30). In addition, this segment had the highest proportion of unmarried respondents. Segment 4 was the healthiest segment, with the fewest number of unhealthy days per month (~1), but also the oldest, most educated, and most likely to be male and married. This segment was also the least likely to be obese, and over 90% have never smoked. Segment 5 was also a younger segment and more likely to be depressed, but not as severe as Segment 3. This segment could likely become Segment 3, as their health declines. Finally, Segment 6 represents a highly educated segment, of average health, with moderate energy and activity levels. This group are less likely to be smokers.

The full collection of graphs is available in Appendix H. A comparison of each method by input scale variation (full-scale vs. dichotomized) is presented in Section 4.1.5.



	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
1	13.14	44.19	42.67	48.69	51.31	81.34	11.51	7.15	60.59	39.41	24.1	36.14	39.76	39.85	60.15
2	15.79	45.11	39.1	40.05	59.95	68.94	19.66	11.4	67.64	32.36	38.24	31.8	29.97	42.48	57.52
3	36.65	28.67	34.69	39.74	60.26	77.25	13.96	8.79	56.18	43.82	33.26	34.58	32.16	38.2	61.8
4	4.15	67.85	28	60.94	39.06	85.11	8.37	6.52	78.07	21.93	15.92	33.8	50.27	47.21	52.79
5	30.04	40.36	29.6	41.7	58.3	73.08	16.95	9.98	60.28	39.72	33.4	34.96	31.64	40.23	59.77
6	5.64	55.44	38.93	56.6	43.4	83.29	9.47	7.24	69.33	30.67	19.76	35.9	44.33	43.09	56.91
Sample	14.99	50.54	34.47	51.28	48.72	80.58	11.62	7.8	67.36	32.64	24.07	34.76	41.17	42.7	57.3

Figure 19. Demographic Distribution of K-Means Dichotomous Scale

4.1.6 Variable importance for the preferred model

The differences in the variables most important for the segmenting of classes is more a function of classification method type (i.e., LCA vs. K-Means) than scale (Table 11). By design, categorical variables of multiple levels must be dummy coded prior to entering a K-Means model; whereas LCA methods can handle multiple categorical levels of a single variable. Thus, multiple levels of a variable were reported in the output of the K-Means model, as compared to a single variable for LCA, and by default, K-Means models could have more significant "important" variables.

The lists of important variables derived from each method, regardless of the treatment of the depression scale, were almost identical within the classification methods. (LCA compared to LCA and K-Means compared to K-Means). This finding was expected, due to the nature of the methods, as described above. The main differences across methods (LCA compared to K-Means),

were a function of the physical health variables. For example, 4/5 (90%) important LCA variables dealt with emotional health; whereas only 4/10 (40%) important K-Means variables involved an emotional health component. The important variables for segmenting respondents using the K-Means methods contained a mix of both physical and emotional health behaviors.

Within Enterprise Miner, variable importance is automatically provided when using the segmentation analysis node. Variables are listed in descending order, according to their worth (discriminating power). In order to identify significant variables in R, the maximum probability of response was examined, which provides an indication of variable worth. Then, variables containing the maximum number of classes in which the probability of response was over .25 were selected. See Appendix I for a visualization of LCA conditional probabilities by segment class.

Table 11. Significant Input Variables for Classification Methods

	Full Scale			
Variables in Order of Importance	*LCA*		*K-Means*	
	Variable Name	*Description*	*Variable Name*	*Description*
1	C15DEP2WK	2 weeks of depression	C15VRPWORK01	Physical health not interfering with work
2	C15DEP2YR	Depressed for 2 years	C15VRPACCL01	None of the time accomplish less due to physical health
3	C15VRMACT	Emotional health interferes with activities	C15VRMACT03	Emotional health limiting activities some of the time
4	C15VRPACCL	Accomplished less due to physical health	C15VRPAIN01	Pain not interfering with work
5	C15VRMWORK	Emotional health interferes with work	C15VRSTAIR03	Not limited in climbing stairs

Table 11 continued

6			C15VRSACT05	None of the time health interferes with social activities
7			C15VRMACCL01	None of the time accomplish less due to emotional health
8			C15VRMWORK01	Emotional health not interfering with work
9			C15VRENERGY02	Most of the time have lots of energy
10			C15VRDOWN06	None of the time felt downhearted and blue
		Dichotomous Scale		
Variables in Order of Importance	*LCA*		*K-Means*	
	Variable Name	*Description*	*Variable Name*	*Description*
1	DEPRESS1	Dichotomous depression scale	C15VRPWORK01	Physical health not interfering with work
2	C15VRMACT	Emotional health interferes with activities	C15VRPACCL01	None of the time accomplish less due to physical health
3	C15VRMWORK	Emotional health interferes with work	C15VRMACT03	Emotional health limiting activities some of the time
4	C15VRSACT	Health interferes with social activities	C15VRSTAIR03	Not limited in climbing stairs
5			C15VRPAIN01	Pain not interfering with work
6			C15VRSACT05	None of the time health interferes with social activities
7			C15VRMACCL01	None of the time accomplish less

Table 11 continued

		due to emotional health
8	C15VRMWORK01	Emotional health not interfering with work
9	C15VRENERGY02	Most of the time have lots of energy
10	C15VRDOWN06	None of the time felt downhearted and blue

4.1.7 Summary comparison of classification models across variable types

4.1.7.1 K-Means full scale vs. dichotomous

When the original depression scale was included, the K-Means analysis resulted in a 9-segment solution; compared to a 6-segment solution when the scale was dichotomized, with similar profiles for 4 segments across the scales. As previously indicated, the 9-segment solution could be reduced to 6 segment classes. Both scales (full scale vs. dichotomous) resulted in identical profiles for Segment 1. For example, when Age=1 (Less than 65), Segment 1 has 13% respondents for both full scale and dichotomous; when BMI=1 (Not obese), both scales resulted in 60% response; Education=1 (Less than high school or GED) has 24% response, when Race =1 (White), both scales resulted in 81% of respondents; 48% of respondents in both groups were married, and when Gender=1 (Male), both the full scale and dichotomous inputs resulted in 40% of respondents. In addition, Segment 1 looked similar on several of the profiling questions. Both full scale and dichotomous resulted in 88% non-smokers and 10.5 (full scale) and 9.2 (dichotomous) mean number of unhealthy days. Similar patterns were identified for Segment 5 (full scale) and Segment 3 (dichotomous), with the exception of depression (respondents in the dichotomous model were more likely to be depressed), as demonstrated in Table 12.

Table 12. Distribution of Demographic Variables for K-Means Models

	Segment 1		*Segment 5/3*	
	Full Scale	***Dichotomous***	***Full Scale***	***Dichotomous***
When AGE=1 (Less than 65)	13.64%	13.14%	47.85%	36.65%
When EDUC=1 (Less than HS/GED)	24.82%	24.1%	34.25%	33.26%
When Gender=1 (Male)	40.35%	39.85%	37.53%	38.2%
When RACE=1 (White)	81.06%	81.34%	75.06%	77.25%
When BMI=1 (Not Obese)	59.79%	60.59%	54.3%	56.18%
When MRSTAT=1 (Married)	48.59%	48.69%	36.79%	39.74%
When Depressed=1 (Depressed)	83.57%	77.41%	2.03%	22.58%
When Smoking=1 (Non-smoker)	7.76%	7.85%	16.85%	13.77%
Mean Number of Unhealthy Days	10.56%	9.22%	23.46	22.47

4.1.7.2 LCA full scale vs. dichotomous

Both the full and dichotomous depression scale resulted in 9 segments, with very similar profiles for 5 out of 9 segments. For example, Segment 1 (full scale) is demographically very similar to Segment 1 (dichotomous), along with Segments 2, 3, 6, and 7, as indicated in Table 13 below. The differences in age, marital status, education, and gender by scaling method differs only between less than 1 and 5%. For example, the proportion of married respondents in Segment 3 for the full-scale K-Means method is 37%; the dichotomous method classifies 38% of respondents as married.

Table 13. Distribution of Age, Marital Status, Race, BMI, Education, and Gender for LCA Models

	Segment 1	*Segment 2*	*Segment 3*	*Segment 6*	*Segment 7*
Age = 1 (Less than 65)					
Full Scale	12.66%	21.03%	44.41%	36.12%	5.86%
Dichotomous Scale	13.75%	21.23%	39.4%	38.93%	4.72%
MRSTAT= 1 (Married)					
Full Scale	47.08%	46.5%	37.42%	41.13%	63.23%
Dichotomous Scale	43.4%	43.7%	38.35%	40.23%	62.5%
RACE= 1 (White)					
Full Scale	75.81%	80.51%	75.8%	71.6%	85.51%
Dichotomous Scale	72.63%	73.24%	76.73%	74.8%	85.51%
BMI= 1 (Not Obese)					
Full Scale	67.01%	58.66%	53.44%	57.49%	77.08%
Dichotomous Scale	71.15%	63.23%	56.97%	55.7%	79.91%
EDUC= 1 (Less than high school/GED)					

Table 13 continued

Full Scale	28.79%	29.65%	34.67%	34.04%	14.2%
Dichotomous Scale	31.06%	33.06%	37.14%	33.1%	14.5%
GENDER= 1 (Male)					
Full Scale	39.24%	40.12%	41.4%	39.37%	47.51%
Dichotomous Scale	40.16%	39.6%	40.23%	38.05%	46.76%

4.1.7.3 K-Means vs. LCA (full scale)

Both K-Means and LCA methods utilizing a full scale resulted in a 9-segment solution, with Segments 1-5 being very similar to each other, according to the method (i.e., K-Means Segment 1 = LCA Segment 1, K-Means Segment 2= LCA Segment 2, etc.). K-Means and LCA yielded similar age, education, marital status, and gender response profiles for Segments 1,2,3,4 & 5 (Table 14). When respondents have less than a high school education, the difference in proportion of respondents classified by K-Means and LCA is a maximum of 3%.

Table 14. Distribution of Demographic Variables for K-Means and LCA Full-Scale Models

When AGE=1 (Less than 65)		
Segment	*K-Means*	*LCA*
1	12.66%	13.64%
2	21.03%	27.9%
3	44.4%	29.65%
4	8.23%	19.15%
5	30.1%	47.85%
When EDUC=1 (Less than HS or GED)		
Segment	*K-Means*	*LCA*
1	28.79%	24.82%
2	35.49%	32.13%
3	34.67%	37.26%
4	28.47%	33.58%
5	34.25%	34.25%
When MRSTAT=1 (Married)		
Segment	*K-Means*	*LCA*
1	48.59%	47.08%
2	39.8%	46.5%
3	34.96%	37.42%
4	43.7%	52.17%
5	36.79%	37.09%
When GENDER=1 (Male)		
Segment	*K-Means*	*LCA*

Table 14 continued

1	40.35%	39.24%
2	38.59%	40.12%
3	38.99%	41.4%
4	39.38%	41.7%
5	37.53%	39.2%

On the other hand, the segments were more differentiated in terms of the profiling questions. For example, LCA resulted in more variability in the mean number of unhealthy days. In addition, LCA resulted in a clear differentiation in depression across segments (90% of respondents were either depressed or not depressed). Both methods were able to identify a healthy segment (Segment 7 in each). This segment never smoked, had the lowest probability of depression, and reported the fewest number of unhealthy days

4.2 COMPARISON OF PREDICTION MODELS

Unlike classification methods, Neural Networks and CHAID models do not allow for pre-specified cluster solutions. Therefore, the best segment class solution is typically produced by the model output within Enterprise Miner. The software chooses a best fitting model by considering gain/lift statistics. The cumulative gains chart represents the respondents in the highest decile of number of unhealthy days, where the greater the area between the lift curve and the baseline, the better the model. Similarly, the cumulative % Response lift chart identifies 10% of the individuals who are most likely to have a high proportion of unhealthy days, plotted against the actual cumulative response rate in each decile. A disadvantage of these models is that they often produce solutions containing many segments (in an upwards of 100), but with very small proportions of respondents

making up the majority of the additional segments. Therefore, classes with <5% membership were combined into a common segment.

The variables from the HOS dataset that were used in the analysis of prediction models included a subset of the full dataset (variables with missingness less than 15%): C15VRGENHTH (general health status), C15VRMACT (moderate activities limited), C15VRSTAIR (climbing stairs limited), C15VRPACCL (accomplishing less due to physical health), C15VRPWORK (work or activities limited by physical health), C15VRMACCL (accomplishing less due to emotional health), C15VRMWORK (work or activities limited by emotional health), C15VRPAIN (degree pain interfered with normal work), C15VRCALM (felt calm and peaceful), C15VRENERGY (energy level), C15VRDOWN (felt downhearted and blue), C15VRSACT (amount of time health interfering with social activities), C15VRPHCMP (physical health compared to 1 year ago), C15VRMHCMP (emotional health compared to 1 year ago), C15ADLBTH (difficulty bathing), C15ADLDRS (difficulty dressing), C15ADLEAT (difficulty eating), C15ADLCHR (difficulty getting in or out of chairs), C15ADLWLK (difficulty walking), C15ADLTLT (difficulty using toilet), C15HDMEN (number of days mental health not good), C15HDACT (number of days poor health interfered w/activities), C15CHSTEX (chest pain-exercise), C15CHSTRST (chest pain-resting), C15SOBFLT (short of breath lying flat), C15SOBSIT (short of breath sitting or resting), C15SOBWLK (short of breath walking less than 1 block), C15SOBSTR (short of breath climbing 1 flight stairs), C15FTNUMB (numbness or loss of feeling in feet), C15FTSENS (tingling burning in feet), C15FTHC (decreased feeling of hot or cold in feet), C15FTSRS (sores that do not heal on feet), C15PNART (arthritis pain), C15READ (can see to read newspaper), C15HEAR (can hear most things), C15CCHBP (hypertension or high blood pressure), C15CC_CAD (angina pectoris or coronary artery disease), C15CC_CHF (congestive heart failure), C15CCMI

(myocardial infarction or heart attack), C15CCHRTOTH (other heart conditions), C15CCSTROKE (stroke), C15CC_COPD (emphysema), C15CCGI (inflammatory bowel diseases), C15CCARTHIP (arthritis of hip or knee), C15CCARTHND (arthritis of hand or wrist), C15CCOSTEO (osteoporosis), C15CCSCIATI (sciatica), C15CCDIABET (diabetes), C15CCANYCA (any cancer, other than skin cancer) , C15PNBACK (back pain), C15DEP2WK (sad/blue for two + weeks in past year), C15DEPYR (depressed for much of past year), C15DEP2YR (depressed for two + years), C15DEPWEEK (depressed during past week), C15CMPHTH (general health compared to peers), C15SMOKE (smoke every day), C15MUILKG (urine leakage) C15PAOTLK (talked with doctor about physical activities), C15PAOADV (advised to increase or maintain activities), C15FRMTLK (talked with doctor about falling or balance problem), C15FRMFALL (fell in past 12 months), C15FRMBAL (problem with walking or balance in past 12 Months), C15FRMPREV (talked with doctor about how to prevent falls), C15OTOTEST (had a bone density test for osteoporosis); with the demographic variables of age, race, gender, marital status, education, BMI, and smoking history used as profiling variables (after the segmentation models were completed).

4.2.1 Selection of the preferred model solution

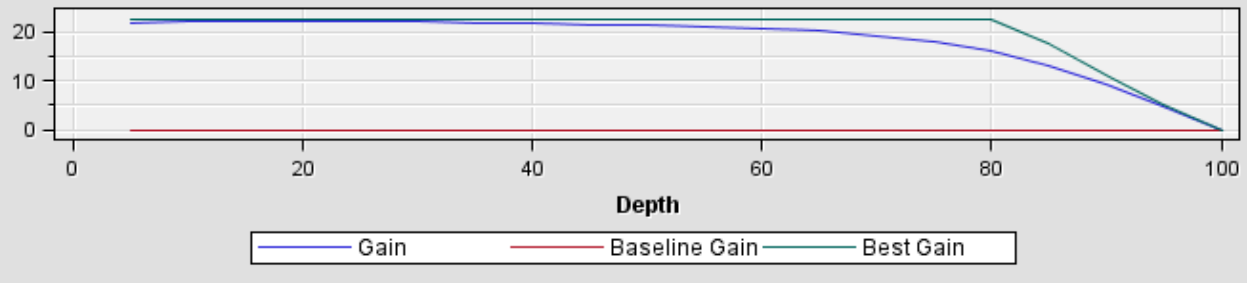
Neural Networks and CHAID models resulted in identical segment class solutions (number of segments= 4) when the outcome was dichotomous. The appropriate solution was chosen based on the model that minimized the differences between the estimator (prediction) and what is being estimated (actual) in terms of the average squared error when the outcome was continuous, and the effectiveness of the predictive model (largest % gain and lift when the outcome was dichotomous), as demonstrated in Table 15 and Figure 20. When the outcome was continuous,

both methods produced additional segments, resulting is a 5-segment (CHAID) and 6-segment (Neural Network) solution. The sizes of the segment classes are similar across methods, in the fact that each method produces a solution with a fairly large segment class (ranging from 34-67% membership), with several smaller segment classes. This is due partially to the fact that both Neural Networks and CHAID models produce solutions of many segments (in an upwards of 100), but with very small proportions of respondents making up the majority of the additional segments. Classes with <5% membership were combined into a common segment.

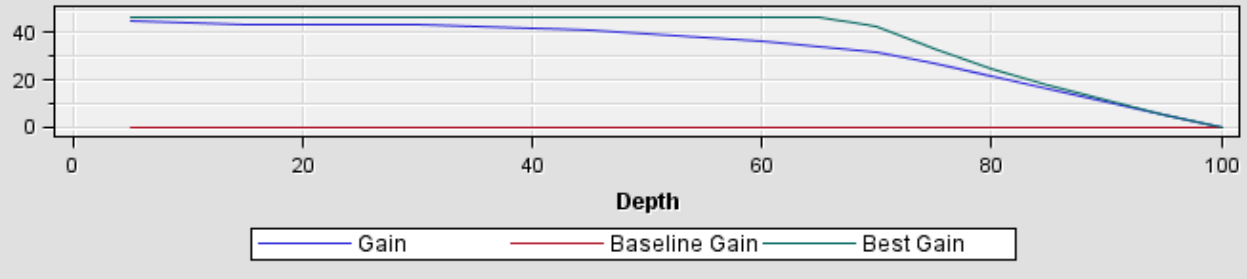
Table 15. Best Segment Class Solution for Prediction Methods

Method	*# of Segments*	*%Gain (Avg. Sq. Error)*
Continuous Outcome		
CHAID	5	(.48)
Neural Networks	6	(.52)
Dichotomous Outcome: Even Split		
CHAID	4	19%
Neural Networks	4	22%
Dichotomous Outcome: Uneven Split		
CHAID	4	43%
Neural Networks	4	45%

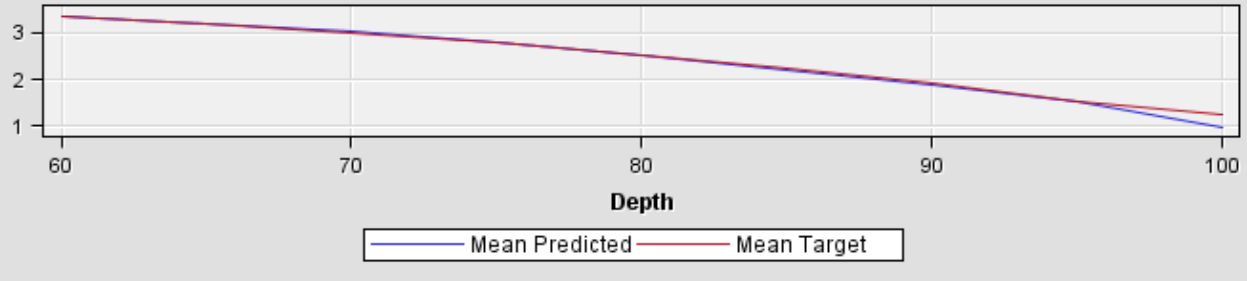
Neural Network Dichotomous: Even Split



Neural Network Dichotomous: Uneven Split



Neural Network Continuous



CHAID Dichotomous: Even Split

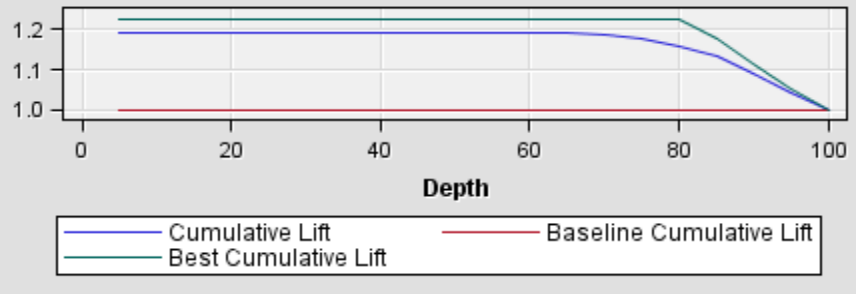
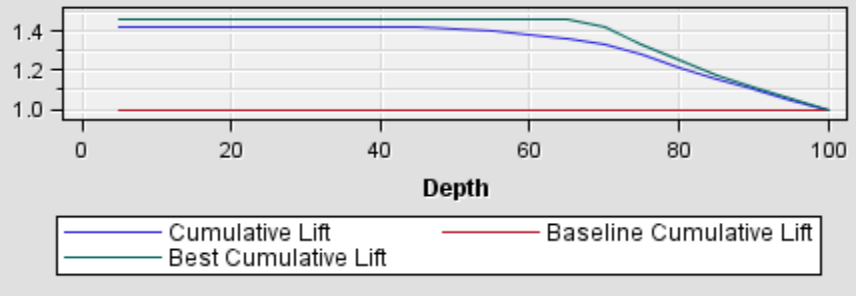


Figure 20. Gain/Lift Charts

CHAID Dichotomous: Uneven Split



CHAID Continuous

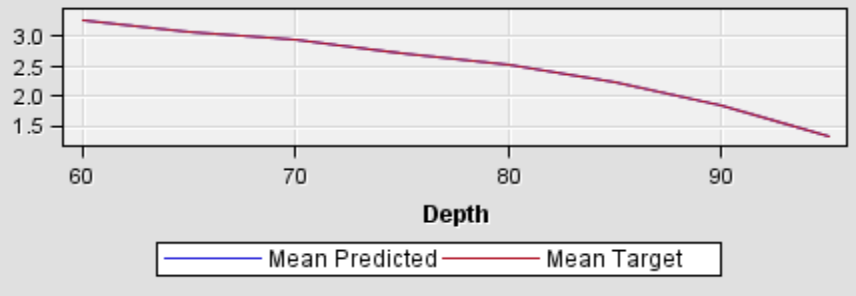


Fig. 20: Gain/Lift Charts continued

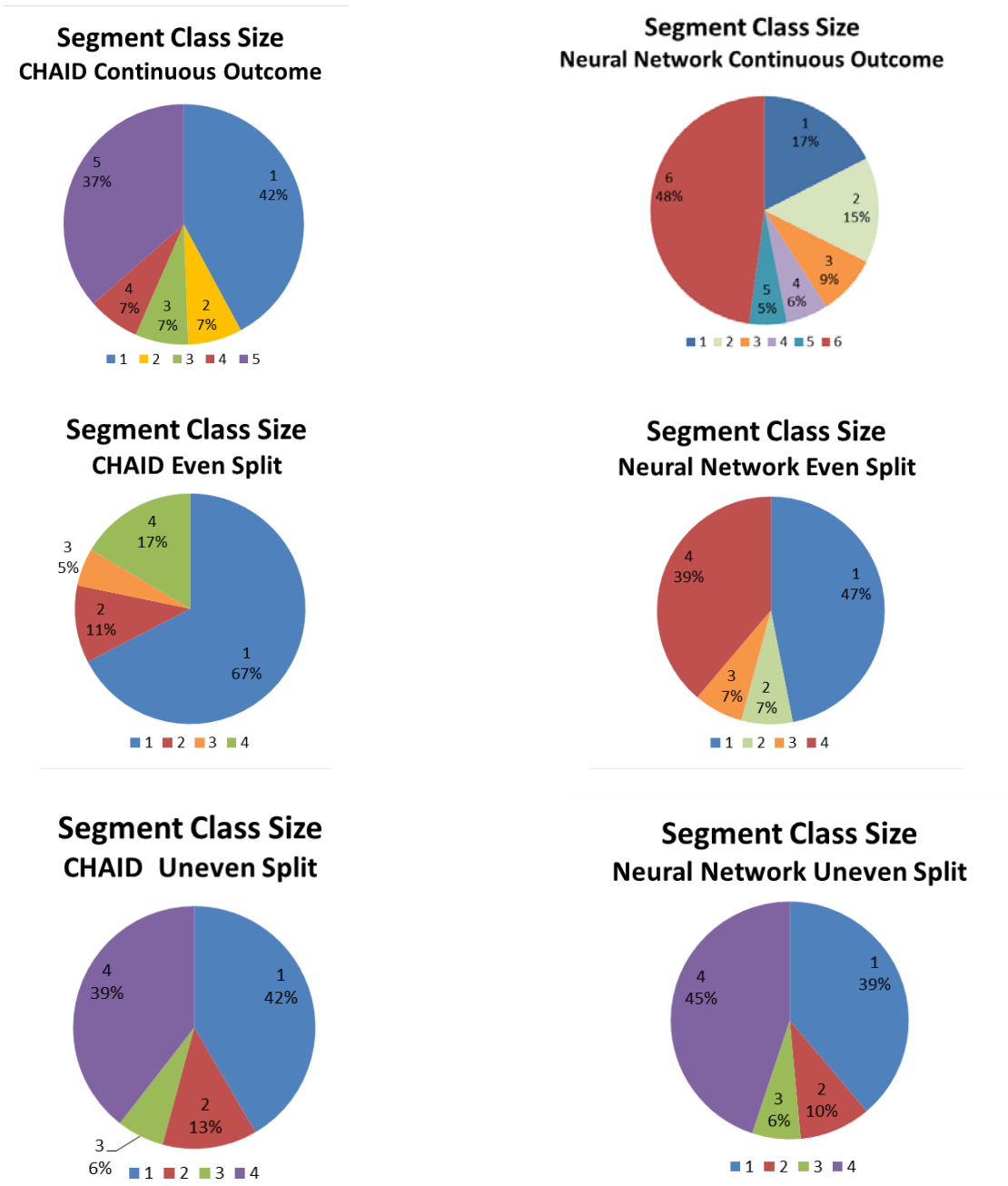


Figure 21. Segment Class Member Distribution for Prediction Models

4.2.2 Segment class profiles

The same method for examining segment class profiles of classification methods (comparing the mean proportion of respondents in each class using heat maps/graphs) was employed in comparing

segments from prediction methods. For example, when the outcome was dichotomous and evenly split (#unhealthy days ≥ 15), the CHAID method was able to identify 4 distinct segments. Segment 1 was again the healthiest segment; older and more male, with higher levels of education. They were predominantly white and most likely to married. This segment was also the least likely to be depressed or have smoked, and reported that pain does not generally interfere with their daily activities. On the other hand, Segment 2 was the unhealthiest segment, with the highest mean number of unhealthy days (~25). This segment also contained the highest proportion of smokers, and those likely to be depressed (almost 80%) and unmarried (59%). Segment 2 was the youngest segment, and more likely to be female and obese. This segment was also the least educated, with 33% having less than a high school education. Segment 3 contained the highest proportion of women (~65%), and while they reported being obese, having less energy, and being in pain, they were less likely to be depressed. They were also the most ethnically diverse. Segment 4 looked similar to Segment 2 in terms of depression (~68% depressed) and mean number of unhealthy days (~17), but had a lot of variability in terms of demographics. This group was evenly distributed across each of the demographic variables and levels.

A full comparison of prediction models across variable types is presented in Section 4.2.6.

4.2.3 **Model fit/accuracy**

Misclassification rates for prediction models with dichotomous outcomes were assessed using the mean proportion of misclassified observations, *predicted – actual*. Neural Networks and CHAID models performed similarly, in terms of model accuracy, when the outcome is dichotomous, as indicated in Table 16 below. Since misclassification rates can't be obtained for continuous

outcomes, the average square error is used to represent model fit. For the continuous outcome, the CHAID model resulted in slightly smaller error (.48), as compared to the Neural Network model (0.52). As indicated in Section 4.2.1, the best segment class solutions were identified according to splitting criteria and model fit. CHAID and Neural Networks had similar gain/lift statistics when the outcome was dichotomous (within +/- 3% points for both splitting criteria).

Table 16. Misclassification Table for Prediction Models

Method	*%Misclassified (Avg. Sq. Error)*
Continuous Outcome	
CHAID	(.48)
Neural Networks	(.52)
Dichotomous Outcome: Even Split	
CHAID	9%
Neural Networks	10%
Dichotomous Outcome: Uneven Split	
CHAID	9%
Neural Networks	10%

4.2.4 Variable importance

For the dichotomous outcome, the same 7 variables were identified as the best predictors across method (CHAID vs. Neural Network) and splitting criteria (even vs. uneven), as reported in Table 17. These variables include: C15VRPAIN (pain interfering with activities), C15ADLWLK (difficulty walking), C15CMPHTH (general health compared to peers), C15HDACT (number of days poor health interfered w/activities), C15VRGENHTH (general health status), C15HDMEN (number of days mental health not good), and C15SRVDISP (baseline survey disposition). Additional variables were identified in the even split condition, such as C15VRPWORK (work or activities limited by physical health) and C15VRPHCMP (physical health compared to a year ago. When the outcome is continuous, six variables were identified as significant in predicting segment

classes for the CHAID model; 8 for the Neural Network. Two variables were unique to the Neural Network model C15SRVDISP (baseline survey disposition) and C15PCTCMP (percent of survey completed); whereas information regarding a respondent's survey style was not important in differentiating respondents according to the CHAID model.

Across methods and IV's, variables related to physical health proved to be the most impactful with respect to segmenting respondents on the dependent outcome- number of unhealthy days. Only two mental health variables made the significance cut- days mental health not good and depressed for 2+ weeks. In addition, the significant input variables did not differ greatly across methods and conditions, indicating similar segmenting components for all cases.

Table 17. Significant Input Variables for Prediction Methods

Dichotomous Outcome				
		Even Spilt		
Variables in Order of Importance	*CHAID*	*Variable Description*	*Neural Network*	*Variable Description*
1	C15VRPAIN	Pain interfering with activities	C15VRPAIN	Pain interfering with activities
2	C15HDACT	Health interfering with activities	C15ADLWLK	Difficulty walking
3	C15HDPHY	Days physical health not good	C15VRPWORK	Work limited by physical health
4	C15HDMEN	Days mental health not good	C15CMPHTH	General health compared to peers
5	C15CMPHTH	General health compared to peers	C15VRGENHTH	General health status
6	C15VRGENHTH	General health status	C15HDACT	Days poor health interfered with activities
7	C15ADLWLK	Difficulty walking	C15VRPHCMP	Physical health compared to a year ago
8	C15VRPHCMP	Physical health compared to a year ago	C15HDMEN	Days mental health not good
9	C15DEP2WK	Depressed for 2 weeks	C15SRVDISP	Survey disposition

Table 17 continued

Uneven Split				
Variables in Order of Importance	*CHAID*	*Variable Description*	*Neural Network*	*Variable Description*
1	C15VRPAIN	Pain interfering with activities	C15VRPAIN	Pain interfering with activities
2	C15ADLWLK	Difficulty walking	C15ADLWLK	Difficulty walking
3	C15HDPHY	Days physical health not good	C15CMPHTH	General health compared to peers
4	C15HDACT	Days poor health interfered with activities	C15HDACT	Days poor health interfered with activities
5	C15CMPHTH	General health compared to peers	C15VRGENHTH	General health status
6	C15HDMEN	Days mental health not good	C15HDMEN	Days mental health not good
7	C15VRGENHTH	General health status	C15SRVDISP	Survey disposition
8	C15VRPHCMP	Physical health compared to a year ago		
9	C15PCTCMP	Percent of survey completed		
Continuous Outcome				
Variables in Order of Importance	*CHAID*	*Variable Description*	*Neural Network*	*Variable Description*
1	C15VRPAIN	Pain interfering with activities	C15VRPWORK	Work limited by physical health
2	C15VRGENHTH	General health status	C15VRPAIN	Pain interfering with activities
3	C15VRPACCL	Accomplishing less due to physical health	C15VRGENTH	General health status
4	C15ADLWLK	Difficulty walking	C15HDACT	Days poor health interfered with activities
5	C15HDACT	Days poor health interfered with activities	C15HDMEN	Days mental health not good
6	C15HDMEN	Days mental health not good	C15VRPHCMP	Physical health compared to a year ago
7			C15SRVDISP	Survey disposition
8			C15PCTCMP	Percent survey completed

4.2.5 Boosting and bagging

Boosting significantly improved the accuracy of the Neural Network model when the outcome was continuous, by reducing the average squared error from 0.52 to 0.16. Bagging also improved accuracy in continuous outcomes of the Neural Network, but to a far less degree (difference of 0.03). Boosting and bagging also made a slight improvement in classification accuracy when a CHAID model was used with a dichotomous outcome, even split. Boosting and/or bagging did not improve model accuracy in any of the other conditions.

Table 18. Misclassification and Average Squared Error Tables for Boosting and Bagging

% Misclassified	CHAID			Neural Network		
	Cont.	**Even**	**Uneven**	**Cont.**	**Even**	**Uneven**
Boosting	.49 *(.48)*	**7%** *(9%)*	9% *(9%)*	**.16** *(.52)*	47% *(10%)*	16% *(10%)*
Bagging	.54 *(.48)*	**8%** *(9%)*	12% *(9%)*	**.49** *(.52)*	12% *(10%)*	15% *(10%)*

4.2.6 Summary comparison of prediction models across variable types

4.2.6.1 Continuous outcome: CHAID vs. neural network

When the outcome was continuous (number of days physical health not good), the CHAID method of segmentation resulted in 5 distinct segments; whereas the Neural Network yielded 6 segments. The two methods shared 2 very similar segments, and were both successful in identifying the healthiest (Segment 1 for CHAID and Neural Network) and unhealthiest (Segment 5 for CHAID and Segment 6 for Neural Network) segments, as illustrated in Table 19 below.

Table 19. Distribution of Demographic Variables for Continuous Outcome: CHAID and Neural Network

When AGE=1 (Less than 65)		
Segment	*CHAID*	*NEURAL NETWORK*
1	4.51%	3.01%

Table 19 continued

5 (6)	24.69%	25.17%
When BMI=1 (Not obese)		
Segment	*CHAID*	*NEURAL NETWORK*
1	77.23%	79.15%
5(6)	60.22%	60.38%
When MRSTAT=1 *(Married)*		
Segment	*CHAID*	*NEURAL NETWORK*
1	60.59%	62.33%
5(6)	43.12%	43.78%
When RACE=1 *(White)*		
Segment	*CHAID*	*NEURAL NETWORK*
1	85.61%	87.12%
5(6)	77.08%	76%
When EDUC=1 (Less than high school or GED)		
Segment	*CHAID*	*NEURAL NETWORK*
1	15.41%	13.49%
5 (6)	29.19%	30.91%
When GENDER=1 (Male)		
Segment	*CHAID*	*NEURAL NETWORK*
1	45.83%	46.83%
5 (6)	39.39%	39.56%
When DEPRESS=1 (Depressed)		
Segment	*CHAID*	*NEURAL NETWORK*
1	10.97%	9.72%
5 (6)	57.65%	57.63%
When SMOKE=1 (Smoke every day)		
Segment	*CHAID*	*NEURAL NETWORK*
1	5.73%	5.47%
5 (6)	10.94%	11.29%
# of Unhealthy Days		
Segment	*CHAID*	*NEURAL NETWORK*
1	0.84	0.49
5 (6)	13.19	14.85

4.2.6.2 Dichotomous outcome even split: CHAID vs. Neural network

The CHAID and Neural Network models both produced a 4-segment solution when the continuous outcome (# of unhealthy days was dichotomized according to the even split). As seen in the continuous outcome, both methods identified an identical healthy segment (Segment 1), and both

produced an unhealthy segment (Segment 4), which was similar but not identical as with Segment 1. The CHAID model picked up an additional unhealthy segment, Segment 2, which was similar to the unhealthy Neural Network Segment 4 in terms of demographics (AGE and BMI).

Table 20. Distribution of Demographic Variables for Dichotomous Outcome (Even): CHAID and Neural Network

When AGE=1 (Less than 65)		
Segment	*CHAID*	*NEURAL NETWORK*
1	7.1%	5.27%
2 (4)	40.97%	29.22%
When MRSTAT=1 (Married)		
Segment	*CHAID*	*NEURAL NETWORK*
1	56%	59.39%
2 (4)	41.01%	42.62%
When RACE=1 (White)		
Segment	*CHAID*	*NEURAL NETWORK*
1	82.96%	85.09%
2 (4)	76.36%	74.8%
When BMI=1 (Not obese)		
Segment	*CHAID*	*NEURAL NETWORK*
1	72.66%	77.01%
2 (4)	54.69%	57.93%
When EDUC=1 (Less than high school or GED)		
Segment	*CHAID*	*NEURAL NETWORK*
1	20.01%	16.22%
2 (4)	33%	33.21%
When GENDER=1 (Male)		
Segment	*CHAID*	*NEURAL NETWORK*
1	44.6%	45.06%
2 (4)	39.53%	39.89%
When DEPRESS=2 (Not depressed)		
Segment	*CHAID*	*NEURAL NETWORK*
1	81%	86%
2 (4)	78.49%	61.87%

4.2.6.3 Dichotomous outcome uneven split: CHAID vs. neural network

When the dichotomous outcome was split according to CDC research (uneven split), the methods produced almost demographically identical four segment solutions, as illustrated in Table 20. For

example, 5% of respondents in Segment 1 were younger than 65 in both CHAID and Neural Network models. Within the same segment, 60% were married according to both methods, 77% were not obese, and 46% were male.

Table 21. Distribution of Demographic Variables for Dichotomous Outcome Uneven: CHAID and Neural Network

When AGE=1 (Less than 65)		
Segment	*CHAID*	*NEURAL NETWORK*
1	5.1%	5.14%
2	37.2%	45.49%
3	5.93%	5.28%
4	21.6%	18.22%
WHEN MRSTAT=1 (Married)		
Segment	*CHAID*	*NEURAL NETWORK*
1	59.95%	60.3%
2	41.68%	41.57%
3	54.61%	54.84%
4	44.16%	46.09%
WHEN RACE=1 (White)		
Segment	*CHAID*	*NEURAL NETWORK*
1	84.31%	86.02%
2	76.82%	75.84%
3	82.93%	85.05%
4	77.32%	76.94%
When BMI=1 (Not obese)		
Segment	*CHAID*	*NEURAL NETWORK*
1	76.66%	77.14%
2	55.44%	52.73%
3	65.39%	65.94%
4	61.63%	63.81%
When EDUC=1 (Less than high school or GED)		
Segment	*CHAID*	*NEURAL NETWORK*
1	17.4%	15.96%
2	32.21%	33.63%
3	22.56%	20.69%
4	38.85%	28.3%
When GENDER=1 (Male)		
Segment	*CHAID*	*NEURAL NETWORK*
1	45.88%	45.66%
2	38.84%	40.37%
3	42.61%	42.6%
4	39.79%	40.47%
When DEPRESS=1 (Depressed)		

Table 21 continued

Segment	*CHAID*	*NEURAL NETWORK*
1	11.01%	10.18%
2	76.75%	82.11%
3	20.76%	21.99%
4	53.2%	47.77%

4.2.6.4 CHAID: Even vs. uneven split

Both an even and uneven split of the dichotomous outcome for the CHAID model resulted in a 4-segment solution with 2 very similar segments, as indicated in Table 22. For the segments that weren't identical (Segments 3&4), the even split included segments with higher proportions of respondents indicating that they were depressed, and a greater mean number of unhealthy days across segments.

Table 22. Distribution of Demographic Variables for Dichotomous Outcome: CHAID and Neural Network

		Even		Uneven	
Segment		**1**	**2**	**1**	**2**
AGE	1	7.1	40.97	5.1	37.2
	2	57.19	31.17	61.28	33.22
	3	35.71	27.86	33.62	29.58
MRSTAT	1	56	41.01	59.95	41.68
	2	44	58.99	40.05	58.32
RACE	1	82.96	76.36	84.31	76.82
	2	9.63	15.17	8.66	14.4
	3	7.42	8.47	7.03	8.78
BMICAT	1	72.66	54.69	76.66	55.44
	2	27.34	45.31	23.34	44.56
EDUC	1	20.01	33	17.4	32.21
	2	34.77	33.63	34.1	34.33
	3	45.22	33.37	48.5	33.46
GENDER	1	44.6	39.53	45.88	38.84
	2	55.4	60.47	54.12	61.16

4.2.6.5 Neural network: Even vs. uneven split

Similarly to the CHAID model, the Neural Network model produced 4 segments with a dichotomous outcome, regardless of the split. Again, Segment 1 was similar in both models

(representing the healthy segment), and Segment 4 (even split) = Segment 2 (uneven split), representing the unhealthy segment. The uneven outcome, resulted in greater variability in segments, with more extreme healthy and unhealthy segments. For example, the mean number of unhealthy days in Segment 1 (even) was 1.24; in Segment 1 (uneven) the mean was 0.88. Similarly, the mean number of unhealthy days in Segment 4 (even) was 16.16; in Segment 1 (uneven) the mean was 24.62. For the even split outcome, the Neural Network produced 2 segments (2&3) that were very similar in terms of demographics and mean number of unhealthy days; whereas the uneven split produced middle of the road segments (not the healthiest/unhealthiest) that exhibited greater variation between each other.

5.0 DISCUSSION

5.1 REVIEW OF STUDY PURPOSE

The primary aim of this study was to examine the differences in outcomes between models for two types of segmentation analysis problems. For classification purposes, latent class analysis and k-means clustering were compared. To illustrate a segmentation study in which there was a defined research question and/or prediction need, outcomes of neural networks and CHAID models were assessed. In addition, a secondary aim was to evaluate the effect of manipulation of the independent (input) and dependent variables within the segmentation models. The types of input variables examined consisted of dichotomous and categorical scales for classification models, and continuous and categorical responses for prediction models. The varieties of dependent variables studied in this analysis included dichotomous (binary) and continuous outcomes for prediction models

5.2 MAJOR FINDINGS BY RESEARCH QUESTION

5.2.1 Research Question #1

Do distinct segmentation methods for classification questions result in different outcomes, such as the number of segment classes, segment class size, and important variables, when the scales of the

inputs (continuous vs. binary) are varied? How can/should a researcher interpret potential differences?

Specifically,

1.1 *How do the segments differ by number of classes, size, and important variables by method, given the original inputs (full depression and RAND scales)?*

Both methods (K-Means and LCA) yielded 9 segment solutions, with similar segment class membership allocations, ranging in size from 5% to 27% of respondents. Segments 1,2,3,4, and 5 were virtually identical across methods. In both K-Means and LCA models, Segment 5 was the youngest, unlikely married, female, obese, most unhealthy, and most likely to be depressed. On the other hand, Segment 1 was the healthiest; least likely to be depressed, and most educated of the segments, given both LCA and K-Means methods. Segments 2 and 3 also had high rates of depression, but reported fewer number of unhealthy days. Segment 4 was the oldest segment, and while reporting higher counts of unhealthy days, frequency of pain interfering with daily activities, etc. they were, on average, less depressed.

On the other hand, the interpretation of Segments 6-9 for K-Means and LCA models is not consistent. For example, Segment 9 of the K-Means model is much older, more likely to be married, and has a low likelihood of depression (but similar health status), as compared to Segment 9 of the LCA model. In addition, the LCA resulted in 2 segments (7 & 8) with very low reporting of depression; whereas the K-Means model did not produce these same segment classes. The differences in the segment class could potentially be problematic for researchers who have not executed a methodological study of classification methods, since the marketing and engagement efforts would take into account the demographic and behavioral characteristics of each segment class.

In Enterprise Miner, the Segment Profile node (which is attached after the K-Means analysis node) yields variable importance statistics, including variable worth and rank. Variable worth represents the logworth of the factors (inputs) that have been identified for each segment. As previously indicated, in order to examine variable worth for the LCA model, the maximum probability of response was examined, and variables containing the maximum number of classes in which the probability of response was over .25 were selected. The K-Means model yielded 10 important variables; whereas the LCA model is limited to 5. The two methods had 3 common differentiating variables, including C15VRMACT (amount of time emotional health interferes with social activities), C15VRPACCL (accomplishing less due to physical health), and C15VRMWORK (work or activities limited by emotional health). Important to LCA were the two depression questions (depressed in the last 2 weeks and depressed in the last 2 years); while K-Means also included the questions related to pain, ability to climb stairs, average energy level, and feeling down.

While both methods yielded similar segment class solutions, upon further examination of the profiles, differences in demographics and health behaviors were discovered. For example, LCA resulted in more differentiation in respondents' BMI categories across segment classes, as compared to K-Means. In addition, LCA identified more extreme responses for the pain question. For example, over 60% of Segment 3 from the LCA method reported being in pain most of the time (the highest proportion in K-Means was 47%). In addition, LCA identified a segment (Segment 8) in which 0% of respondents reported being in pain all of the time.

1.2. *How do the segments vary by number of classes, size, and important variables by method, given the dichotomization of the depression scale?*

When the depression scale was dichotomized, K-means and LCA methods for segmentation produced very different results. The main inconsistency is the best class segment solution; 9 segment classes for LCA and 6 segment classes for K-Means. While Segment 3 is nearly identical in both models and LCA Segment 5 is very similar to K-Means Segment 4, the rest of the classes do not exhibit the same similarities as in the full-scale case. K-Means models had greater variability of number of unhealthy days, which ranged from 1.1 days (Segment 4) to 22.5 days (Segment 3), while LCA segments varied from 4.2 days (Segment 8) to 24.3 days (Segment 3). Although the LCA method produced more segment classes, the segments themselves contained a high proportion of demographic variability, whereas one of the six K-Means segments (Segment 2) had demographic response proportions that were almost identical to the sample mean, and provided little variation in response.

As noted with the full-scale case, the K-Means model contained many more important variables (10), as compared to the LCA model (4) when the dichotomous depression scale was used. The two methods had only 2 common differentiating variables, including C15VRMACT (amount of time emotional health interferes with social activities) and C15VRMWORK (work or activities limited by emotional health). Important to LCA was the dichotomized depression scale (DEPRESS1) and physical health interferes with social activities (C15VRSACT) ; while K-Means also included the questions related to pain, ability to climb stairs, average energy level, and feeling down.

1.3. *What are the potential implications resulting from different outcomes across methods and inputs?*

A possible issue, resulting from the inconsistency of results across methods, would be the reduction in survey respondents assigned to a segment class under the LCA model. Although, the

mean response frequencies were comparable across LCA and K-Means model, The LCA model identified less approximated 10% survey respondents, as compared to the K-Means model, as presented in Section 3.4.1. Furthermore, both LCA and K-Means models utilized less than 50% of the total sample.

In addition, the identification of differences in segment classes (i.e. Segments 6-9 of the full-scale models and completely different solutions for dichotomized-scale models) could present problems for researchers. Ideally, the methods would triangulate, or result in similar segment class solutions, sizes, important variables and interpretations of classes for both methods. When this isn't the case, the researcher is forced to choose a method, provided that the methodological study has been conducted, or determine the solution that makes the most practical or business sense. When methods do not align, there is always a threat to the validity of the interpretation of findings, in which the researcher must provide justification.

The inconsistency of results across the treatment of the input scale, also has implications for researchers. This finding suggests that additional research may need to be conducted on the depression scale itself, and that the dichotomization procedure selected may not have been sufficient at identifying varying states of depression. In addition, it is interesting to note that the range of important input variables differed greatly by method, with more variables involved in the separation of segment classes for the K-Means model. It could be interpreted that the LCA model needs fewer inputs to differentiate segment classes, than the K-Means, and could potentially be useful for researchers designing a survey, when time, cost, etc. are of concern.

5.2.2 **Research Question #2**

Do distinct segmentation methods for prediction questions result in different outcomes, such as the number of segment classes, segment class size, and important variables, when the dependent variable is continuous vs. binary? How can/should a researcher interpret potential differences?

2.1. *How do the segments differ by number of classes, size, and important predictor variables by method when the outcome is binary?*

When the outcome is dichotomous, both CHAID and Neural Networks result in a 4 segment class solution, irrespective of the dichotomous splitting criteria. Both methods yielded a dominant segment class (Segment 1) which contained anywhere from 39-67% of total membership; with Segment 4 of all models being the next largest segment (17-45%) and Segments 2 and 3 containing the fewest proportion of respondents (5-11%). The methods were also similar in the proportion of misclassified cases (9% CHAID and 10% Neural Network).

In addition to accuracy, important variables were nearly identical across splitting criteria (even vs. uneven) and method (CHAID vs. Neural Network). C15VRPAIN (degree pain interfered with normal work) was the most important predictor in all 4 cases. Moreover, C15HDACT (number of days poor health interfered with activates), C15ADLWALK (difficulty walking), C15HDMEN (number of days mental health not good), C15CMPHTH (general health compared to peers), and C15VRGENHTH (general health status) were common important variables across method and splitting criteria. Seven variables were responsible for the segment class differentiation in the Neural Network model with an uneven split; whereas it took nine variable to split the segments using the CHAID method (even and uneven split) and Neural Network with an even split outcome.

As with classification methods, the segment class profiling uncovered differences in outcomes by means of categorizing the dependent variable. For example, when the outcome was dichotomous and evenly split, the CHAID model resulted in higher differentiation of the age variable across segments. Forty percent of Segment 2 of the CHAID model were classified as "AGE=1"; whereas the highest proportion of respondents in the AGE=1 category for the neural network model was 29%. For the CHAID model, the next highest AGE=1 segment was Segment 4 (28.4%), while the Neural Network model had only 7.9% of Segment 2 classified as AGE=1 (representing the second highest AGE=1 segment, where age < 65). In addition to age, the methods differed in the differentiation of number of unhealthy days and the pain question. For example, the highest average number of unhealthy days for the CHAID model was 25 (Segment 2); for the CHAID model, 16 (Segment 4). The Neural Network model had 3 segments with 0 respondents reporting a "4" or "5" on the pain scale; while the CHAID model had 2 (indicating the CHAID model produced segments that were more different in their responses to pain).

2.2. *How do the segments differ by number of classes, size, and important predictor variables by method when the outcome is continuous?*

While the models for the continuous outcome exhibited similar accuracy (as indicated by the average square error), the CHAID model yielded a 5-segment solution; the Neural Network model resulted in 6 segments. The segment class sizes were less consistent with the continuous outcome, as identified with the dichotomous models. The CHAID model resulted in 2 large segments, with Segments 1 and 2 representing 42% and 37% of total membership, respectively. On the other hand, Segment 6 of the Neural Network model was comprised of 48% of respondents, with Segments 1-5 ranging from 5-17% of total membership. In addition, to the number of segment classes, there were differences in the interpretations of the segment classes. Segments 1, 3, and 5

(CHAID)/6 (Neural Network) were almost identical across method in terms of demographic variables when the target (# of unhealthy days) was continuous. On the other hand, Segments 2 and 4 of the CHAID model did not align with Segments 2, 4, and 5 from the Neural Network. For example, the CHAID model's segment 2 was older, more ethnically diverse, less educated, but healthy. Segment 2 within the Neural Network model was predominantly white, highly educated, healthy and least likely to be depressed. There was not a match for CHAID Segment 2 in the Neural Network segment class solution.

Four out of the six important predictor variables for the CHAID model were also included in the Neural Network solution: pain interfering with activities (C15VRPAIN), general health status (C15VRGENHTH), days poor health interfered with activities (C15HDACT), and days mental health not good (C15HDMEN). Physical health limiting work, physical health compared to a year ago, survey disposition and percent survey completed were also important to differentiating segment classes for the Neural Network continuous model.

2.3. *What are the potential implications resulting from different outcomes across methods and treatment of dependent variable?*

As with classification models, the methods for prediction differed (but only slightly) in the number of respondents who fell into a segment class. The CHAID models were able to classify 182,735 respondents, while the Neural Network models identified 207,423 respondents. The differences and nuances between methods are slight; with an additional segment class generated by the Neural Network model for the continuous outcome.

The differences in segment class solutions as a result of method were not as pronounced for prediction models, as compared to classification models when the outcome was dichotomous. On the other hand, the continuous outcome forced an additional segment with the Neural Network

model and differences between segments across methods. Also, survey disposition (mode of survey administration + proportion completed) was only important in differentiating segment classes for Neural Network models, which could indicate that the Neural Network is identifying an additional component that CHAID models do not. As discussed in Section 5.2.1, the researcher must make a decision as to which segment solution to use, based on not only model performance, but interpretability and business case use.

5.2.3 Research Question #3

How does the treatment of the dependent and independent variable affect model results?

Specifically,

3.1. *What are the differences in model fit/accuracy, number of segment classes, and important predictor variables when the dependent variable is dichotomous, as compared to a continuous outcome for prediction models?*

When the outcome is dichotomous, uneven splitting criteria resulted in twice the gain %, as compared to the even split outcome, for both CHAID and Neural Network models. Since the continuous outcome does not yield a gain percentage, comparison across continuous vs. dichotomous outcomes for model fit is challenging. Both CHAID and Neural Network models yielded four very similar segment classes, for the dichotomous outcome, regardless of splitting criteria. Use of a continuous outcome resulted in an additional segment for the CHAID model, and an additional 2 segments for the Neural Network model. The lists of important predictor variables for each model and IV type were comparable, especially in the dichotomous cases. When the outcome was dichotomous, the same 7 variables were identified as the best predictors across method and splitting criteria. On the other hand, the CHAID model required fewer predictor

variables to classify respondents into segment groups (6) when the outcome was continuous, as compared to the dichotomous outcomes, which required 9 predictor variables. For the Neural Network model, the important variables were analogous for 8/9 variables when the outcome was continuous as compared to the even split dichotomous condition. For the uneven split, the Neural Network model required only 7 important variables, with physical health compared to a year ago (C15VRPHCMP) and health interfering with work (C15VRPWORK) dropping off the list.

It should be noted that the splitting criteria has additional implications for interpretations of the outcomes. The uneven split, by nature, has more severe criteria for identifying respondents that are considered to be "unhealthy" (5 or more unhealthy days as compared to 15 for the even split). In these cases, it is possible that the respondent could have been experiencing an acute event, i.e. an infection or virus; whereas in the evenly split condition, it is likely that the respondent is experiencing a more chronic condition.

3.2. *What are the differences in model fit/accuracy, number of segment classes, and important input variables when a dichotomous vs. a continuous scale is used for classification models?*

The dichotomous scale yielded better model fit across methods, as indicated by a smaller BIC (LCA) and larger gap statistic (K-Means). On the other hand, misclassification appears to be method, not scale- dependent. The LCA model resulted in much lower misclassification overall, with 4% for full scale and 8% for the dichotomous scale. The misclassification rates for the K-means models were much higher, with the dichotomous rate being much lower (29%) than the full scale misclassification rate (43%). All models, with the exception of the K-Means model for the dichotomous depression scale, yielded 9 segment classes (while there were differences in the interpretations of the segment classes, as discussed in Section 5.2.1). Important variables for the

LCA model did not change significantly as a function of the treatment of the input depression scale. Accordingly, the same 10 variables were important in differentiating segment classes for both K-Means models (original and dichotomous depression scale).

3.3. *What are the differences in model fit and accuracy when boosting vs. bagging algorithms are used in prediction models?*

Boosting only enhanced model accuracy for the dichotomous even-split CHAID model; whereas boosting improved model accuracy for the Neural Network continuous outcome. Bagging also improved classification for the dichotomous even-split CHAID model, while also improving accuracy for the Neural Network continuous model. Boosting and/or bagging algorithms did not improve model performance for uneven dichotomous outcomes for either method.

5.3 LIMITATIONS AND ADDITIONAL CONSIDERATIONS

5.3.1 Survey data and scaling

Survey data is inherently unreliable. With any survey, there is the potential for respondents to misrepresent themselves, engage in response style patterns and biases, and complete the survey in a less-than-thoughtful manner. All of these potential problems with survey response were not evaluated in this study, and could potentially skew the results. Future research could attempt to identify any potential response bias in the response data, and account/correct for it in future analyses. Of note, the survey developers do adjust scores as a result of telephone administrations, since research has shown that health scores and status tend to be higher with “live” surveys.

Future research could also include multiple scaling procedures. For example, the original depression scale, which contained four separate depression screening questions, was dichotomized into “depressed” if the respondent answered “Yes” to any of the four depression questions. Additional research could be done to verify that this is an appropriate method, and if there are differences in weights or values of the individual depression screening questions. With respect to the prediction models, the splitting criteria for the dichotomization of the number of healthy days (outcome variable) could be further tested. As previously discussed in 3.3.3, splitting criteria was based on CDC research from the early 1990’s. The documentation was not very specific regarding the methodology or validation procedures, and future work could examine the appropriateness of the splitting criteria.

5.3.2 Real data

For the purpose of this study, actual data from the Medicare Health Outcomes Survey was used to compare the models. The limitation of using a real data set is that no determination can made as to which approach is better; LCA vs. K-Means and CHAID vs. Neural Networks. This study explored whether different interpretations are found by different model. Since real data was used the true underlying model (number of segments) is unknown.. A simulation study could be used to determine which model more accurately recovers a simulated segmented structure.

5.3.3 Additional testing conditions

In the present study, default values for K-Means, LCA, CHAID, and Neural Network models were primarily used, and future studies could evaluate the effects of the following conditions:

For the CHAID methods, the following alternatives could be examined:

- Adjust splitting rules (significance level, maximum branch, depth, categorical size, pruning)
- Node: leaf size (default is 5; 0 surrogate rules)

As with CHAID models, the default selection criteria were maintained with methods using Neural Networks. The following options could be considered in additional research:

- Multiple vs. single layer architecture
- Increase # of hidden layers (default is 3)
- Target activation function: default is Identity (could use exponential, sine, tanh)
- Change # of tries to train the network (default is 2)
- Change maximum number of iterations (default is 300)

For the K-Means model, the criteria in Enterprise Miner is very specific to allow for a K-Means clustering approach (other classification approaches are available), but there is an option to vary the missing data imputation method (for this study and the default, mean imputation was used).

In addition to model building criteria, additional inputs could be used for both methods. For classification methods, only the depression scale and Veteran's RAND 12 Item Health Survey (VR-12) were utilized. While both methods perform better with fewer inputs, additional survey questions (presence/absence of diseases, treatments received, etc.) could be used. All inputs were available for the prediction models, but in order for the models to converge, variable selection procedures were implemented. Future research could take a more deliberate, investigator-led approach to variable selection. Some of the most conventional and well-documented variable selection methods include: factor analysis, mean square error of prediction, clustering, variable

ranking and subset selection, Gibbs sampling, Bayesian modeling, and regression shrinkage techniques

5.3.4 Alternative software programs

Alternative software programs could be considered for future research, which may include additional model building options. For example, the open-source programs (i.e., R) allow for additional flexibility of model building conditions. More robust LCA software could also be tested. For example, Latent Gold® contains separate modules for estimating three different model structures: LC Cluster models, DFactor models, and LC Regression models. In future studies, alternative programs and programming languages could be used to test different segment class solutions for the predictive models (Neural Network and CHAID), that previously did not allow for a pre-specification of the desired number of clusters or segment class solution.

5.3.5 Implications for healthcare and marketing researchers

The ability to identify distinct attitudinal, behavioral, and demographic groupings of Medicare survey respondents has multiple benefits for healthcare and marketing researchers. As previously discussed, the literature surrounding the uses of the Medicare Health Outcomes survey has been primarily focused on quality care ratings, physician-patient communication outcomes, and the comparison of health plans within the Medicare population. The results of this study can be used in several ways. First, researchers can use the segmentation analysis to identify high risk segments (i.e., Segment 3 of K-Means dichotomous, which included those who were obese, smokers, likely to be depressed and in pain) and those likely to become high risk as health declines (i.e. Segment

5). Consequently, the identification of an at-risk (vs. a high-risk segment) introduces an opportunity for an intervention for the at-risk segment, before they become high-risk. Researchers and healthcare managers can continue to monitor the healthier segments and offer incentives, programs, and wellness initiatives to keep them healthy and reduce the likelihood of becoming depressed. Having distinct segment classes offers marketing researchers the ability to design targeted campaigns and messaging, that are unique to that groups' needs and wants, with the goal of increasing (or sustaining) engagement with the health plan and improving health outcomes.

A few other interesting features were teased out of the segmentation analyses. While examining the profiles of the CHAID model (even-split outcome), it was discovered that the most likely to be depressed were younger, female, and less educated. And although there were two segments that were very similar in terms of obesity, pain, and energy levels, one was less likely to be depressed. Nuances like these could warrant a deeper dive into the data; for an additional level of discovery could be conducted in order to determine why one of two very similar groups was more likely to be depressed. Learnings from this analysis could be used by practitioners and healthcare providers that specifically treat the elderly population.

5.3.6 Recommendations for additional research

Additional research could enhance the current study in multiple ways. First, a methodological study could include several of the alternative testing conditions, as described in Section 5.3.2. This follow-up study could include a simulation component, plus additional variations of components of both the classification and prediction models. For example, a researcher may have interest in specific chronic conditions (i.e., diabetes, heart disease) and choose to include these in the classification models, or use the presence/absence of the condition as an outcome for prediction

models. Researchers may only be equipped to handle a small number of segments, and therefore could restrict the splitting criteria for CHAID models, or further reduce the number of segments produced by neural networks by conducting ad-hoc aggregation.

In addition, survey data could be combined with utilization (i.e., insurance claims), third-party demographic, and/or public health data to introduce a behavioral component to the segments (via claims), and further refine and differentiate the current segment class groupings. By including both inpatient and outpatient claims data, researchers could identify patterns of utilization, potential care gaps, and create additional derived variables such as healthcare engagement and future service needs. Additional third-party data sources, such as Acxiom and Experian, provide important, potentially differentiating information on individuals, such as income, household size, occupation, personal interests, and shopping habits. These additional variables can help researchers or marketers with engagement strategies and/or preferences (i.e., identifying groups that may be more inclined to respond to electronic vs. paper communications). Finally, the inclusion of public health data such as flu statistics, inpatient discharges and ER visits by region can account for external variability (seasonality, regional and population health, availability of medical care facilities and resources, etc.) not otherwise accounted for.

5.4 FINAL CONCLUSIONS

The results of this study can be used to enhance the current literature in multiple ways. First, very few studies of classification and/or prediction methods for segmentation address the issues surrounding the treatment of outcomes (i.e., dichotomous vs. continuous, levels of dichotomous outcome (even vs. uneven split). By design, this study has allowed for a direct comparison of

methods by problem, while utilizing a common data set. And while the non-theoretical nature of the data does not allow for the selection of a "winning" method, results of this study are consistent with the literature. The LCA segments, under the full-depression scale condition, tended to exhibit greater differentiation on some of the demographic and non-segmentation variables, such as BMI and pain. As demonstrated in Sarstedt & Ringle's (2010) research, the LCA models in this study resulted in far greater model accuracy and fewer misclassified cases, as compared to the K-Means models. In addition, the similarity of results for Neural Network and CHAID models with a dichotomous outcome is consistent with earlier findings by Olsen et al., 2009. Yet, few papers have addressed both types of segmentation problems, and discussed the requirements and necessary conditions for both classification and prediction problems. The current study also demonstrated the "art vs. science" juxtaposition of choosing the best segment class solution. As seen in the case of the LCA methods, while the model identified a particular best-segment-class solution (9 classes), a stakeholder might argue that this is too many segments and decide to combine to reduce efforts in designing separate marketing campaigns, programs, etc. for each individual segment class.

Results of this study can help and inform healthcare and marketing researchers, practitioners, and managed care organizations. They can be used to design specific interventions, products, and messaging to treat and reach various populations. They can also serve as a springboard for additional research initiatives around the relationship between physical and mental health, and help to begin to identify potential risk factors for both types of illnesses.

Adding to the context and knowledge base around the survey itself, this study has been able to identify predictors of depression (physical health) in classification models. For prediction models, both physical *and* mental health were predictive of the outcome- number of unhealthy

days. The survey developers could use the results of this study to potential eliminate low-performing or uninformative questions.

APPENDIX A

MEDICARE HEALTH OUTCOMES SURVEY

1. In general, would you say your health is:
1 Excellent
2 Very good
3 Good
4 Fair
5 Poor

2. The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities? If so, how much?

a. **Moderate activities,** such as moving a table, pushing a vacuum cleaner, bowling, or playing golf
1 Yes, limited a lot
2 Yes, limited a little
3 No, not limited at all

b. Climbing **several** flights of stairs
1 Yes, limited a lot
2 Yes, limited a little
3 No, not limited at all

3. During the **past 4 weeks,** have you had any of the following problems with your work or other regular daily activities **as a result of your physical health?**

a. **Accomplished less** than you would like **as a result of your physical health?**
1 No, none of the time
2 Yes, a little of the time
3 Yes, some of the time
4 Yes, most of the time
5 Yes, all of the time

b. Were limited in the **kind** of work or other activities **as result of your physical health?**
1 No, none of the time
2 Yes, a little of the time
3 Yes, some of the time
4 Yes, most of the time
5 Yes, all of the time

4. During the **past 4 weeks,** have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)?

a. **Accomplished less** than you would like **as a result of any emotional problems**
1 No, none of the time
2 Yes, a little of the time
3 Yes, some of the time
4 Yes, most of the time
5 Yes, all of the time

b. Didn't do work or other activities as **carefully** as usual **as a result of any emotional problems**
1 No, none of the time
2 Yes, a little of the time
3 Yes, some of the time
4 Yes, most of the time
5 Yes, all of the time

5. During the **past 4 weeks,** how much did **pain** interfere with your normal work (including both work outside the home and housework)?
1 Not at all
2 A little bit
3 Moderately
4 Quite a bit
5 Extremely

These questions are about how you feel and how things have been with you during the **past 4 weeks.** For each question, please give the one answer that comes closest to the way you have been feeling.

6. How much of the time during the **past 4 weeks:**

a. Have you felt calm and peaceful?
1 All of the time
2 Most of the time
3 A good bit of the time
4 Some of the time
5 A little of the time
6 None of the time

b. Did you have a lot of energy?
1 All of the time
2 Most of the time
3 A good bit of the time
4 Some of the time
5 A little of the time
6 None of the time

c. Have you felt downhearted and blue?
1 All of the time
2 Most of the time
3 A good bit of the time
4 Some of the time
5 A little of the time
6 None of the time

7. During the **past 4 weeks,** how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting with friends, relatives, etc.)?
1 All of the time
2 Most of the time
3 Some of the time
4 A little of the time
5 None of the time

Now, we'd like to ask you some questions about how your health may have changed.

8. **Compared to one year ago,** how would you rate your **physical health** in general **now?**
1 Much better
2 Slightly better
3 About the same
4 Slightly worse
5 Much worse

9. **Compared to one year ago,** how would you rate your **emotional problems** (such as feeling anxious, depressed or irritable) in general **now?**
1 Much better
2 Slightly better
3 About the same
4 Slightly worse
5 Much worse

Earlier in the survey you were asked to indicate whether you have any limitations in your activities. We are now going to ask a few additional questions in this area.

10. Because of a health or physical problem, do you have any difficulty doing the following activities **without special equipment or help from another person?**

a. Bathing
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I am unable to do this activity

b. Dressing
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I am unable to do this activity

c. Eating
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I am unable to do this activity

d. Getting in or out of chairs
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I am unable to do this activity

e. Walking
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I am unable to do this activity

f. Using the toilet
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I am unable to do this activity

11. Because of a health or physical problem, do you have any difficulty doing the following activities?

a. Preparing meals
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I don’t do this activity

b. Managing money
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I don’t do this activity

c. Taking medication as prescribed
1 No, I do not have difficulty
2 Yes, I have difficulty
3 I don’t do this activity

These next questions ask about your physical and mental health during the past 30 days.

12. Now, thinking about your physical health, which includes physical illness and injury, for how many days during the **past 30 days** was your **physical health not** good?

Please enter a number between "0" and "30" days. If no days, please enter “0” days. Your best estimate would be fine.

days

13. Now, thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the **past 30 days** was your **mental health not** good?

14. During the **past 30 days,** for about how many days did **poor physical or mental health** keep you from doing your usual activities, such as self-care, work, or recreation?
Please enter a number between "0" and "30" days. If no days, please enter “0” days. Your best estimate would be fine.
days

Now we are going to ask some questions about specific medical conditions.

15. Are you blind or do you have serious difficulty seeing, even when wearing glasses?
1 Yes
2 No

16. Are you deaf or do you have serious difficulty hearing, even with a hearing aid?
1 Yes
2 No

17. **Because of a physical, mental, or emotional condition**, do you have **serious** difficulty concentrating, remembering or making decisions?
1 Yes
2 No

18. **Because of a physical, mental, or emotional condition**, do you have difficulty doing errands alone such as visiting a doctor’s office or shopping?
1 Yes
2 No

19. In the **past month**, how often did memory problems interfere with your daily activities?
1 Every day (7 days a week)
2 Most days (5-6 days a week)
3 Some days (2-4 days a week)
4 Rarely (once a week or less)
5 Never

Has a doctor ever told you that you had:

20. Hypertension or high blood pressure
1 Yes
2 No

21. Angina pectoris or coronary artery disease
1 Yes
2 No

22. Congestive heart failure
1 Yes
2 No

23. A myocardial infarction or heart attack
1 Yes
2 No

24. Other heart conditions, such as problems with heart valves or the rhythm of your heartbeat
1 Yes
2 No

25. A stroke
1 Yes
2 No

Has a doctor ever told you that you had:

26. Emphysema, or asthma, or COPD (chronic obstructive pulmonary disease)
1 Yes
2 No

27. Crohn's disease, ulcerative colitis, or inflammatory bowel disease
1 Yes
2 No

28. Arthritis of the hip or knee
1 Yes
2 No

29. Arthritis of the hand or wrist
1 Yes
2 No

30. Osteoporosis, sometimes called thin or brittle bones
1 Yes
2 No

31. Sciatica (pain or numbness that travels down your leg to below your knee)
1 Yes
2 No

32. Diabetes, high blood sugar, or sugar in the urine
1 Yes
2 No

33. Depression
1 Yes
2 No

34. Any cancer (other than skin cancer)
1 Yes ***Go to Question* 35**
➔ ➔
2 No ***Go to Question* 36**

35. Are you **currently** under treatment for:

a. Colon or rectal cancer
1 Yes
2 No

b. Lung cancer
1 Yes
2 No

c. Breast cancer
1 Yes
2 No

d. Prostate cancer
1 Yes
2 No

e. Other cancer (other than skin cancer)
1 Yes
2 No

36. In the **past 7 days,** how much did pain interfere with your day to day activities?
1 Not at all
2 A little bit
3 Somewhat
4 Quite a bit
5 Very much

37. In the **past 7 days**, how often did pain keep you from socializing with others?
1 Never
2 Rarely
3 Sometimes
4 Often
5 Always

38. In the **past 7 days**, how would you rate your pain **on average**?
01 1 No pain
02 2
03 3
04 4
05 5
06 6
07 7
08 8
09 9
10 10 Worst imaginable pain

39. Over the **past 2 weeks**, how often have you been bothered by any of the following problems?

a. Little interest or pleasure in doing things
1 Not at all
2 Several days
3 More than half the days
4 Nearly every day

b. Feeling down, depressed or hopeless
1 Not at all
2 Several days
3 More than half the days
4 Nearly every day

40. In general, compared to other people your age, would you say that your health is:
1 Excellent
2 Very good
3 Good
4 Fair
5 Poor

41. Do you **now** smoke every day, some days, or not at all?
1 Every day
2 Some days
3 Not at all
4 Don't know

42. Many people experience leakage of urine, also called urinary incontinence. In the **past six months**, have you experienced leaking of urine?
1 Yes ***Go to Question* 43**
➔
2 No ***Go to Question* 46**
➔

43. During the **past six months**, how much did leaking of urine make you change your daily activities or interfere with your sleep?
1 A lot
2 Somewhat
3 Not at all

44. Have you **ever** talked with a doctor, nurse, or other health care provider about leaking of urine?
1 Yes
2 No

45. There are many ways to control or manage the leaking of urine, including bladder training exercises, medication and surgery. Have you **ever** talked with a doctor, nurse, or other health care provider about any of these approaches?
1 Yes
2 No

46. In the **past 12 months,** did you talk with a doctor or other health provider about your level of exercise or physical activity? For example, a doctor or other health provider may ask if you exercise regularly or take part in physical exercise.
1 Yes ***Go to Question* 47**
➔
2 No ***Go to Question* 47**
➔
3 I had no visits in the past 12 months ***Go to Question* 48**
➔
47. In the **past 12 months,** did a doctor or other health provider advise you to start, increase or maintain your level of exercise or physical activity? For example, in order to improve your health, your doctor or other health provider may advise you to start taking the stairs, increase walking from 10 to 20 minutes every day or to maintain your current exercise program.
1 Yes
2 No
48. A fall is when your body goes to the ground without being pushed. In the **past 12 months,** did you talk with your doctor or other health provider about falling or problems with balance or walking?
1 Yes
2 No
3 I had no visits in the past 12 months
49. Did you fall in the **past 12 months?**
1 Yes
2 No
50. In the **past 12 months,** have you had a problem with balance or walking?
1 Yes
2 No
51. Has your doctor or other health provider done anything to help prevent falls or treat problems with balance or walking? Some things they might do include:
• Suggest that you use a cane or walker.
• Check your blood pressure lying or standing.
• Suggest that you do an exercise or physical therapy program.
• Suggest a vision or hearing testing.

1 Yes
2 No
3 I had no visits in the past 12 months
52. Have you **ever** had a **bone density test** to check for **osteoporosis,** sometimes thought of as “brittle bones”? This test would have been done to your back or hip.
1 Yes
2 No
53. During the **past month**, on average, how many hours of actual sleep did you get at night? (This may be different from the number of hours you spent in bed.)
1 Less than 5 hours
2 5 – 6 hours
3 7 – 8 hours
4 9 or more hours
54. During the **past month**, how would you rate your overall sleep quality?
1 Very Good
2 Fairly Good
3 Fairly Bad
4 Very Bad
55. How much do you weigh in pounds (lbs.)?
lbs.

56. How tall are you without shoes on in feet (ft.) and inches (in.)? Please remember to fill in both feet and inches (for example, 5 ft. 00 in.) If 1/2 in., please round up.

ft.

in.

57. Are you male or female?

1 Male
2 Female

58. Are you Hispanic, Latino/a or Spanish Origin? (One or more categories may be selected)

1 No, not of Hispanic, Latino/a or Spanish origin
2 Yes, Mexican, Mexican American, Chicano/a
3 Yes, Puerto Rican
4 Yes, Cuban
5 Yes, Another Hispanic, Latino/a or Spanish origin

59. What is your race? (One or more categories may be selected)

01 White
02 Black or African American
03 American Indian or Alaska Native
04 Asian Indian
05 Chinese
06 Filipino
07 Japanese
08 Korean
09 Vietnamese
10 Other Asian
11 Native Hawaiian
12 Guamanian or Chamorro
13 Samoan
14 Other Pacific Islander

60. What language do you **mainly** speak at home?

1 English
2 Spanish
3 Chinese
4 Some other language (please specify)

61. What is your current marital status?

1 Married
2 Divorced
3 Separated
4 Widowed
5 Never married

62. What is the highest grade or level of school that you have completed?

1 8th grade or less
2 Some high school, but did not graduate
3 High school graduate or GED
4 Some college or 2 year degree
5 4 year college graduate
6 More than a 4 year college degree

63. Do you live alone or with others? (One or more categories may be selected)

1 Alone
2 With spouse/significant other
3 With children/other relatives
4 With non-relatives
5 With paid caregiver

64. Where do you live?

1 House, apartment, condominium or mobile home ***Go to Question 65*** ➔
2 Assisted living or board and care home ***Go to Question 65*** ➔
3 Nursing home ***Go to Question 66*** ➔
4 Other ***Go to Question 66*** ➔

65. Is the house or apartment you currently live in:
1 Owned or being bought by you
2 Owned or being bought by someone in your family other than you
3 Rented for money
4 Not owned and one in which you live without payment of rent
5 None of the above
66. Who completed this survey form?
1 Person to whom survey was addressed ***Go to Question 68***
➔
2 Family member or relative of person to whom the survey was addressed
3 Friend of person to whom the survey was addressed
4 Professional caregiver of person to whom the survey was addressed
67. If you completed the survey for someone else, please fill in your name. **DO NOT** complete this question if you completed the survey for yourself. Please **print** clearly.
First Name
Last Name
68. Which of the following categories best represents the **combined income for all family members in your household** for the **past 12 months**?
01 Less than $5,000
02 $5,000–$9,999
03 $10,000–$19,999
04 $20,000–$29,999
05 $30,000–$39,999
06 $40,000–$49,999
07 $50,000–$79,999
08 $80,000–$99,999
09 $100,000 or more
10 Don't know
YOU HAVE COMPLETED THE SURVEY. THANK YOU.
: _____
: _____

Insert Survey Vendor Contact Information Here

APPENDIX B

FREQUENCY OF RESPONSE VARIABLES

Variable	*Frequency*	*Percent*
AGE		
Less than 65	44266	14.94
65 to 74	149961	50.61
75 and older	102093	34.45
RACE		
White	205534	80.56
Black	29591	11.6
Other	19996	7.84
GENDER		
		+
Male	121057	42.45
Female	164112	57.55
MARRIAGE STATUS		
Married	135192	51.12
Not Married	129244	48.88
EDUCATION		
Less than HS or GED	63303	24.08
High school or GED	91232	34.71
Greater than HS or GED	108303	41.21
BMI CATEGORY		
Not obese (BMI<30)	173375	67.39
Obese (BMI>30)	83911	32.61
GENERAL HEALTH		
Excellent	17491	6.29
Very Good	64323	23.13
Good	99681	35.85
Fair	72385	26.03
Poor	24179	8.7
HEALTH LIMITATION:MODERATE ACTIVITIES		
Limited a lot	71048	25.88
Limited a little	92582	33.72
Not limited at all	110904	40.4
HEALTH LIMITATION: CLIMBING STAIRS		

Limited a lot	86729	32.44
Limited a little	91747	34.31
Not limited at all	88916	33.25
PHYSICAL: ACCOMPLISH LESS THAN WOULD LIKE		
None of the time	88556	32.45
A little of the time	51711	18.95
Some of the time	61358	22.48
Most of the time	45490	16.67
All of the time	25821	9.46
PHYSICAL: LIMITED IN WORK/ACTIVITIES		
None of the time	90978	34.06
A little of the time	47990	17.97
Some of the time	56351	21.1
Most of the time	43124	16.15
All of the time	28634	10.72
EMOTIONAL: ACCOMPLISH LESS		
None of the time	151180	55.4
A little of the time	41386	15.17
Some of the time	41195	15.1
Most of the time	25590	9.38
All of the time	13521	4.96
EMOTIONAL: DID NOT DO WORK AS CAREFULLY		
None of the time	156099	58.72
A little of the time	39768	14.96
Some of the time	36180	13.61
Most of the time	21425	8.06
All of the time		
PAIN INTERFERED WITH NORMAL WORK		
Not at all	77182	28.36
A little bit	70065	25.74
Moderately	49800	18.3
Quite a bit	52542	19.3
Extremely	22600	8.3
FELT CALM/ PEACEFUL		
All of the time	42221	15.55
Most of the time	109290	40.24
A good bit of the time	35858	13.2
Some of the time	49241	18.13
A little of the time	25718	9.47
HAVE A LOT OF ENERGY		
All of the time	19161	7.09
Most of the time	71294	26.39
A good bit of the time	41329	15.3
Some of the time	63539	23.52
A little of the time	47481	17.57
None of the time	27396	10.14

FELT DOWNHEARTED/ BLUE		
All of the time	9715	3.61
Most of the time	14084	5.24
A good bit of the time	15886	5.91
Some of the time	53625	19.95
A little of the time	70548	26.24
None of the time	104997	39.05
SOCIAL ACTIVITIES		
All of the time	16187	5.96
Most of the time	30100	11.08
Some of the time	52935	19.48
A little bit of the time	43133	15.87
None of the time	129359	47.61
PHYSICAL HEALTH COMPARED TO 1YR AGO		
Much better	18719	6.88
Slightly better	24764	9.1
About the same	146533	53.85
Slightly worse	61066	22.44
Much worse	21010	7.72
EMOTIONAL HEALTH COMPARED TO 1YR AGO		
Much better	26317	9.8
Slightly better	25100	9.35
About the same	171305	63.81
Slightly worse	33847	12.61
Much worse	11891	4.43
DIFFICULTY BATHING		
No difficulty	214964	79.61
Difficulty	42985	15.92
Unable to do	12085	4.48
DIFFICULTY DRESSING		
No difficulty	224532	83.17
Difficulty	37700	13.96
Unable to do	7734	2.86
DIFFICULTY EATING		
No difficulty	249652	92.67
Difficulty	16659	6.18
Unable to do	3082	1.14
DIFFICULTY GETTING IN/ OUT OF CHAIRS		
No difficulty	197436	73.27
Difficulty	65751	24.4
Unable to do	6274	2.33
DIFFICULTY WALKING		
No difficulty	169886	63.08
Difficulty	88100	32.71
Unable to do	11326	4.21
DIFFICULTY USING TOILET		

No difficulty	235974	87.62
Difficulty	27660	10.27
Unable to do	5690	2.11
CHEST PAIN DURING EXERCISE		
All of the time	5880	2.21
Most of the time	9391	3.53
Some of the time	26828	10.08
A little bit of the time	33400	12.54
None of the time	190767	71.65
CHEST PAIN WHILE RESTING		
All of the time	2442	0.91
Most of the time	4248	1.59
Some of the time	20603	7.72
A little bit of the time	29181	10.93
None of the time	210424	78.84
SHORT OF BREATH LYING FLAT		
All of the time	5765	2.16
Most of the time	7491	2.81
Some of the time	25155	9.45
A little bit of the time	29739	11.17
None of the time	198136	74.41
SHORT OF BREATH SITTING OR RESTING		
All of the time	2932	1.1
Most of the time	5655	2.13
Some of the time	24995	9.41
A little bit of the time	30410	11.45
None of the time	201649	75.91
SHORT OF BREATH WALKING LESS THAN 1 BLOCK		
All of the time	19910	7.52
Most of the time	21756	8.22
Some of the time	32729	12.37
A little bit of the time	38109	14.4
None of the time	152157	57.49
SHORT OF BREATH CLIMBING STAIRS		
All of the time	30498	11.6
Most of the time	25077	9.54
Some of the time	32143	12.22
A little bit of the time	46926	17.84
None of the time	128326	48.8
NUMBNESS IN FEET		
All of the time	20044	7.5
Most of the time	18071	6.76
Some of the time	34352	12.85
A little bit of the time	31518	11.79
None of the time	163340	61.1
TINGLING/ BURNING INT FEET		

All of the time	16109	6.03
Most of the time	17808	6.67
Some of the time	35035	13.12
A little bit of the time	33446	12.53
None of the time	164537	61.64
DECREASED FEELING OF HOT OR COLD IN FEET		
All of the time	10434	3.93
Most of the time	10698	4.03
Some of the time	22422	8.45
A little bit of the time	21002	7.92
None of the time	200780	75.67
SORES THAT DO NOT HEAL ON FEET		
All of the time	3729	1.4
Most of the time	2859	1.07
Some of the time	6712	2.52
A little bit of the time	9294	3.49
None of the time	243901	91.52
ARTHRITIS PAIN		
None	63889	24.13
Very mild	45133	17.05
Mild	51864	19.59
Moderate	71510	27.01
Severe	32377	12.23
Can see to read newspaper	243512	91.25
Can hear most things	221550	84.95
Hypertension or High Blood Pressure	178168	66.62
Angina Pectoris or Coronary Artery Disease	37523	14.23
Congestive Heart Failure	25220	9.52
Myocardial Infarction or Heart Attack	27667	10.42
Other Heart Conditions	59552	22.47
Stroke	24378	9.16
Emphysema	49897	18.74
Inflammatory Bowel Diseases	15777	5.96
Arthritis of hip or knee	116502	43.87
Arthritis of hand or wrist	102371	38.6
Osteoporosis	54039	20.46
Sciatica	68883	26.04
Diabetes	75016	28.13
Any Cancer (other than skin cancer)	38544	14.43
Under Treatment for Colon Cancer	2670	2.28
Under Treatment for Lung Cancer	1815	1.56
Under Treatment for Breast Cancer	4993	4.32
Under Treatment for Prostate Cancer	6839	6.12
BACK PAIN INTERFERED W/ACTIVITIES PAST 4 WEEKS		
All of the time	27135	10.18
Most of the time	29751	11.16

Some of the time	49410	18.54
A little of the time	53995	20.26
None of the time	106232	39.86
Sad/Blue for Two + Weeks in Past Year	78036	29.34
Depressed for Much of Past Year	54419	20.47
Depressed for Two + Years in Life	59659	22.58
DEPRESSED DURING PAST WEEK		
Rarely or none of the time	163164	61.49
Some or a little of the time	51505	19.41
Occasionally or a moderate amount of time	32571	12.27
Most or all of the time	18124	6.83
GENERAL HEALTH COMPARED TO PEERS		
Excellent	32239	12.1
Very good	68785	25.82
Good	82235	30.87
Fair	59332	22.28
Poor	23762	8.92
SMOKE		
Every day	23135	8.67
Some days	11373	4.26
Not at all	230563	86.4
Don’t know	1778	0.67
Urine Leakage in Past 6 Months	101176	38.5
MAGNITUDE OF URINE LEAKAGE PROBLEM		
Big problem	22482	16.63
Small problem	59059	43.69
Not a problem	53650	39.68
Talked with Doctor About Urine Leakage	50600	42.2
Received Treatment for Urine Leakage	31018	26.13
TALKED WITH DOCTOR ABOUT PHYSICAL ACTIVITES		
Yes	140704	54.13
No	109563	42.15
No visits in past 12 months	9693	3.73
Advised to Increase or Maintain Activities	128852	50.35
TAKLED TO DOCTOR ABOUT FALLING PROBLEM		
Yes	61122	23.2
No	193628	73.49
No visits in past 12 months	8719	3.31
Fell in Past 12 Months	68897	25.88
Problem with Walking or Balance in Past 12 Months	100282	37.83
TALKED WITH DOCTOR ABOUT HOW TO PREVENT FALLS		
Yes	90870	35.03
No	157620	60.77
No visits in past 12 months	10894	4.2
Had a Bone Density Test for Osteoporosis	126475	48.25

WHO COMPLETED SURVEY		
Person to whom survey was addressed	249206	87.39
Family member or relative	31167	10.93
Friend	2404	0.84
Professional caregiver	2399	0.84
BASELINE SURVEY DISPOSITION		
Mail 79.5-100% complete & 6 ADLs	203892	68.81
Mail 50-79.5% complete and <6 ADL's	7532	2.54
Mail 0-49% complete	595	0.2
Tele 79.5-100% complete & 6 ADLs	54762	18.48
Tele 50-79.5% complete and <6 ADL's	4054	1.37
Mail 0-49% complete	25485	8.6
BASELINE SURVEY ROUND		
1st mailing	149170	50.34
2nd mailing	62849	21.21
Partially completed by mail and converted to tele	1975	0.67
1st telephone	21812	7.36
2nd telephone	14710	4.96
3rd telephone	10735	3.62
4th telephone	7588	2.56
5th telephone	5711	1.93
6th telephone	4094	1.38
7th telephpne	2888	0.97
8th telephone	1991	0.67
9th telephone	12797	4.32
BASELINE SURVEY LANGUAGE		
English	282211	95.24
Spanish	13263	4.48
Chinese	846	0.29
CMS REGION CODE		
Boston	15311	5.17
NY	33970	11.46
Philadelphia	21670	7.31
Atlanta	42091	14.2
Chicago	65357	22.06
Dallas	32125	10.84
Kansas City	16211	5.47
Denver	10353	3.49
San Francisco	39153	13.21
Seattle	20079	6.78
TIME OF FOLLOW-UP		
Respondent	125548	42.37
Non-respondent	44692	15.08
Ineligible	1280	0.43
Disenrolled	102350	34.54
Dead	22450	7.58

Number of Days Physical Health Not Good	8.07 (*M*)	11.88 (*SD*)
Number of Days Mental Health Not Good	5.58 (*M*)	10.33 (*SD*)
Number of Days Poor Health interfered w/activities	6.74 (*M*)	

5.23 *SD*)

APPENDIX C

DISTRIBUTION OF DEPENDENT VARIABLES

Variable	***Frequency***	***Percent***
GENERAL HEALTH		
Excellent	17491	6.29
Very Good	64323	23.13
Good	99681	35.85
Fair	72385	26.03
Poor	24179	8.7
NUMBER OF DAYS PHYSICAL HEALTH NOT GOOD		
0	129907	49.78
1-5	36398	13.95
6-10	23284	8.92
11-15	16637	6.37
16-20	11926	4.57
21+	42832	16.41
TIME OF FOLLOW-UP		
Respondent	125548	42.37
Non-respondent	44692	15.08
Ineligible	1280	0.43
Disenrolled	102350	34.54
Dead	22450	7.58
EXCELLENT HEALTH	17491	6.29
DEAD AT FOLLOWUP	22450	7.58
Number of Days Physical Health Not Good	8.07 (*M*)	11.88 (*SD*)

APPENDIX D

FULL SAS CODE

```
*************************************************************************
************************
*************************************************************************
************************;
*  Program Name: C15A_PUF_IMPORT_CODE.sas
*  Created By: Health Services Advisory Group on 06/30/2015
*  Edited By: Gina McKernan, PhD candidate on 5/01/2017
*  Purpose: SAS code for importing a data file from C15A_PUF.txt file
and analytic processes
*
*  Source Data: 1) C15A_PUF.txt, created 09/17/2015
*************************************************************************
************************;
* Create cohort specific prefix for each field;
%Let c = C15;       * Baseline fields;
%Let r = R15;       * Follow Up fields;
%Let p = P15;       * Fields created at time of Performance Measurement
analysis;

* Create a libname with a path that will contain the imported PUF sas
data file;
libname HOS '/n04/data/entactin/EA_GINA/HOS';

/* Import the puf text file */
filename &c.A "/n04/data/entactin/EA_GINA/HOS/C15A_PUF.TXT"; *Specify
location path for the PUF text file;

/*************************************************************************
************************/
data HOS.&c.A_PUF;
        infile &c.A;
    input
      /* demographics and ID
    /*  var         position    varnum  */
        CASE_ID     $ 1-9     /* 1 Random ID number -- Different across
years for all                    */
        AGE           10      /* 2 Rolled-up Age groups
           */
```

```
        RACE          11    /* 3 Rolled-up Race groups - baseline race
codes used                    */
        GENDER        12    /* 4 Gender
           */
        MRSTAT        13    /* 5 Rolled-up Marriage status
           */
        EDUC          14    /* 6 Rolled-up Education level
           */
        BMICAT        15    /* 7 Rolled-up BMICAT - baseline data used
- variable is output          */

    /* Survey - Baseline data  */

    /* Baseline Survey Fields  - column 16 - 91 */
    /* var         position      varnum  */
       &c.VRGENHTH   16    /*   8  Q1 General Health Question
           */
       &c.VRMACT     17    /*   9  Q2a Health Limitation-In moderate
activities                    */
       &c.VRSTAIR    18    /*  10  Q2b Health Limitation-Climbing
several flights                  */
       &c.VRPACCL    19    /*  11  Q3a Physical-Accomplished less than
you would like                */
       &c.VRPWORK    20    /*  12  Q3b Physical-Limited in work or
activities                    */
       &c.VRMACCL    21    /*  13  Q4a Emotional-Accomplished less
than you would like           */
       &c.VRMWORK    22    /*  14  Q4b Emotional-Did not do work /
activities as carefully       */
       &c.VRPAIN     23    /*  15  Q5 Pain-Interfered with normal work
           */
       &c.VRCALM     24    /*  16  Q6a Felt calm and peaceful
           */
       &c.VRENERGY   25    /*  17  Q6b Have a lot of energy
           */
       &c.VRDOWN     26    /*  18  Q6c Felt downhearted and blue
           */
       &c.VRSACT     27    /*  19  Q7 Social Activities
           */
       &c.VRPHCMP    28    /*  20  Q8 Physical Health compared to 1
Year Ago                      */
       &c.VRMHCMP    29    /*  21  Q9 Emotional Health compared to 1
Year Ago                      */
       &c.ADLBTH     30    /*  22  Q10a Difficulty Bathing
           */
       &c.ADLDRS     31    /*  23  Q10b Difficulty Dressing
           */
       &c.ADLEAT     32    /*  24  Q10c Difficulty Eating
           */
       &c.ADLCHR     33    /*  25  Q10d Difficulty Getting in or out
of Chairs                     */
```

```
       &c.ADLWLK      34     /*  26  Q10e Difficulty Walking
           */
       &c.ADLTLT      35     /*  27  Q10f Difficulty Using Toilet
           */
       &c.HDPHY       36-37  /*  28  Q11 Number of Days Physical
Health Not Good                    */
       &c.HDMEN       38-39  /*  29  Q12 Number of Days Mental Health
Not Good                         */
       &c.HDACT       40-41     /*  30  Q13 Number of Days Poor Health
interfered w/activities             */
       &c.CHSTEX      42     /*  31  Q14a Chest Pain-Exercise
           */
       &c.CHSTRST     43     /*  32  Q14b Chest Pain-Resting
           */
       &c.SOBFLT      44     /*  33  Q15a Short of Breath lying flat
           */
       &c.SOBSIT      45     /*  34  Q15b Short of Breath sitting or
resting                          */
       &c.SOBWLK      46     /*  35  Q15c Short of Breath walking less
than 1 block                     */
       &c.SOBSTR      47     /*  36  Q15d Short of Breath climbing 1
flight stairs                    */
       &c.FTNUMB      48     /*  37  Q16a Numbness or Loss of feeling in
feet                             */
       &c.FTSENS      49     /*  38  Q16b Tingling burning in feet
           */
       &c.FTHC        50     /*  39  Q16c Decreased feeling of hot or
cold in feet                     */
       &c.FTSRS       51     /*  40  Q16d Sores that do not heal on feet
           */
       &c.PNART       52     /*  41  Q17 Arthritis pain
           */
       &c.READ        53     /*  42  Q18 See to read newspaper
           */
       &c.HEAR        54     /*  43  Q19 Hear most things
           */
       &c.CCHBP       55     /*  44  Q20 Hypertension or High Blood
Pressure                            */
       &c.CC_CAD      56     /*  45  Q21 Angina Pectoris or Coronary
Artery Disease                   */
       &c.CC_CHF      57     /*  46  Q22 Congestive Heart Failure
           */
       &c.CCMI        58     /*  47  Q23 A Myocardial Infarction or
Heart Attack                        */
       &c.CCHRTOTH    59     /*  48  Q24 Other Heart Conditions
           */
       &c.CCSTROKE    60     /*  49  Q25 Stroke
           */
       &c.CC_COPD     61     /*  50  Q26 Emphysema
           */
       &c.CCGI        62     /*  51  Q27 Inflammatory Bowel Diseases
           */
```

```
        &c.CCARTHIP    63     /*  52  Q28 Arthritis of hip or knee
            */
        &c.CCARTHND    64     /*  53  Q29 Arthritis of hand or wrist
            */
        &c.CCOSTEO     65     /*  54  Q30 Osteoporosis
            */
        &c.CCSCIATI    66      /*  55  Q31 Sciatica
            */
        &c.CCDIABET    67     /*  56  Q32 Diabetes
            */
        &c.CCANYCA     68      /*  57  Q33 Any Cancer (other than skin
cancer)                          */
        &c.CACOLON     69      /*  58  Q34a Under Treatment for Colon
Cancer                             */
        &c.CALUNG      70      /*  59  Q34b Under Treatment for Lung
Cancer                             */
        &c.CABRST      71      /*  60  Q34c Under Treatment for Breast
Cancer                           */
        &c.CAPROS      72      /*  61  Q34d Under Treatment for Prostate
Cancer                           */
        &c.PNBACK      73      /*  62  Q35 Back Pain Interfered
w/Activities in Past 4 Weeks                */
        &c.DEP2WK      74      /*  63  Q36 Sad/Blue for Two + Weeks in
Past Year                        */
        &c.DEPYR       75      /*  64  Q37 Depressed for Much of Past
Year                               */
        &c.DEP2YR      76      /*  65  Q38 Depressed for Two + Years in
Life                             */
        &c.DEPWEEK     77      /*  66  Q39 Depressed for How Much of the
time in Past Week                */
        &c.CMPHTH      78      /*  67  Q40 General Health compared to
peers                              */
        &c.SMOKE       79      /*  68  Q41 Smoke every day
            */
        &c.MUILKG      80      /*  69  Q42 Urine Leakage in Past 6 Months
            */
        &c.MUIMAG      81      /*  70  Q43 Magnitude of Urine Leakage
Problem                            */
        &c.MUITLK      82      /*  71  Q44 Talked with Doctor About Urine
Leakage                          */
        &c.MUITRT      83      /*  72  Q45 Received Treatment for Urine
Leakage                          */
        &c.PAOTLK      84      /*  73  Q46 Talked with Doctor About
Physical Activities                */
        &c.PAOADV      85      /*  74  Q47 Advised to Increase or
Maintain Activities                */
        &c.FRMTLK      86      /*  75  Q48 Talked with Doctor about
Falling or Balance Problem         */
        &c.FRMFALL     87      /*  76  Q49 Fell in Past 12 Months
            */
        &c.FRMBAL      88      /*  77  Q50 Problem with Walking or
Balance in Past 12 Months          */
```

```
       &c.FRMPREV    89     /*  78  Q51 Talked with Doctor about How
to Prevent Falls                 */
       &c.OTOTEST    90     /*  79  Q52 Bone Density Test for
Osteoporosis                              */
       &c.CMPWHO     91    /*  80  Q62 Who completed Survey
          */

     /* Baseline Survey/Plan-Related Fields */
       &c.SRVDISP  $ 92-94  /*  81  Baseline survey disposition
                                  */
         &c.SRVMODE  $ 95-96  /*  82  Baseline survey round
                                  */
       &c.PCTCMP   $ 97-101 /*  83  Use character form of variable for
baseline percent complete  */
       &c.SRVLANG   102     /*  84  Baseline survey language
                                  */

          /* Follow Up Survey - column 103-176 */
          /* var      position      varnum  */
         &R.VRGENHTH   103        /*    85    Q1 General Health Question
                                       */
         &R.VRMACT     104        /*    86    Q2a Health Limitation-In
moderate activities                    */
         &R.VRSTAIR    105        /*    87    Q2b Health Limitation-
Climbing several flights                    */
         &R.VRPACCL    106        /*    88    Q3a Physical-Accomplished
less than you would like               */
         &R.VRPWORK    107        /*    89    Q3b Physical-Limited in
work or activities                          */
         &R.VRMACCL    108        /*    90    Q4a Emotional-Accomplished
less than you would like               */
         &R.VRMWORK    109        /*    91    Q4b Emotional-Did not do
work / activities as carefully         */
         &R.VRPAIN     110        /*    92    Q5 Pain-Interfered with
normal work                                 */
         &R.VRCALM     111        /*    93    Q6a Felt calm and peaceful
                                       */
         &R.VRENERGY   112        /*    94    Q6b Have a lot of energy
                                       */
         &R.VRDOWN     113        /*    95    Q6c Felt downhearted and
blue                                        */
         &R.VRSACT     114        /*    96    Q7 Social Activities
                                       */
         &R.VRPHCMP    115        /*    97    Q8 Physical Health compared
to 1 Year Ago                          */
         &R.VRMHCMP    116        /*    98    Q9 Emotional Health
compared to 1 Year Ago                      */
         &R.ADLBTH     117        /*    99    Q10a Difficulty Bathing
                                            */
         &R.ADLDRS     118        /* 100  Q10b Difficulty Dressing
                                       */
```

```
        &R.ADLEAT     119        /* 101     Q10c Difficulty Eating
                                                          */
        &R.ADLCHR     120        /* 102     Q10d Difficulty Getting in
or out of Chairs                                          */
        &R.ADLWLK     121        /* 103     Q10e Difficulty Walking
                                                          */
        &R.ADLTLT     122        /* 104     Q10f Difficulty Using
Toilet                                                    */
        &R.DIFMEALS   123        /* 105     Q11a Difficulty Preparing
Meals                                                     */
        &R.DIFMONEY   124        /* 106     Q11b Difficulty Managing
Money                                                     */
        &R.DIFMEDS    125        /* 107     Q11c Difficulty Taking
Medication as Prescribed                                  */
        &R.HDPHY           126-127    /* 108     Q12 Number of Days
Physical Health Not Good                                  */
        &R.HDMEN           128-129    /* 109     Q13 Number of Days
Mental Health Not Good                                    */
        &R.HDACT           130-131    /* 110     Q14 Number of Days
Poor Health interfered w/activities                       */
        &R.DIFSEE     132        /* 111     Q15 Blind or Serious
Difficulty Seeing                                         */
        &R.DIFHEAR    133        /* 112     Q16 Deaf or Serious
Difficulty Hearing                                        */
        &R.DIFREMEM   134        /* 113     Q17 Difficulty
concentrating, remembering, or making decisions  */
        &R.DIFERRND   135        /* 114     Q18 Difficulty doing
errands                                                   */
        &R.DIFMPROB   136        /* 115     Q19 Memory problems
interfered with activities in past month    */
        &R.CCHBP           137        /* 116     Q20 Hypertension or
High Blood Pressure                                       */
        &R.CC_CAD     138        /* 117     Q21 Angina Pectoris or
Coronary Artery Disease                                   */
        &R.CC_CHF     139        /* 118     Q22 Congestive Heart
Failure                                                   */
        &R.CCMI       140        /* 119     Q23 Myocardial Infarction
or Heart Attack                                           */
        &R.CCHRTOTH   141        /* 120     Q24 Other Heart Conditions
                                                          */
        &R.CCSTROKE   142        /* 121     Q25 Stroke
                                                          */
        &R.CC_COPD    143        /* 122     Q26 Emphysema, or Asthma,
or COPD                                                   */
        &R.CCGI       144        /* 123     Q27 Inflammatory Bowel
Diseases                                                  */
        &R.CCARTHIP   145        /* 124     Q28 Arthritis of hip or
knee                                                      */
        &R.CCARTHND   146        /* 125     Q29 Arthritis of hand or
wrist                                                     */
        &R.CCOSTEO    147        /* 126     Q30 Osteoporosis, or
thin/brittle bones                                        */
```

```
        &R.CCSCIATI   148        /* 127     Q31 Sciatica, or
pain/numbness traveling down leg                    */
        &R.CCDIABET   149        /* 128     Q32 Diabetes, or high blood
sugar, or sugar in the urine          */
        &R.CCDEP           150        /* 129     Q33 Depression
                                                    */
        &R.CCANYCA    151        /* 130     Q34 Any Cancer (other than
skin cancer)                                        */
        &R.CACOLON    152        /* 131     Q35a Under Treatment for
Colon Cancer                                        */
        &R.CALUNG     153        /* 132     Q35b Under Treatment for
Lung Cancer                                         */
        &R.CABRST     154        /* 133     Q35c Under Treatment for
Breast Cancer                                       */
        &R.CAPROS     155        /* 134     Q35d Under Treatment for
Prostate Cancer                                     */
        &R.CAOTHER    156        /* 135     Q35e Under Treatment for
Other Cancer                                        */
        &R.PAINDACT   157        /* 136     Q36 Pain interfered with
activities in past 7 days                           */
        &R.PAINSACT   158        /* 137     Q37 Pain kept you from
socializing in past 7 days                          */
        &R.PAINRATE   159-160    /* 138     Q38 Average pain rating in
past 7 days (1=No pain,10=Worst pain)*/
        &R.DEPNOPLS   161        /* 139     Q39a Little interest or
pleasure in doing things in past 2 weeks*/
        &R.DEPDOWN    162        /* 140     Q39b Feeling down,
depressed, or hopeless in past 2 weeks      */
        &R.CMPHTH     163        /* 141     Q40 General Health compared
to peers                                            */
        &R.SMOKE           164        /* 142     Q41 Smoke every day,
some days, or not at all                            */
        &R.MUILKG     165        /* 143     Q42 Urine Leakage in Past 6
Months                                              */
        &R.MUIMAG     166        /* 144     Q43 Magnitude of Urine
Leakage Problem                                     */
        &R.MUITLK     167        /* 145     Q44 Talked with Doctor
About Urine Leakage                                 */
        &R.MUITRT     168        /* 146     Q45 Received Treatment for
Urine Leakage                                       */
        &R.PAOTLK     169        /* 147     Q46 Talked with Doctor
About Physical Activities                           */
        &R.PAOADV     170        /* 148     Q47 Advised to Increase or
Maintain Activities                                 */
        &R.FRMTLK     171        /* 149     Q48 Talked with Doctor
about Falling or Balance Problem                    */
        &R.FRMFALL    172        /* 150     Q49 Fell in Past 12 Months
                                                    */
        &R.FRMBAL     173        /* 151     Q50 Problem with Walking or
Balance in Past 12 Months             */
        &R.FRMPREV    174        /* 152     Q51 Talked with Doctor
about How to Prevent Falls                          */
```

```
        &R.OTOTEST    175        /* 153     Q52 Bone Density Test for
Osteoporosis                                  */
        &R.CMPWHO     176        /* 154     Q64 Who completed Survey
                                              */

       /* Follow Up Survey/Plan-Related Fields */
       &r.SRVDISP $ 177-179/* 155  Follow Up survey disposition
           */
       &r.SRVMODE $ 180-181 /* 156  Follow Up survey round
           */
       &r.PCTCMP  $ 182-186 /* 157  Use character form of var for
follow up percent complete            */
       &r.SRVLANG   187    /* 158  Follow Up survey language
           */
       COHORT     $ 188-190/* 159  Cohort for this PUF
           */
       &p.PLREGCDE  191-192 /* 160  Follow Up CMS region for this
combined PUF                          */
       SAMPLED      193     /* 161  Sampled at Follow Up Flag
           */
       SFLAG        194     /* 162  Dead, Disenroll, Inval, Resp,
Nonresp Categories Flag               */
;

/*****************************************************************
************************/
label
          CASE_ID          ="&c. Random Id number
                                    "
          AGE              ="&c. Rolled-up Age groups
                                    "
          RACE       ="&c. Rolled-up Survey Race groups
                              "
          GENDER           ="&c. Survey Gender
                                          "
          MRSTAT           ="&c. Rolled-up Marriage status
                                    "
          EDUC       ="&c. Rolled-up Education level
                                "
          BMICAT           ="&c. Rolled-up BMICAT
                                          "
          &c.VRGENHTH      ="&c. Q1 General Health Question
                                "
          &c.VRMACT  ="&c. Q2a Health Limitation-In moderate
activities                    "
          &c.VRSTAIR ="&c. Q2b Health Limitation-Climbing several
flights                       "
          &c.VRPACCL ="&c. Q3a Physical-Accomplished less than you
would like                   "
          &c.VRPWORK ="&c. Q3b Physical-Limited in work or activities
                      "
```

```
            &c.VRMACCL ="&c. Q4a Emotional-Accomplished less than you
would like                  "
            &c.VRMWORK ="&c. Q4b Emotional-Did not do work / activities
as carefully                "
            &c.VRPAIN  ="&c. Q5 Pain-Interfered with normal work
                            "
            &c.VRCALM  ="&c. Q6a Felt calm and peaceful
                            "
            &c.VRENERGY     ="&c. Q6b Have a lot of energy
                                  "
            &c.VRDOWN  ="&c. Q6c Felt downhearted and blue
                                "
            &c.VRSACT  ="&c. Q7 Social Activities
            "
            &c.VRPHCMP ="&c. Q8 Physical Health compared to 1 Year Ago
                            "
            &c.VRMHCMP ="&c. Q9 Emotional Health compared to 1 Year Ago
                           "
            &c.ADLBTH  ="&c. Q10a Difficulty Bathing
                                "
            &c.ADLDRS  ="&c. Q10b Difficulty Dressing
                                "
            &c.ADLEAT  ="&c. Q10c Difficulty Eating
                                "
            &c.ADLCHR  ="&c. Q10d Difficulty Getting in or out of Chairs
                           "
            &c.ADLWLK  ="&c. Q10e Difficulty Walking
                                "
            &c.ADLTLT  ="&c. Q10f Difficulty Using Toilet
                                "
            &c.HDPHY   ="&c. Q11 Number of Days Physical Health Not Good
                           "
            &c.HDMEN   ="&c. Q12 Number of Days Mental Health Not Good
                            "
            &c.HDACT   ="&c. Q13 Number of Days Poor Health interfered
w/activities                "
            &c.CHSTEX  ="&c. Q14a Chest Pain-Exercise
                                "
            &c.CHSTRST ="&c. Q14b Chest Pain-Resting
                                "
            &c.SOBFLT  ="&c. Q15a Short of Breath lying flat
                            "
            &c.SOBSIT  ="&c. Q15b Short of Breath sitting or resting
                            "
            &c.SOBWLK  ="&c. Q15c Short of Breath walking less than 1
block                       "
            &c.SOBSTR  ="&c. Q15d Short of Breath climbing 1 flight
stairs                      "
            &c.FTNUMB  ="&c. Q16a Numbness or Loss of feeling in feet
                            "
            &c.FTSENS  ="&c. Q16b Tingling burning in feet
                                "
```

```
            &c.FTHC          ="&c. Q16c Decreased feeling of hot or cold
in feet                                 "
            &c.FTSRS    ="&c. Q16d Sores that do not heal on feet
                                "
            &c.PNART    ="&c. Q17 Arthritis pain
                  "
            &c.READ          ="&c. Q18 See to read newspaper
                                        "
            &c.HEAR          ="&c. Q19 Hear most things
                                        "
            &c.CCHBP    ="&c. Q20 Hypertension or High Blood Pressure
                                "
            &c.CC_CAD   ="&c. Q21 Angina Pectoris or Coronary Artery
Disease                         "
            &c.CC_CHF   ="&c. Q22 Congestive Heart Failure
                                "
            &c.CCMI          ="&c. Q23 A Myocardial Infarction or Heart
Attack                          "
            &c.CCHRTOTH      ="&c. Q24 Other Heart Conditions
                                "
            &c.CCSTROKE      ="&c. Q25 Stroke
            "
            &c.CC_COPD ="&c. Q26 Emphysema
                  "
            &c.CCGI          ="&c. Q27 Inflammatory Bowel Diseases
                                "
            &c.CCARTHIP      ="&c. Q28 Arthritis of hip or knee
                                        "
            &c.CCARTHND      ="&c. Q29 Arthritis of hand or wrist
                                "
            &c.CCOSTEO ="&c. Q30 Osteoporosis
                                        "
            &c.CCSCIATI      ="&c. Q31 Sciatica
                                                "
            &c.CCDIABET      ="&c. Q32 Diabetes
            "
            &c.CCANYCA ="&c. Q33 Any Cancer (other than skin cancer)
            "
            &c.CACOLON ="&c. Q34a Under Treatment for Colon Cancer
                                "
            &c.CALUNG   ="&c. Q34b Under Treatment for Lung Cancer
                                "
            &c.CABRST   ="&c. Q34c Under Treatment for Breast Cancer
                                "
            &c.CAPROS   ="&c. Q34d Under Treatment for Prostate Cancer
                                "
            &c.PNBACK   ="&c. Q35 Back Pain Interfered w/Activities in
Past 4 Weeks                    "
            &c.DEP2WK   ="&c. Q36 Sad/Blue for Two + Weeks in Past Year
                                "
            &c.DEPYR    ="&c. Q37 Depressed for Much of Past Year
                                "
```

```
            &c.DEP2YR  ="&c. Q38 Depressed for Two + Years in Life
            "
            &c.DEPWEEK ="&c. Q39 Depressed for How Much of the time in
Past Week                  "
            &c.CMPHTH  ="&c. Q40 General Health compared to peers
                            "
            &c.SMOKE   ="&c. Q41 Smoke every day
                              "
            &c.MUILKG  ="&c. Q42 Urine Leakage in Past 6 Months
            "
            &c.MUIMAG  ="&c. Q43 Magnitude of Urine Leakage Problem
                            "
            &c.MUITLK  ="&c. Q44 Talked with Doctor About Urine Leakage
                           "
            &c.MUITRT  ="&c. Q45 Received Treatment for Urine Leakage
                            "
            &c.PAOTLK  ="&c. Q46 Talked with Doctor About Physical
Activities                  "
            &c.PAOADV  ="&c. Q47 Advised to Increase or Maintain
Activities                  "
            &c.FRMTLK  ="&c. Q48 Talked with Doctor about Falling or
Balance Problem            "
            &c.FRMFALL ="&c. Q49 Fell in Past 12 Months
                            "
            &c.FRMBAL  ="&c. Q50 Problem with Walking or Balance in Past
12 Months                  "
            &c.FRMPREV ="&c. Q51 Talked with Doctor about How to Prevent
Falls                      "
            &c.OTOTEST ="&c. Q52 Bone Density Test for Osteoporosis
                            "
            &c.CMPWHO  ="&c. Q62 Who completed Survey
                              "
            &c.SRVDISP ="&c. Baseline Survey Disposition
                              "
            &c.SRVMODE ="&c. Baseline Survey Round
                              "
            &c.PCTCMP  ="&c. Baseline Percent of Survey Completed
                            "
            &c.SRVLANG ="&c. Baseline Survey Language
                            "
            &r.VRGENHTH     ="&r. Q1 General Health Question
                              "
            &r.VRMACT  ="&r. Q2a Health Limitation-In moderate
activities                  "
            &r.VRSTAIR ="&r. Q2b Health Limitation-Climbing several
flights                     "
            &r.VRPACCL ="&r. Q3a Physical-Accomplished less than you
would like                 "
            &r.VRPWORK ="&r. Q3b Physical-Limited in work or activities
                           "
            &r.VRMACCL ="&r. Q4a Emotional-Accomplished less than you
would like                 "
```

```
            &r.VRMWORK ="&r. Q4b Emotional-Did not do work / activities
as carefully                    "
            &r.VRPAIN  ="&r. Q5 Pain-Interfered with normal work
                                "
            &r.VRCALM  ="&r. Q6a Felt calm and peaceful
                                "
            &r.VRENERGY     ="&r. Q6b Have a lot of energy
                                        "
            &r.VRDOWN  ="&r. Q6c Felt downhearted and blue
                                "
            &r.VRSACT  ="&r. Q7 Social Activities
                                "
            &r.VRPHCMP ="&r. Q8 Physical Health compared to 1 Year Ago
                                "
            &r.VRMHCMP ="&r. Q9 Emotional Health compared to 1 Year Ago
                               "
            &r.ADLBTH  ="&r. Q10a Difficulty Bathing
                                "
            &r.ADLDRS  ="&r. Q10b Difficulty Dressing
                                "
            &r.ADLEAT  ="&r. Q10c Difficulty Eating
                                "
            &r.ADLCHR  ="&r. Q10d Difficulty Getting in or out of Chairs
                               "
            &r.ADLWLK  ="&r. Q10e Difficulty Walking
                                "
            &r.ADLTLT  ="&r. Q10f Difficulty Using Toilet
                                "
            &r.DIFMEALS     ="&r. Q11a Difficulty Preparing Meals
                                "
            &r.DIFMONEY     ="&r. Q11b Difficulty Managing Money
                                "
            &r.DIFMEDS ="&r. Q11c Difficulty Taking Medication as
Prescribed                      "
            &r.HDPHY   ="&r. Q12 Number of Days Physical Health Not Good
                               "
            &r.HDMEN   ="&r. Q13 Number of Days Mental Health Not Good
                                "
            &r.HDACT   ="&r. Q14 Number of Days Poor Health interfered
w/activities                    "
            &r.DIFSEE  ="&r. Q15 Blind or Serious Difficulty Seeing
                                "
            &r.DIFHEAR ="&r. Q16 Deaf or Serious Difficulty Hearing
                                "
            &r.DIFREMEM     ="&r. Q17 Difficulty concentrating,
remembering, or making decisions "
            &r.DIFERRND     ="&r. Q18 Difficulty doing errands
                                        "
            &r.DIFMPROB     ="&r. Q19 Memory problems interfered with
activities in past month        "
            &r.CCHBP   ="&r. Q20 Hypertension or High Blood Pressure
                                "
```

```
            &r.CC_CAD  ="&r. Q21 Angina Pectoris or Coronary Artery
Disease                                  "
            &r.CC_CHF  ="&r. Q22 Congestive Heart Failure
                                         "
            &r.CCMI          ="&r. Q23 Myocardial Infarction or Heart
Attack                                                 "
            &r.CCHRTOTH      ="&r. Q24 Other Heart Conditions
                                         "
            &r.CCSTROKE      ="&r. Q25 Stroke
                                                       "
            &r.CC_COPD ="&r. Q26 Emphysema, or Asthma, or COPD
                                         "
            &r.CCGI          ="&r. Q27 Inflammatory Bowel Diseases
                                         "
            &r.CCARTHIP      ="&r. Q28 Arthritis of hip or knee
                                                       "
            &r.CCARTHND      ="&r. Q29 Arthritis of hand or wrist
                                         "
            &r.CCOSTEO ="&r. Q30 Osteoporosis, or thin/brittle bones
                                         "
            &r.CCSCIATI      ="&r. Q31 Sciatica, or pain/numbness
traveling down leg                                     "
            &r.CCDIABET      ="&r. Q32 Diabetes, or high blood sugar, or
sugar in the urine                       "
            &r.CCDEP   ="&r. Q33 Depression
                                         "
            &r.CCANYCA ="&r. Q34 Any Cancer (other than skin cancer)
                                         "
            &r.CACOLON ="&r. Q35a Under Treatment for Colon Cancer
                                         "
            &r.CALUNG  ="&r. Q35b Under Treatment for Lung Cancer
                                         "
            &r.CABRST  ="&r. Q35c Under Treatment for Breast Cancer
                                         "
            &r.CAPROS  ="&r. Q35d Under Treatment for Prostate Cancer
                                         "
            &r.CAOTHER ="&r. Q35e Under Treatment for Other Cancer
                                         "
            &r.PAINDACT      ="&r. Q36 Pain interfered with activities
in past 7 days                           "
            &r.PAINSACT      ="&r. Q37 Pain kept you from socializing in
past 7 days                                            "
            &r.PAINRATE      ="&r. Q38 Average pain rating in past 7
days (1=No pain,10=Worst pain)   "
            &r.DEPNOPLS      ="&r. Q39a Little interest or pleasure in
doing things in past 2 weeks     "
            &r.DEPDOWN ="&r. Q39b Feeling down, depressed, or hopeless
in past 2 weeks                  "
            &r.CMPHTH  ="&r. Q40 General Health compared to peers
                                         "
            &r.SMOKE   ="&r. Q41 Smoke every day, some days, or not at
all                                      "
```

```
            &r.MUILKG  ="&r. Q42 Urine Leakage in Past 6 Months
                         "
            &r.MUIMAG  ="&r. Q43 Magnitude of Urine Leakage Problem
                         "
            &r.MUITLK  ="&r. Q44 Talked with Doctor About Urine Leakage
                        "
            &r.MUITRT  ="&r. Q45 Received Treatment for Urine Leakage
                         "
            &r.PAOTLK  ="&r. Q46 Talked with Doctor About Physical
Activities
            &r.PAOADV  ="&r. Q47 Advised to Increase or Maintain
Activities
            &r.FRMTLK  ="&r. Q48 Talked with Doctor about Falling or
Balance Problem
            &r.FRMFALL ="&r. Q49 Fell in Past 12 Months
                         "
            &r.FRMBAL  ="&r. Q50 Problem with Walking or Balance in Past
12 Months
            &r.FRMPREV ="&r. Q51 Talked with Doctor about How to Prevent
Falls
            &r.OTOTEST ="&r. Q52 Bone Density Test for Osteoporosis
                         "
            &r.CMPWHO  ="&r. Q64 Who completed Survey
                              "
            &r.SRVDISP ="&r. Follow Up Survey Disposition
                  "
            &r.SRVMODE  ="&r. Follow Up Survey Round
                  "
            &r.PCTCMP   ="&r. Follow Up Percent of Survey Completed
                  "
            &r.SRVLANG  ="&r. Follow Up Survey Language
                  "
            COHORT      ="&r. COHORT ID
                  "
            &p.PLREGCDE ="&r. Reported Plan CMS Region Code
                  "
            SAMPLED     ="&r. Follow Up Sample Indicator
                  "
            SFLAG       ="&r. Dead, Disenroll, Inval, Resp, Nonresp
(Analytic Sample)
                         "
          ;

run;


proc contents varnum data=HOS.&c.A_PUF;
          title " &c. Analytic PUF Imported Data";
run;
/* ---------------------------------------------------------------------
   Code generated by SAS Task

   Generated on: Monday, November 07, 2016 at 11:41:11 AM
```

```
   By task: One-Way Frequencies (2)

   Input Data: EMApp:HOS.C15A_PUF
   Server:  EMApp
   ---------------------------------------------------------------------
*/

%_eg_conditional_dropds(WORK.SORT);
/* ---------------------------------------------------------------------
   Sort data set EMApp:HOS.C15A_PUF
   ---------------------------------------------------------------------
*/

PROC SQL;
     CREATE VIEW WORK.SORT AS
          SELECT T.AGE, T.RACE, T.GENDER, T.MRSTAT, T.EDUC, T.BMICAT,
T.C15VRGENHTH, T.C15VRMACT, T.C15VRSTAIR, T.C15VRPACCL, T.C15VRPWORK,
T.C15VRMACCL, T.C15VRMWORK, T.C15VRPAIN, T.C15VRCALM, T.C15VRENERGY,
T.C15VRDOWN, T.C15VRSACT, T.C15VRPHCMP
                , T.C15VRMHCMP, T.C15ADLBTH, T.C15ADLDRS, T.C15ADLEAT,
T.C15ADLCHR, T.C15ADLWLK, T.C15ADLTLT, T.C15HDPHY, T.C15HDMEN,
T.C15HDACT, T.C15CHSTEX, T.C15CHSTRST, T.C15SOBFLT, T.C15SOBSIT,
T.C15SOBWLK, T.C15SOBSTR, T.C15FTNUMB
                , T.C15FTSENS, T.C15FTHC, T.C15FTSRS, T.C15PNART,
T.C15READ, T.C15HEAR, T.C15CCHBP, T.C15CC_CAD, T.C15CC_CHF, T.C15CCMI,
T.C15CCHRTOTH, T.C15CCSTROKE, T.C15CC_COPD, T.C15CCGI, T.C15CCARTHIP,
T.C15CCARTHND, T.C15CCOSTEO
                , T.C15CCSCIATI, T.C15CCDIABET, T.C15CCANYCA,
T.C15CACOLON, T.C15CALUNG, T.C15CABRST, T.C15CAPROS, T.C15PNBACK,
T.C15DEP2WK, T.C15DEPYR, T.C15DEP2YR, T.C15DEPWEEK, T.C15CMPHTH,
T.C15SMOKE, T.C15MUILKG, T.C15MUIMAG, T.C15MUITLK
                , T.C15MUITRT, T.C15PAOTLK, T.C15PAOADV, T.C15FRMTLK,
T.C15FRMFALL, T.C15FRMBAL, T.C15FRMPREV, T.C15OTOTEST, T.C15CMPWHO,
T.C15SRVDISP, T.C15SRVMODE, T.C15PCTCMP, T.C15SRVLANG, T.R15VRGENHTH,
T.R15VRMACT, T.R15VRSTAIR
                , T.R15VRPACCL, T.R15VRPWORK, T.R15VRMACCL,
T.R15VRMWORK, T.R15VRPAIN, T.R15VRCALM, T.R15VRENERGY, T.R15VRDOWN,
T.R15VRSACT, T.R15VRPHCMP, T.R15VRMHCMP, T.R15ADLBTH, T.R15ADLDRS,
T.R15ADLEAT, T.R15ADLCHR, T.R15ADLWLK, T.R15ADLTLT
                , T.R15DIFMEALS, T.R15DIFMONEY, T.R15DIFMEDS,
T.R15HDPHY, T.R15HDMEN, T.R15HDACT, T.R15DIFSEE, T.R15DIFHEAR,
T.R15DIFREMEM, T.R15DIFERRND, T.R15DIFMPROB, T.R15CCHBP, T.R15CC_CAD,
T.R15CC_CHF, T.R15CCMI, T.R15CCHRTOTH
                , T.R15CCSTROKE, T.R15CC_COPD, T.R15CCGI,
T.R15CCARTHIP, T.R15CCARTHND, T.R15CCOSTEO, T.R15CCSCIATI,
T.R15CCDIABET, T.R15CCDEP, T.R15CCANYCA, T.R15CACOLON, T.R15CALUNG,
T.R15CABRST, T.R15CAPROS, T.R15CAOTHER, T.R15PAINDACT
                , T.R15PAINSACT, T.R15PAINRATE, T.R15DEPNOPLS,
T.R15DEPDOWN, T.R15CMPHTH, T.R15SMOKE, T.R15MUILKG, T.R15MUIMAG,
T.R15MUITLK, T.R15MUITRT, T.R15PAOTLK, T.R15PAOADV, T.R15FRMTLK,
T.R15FRMFALL, T.R15FRMBAL, T.R15FRMPREV, T.R15OTOTEST
```

```
                , T.R15CMPWHO, T.R15SRVDISP, T.R15SRVMODE,
T.R15PCTCMP, T.R15SRVLANG, T.COHORT, T.P15PLREGCDE, T.SAMPLED, T.SFLAG
     FROM HOS.C15A_PUF as T
;
QUIT;

TITLE;
TITLE1 "One-Way Frequencies";
TITLE2 "Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL) on
%TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";
PROC FREQ DATA=WORK.SORT
     ORDER=INTERNAL
;
     TABLES AGE /  SCORES=TABLE;
     TABLES RACE /  SCORES=TABLE;
     TABLES GENDER /  SCORES=TABLE;
     TABLES MRSTAT /  SCORES=TABLE;
     TABLES EDUC /  SCORES=TABLE;
     TABLES BMICAT /  SCORES=TABLE;
     TABLES C15VRGENHTH /  SCORES=TABLE;
     TABLES C15VRMACT /  SCORES=TABLE;
     TABLES C15VRSTAIR /  SCORES=TABLE;
     TABLES C15VRPACCL /  SCORES=TABLE;
     TABLES C15VRPWORK /  SCORES=TABLE;
     TABLES C15VRMACCL /  SCORES=TABLE;
     TABLES C15VRMWORK /  SCORES=TABLE;
     TABLES C15VRPAIN /  SCORES=TABLE;
     TABLES C15VRCALM /  SCORES=TABLE;
     TABLES C15VRENERGY /  SCORES=TABLE;
     TABLES C15VRDOWN /  SCORES=TABLE;
     TABLES C15VRSACT /  SCORES=TABLE;
     TABLES C15VRPHCMP /  SCORES=TABLE;
     TABLES C15VRMHCMP /  SCORES=TABLE;
     TABLES C15ADLBTH /  SCORES=TABLE;
     TABLES C15ADLDRS /  SCORES=TABLE;
     TABLES C15ADLEAT /  SCORES=TABLE;
     TABLES C15ADLCHR /  SCORES=TABLE;
     TABLES C15ADLWLK /  SCORES=TABLE;
     TABLES C15ADLTLT /  SCORES=TABLE;
     TABLES C15HDPHY /  SCORES=TABLE;
     TABLES C15HDMEN /  SCORES=TABLE;
     TABLES C15HDACT /  SCORES=TABLE;
     TABLES C15CHSTEX /  SCORES=TABLE;
     TABLES C15CHSTRST /  SCORES=TABLE;
     TABLES C15SOBFLT /  SCORES=TABLE;
     TABLES C15SOBSIT /  SCORES=TABLE;
     TABLES C15SOBWLK /  SCORES=TABLE;
     TABLES C15SOBSTR /  SCORES=TABLE;
     TABLES C15FTNUMB /  SCORES=TABLE;
```

```
TABLES C15FTSENS /  SCORES=TABLE;
TABLES C15FTHC /  SCORES=TABLE;
TABLES C15FTSRS /  SCORES=TABLE;
TABLES C15PNART /  SCORES=TABLE;
TABLES C15READ /  SCORES=TABLE;
TABLES C15HEAR /  SCORES=TABLE;
TABLES C15CCHBP /  SCORES=TABLE;
TABLES C15CC_CAD /  SCORES=TABLE;
TABLES C15CC_CHF /  SCORES=TABLE;
TABLES C15CCMI /  SCORES=TABLE;
TABLES C15CCHRTOTH /  SCORES=TABLE;
TABLES C15CCSTROKE /  SCORES=TABLE;
TABLES C15CC_COPD /  SCORES=TABLE;
TABLES C15CCGI /  SCORES=TABLE;
TABLES C15CCARTHIP /  SCORES=TABLE;
TABLES C15CCARTHND /  SCORES=TABLE;
TABLES C15CCOSTEO /  SCORES=TABLE;
TABLES C15CCSCIATI /  SCORES=TABLE;
TABLES C15CCDIABET /  SCORES=TABLE;
TABLES C15CCANYCA /  SCORES=TABLE;
TABLES C15CACOLON /  SCORES=TABLE;
TABLES C15CALUNG /  SCORES=TABLE;
TABLES C15CABRST /  SCORES=TABLE;
TABLES C15CAPROS /  SCORES=TABLE;
TABLES C15PNBACK /  SCORES=TABLE;
TABLES C15DEP2WK /  SCORES=TABLE;
TABLES C15DEPYR /  SCORES=TABLE;
TABLES C15DEP2YR /  SCORES=TABLE;
TABLES C15DEPWEEK /  SCORES=TABLE;
TABLES C15CMPHTH /  SCORES=TABLE;
TABLES C15SMOKE /  SCORES=TABLE;
TABLES C15MUILKG /  SCORES=TABLE;
TABLES C15MUIMAG /  SCORES=TABLE;
TABLES C15MUITLK /  SCORES=TABLE;
TABLES C15MUITRT /  SCORES=TABLE;
TABLES C15PAOTLK /  SCORES=TABLE;
TABLES C15PAOADV /  SCORES=TABLE;
TABLES C15FRMTLK /  SCORES=TABLE;
TABLES C15FRMFALL /  SCORES=TABLE;
TABLES C15FRMBAL /  SCORES=TABLE;
TABLES C15FRMPREV /  SCORES=TABLE;
TABLES C15OTOTEST /  SCORES=TABLE;
TABLES C15CMPWHO /  SCORES=TABLE;
TABLES C15SRVDISP /  SCORES=TABLE;
TABLES C15SRVMODE /  SCORES=TABLE;
TABLES C15PCTCMP /  SCORES=TABLE;
TABLES C15SRVLANG /  SCORES=TABLE;
TABLES R15VRGENHTH /  SCORES=TABLE;
TABLES R15VRMACT /  SCORES=TABLE;
TABLES R15VRSTAIR /  SCORES=TABLE;
TABLES R15VRPACCL /  SCORES=TABLE;
TABLES R15VRPWORK /  SCORES=TABLE;
```

```
TABLES R15VRMACCL /  SCORES=TABLE;
TABLES R15VRMWORK /  SCORES=TABLE;
TABLES R15VRPAIN /  SCORES=TABLE;
TABLES R15VRCALM /  SCORES=TABLE;
TABLES R15VRENERGY /  SCORES=TABLE;
TABLES R15VRDOWN /  SCORES=TABLE;
TABLES R15VRSACT /  SCORES=TABLE;
TABLES R15VRPHCMP /  SCORES=TABLE;
TABLES R15VRMHCMP /  SCORES=TABLE;
TABLES R15ADLBTH /  SCORES=TABLE;
TABLES R15ADLDRS /  SCORES=TABLE;
TABLES R15ADLEAT /  SCORES=TABLE;
TABLES R15ADLCHR /  SCORES=TABLE;
TABLES R15ADLWLK /  SCORES=TABLE;
TABLES R15ADLTLT /  SCORES=TABLE;
TABLES R15DIFMEALS /  SCORES=TABLE;
TABLES R15DIFMONEY /  SCORES=TABLE;
TABLES R15DIFMEDS /  SCORES=TABLE;
TABLES R15HDPHY /  SCORES=TABLE;
TABLES R15HDMEN /  SCORES=TABLE;
TABLES R15HDACT /  SCORES=TABLE;
TABLES R15DIFSEE /  SCORES=TABLE;
TABLES R15DIFHEAR /  SCORES=TABLE;
TABLES R15DIFREMEM /  SCORES=TABLE;
TABLES R15DIFERRND /  SCORES=TABLE;
TABLES R15DIFMPROB /  SCORES=TABLE;
TABLES R15CCHBP /  SCORES=TABLE;
TABLES R15CC_CAD /  SCORES=TABLE;
TABLES R15CC_CHF /  SCORES=TABLE;
TABLES R15CCMI /  SCORES=TABLE;
TABLES R15CCHRTOTH /  SCORES=TABLE;
TABLES R15CCSTROKE /  SCORES=TABLE;
TABLES R15CC_COPD /  SCORES=TABLE;
TABLES R15CCGI /  SCORES=TABLE;
TABLES R15CCARTHIP /  SCORES=TABLE;
TABLES R15CCARTHND /  SCORES=TABLE;
TABLES R15CCOSTEO /  SCORES=TABLE;
TABLES R15CCSCIATI /  SCORES=TABLE;
TABLES R15CCDIABET /  SCORES=TABLE;
TABLES R15CCDEP /  SCORES=TABLE;
TABLES R15CCANYCA /  SCORES=TABLE;
TABLES R15CACOLON /  SCORES=TABLE;
TABLES R15CALUNG /  SCORES=TABLE;
TABLES R15CABRST /  SCORES=TABLE;
TABLES R15CAPROS /  SCORES=TABLE;
TABLES R15CAOTHER /  SCORES=TABLE;
TABLES R15PAINDACT /  SCORES=TABLE;
TABLES R15PAINSACT /  SCORES=TABLE;
TABLES R15PAINRATE /  SCORES=TABLE;
TABLES R15DEPNOPLS /  SCORES=TABLE;
TABLES R15DEPDOWN /  SCORES=TABLE;
TABLES R15CMPHTH /  SCORES=TABLE;
```

```
      TABLES R15SMOKE /  SCORES=TABLE;
      TABLES R15MUILKG /  SCORES=TABLE;
      TABLES R15MUIMAG /  SCORES=TABLE;
      TABLES R15MUITLK /  SCORES=TABLE;
      TABLES R15MUITRT /  SCORES=TABLE;
      TABLES R15PAOTLK /  SCORES=TABLE;
      TABLES R15PAOADV /  SCORES=TABLE;
      TABLES R15FRMTLK /  SCORES=TABLE;
      TABLES R15FRMFALL /  SCORES=TABLE;
      TABLES R15FRMBAL /  SCORES=TABLE;
      TABLES R15FRMPREV /  SCORES=TABLE;
      TABLES R15OTOTEST /  SCORES=TABLE;
      TABLES R15CMPWHO /  SCORES=TABLE;
      TABLES R15SRVDISP /  SCORES=TABLE;
      TABLES R15SRVMODE /  SCORES=TABLE;
      TABLES R15PCTCMP /  SCORES=TABLE;
      TABLES R15SRVLANG /  SCORES=TABLE;
      TABLES COHORT /  SCORES=TABLE;
      TABLES P15PLREGCDE /  SCORES=TABLE;
      TABLES SAMPLED /  SCORES=TABLE;
      TABLES SFLAG /  SCORES=TABLE;
RUN;
/* ---------------------------------------------------------------------
   End of task code
   ---------------------------------------------------------------------
*/
RUN; QUIT;
%_eg_conditional_dropds(WORK.SORT);
TITLE; FOOTNOTE;

/* Data Exploration, Cleaning, etc..*/;

PROC FREQ DATA=hos.C15A_PUF;

      TABLES AGE /  SCORES=TABLE;
      TABLES RACE /  SCORES=TABLE;
      TABLES GENDER /  SCORES=TABLE;
      TABLES MRSTAT /  SCORES=TABLE;
      TABLES EDUC /  SCORES=TABLE;
      TABLES BMICAT /  SCORES=TABLE;
      TABLES C15VRGENHTH /  SCORES=TABLE;
      TABLES C15VRMACT /  SCORES=TABLE;
      TABLES C15VRSTAIR /  SCORES=TABLE;
      TABLES C15VRPACCL /  SCORES=TABLE;
      TABLES C15VRPWORK /  SCORES=TABLE;
      TABLES C15VRMACCL /  SCORES=TABLE;
      TABLES C15VRMWORK /  SCORES=TABLE;
      TABLES C15VRPAIN /  SCORES=TABLE;
      TABLES C15VRCALM /  SCORES=TABLE;
      TABLES C15VRENERGY /  SCORES=TABLE;
      TABLES C15VRDOWN /  SCORES=TABLE;
      TABLES C15VRSACT /  SCORES=TABLE;
```

```
TABLES C15VRPHCMP /  SCORES=TABLE;
TABLES C15VRMHCMP /  SCORES=TABLE;
TABLES C15ADLBTH /  SCORES=TABLE;
TABLES C15ADLDRS /  SCORES=TABLE;
TABLES C15ADLEAT /  SCORES=TABLE;
TABLES C15ADLCHR /  SCORES=TABLE;
TABLES C15ADLWLK /  SCORES=TABLE;
TABLES C15ADLTLT /  SCORES=TABLE;

TABLES C15CHSTEX /  SCORES=TABLE;
TABLES C15CHSTRST /  SCORES=TABLE;
TABLES C15SOBFLT /  SCORES=TABLE;
TABLES C15SOBSIT /  SCORES=TABLE;
TABLES C15SOBWLK /  SCORES=TABLE;
TABLES C15SOBSTR /  SCORES=TABLE;
TABLES C15FTNUMB /  SCORES=TABLE;
TABLES C15FTSENS /  SCORES=TABLE;
TABLES C15FTHC /  SCORES=TABLE;
TABLES C15FTSRS /  SCORES=TABLE;
TABLES C15PNART /  SCORES=TABLE;
TABLES C15READ /  SCORES=TABLE;
TABLES C15HEAR /  SCORES=TABLE;
TABLES C15CCHBP /  SCORES=TABLE;
TABLES C15CC_CAD /  SCORES=TABLE;
TABLES C15CC_CHF /  SCORES=TABLE;
TABLES C15CCMI /  SCORES=TABLE;
TABLES C15CCHRTOTH /  SCORES=TABLE;
TABLES C15CCSTROKE /  SCORES=TABLE;
TABLES C15CC_COPD /  SCORES=TABLE;
TABLES C15CCGI /  SCORES=TABLE;
TABLES C15CCARTHIP /  SCORES=TABLE;
TABLES C15CCARTHND /  SCORES=TABLE;
TABLES C15CCOSTEO /  SCORES=TABLE;
TABLES C15CCSCIATI /  SCORES=TABLE;
TABLES C15CCDIABET /  SCORES=TABLE;
TABLES C15CCANYCA /  SCORES=TABLE;
TABLES C15CACOLON /  SCORES=TABLE;
TABLES C15CALUNG /  SCORES=TABLE;
TABLES C15CABRST /  SCORES=TABLE;
TABLES C15CAPROS /  SCORES=TABLE;
TABLES C15PNBACK /  SCORES=TABLE;
TABLES C15DEP2WK /  SCORES=TABLE;
TABLES C15DEPYR /  SCORES=TABLE;
TABLES C15DEP2YR /  SCORES=TABLE;
TABLES C15DEPWEEK /  SCORES=TABLE;
TABLES C15CMPHTH /  SCORES=TABLE;
TABLES C15SMOKE /  SCORES=TABLE;
TABLES C15MUILKG /  SCORES=TABLE;
TABLES C15MUIMAG /  SCORES=TABLE;
TABLES C15MUITLK /  SCORES=TABLE;
TABLES C15MUITRT /  SCORES=TABLE;
TABLES C15PAOTLK /  SCORES=TABLE;
```

```
      TABLES C15PAOADV /  SCORES=TABLE;
      TABLES C15FRMTLK /  SCORES=TABLE;
      TABLES C15FRMFALL /  SCORES=TABLE;
      TABLES C15FRMBAL /  SCORES=TABLE;
      TABLES C15FRMPREV /  SCORES=TABLE;
      TABLES C15OTOTEST /  SCORES=TABLE;
      TABLES C15CMPWHO /  SCORES=TABLE;
      TABLES C15SRVDISP /  SCORES=TABLE;
      TABLES C15SRVMODE /  SCORES=TABLE;
      TABLES C15PCTCMP / SCORES=TABLE;
      TABLES C15SRVLANG /  SCORES=TABLE;
      TABLES COHORT /  SCORES=TABLE;
      TABLES P15PLREGCDE /  SCORES=TABLE;
      TABLES SAMPLED /  SCORES=TABLE;
      TABLES SFLAG /  SCORES=TABLE;
RUN;


proc means data= hos.c15a_puf;
var C15HDPHY C15HDMEN C15HDACT;
run;


* Transformation of DV *;

* List of DV's
      C15VRGENHTH = General Health Status
      EXCELLENTHLTH = Respondent is in Excellent Health
      C15HDPHY = Number of Days Physical Health Not Good
      DAYSNOTGOOD = Categorization of Days Physical Health Not Good
      SFLAG = Status at Followup
      DEATHATFU= Dead at Followup;


/* 1=Healthy; 0= Unhealthy */;
data HOS.DV1;
set HOS.c15a_puf;

if  C15HDPHY <= 15 then DAYSNOTGOOD_E= 1;
if  C15HDPHY > 15 then DAYSNOTGOOD_E= 0;

if  C15HDPHY <= 5 then DAYSNOTGOOD_UE= 1;
if  C15HDPHY > 5 then DAYSNOTGOOD_UE= 0;

run;

PROC FREQ DATA=hos.DV1;
      TABLES DAYSNOTGOOD_E /  SCORES=TABLE;
      TABLES DAYSNOTGOOD_UE /  SCORES=TABLE;
run;

proc means data= hos.DV1;
```

```
var C15HDPHY;
run;

/*Scales for LCA and K-Means*/;

  proc freq DATA=hos.c15a_puf;
  tables C15VRDOWN;
  run;

  data hos.hos_scale;
  set hos.DV1;
  if C15DEP2WK= 1 or C15DEPYR= 1 or C15DEP2YR=1 or C15DEPWEEK=4 or
C15DEPWEEK=3 then  DEPRESS1= 1; else DEPRESS1=0;
  run;

 PROC FREQ DATA=HOS.hos_scale;
    TABLES DEPRESS1;
  RUN;
  /*Approx 35% Depressed*/


/*Need to run LCA in PC SAS and convert 0's to 2's- just for
DEPRESS1*/;
/*LCA v1 original scales*/;


proc freq data= hos.hos_scale;
tables C15DEP2WK C15DEPYR C15DEP2YR C15DEPWEEK
           C15VRGENHTH C15VRMACT C15VRSTAIR C15VRPACCL
                     C15VRPWORK C15VRMACCL C15VRMWORK C15VRPAIN
                     C15VRCALM C15VRENERGY C15VRDOWN C15VRSACT
DEPRESS1;
run;

%macro latent(iter);
proc lca data = HOS.hos_scale
     outest = HOS.HOSLCA2&iter
     outparam = HOS.outparam2&iter
     outpost = HOS.outpost2&iter;
     title "MHOS Segmentation: &iter Classes";
     nclass &iter;
     id CASE_ID;
     items           C15DEP2WK C15DEPYR C15DEP2YR C15DEPWEEK
                     C15VRGENHTH C15VRMACT C15VRSTAIR C15VRPACCL
                     C15VRPWORK C15VRMACCL C15VRMWORK C15VRPAIN
                     C15VRCALM C15VRENERGY C15VRDOWN C15VRSACT;


     categories  2 2 2 4
                     5 3 3 5
                     5 5 5 5
                     6 6 6 5;
```

```
      seed 6632; rho prior=1; cores 4; maxiter 20000;
run;

proc export data = hos.outparam2&iter
      outfile =
'\n04\data\entactin\EA_GINA\HOS\data\outparam2&iter..csv'
      dbms = csv replace;
run;

/* proc dataset rename, proc sort */
%mend latent;

%latent(2); %latent(3); %latent(4); %latent(5); %latent(6);


quit;
proc lca data = HOS.hos_scale
      outest = HOSLCA2
      outparam = outparam2
      outpost = outpost2;
id CASE_ID;
nclass 6;

      items            C15DEP2WK C15DEPYR C15DEP2YR C15DEPWEEK
                       /*C15VRGENHTH C15VRMACT C15VRSTAIR C15VRPACCL
                       C15VRPWORK C15VRMACCL C15VRMWORK C15VRPAIN
                       C15VRCALM C15VRENERGY C15VRDOWN C15VRSACT*/;

      categories  2 2 3 4
                       /*5 3 3 5
                       5 5 5 5
                       6 6 6 5*/;


      seed 6632; rho prior=1; cores 4; maxiter 20000;
run;

DATA test;
INPUT it1 it2 it3 it4 count;
DATALINES;
1  1  1  1  5
1  1  1  2  5
1  1  2  1  9
1  1  2  2  8
1  2  1  2  5
1  2  2  1  8
1  2  2  2  4
2  1  1  1  5
2  1  1  2  3
2  1  2  1  6
```

```
2  1  2  2  8
2  2  1  1  3
2  2  1  2  7
2  2  2  1  5
2  2  2  2  10
;
RUN;
PROC LCA DATA=test;
NCLASS 2;
ITEMS it1 it2 it3 it4;
CATEGORIES 2 2 2 2;
FREQ count;
SEED 100000;
RHO PRIOR=1;
RUN;

/*Transform data for K-Means*/;

data HOS.hos_scale2;
set HOS.hos_scale;
if C15DEPWEEK= 1 then C15DEPWEEK01=1; else C15DEPWEEK01=0;
if C15DEPWEEK= 2 then C15DEPWEEK02=1; else C15DEPWEEK02=0;
if C15DEPWEEK= 3 then C15DEPWEEK03=1; else C15DEPWEEK03=0;
if C15DEPWEEK= 4 then C15DEPWEEK04=1; else C15DEPWEEK04=0;

if C15DEP2WK= 1 then C15DEP2WK01=1; else C15DEP2WK01=0;
if C15DEP2WK= 2 then C15DEP2WK02=1; else C15DEP2WK02=0;

if C15DEPYR= 1 then C15DEPYR01=1; else C15DEPYR01=0;
if C15DEPYR= 2 then C15DEPYR02=1; else C15DEPYR02=0;

if C15DEP2YR= 1 then C15DEP2YR01=1; else C15DEP2YR01=0;
if C15DEP2YR= 2 then C15DEP2YR02=1; else C15DEP2YR02=0;

if C15VRGENHTH= 1 then C15VRGENHTH01=1; else C15VRGENHTH01=0;
if C15VRGENHTH= 2 then C15VRGENHTH02=1; else C15VRGENHTH02=0;
if C15VRGENHTH= 3 then C15VRGENHTH03=1; else C15VRGENHTH03=0;
if C15VRGENHTH= 4 then C15VRGENHTH04=1; else C15VRGENHTH04=0;
if C15VRGENHTH= 5 then C15VRGENHTH05=1; else C15VRGENHTH05=0;

if C15VRMACT= 1 then C15VRMACT01=1; else C15VRMACT01=0;
if C15VRMACT= 2 then C15VRMACT02=1; else C15VRMACT02=0;
if C15VRMACT= 3 then C15VRMACT03=1; else C15VRMACT03=0;

if C15VRSTAIR= 1 then C15VRSTAIR01=1; else C15VRSTAIR01=0;
if C15VRSTAIR= 2 then C15VRSTAIR02=1; else C15VRSTAIR02=0;
if C15VRSTAIR= 3 then C15VRSTAIR03=1; else C15VRSTAIR03=0;

if C15VRPACCL= 1 then C15VRPACCL01=1; else C15VRPACCL01=0;
if C15VRPACCL= 2 then C15VRPACCL02=1; else C15VRPACCL02=0;
if C15VRPACCL= 3 then C15VRPACCL03=1; else C15VRPACCL03=0;
if C15VRPACCL= 4 then C15VRPACCL04=1; else C15VRPACCL04=0;
```

```
if C15VRPACCL= 5 then C15VRPACCL05=1; else C15VRPACCL05=0;

if C15VRPWORK= 1 then C15VRPWORK01=1; else C15VRPWORK01=0;
if C15VRPWORK= 2 then C15VRPWORK02=1; else C15VRPWORK02=0;
if C15VRPWORK= 3 then C15VRPWORK03=1; else C15VRPWORK03=0;
if C15VRPWORK= 4 then C15VRPWORK04=1; else C15VRPWORK04=0;
if C15VRPWORK= 5 then C15VRPWORK05=1; else C15VRPWORK05=0;

if C15VRMACCL= 1 then C15VRMACCL01=1; else C15VRMACCL01=0;
if C15VRMACCL= 2 then C15VRMACCL02=1; else C15VRMACCL02=0;
if C15VRMACCL= 3 then C15VRMACCL03=1; else C15VRMACCL03=0;
if C15VRMACCL= 4 then C15VRMACCL04=1; else C15VRMACCL04=0;
if C15VRMACCL= 5 then C15VRMACCL05=1; else C15VRMACCL05=0;

if C15VRMWORK= 1 then C15VRMWORK01=1; else C15VRMWORK01=0;
if C15VRMWORK= 2 then C15VRMWORK02=1; else C15VRMWORK02=0;
if C15VRMWORK= 3 then C15VRMWORK03=1; else C15VRMWORK03=0;
if C15VRMWORK= 4 then C15VRMWORK04=1; else C15VRMWORK04=0;
if C15VRMWORK= 5 then C15VRMWORK05=1; else C15VRMWORK05=0;

if C15VRPAIN= 1 then C15VRPAIN01=1; else C15VRPAIN01=0;
if C15VRPAIN= 2 then C15VRPAIN02=1; else C15VRPAIN02=0;
if C15VRPAIN= 3 then C15VRPAIN03=1; else C15VRPAIN03=0;
if C15VRPAIN= 4 then C15VRPAIN04=1; else C15VRPAIN04=0;
if C15VRPAIN= 5 then C15VRPAIN05=1; else C15VRPAIN05=0;

if C15VRCALM= 1 then C15VRCALM01=1; else C15VRCALM01=0;
if C15VRCALM= 2 then C15VRCALM02=1; else C15VRCALM02=0;
if C15VRCALM= 3 then C15VRCALM03=1; else C15VRCALM03=0;
if C15VRCALM= 4 then C15VRCALM04=1; else C15VRCALM04=0;
if C15VRCALM= 5 then C15VRCALM05=1; else C15VRCALM05=0;
if C15VRCALM= 6 then C15VRCALM06=1; else C15VRCALM06=0;

if C15VRENERGY= 1 then C15VRENERGY01=1; else C15VRENERGY01=0;
if C15VRENERGY= 2 then C15VRENERGY02=1; else C15VRENERGY02=0;
if C15VRENERGY= 3 then C15VRENERGY03=1; else C15VRENERGY03=0;
if C15VRENERGY= 4 then C15VRENERGY04=1; else C15VRENERGY04=0;
if C15VRENERGY= 5 then C15VRENERGY05=1; else C15VRENERGY05=0;
if C15VRENERGY= 6 then C15VRENERGY06=1; else C15VRENERGY06=0;

if C15VRDOWN= 1 then C15VRDOWN01=1; else C15VRDOWN01=0;
if C15VRDOWN= 2 then C15VRDOWN02=1; else C15VRDOWN02=0;
if C15VRDOWN= 3 then C15VRDOWN03=1; else C15VRDOWN03=0;
if C15VRDOWN= 4 then C15VRDOWN04=1; else C15VRDOWN04=0;
if C15VRDOWN= 5 then C15VRDOWN05=1; else C15VRDOWN05=0;
if C15VRDOWN= 6 then C15VRDOWN06=1; else C15VRDOWN06=0;

if C15VRSACT= 1 then C15VRSACT01=1; else C15VRSACT01=0;
if C15VRSACT= 2 then C15VRSACT02=1; else C15VRSACT02=0;
if C15VRSACT= 3 then C15VRSACT03=1; else C15VRSACT03=0;
if C15VRSACT= 4 then C15VRSACT04=1; else C15VRSACT04=0;
if C15VRSACT= 5 then C15VRSACT05=1; else C15VRSACT05=0;
```

```
RUN;

proc contents data= HOS.hos_scale2;
run;

/*K-Means original questions within scale*/
proc fastclus data= HOS.hos_scale2 radius=0 replace=full maxclusters=6
maxiter=20 list distance;
id CASE_ID;
var
C15DEP2WK01
C15DEP2WK02
C15DEP2YR01
C15DEP2YR02
C15DEPWEEK01
C15DEPWEEK02
C15DEPWEEK03
C15DEPWEEK04
C15DEPYR01
C15DEPYR02
C15VRCALM01
C15VRCALM02
C15VRCALM03
C15VRCALM04
C15VRCALM05
C15VRCALM06
C15VRDOWN01
C15VRDOWN02
C15VRDOWN03
C15VRDOWN04
C15VRDOWN05
C15VRDOWN06
C15VRMACCL01
C15VRMACCL02
C15VRMACCL03
C15VRMACCL04
C15VRMACCL05
C15VRMACT01
C15VRMACT02
C15VRMACT03
C15VRMHCMP
C15VRMWORK01
C15VRMWORK02
C15VRMWORK03
C15VRMWORK04
C15VRMWORK05
C15VRPACCL01
C15VRPACCL02
C15VRPACCL03
C15VRPACCL04
C15VRPACCL05
```

```
C15VRPAIN01
C15VRPAIN02
C15VRPAIN03
C15VRPAIN04
C15VRPAIN05
C15VRPWORK01
C15VRPWORK02
C15VRPWORK03
C15VRPWORK04
C15VRPWORK05
C15VRSACT01
C15VRSACT02
C15VRSACT03
C15VRSACT04
C15VRSACT05
C15VRSTAIR01
C15VRSTAIR02
C15VRSTAIR03;
run;


/*K-Means dichotomization of depression scale*/
proc fastclus data= HOS.hos_scale2 radius=0 replace=full maxclusters=6
maxiter=20 list distance;
id CASE_ID;
var
DEPRESS1
C15VRCALM01
C15VRCALM02
C15VRCALM03
C15VRCALM04
C15VRCALM05
C15VRCALM06
C15VRDOWN01
C15VRDOWN02
C15VRDOWN03
C15VRDOWN04
C15VRDOWN05
C15VRDOWN06
C15VRMACCL01
C15VRMACCL02
C15VRMACCL03
C15VRMACCL04
C15VRMACCL05
C15VRMACT01
C15VRMACT02
C15VRMACT03
C15VRMHCMP
C15VRMWORK01
C15VRMWORK02
C15VRMWORK03
C15VRMWORK04
```

```
C15VRMWORK05
C15VRPACCL01
C15VRPACCL02
C15VRPACCL03
C15VRPACCL04
C15VRPACCL05
C15VRPAIN01
C15VRPAIN02
C15VRPAIN03
C15VRPAIN04
C15VRPAIN05
C15VRPWORK01
C15VRPWORK02
C15VRPWORK03
C15VRPWORK04
C15VRPWORK05
C15VRSACT01
C15VRSACT02
C15VRSACT03
C15VRSACT04
C15VRSACT05
C15VRSTAIR01
C15VRSTAIR02
C15VRSTAIR03;
run;

/*This works!*/;
ods output ClusterSum=Lv1ClusterSum  CCC=Lv1CC;
proc fastclus data=HOS.hos_scale2 out=clusout outseed=clusterseed
maxclusters=6 outstat=cluST1;
id CASE_ID;
var
DEPRESS1
C15VRCALM01
C15VRCALM02
C15VRCALM03
C15VRCALM04
C15VRCALM05
C15VRCALM06
C15VRDOWN01
C15VRDOWN02
C15VRDOWN03
C15VRDOWN04
C15VRDOWN05
C15VRDOWN06
C15VRMACCL01
C15VRMACCL02
C15VRMACCL03
C15VRMACCL04
C15VRMACCL05
C15VRMACT01
C15VRMACT02
```

```
C15VRMACT03
C15VRMHCMP
C15VRMWORK01
C15VRMWORK02
C15VRMWORK03
C15VRMWORK04
C15VRMWORK05
C15VRPACCL01
C15VRPACCL02
C15VRPACCL03
C15VRPACCL04
C15VRPACCL05
C15VRPAIN01
C15VRPAIN02
C15VRPAIN03
C15VRPAIN04
C15VRPAIN05
C15VRPWORK01
C15VRPWORK02
C15VRPWORK03
C15VRPWORK04
C15VRPWORK05
C15VRSACT01
C15VRSACT02
C15VRSACT03
C15VRSACT04
C15VRSACT05
C15VRSTAIR01
C15VRSTAIR02
C15VRSTAIR03;
run ;


/*try LCA with proc catmod*/

 ods output
 anova=mlr MaxLikelihood=iters estimates=mu covb=covb;
 proc catmod data=hos.hos_scale order=data;
 model C15DEP2WK*C15DEPYR*C15DEP2YR*C15DEPWEEK*
                      C15VRGENHTH*C15VRMACT*C15VRSTAIR*C15VRPACCL*
                      C15VRPWORK*C15VRMACCL*C15VRMWORK*C15VRPAIN*
                      C15VRCALM*C15VRENERGY*C15VRDOWN*C15VRSACT*x =
_response_ / wls covb addcell=.1;
 loglin C15DEP2WK C15DEPYR C15DEP2YR C15DEPWEEK
                      C15VRGENHTH C15VRMACT C15VRSTAIR C15VRPACCL
                      C15VRPWORK C15VRMACCL C15VRMWORK C15VRPAIN
                      C15VRCALM C15VRENERGY C15VRDOWN C15VRSACT x
C15DEP2WK*x C15DEPYR*x C15DEP2YR*x C15DEPWEEK*x
                      C15VRGENHTH*x C15VRMACT*x C15VRSTAIR*x
C15VRPACCL*x
                      C15VRPWORK*x C15VRMACCL*x C15VRMWORK*x
C15VRPAIN*x
```

```
                        C15VRCALM*x C15VRENERGY*x C15VRDOWN*x
C15VRSACT*x;
 run;
 quit;

 proc catmod data = hos.hos_scale ;
  direct C15DEP2WK C15DEPYR C15DEP2YR C15DEPWEEK
                        C15VRGENHTH C15VRMACT C15VRSTAIR C15VRPACCL
                        C15VRPWORK C15VRMACCL C15VRMWORK C15VRPAIN
                        C15VRCALM C15VRENERGY C15VRDOWN C15VRSACT;
  response logits;
  model C15DEP2WK C15DEPYR C15DEP2YR C15DEPWEEK
                        C15VRGENHTH C15VRMACT C15VRSTAIR C15VRPACCL
                        C15VRPWORK C15VRMACCL C15VRMWORK C15VRPAIN
                        C15VRCALM C15VRENERGY C15VRDOWN C15VRSACT;
run;

proc freq data=HOS.hos_scale2;
tables DEPRESS1;
run;

Data HOS.hos_scale2b;
set HOS.hos_scale2;

if DEPRESS1= 0 then DEPRESS1=2; else DEPRESS1=1;
run;
proc freq data=HOS.hos_scale2b;
tables DEPRESS1;
run;

**************************************************************************
**********************;
/* Profiling of segment classes */
**************************************************************************
**********************;

/* K-Means Full Scale*/
libname KM '/n04/data/entactin/EA_GINA/HOS/Results/Kmeans';
libname CH '/n04/data/entactin/EA_GINA/HOS/Results/CHAID';

/*need to merge files back to HOS.hos_scale2b for full profiling data
set
/*ID= Case_ID*/;


data KM.KMFS;
set Work.kmfullscale;
run;

/*Don't need to merge back*/

proc freq data= KM.KMFS;
```

```
tables _CLUSTER_ID_;
run;

proc sort data=KM.KMFS; by _CLUSTER_ID_; run;

proc freq data=KM.KMFS;
tables age bmicat educ gender;
*by _CLUSTER_ID_;
run;

proc freq data=KM.KMFS;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
*by _CLUSTER_ID_;
run;

proc means data= num;
var C15HDPHY1;

run;

data num;
set KM.KMFS;
C15HDPHY1= C15HDPHY*1;
run;

/* K-Means Dichotomous Scale Transformation*/

data KM.KMDS;
set work.kmdichotomous;
run;

/*Don't need to merge back*/

proc freq data= KM.KMDS;
tables _CLUSTER_ID_;
run;

proc sort data=KM.KMDS; by _CLUSTER_ID_; run;

proc freq data=KM.KMDS;
tables age bmicat educ gender;
*by _CLUSTER_ID_;
run;

proc freq data=KM.KMDS;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
*by _CLUSTER_ID_;
run;

data num2;
set KM.KMDS;
C15HDPHY1= C15HDPHY*1;
```

```
run;

proc means data= num2;
var C15HDPHY1;
*by _CLUSTER_ID_;
run;


/* CHAID Continuous*/
/*use leaf as segment class*/;

proc contents data=CH.'chaid cont'n;
run;

proc contents data= work.chaidcont;
run;

data CH.CHCont;
set work.chaidcont;
run;

proc contents data=CH.chcont;run;
proc contents data=HOS.hos_scale2b; run;
proc freq data= CH.chcont;
tables _Leaf_;
run;


data CH.CHCont2 (keep= CASE_ID Segment);
set CH.CHCont;
if _Leaf_= 20 then Segment=1;
else if _Leaf_= 9 then Segment=2;
else if _Leaf_= 19 then Segment=3;
else if _Leaf_= 18 then Segment=4;
else  Segment= 5;
run;

proc freq data= CH.CHCONT2;
tables segment;
run;

proc sort data=CH.chcont2; by CASE_ID; run;
proc sort data=HOS.hos_scale2b; by CASE_ID; run;

data CH.CHCONTMerge;
merge HOS.hos_scale2b (in=a) CH.chcont2;
by CASE_ID;
if a=1;
run;

proc sort data= ch.chcontmerge; by segment;
run;
```

```
proc freq data=CH.chcontmerge;
tables age bmicat educ gender;
by segment;
run;

proc freq data=CH.chcontmerge;
tables age bmicat educ gender;
run;

proc freq data=CH.chcontmerge;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
by segment;
run;

proc means data= CH.chcontmerge;
var C15HDPHY;
by segment;
run;

proc freq data=CH.chcontmerge;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
run;

proc means data= CH.chcontmerge;
var C15HDPHY;
run;

/*CHAID EVEN*/


data CH.CHeven (keep= CASE_ID Segment);
set Work.CHAIDEVEN;
if _Leaf_= 1 then Segment=1;
else if _Leaf_= 0 then Segment=2;
else if _Leaf_= 6 then Segment=3;
else  Segment= 4;
run;

proc freq data= CH.CHeven;
tables segment;
run;

proc sort data=CH.CHeven; by CASE_ID; run;
proc sort data=HOS.hos_scale2b; by CASE_ID; run;

data CH.CHevenMerge;
merge HOS.hos_scale2b (in=a) CH.CHeven;
by CASE_ID;
if a=1;
run;
```

```
proc sort data= CH.CHevenMerge; by segment;
run;

proc freq data=CH.CHevenMerge;
tables age bmicat educ gender;
by segment;
run;

proc freq data=CH.CHevenMerge;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
by segment;
run;

proc means data= CH.CHevenMerge;
var C15HDPHY;
by segment;
run;

proc freq data=CH.CHevenMerge;
tables age bmicat educ gender;
run;

proc freq data=CH.CHevenMerge;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
run;

proc means data= CH.CHevenMerge;
var C15HDPHY;
run;



/*CHAID UNEVEN*/


data CH.CHUneven (keep= CASE_ID Segment);
set Work.CHAIDUneven;
if _Leaf_= 4 then Segment=1;
else if _Leaf_= 0 then Segment=2;
else if _Leaf_= 7 then Segment=3;
else  Segment= 4;
run;

proc freq data= CH.CHUneven;
tables segment;
run;

proc sort data=CH.CHUneven; by CASE_ID; run;
proc sort data=HOS.hos_scale2b; by CASE_ID; run;

data CH.CHUnevenMerge;
merge HOS.hos_scale2b (in=a) CH.CHUneven;
```

```
by CASE_ID;
if a=1;
run;

proc sort data= CH.CHUnevenMerge; by segment;
run;

proc freq data=CH.CHUnevenMerge;
tables age bmicat educ gender;
by segment;
run;

proc freq data=CH.CHUnevenMerge;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
by segment;
run;

proc means data= CH.CHUnevenMerge;
var C15HDPHY;
by segment;
run;

proc freq data=CH.CHUnevenMerge;
tables age bmicat educ gender;
run;

proc freq data=CH.CHUnevenMerge;
tables DEPRESS1 C15VRGENHTH C15VRPAIN C15VRENERGY C15SMOKE;
run;

proc means data= CH.CHUnevenMerge;
var C15HDPHY;
```

APPENDIX E

R CODE

```
install.packages("poLCA", dependencies = TRUE);
install.packages("car");
install.packages("scatterplot3d");
install.packages("MASS");
install.packages("gbm");
install.packages("caret");
library(caret)
library(gbm)
library(car)
library(poLCA)
library("reshape2")
library("plyr")
library("dplyr")
library("poLCA")
library("ggplot2")
library("ggparallel")
library("igraph")
library("tidyr")
library("knitr")

ds <- read.csv("Y:/EA_GINA/HOS/Data/HOS_SCALE2.csv", header=TRUE, sep=",")

#need to specify a covariate to get parameter estimates
####Dichotomized Scale

library(poLCA)
m2 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
                  C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
                  C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
                  C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
           maxiter=50000, nclass=2, na.rm=FALSE, graphs=TRUE,
           nrep=10, data=ds)
m3 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
                  C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
                  C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
                  C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
```

```
        maxiter=50000, nclass=3, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)
m4 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=4, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)
m5 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=5, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)

m6 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=6, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)



m7 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=7, na.rm=FALSE,
        nrep=10, data=ds)

m8 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=8, na.rm=FALSE,
        nrep=10, data=ds)

m9 = poLCA(cbind( C15DEP2WK, C15DEPYR, C15DEP2YR, C15DEPWEEK,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=9, na.rm=FALSE,
        nrep=10, data=ds)
```

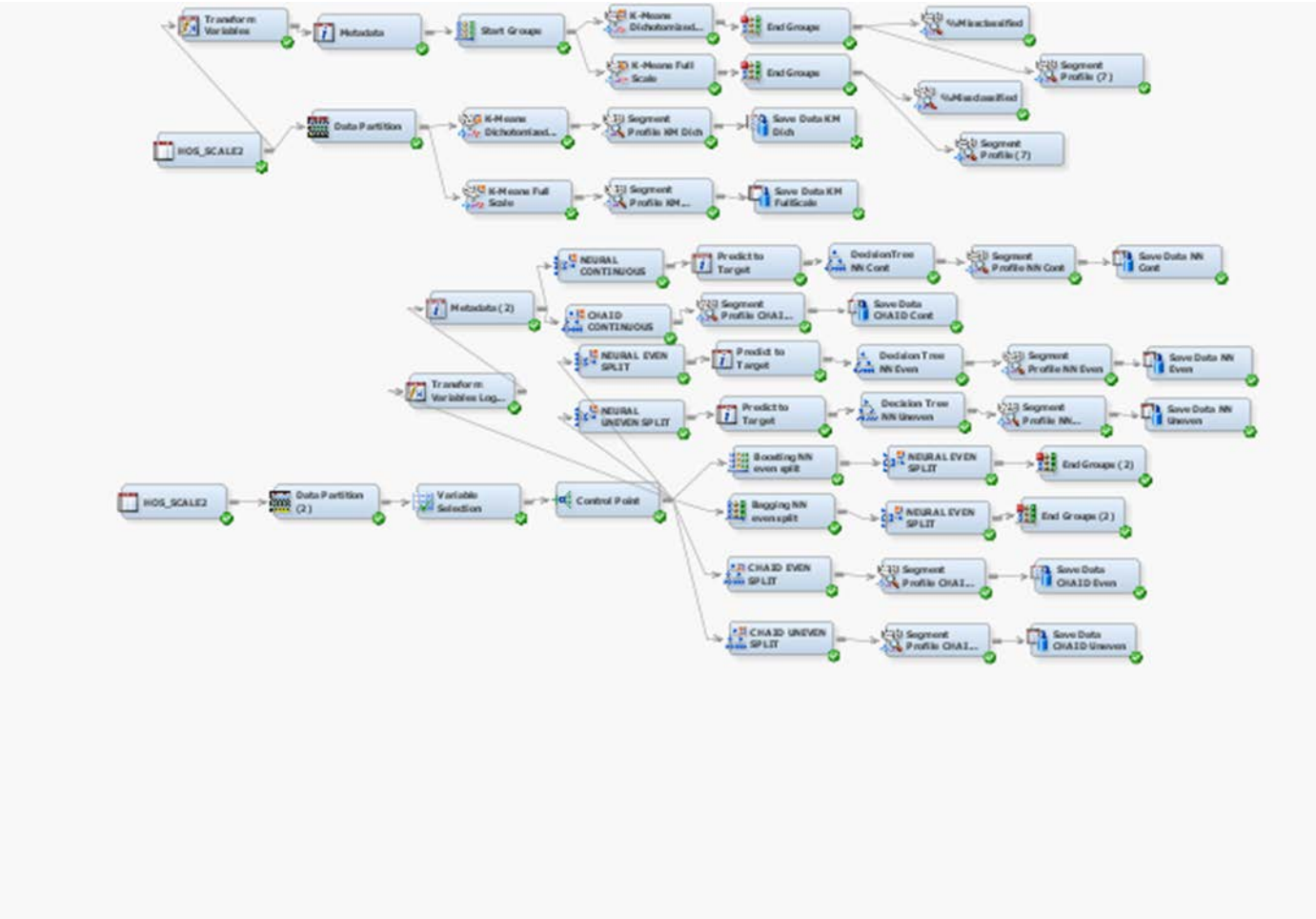
```
#merge segment membership back to data

ds$m2 <- m2$predclass
ds$m3 <- m3$predclass
ds$m4 <- m4$predclass
ds$m5 <- m5$predclass
ds$m6 <- m6$predclass
ds$m7 <- m7$predclass

ds$m8 <- m8$predclass
ds$m9 <- m9$predclass

#export as csv
write.table(ds, file= "RLCA2.csv", sep=",", col.names=TRUE, qmethod="double", na="", row.names=FALSE)

####Full Scale
ds <- read.csv("Y:/EA_GINA/HOS/Data/HOS_SCALE2b.csv", header=TRUE, sep=",")

#need to specify a covariate to get parameter estimates

library(poLCA)
c2 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=2, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)
c3 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=3, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)
c4 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=4, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)
c5 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=5, na.rm=FALSE, graphs=TRUE,
```

```
        nrep=10, data=ds)

c6 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=6, na.rm=FALSE, graphs=TRUE,
        nrep=10, data=ds)



c7 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=7, na.rm=FALSE,
        nrep=10, data=ds)

c8 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=8, na.rm=FALSE,
        nrep=10, data=ds)

c9 = poLCA(cbind( DEPRESS1,
            C15VRGENHTH, C15VRMACT, C15VRSTAIR, C15VRPACCL,
            C15VRPWORK, C15VRMACCL, C15VRMWORK, C15VRPAIN,
            C15VRCALM, C15VRENERGY, C15VRDOWN, C15VRSACT) ~ 1,
        maxiter=50000, nclass=9, na.rm=FALSE,
        nrep=10, data=ds)



#merge segment membership back to data

ds$c2 <- c2$predclass
ds$c3 <- c3$predclass
ds$c4 <- c4$predclass
ds$c5 <- c5$predclass
ds$c6 <- c6$predclass
ds$c7 <- c7$predclass

ds$c8 <- c8$predclass
ds$c9 <- c9$predclass
```

```
#export as csv
write.table(ds, file= "RLCAFS.csv", sep=",", col.names=TRUE, qmethod="double", na="", row.names=FALSE)
```

APPENDIX F

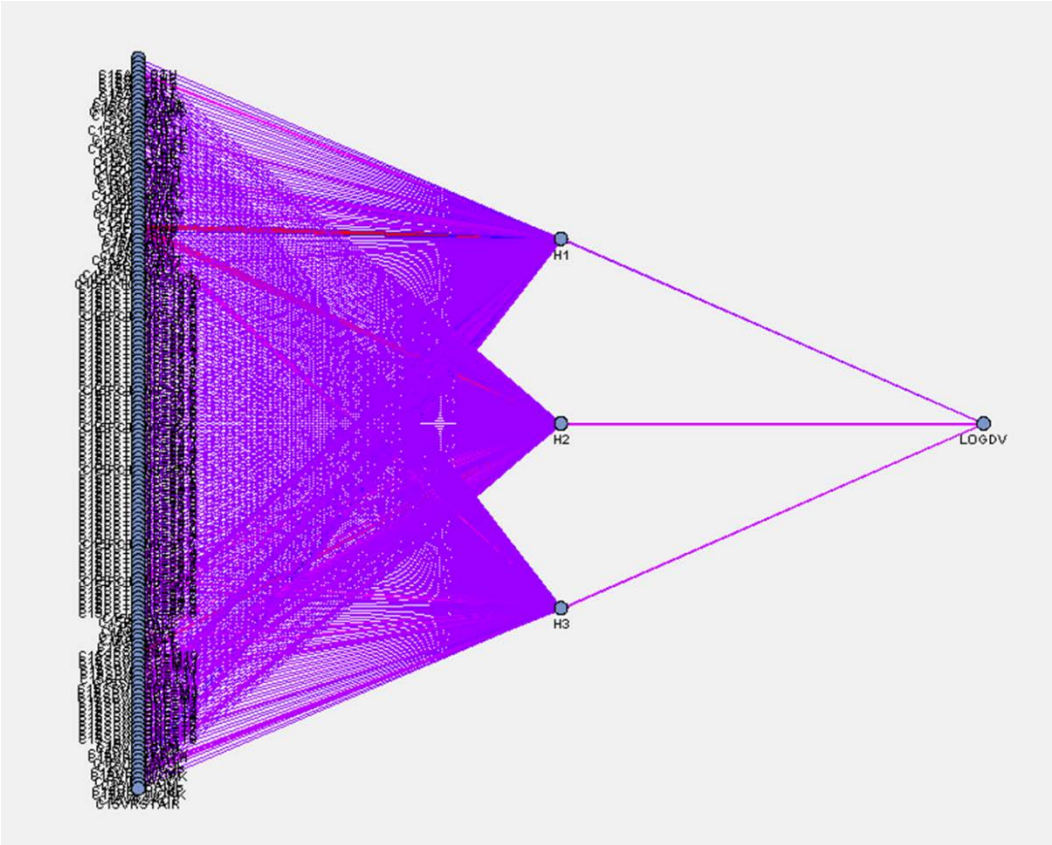
MINER DIAGRAM



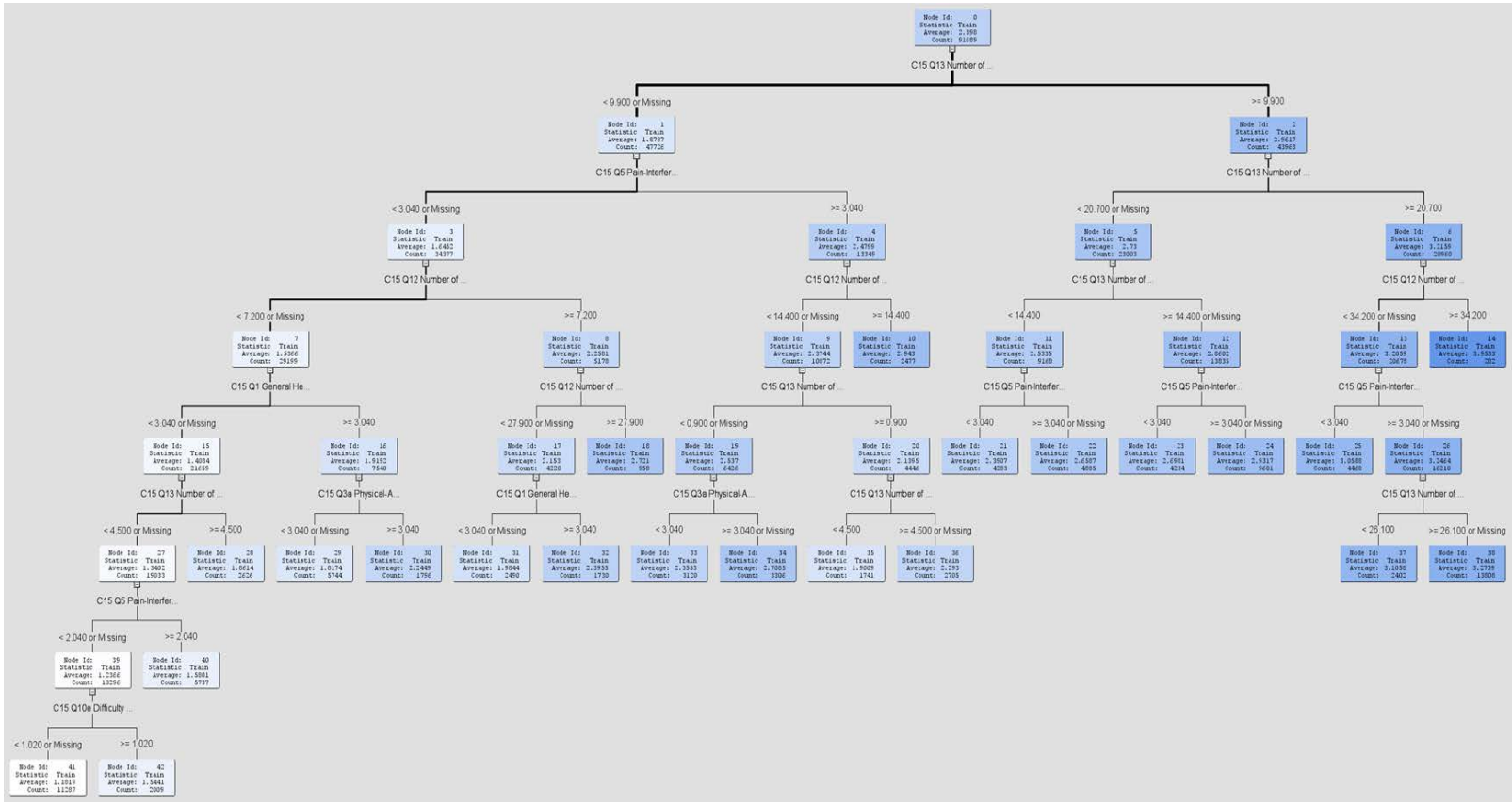
APPENDIX G

CHAID AND NEURAL NETWORK DIAGRAMS

G. 1 EURAL NETWORK DIAGRAM: CONTINUOUS OUTCOME



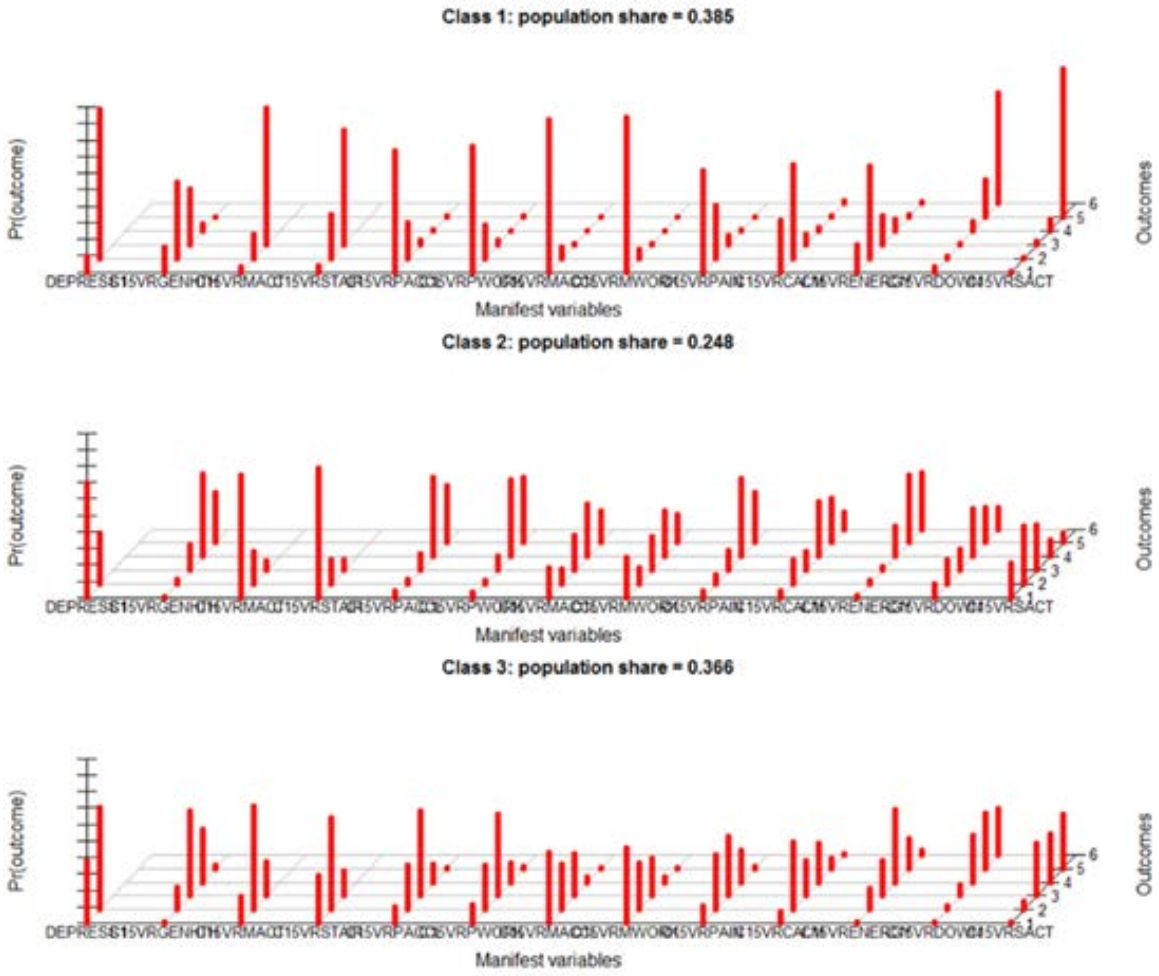
G. 2 DIAGRAM: CONTINUOUS OUTCOME



APPENDIX H

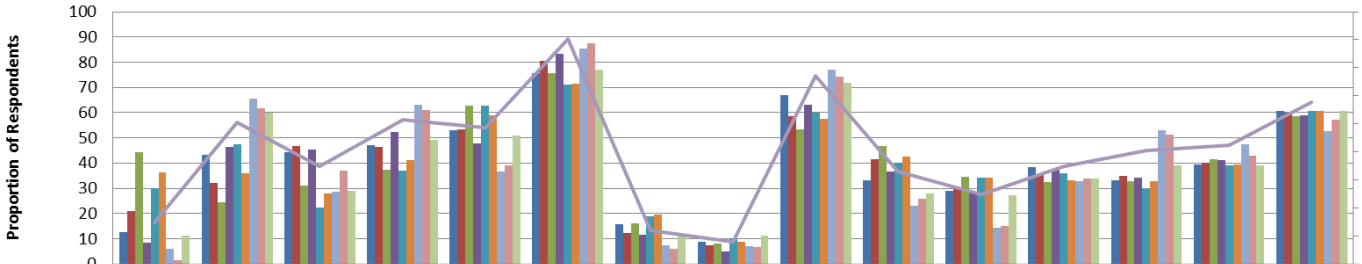
ADDITIONAL VISUALIZATIONS

H. 1 LCA CONDITIONAL PROBABILITIES



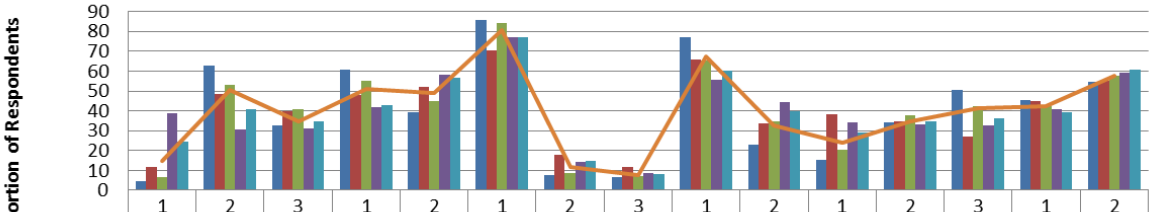
H. 2 DEMOGRAPHIC PROFILES

Demographic Distribution of Segment Classes
LCA Full Scale



	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
1	12.66	43.11	44.23	47.08	52.92	75.81	15.61	8.58	67.01	32.99	28.79	38.22	33	39.24	60.76
2	21.03	32.12	46.85	46.5	53.5	80.51	12.08	7.42	58.66	41.34	29.65	35.49	34.86	40.12	59.88
3	44.41	24.58	31.01	37.42	62.58	75.8	16.24	7.96	53.44	46.56	34.67	32.51	32.82	41.4	58.6
4	8.23	46.29	45.48	52.17	47.83	83.27	11.69	5.04	63.21	36.79	28.47	37.43	34.09	41.17	58.83
5	30.1	47.38	22.51	37.09	62.91	71.09	18.88	10.03	60	40	34.25	35.91	29.83	39.2	60.8
6	36.12	35.9	27.97	41.13	58.87	71.6	19.51	8.89	57.49	42.51	34.04	33.1	32.86	39.37	60.63
7	5.86	65.42	28.72	63.23	36.77	85.51	7.46	7.02	77.08	22.92	14.2	32.93	52.87	47.51	52.49
8	1.39	61.67	36.94	60.86	39.14	87.5	5.93	6.57	74.15	25.85	14.92	33.97	51.11	42.99	57.01
9	11.14	59.95	28.91	49.11	50.89	76.92	11.81	11.26	71.92	28.08	27.11	33.76	39.13	39.22	60.78
Sample	14.83	50.48	34.69	51.45	48.55	80.34	11.92	7.73	67.17	32.83	24.78	34.74	40.49	42.34	57.66

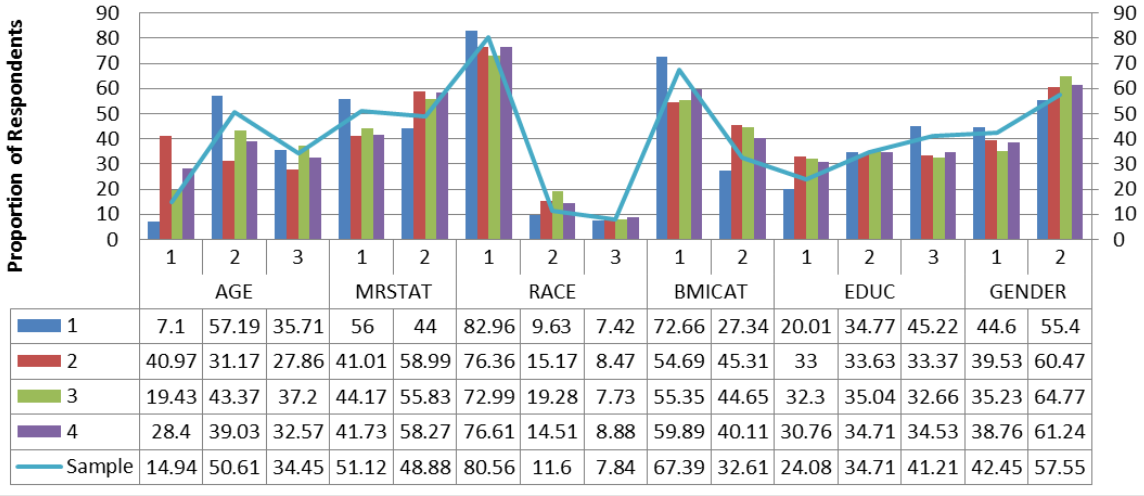
Demographic Distribution of Segment Classes
CHAID Continuous Outcome



	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
1	4.51	62.6	32.9	60.59	39.41	85.61	7.77	6.62	77.23	22.77	15.41	34.17	50.42	45.38	54.62
2	11.94	48.49	39.57	47.9	52.1	70.25	18.04	11.71	66.11	33.89	38.44	34.5	27.05	44.86	55.14
3	6.38	52.93	40.69	55.05	44.95	84.23	8.9	6.88	65.4	34.6	20.19	37.59	42.22	42.23	57.77
4	38.69	30.37	30.94	41.68	58.32	76.95	14.43	8.62	55.85	44.15	33.99	33.4	32.61	40.66	59.34
5	24.69	40.72	34.59	43.12	56.88	77.08	14.53	8.39	60.22	39.78	29.19	34.8	36.01	39.39	60.61
Sample	14.94	50.61	34.45	51.12	48.88	80.56	11.6	7.84	67.39	32.61	24.08	34.71	41.21	42.45	57.55

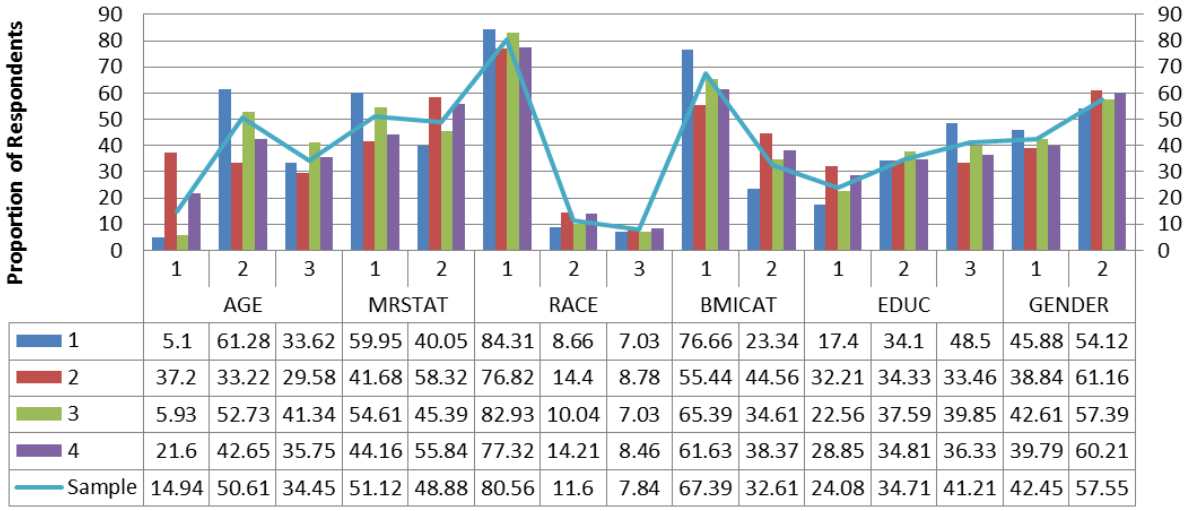
															59.39
															60.4
3	39.4	26.66	33.94	38.35	61.65	76.73	14.25	9.02	56.97	43.03	37.14	32.69	30.17	40.23	59.77
4	20.61	35.79	43.59	44.26	55.74	80.31	12.44	7.25	57.31	42.69	26.86	35.48	37.66	40.04	59.96
5	2.96	60.4	36.64	61.98	38.02	88.21	6.45	5.35	71.25	28.75	13.85	34.89	51.26	45.06	54.94
6	38.93	32.1	28.97	40.23	59.77	74.8	16.46	8.74	55.7	44.3	33.1	33.55	33.35	38.05	61.95
7	4.72	66.31	28.97	62.5	37.5	85.51	7.91	6.58	79.91	20.09	14.45	33.4	52.14	46.76	53.24
8	7.71	53.53	38.76	51.28	48.72	80.03	10.18	9.79	72.25	27.75	22.83	35.28	41.89	42.55	57.45
9	8.18	48.77	43.05	54.01	45.99	83.01	11.18	5.81	62.45	37.55	22.26	37.24	40.5	42.25	57.75
Sample	15	50.61	34.39	51.1	48.9	80.5	11.66	7.84	67.23	32.77	24.19	34.57	41.24	42.51	57.49

Demographic Distribution of Segment Classes
CHAID Dichotomous Outcome: Even Split



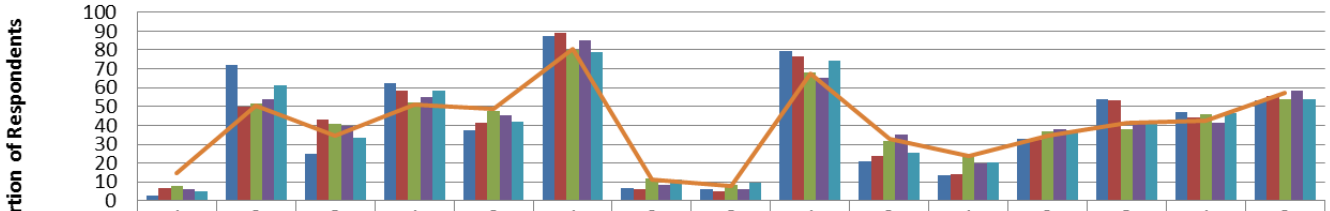
	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
1	7.1	57.19	35.71	56	44	82.96	9.63	7.42	72.66	27.34	20.01	34.77	45.22	44.6	55.4
2	40.97	31.17	27.86	41.01	58.99	76.36	15.17	8.47	54.69	45.31	33	33.63	33.37	39.53	60.47
3	19.43	43.37	37.2	44.17	55.83	72.99	19.28	7.73	55.35	44.65	32.3	35.04	32.66	35.23	64.77
4	28.4	39.03	32.57	41.73	58.27	76.61	14.51	8.88	59.89	40.11	30.76	34.71	34.53	38.76	61.24
Sample	14.94	50.61	34.45	51.12	48.88	80.56	11.6	7.84	67.39	32.61	24.08	34.71	41.21	42.45	57.55

Demographic Distribution of Segment Classes
CHAID Dichotomous Outcome: Uneven Split



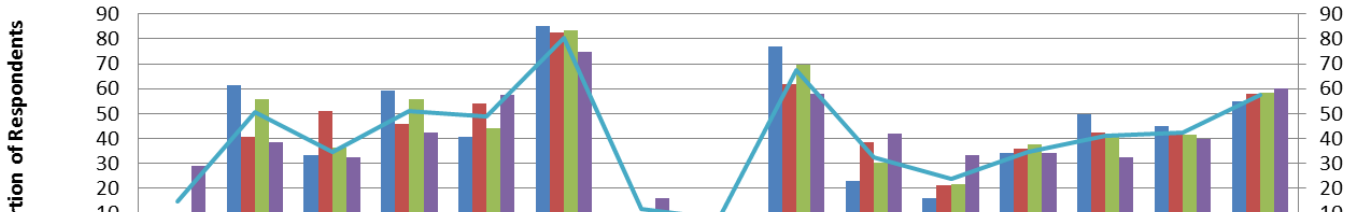
	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
1	5.1	61.28	33.62	59.95	40.05	84.31	8.66	7.03	76.66	23.34	17.4	34.1	48.5	45.88	54.12
2	37.2	33.22	29.58	41.68	58.32	76.82	14.4	8.78	55.44	44.56	32.21	34.33	33.46	38.84	61.16
3	5.93	52.73	41.34	54.61	45.39	82.93	10.04	7.03	65.39	34.61	22.56	37.59	39.85	42.61	57.39
4	21.6	42.65	35.75	44.16	55.84	77.32	14.21	8.46	61.63	38.37	28.85	34.81	36.33	39.79	60.21
Sample	14.94	50.61	34.45	51.12	48.88	80.56	11.6	7.84	67.39	32.61	24.08	34.71	41.21	42.45	57.55

Demographic Distribution of Segment Classes
Neural Network Continuous Outcome



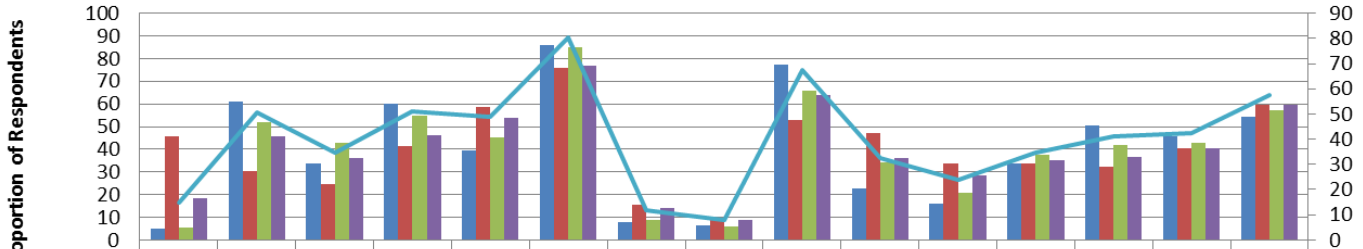
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
1	3.01	72.32	24.67	62.33	37.67	87.12	6.86	6.02	79.15	20.85	13.49	32.73	53.78	46.83	53.17
2	6.97	50.03	43	58.47	41.53	88.98	6.13	4.89	76.44	23.56	14.33	32.45	53.22	44.4	55.6
3	7.9	51.47	40.63	52.41	47.59	79.85	11.67	8.47	68.09	31.91	25.06	37.13	37.81	45.94	54.06
4	6.38	53.89	39.73	54.75	45.25	85.08	8.72	6.21	65.06	34.94	19.77	38.08	42.15	41.49	58.51
5	5.32	61.24	33.44	58.16	41.84	79.04	11.38	9.58	74.55	25.45	20.15	37.18	42.67	46.33	53.67
Sample	14.94	50.61	34.45	51.12	48.88	80.56	11.6	7.84	67.39	32.61	24.08	34.71	41.21	42.45	57.55

Demographic Distribution of Segment Classes
Neural Network Dichotomous Outcome: Even Split



	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
1	5.27	61.5	33.23	59.39	40.61	85.09	8.13	6.78	77.01	22.99	16.22	34.18	49.6	45.06	54.94
2	7.96	40.81	51.22	45.92	54.08	82.73	10.36	6.91	61.65	38.35	21.49	36.12	42.39	42.13	57.87
3	7.71	55.82	36.47	55.7	44.3	83.28	9.42	7.31	69.77	30.23	21.62	37.85	40.52	41.67	58.33
4	29.22	38.45	32.32	42.62	57.38	74.8	16.02	9.17	57.93	42.07	33.21	34.27	32.53	39.89	60.11
Sample	14.94	50.61	34.45	51.12	48.88	80.56	11.6	7.84	67.39	32.61	24.08	34.71	41.21	42.45	57.55

Demographic Distribution of Segment Classes
Neural Network Dichotomous Outcome: Uneven Split



	1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
	AGE			MRSTAT		RACE			BMICAT		EDUC			GENDER	
1	5.14	61.24	33.63	60.3	39.7	86.02	7.65	6.33	77.14	22.86	15.96	33.63	50.41	45.66	54.34
2	45.49	30.14	24.38	41.57	58.43	75.84	15.5	8.66	52.73	47.27	33.63	33.94	32.43	40.37	59.63
3	5.28	51.75	42.98	54.84	45.16	85.05	8.83	6.12	65.94	34.06	20.69	37.48	41.83	42.6	57.4
4	18.22	45.77	36.01	46.09	53.91	76.94	14.12	8.94	63.81	36.19	28.3	35.1	36.6	40.47	59.53
Sample	14.94	50.61	34.45	51.12	48.88	80.56	11.6	7.84	67.39	32.61	24.08	34.71	41.21	42.45	57.55

BIBLIOGRAPHY

Agrawal, D., & Schorling, C. (1997). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing*, *72*(4), 383-407.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, *13*(3), 469-475.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, *16*(1), 125-127.

Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2007). A comparison of decision tree ensemble creation techniques. *IEEE transactions on pattern analysis and machine intelligence*, *29*(1).

Barrett, P. (2007). Structural equation modelling: Adjudging model fit.*Personality and Individual differences*, *42*(5), 815-824.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, *36*(1-2), 105-139.

Bennette, C., & Vickers, A. (2012). Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC medical research methodology*, *12*(1), 21.

Becher, H., Grau, A., Steindorf, K., Buggle, F., & Hacke, W. (1999). Previous infection and other risk factors for acute cerebrovascular ischaemia: attributable risks and the characterisation of high risk groups. *Journal of epidemiology and biostatistics*, *5*(5), 277-283.

Bhatnagar, A., & Ghose, S. (2004). A latent class segmentation analysis of e-shoppers. *Journal of Business Research*, *57*(7), 758-767.

Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, *64*(1), 115-123.

Bonoma, T. V., & Shapiro, B. P. (1984). Evaluating market segmentation approaches. *Industrial Marketing Management*, *13*(4), 257-268.

Centers for Disease Control and Prevention. Measuring Healthy Days. Atlanta, Georgia: CDC, November 2000.

Chang, P. C., & Afifi, A. A. (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, *69*(346), 336-339.

Chaturvedi, A., Carroll, J. D., Green, P. E., & Rotondo, J. A. (1997). A feature-based approach to market segmentation via overlapping K-centroids clustering. *Journal of Marketing Research*, 370-377.

Chaturvedi, A., Green, P. E., & Caroll, J. D. (2001). K-modes clustering. *Journal of Classification*, *18*(1), 35-55.

Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus*, *5*(1), 89.

Chen, C. H., Khoo, L. P., & Yan, W. (2002). A strategy for acquiring customer requirement patterns using laddering technique and ART2 neural network. *Advanced Engineering Informatics*, *16*(3), 229-240.

Cohen, J. (1983). The cost of dichotomization,. *Applied Psychological Measurement, 7(3),* 249-253.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(3), 131-156.

De Keyser, A., Schepers, J., & Konuş, U. (2015). Multichannel customer segmentation: Does the after-sales channel matter? A replication and extension. *International Journal of Research in Marketing*, *32*(4), 453-456.

Dennis, C., Marsland, D., & Cockett, T. (2001). Data mining for shopping centres-customer knowledge-management framework. Journal of Knowledge Management, 5(4), 368-374.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*,*45*(2), 265-282.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, *40*(2), 139-157.

Du Jardin, P. (2010). Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. Neurocomputing, 73(10), 2047-2060.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research.Psychological methods, 4(3), 272.

Fedorov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics*, *8*(1), 50-61.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. Journal of classification, 5(2), 205-228.

Gandek, B., Sinclair, S. J., Kosinski, M., & Ware Jr, J. E. (2004). Psychometric evaluation of the SF-36 health survey in Medicare managed care.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881-889.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 337-348.

Grace, S. C., Elliott, M. N., Giordano, L. A., Burroughs, J. N., & Malinoff, R. L. (2013). Health-related quality of life and quality of care in specialized medicare-managed care plans. The Journal of ambulatory care management, 36(1), 72-84.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.

Haffer, S. C. C., & Bowen, S. E. (2004). Measuring and improving health outcomes in Medicare: the Medicare HOS program. *Health care financing review*, *25*(4), 1.

Hocking, R. R. (1972). Criteria for selection of a subset regression: which one should be used?. Technometrics, 14(4), 967-976.

Hruschka, H., & Natter, M. (1999). Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *European Journal of Operational Research*, *114*(2), 346-353.

Hwang, H., Jung, T., & Suh, E. (2004). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert systems with applications*, *26*(2), 181.

Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, *40*(3), 366-371.

Jonker, J. J., Piersma, N., & Van den Poel, D. (2004). Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications*, *27*(2), 159-168.-188.

Kano, Y., & Harada, A. (2000). Stepwise variable selection in factor analysis. *Psychometrika*, *65*(1), 7-22.

Kim, K. J., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. *Expert systems with applications*,*34*(2), 1200-1209.

Kim, S., Fong, D. K., & Desarbo, W. S. (2012). Model-based segmentation featuring simultaneous segment-level variable selection. *Journal of Marketing Research*, *49*(5), 725-736.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Kumar, A., Rao, V. R., & Soni, H. (1995). An empirical comparison of neural network and logistic regression models. *Marketing Letters*, *6*(4), 251-263.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65-81.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, *29*(11), 1475-1493.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. Structural Equation Modeling, 14(4), 671-694.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, *6*(6), 861-867.

Levin, N., & Zahavi, J. (2001). Predictive modeling using segmentation. *Journal of Interactive Marketing*, *15*(2), 2-22.

MacCallum, R.C., Zhang, S. Preacher, K.J. & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7(1),* 19-40.

Maclin, R., & Opitz, D. (1997). An empirical evaluation of bagging and boosting. *AAAI/IAAI*, *1997*, 546-551.

Manchanda, P., Rossi, P. E., & Chintagunta, P. K. (2004). Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, *41*(4), 467-478.

McCall, N., Khatutsky, G., Smith, K., & Pope, G. C. (2004). Estimation of non-response bias in the Medicare FFS HOS. *Health care financing review*, *25*(4), 27.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of business research*, *60*(6), 656-662.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, *4*(1), 103-120

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023-1032.

Mittal, V., & Kamakura, W. A. (2001). Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of marketing research*, *38*(1), 131-142.

Muthén, L. K., & Muthén, B. O. (2007). Mplus. *Statistical analysis with latent variables. Version, 3*.

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., & Altman, D. G. (2011). Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, *32*(3), 437-440.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, *43*(2), 204-211.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling*, *14*(4), 535-569.

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, *4*(1), 85-117.

Olson, D. L., Cao, Q., Gu, C., & Lee, D. (2009). Comparison of customer response models. *Service Business*, *3*(2), 117-130.

Patterson, B. H., Dayton, C. M., & Graubard, B. I. (2002). Latent class analysis of complex sample survey data: application to dietary data. *Journal of the American Statistical Association*, *97*(459), 721-741.

PROC LCA & PROC LTA (Version 1.3.2) [Software]. (2015). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301-323.

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, *101*(473), 168-178.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *The Journal of Machine Learning Research*, *3*, 1371-1382.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, *17*(3), 354.

Ringle, C. M., Sarstedt, M., Schlittgen, R., & Taylor, C. R. (2013). PLS path modeling and evolutionary segmentation. *Journal of Business Research*, *66*(9), 1318-1324.

Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, *22*(3), 304-328.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in society*, *24*(4), 483-502.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, *23*(19), 2507-2517.

Schreiber, J. B., & Pekarik, A. J. (2014). Technical Note: Using Latent Class Analysis versus K-means or Hierarchical Clustering to Understand Museum Visitors. *Curator: The Museum Journal*, *57*(1), 45-59.

Steenkamp, J. B. E., & Baumgartner, H. (2000). On the use of structural equation models for marketing modeling. *International Journal of Research in Marketing*, *17*(2), 195-202.

Sun, B., & Morwitz, V. G. (2010). Stated intentions and purchase behavior: A unified model. *International Journal of Research in Marketing*, *27*(4), 356-366.

Teichert, T., Shehu, E., & von Wartburg, I. (2008). Customer segmentation revisited: The case of the airline industry. *Transportation Research Part A: Policy and Practice*, *42*(1), 227-242.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *.Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Thompson, D. (2007). Latent Class Analysis in SAS: Promise, Problems, and Programming. In *SAS Global Forum 2007. Orlando, FL*.

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761-1768.

Tuma, M., & Decker, R. (2013). Finite mixture models in market segmentation: a review and suggestions for best practices. *Electronic Journal of Business Research Methods*, *11*(1), 2-15.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*,*14*(10), 2439-2468.

Vellido, A., Lisboa, P. J. G., & Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert systems with applications*, *17*(4), 303-314.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*, *32*(10), 2543-2559.

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal*, *6*(1), 58.

Xu, P., Burgess, J. F., Cabral, H., Soria-Saucedo, R., & Kazis, L. E. (2015). Relationships between Medicare Advantage Contract Characteristics and Quality-Of-Care Ratings: An Observational Analysis of Medicare Advantage Star Ratings. *Annals of internal medicine*, *162*(5), 353-358.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49-67

Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (2001). A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers & Operations Research*, *28*(4), 381-396.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*, *137*(1-2), 239-263.