Characteristics associated with self-rated health:
The CARDIA Study

by

Shilpa Nayak

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor S Leonard Syme, Chair
Professor Alan Hubbard
Professor Meredith Minkler
Professor William Satariano
Dr Steve Sidney

Fall 2012

Abstract

Characteristics associated with self-rated health: The CARDIA Study
by
Shilpa Nayak

Doctor of Philosophy in Epidemiology

University of California, Berkeley
Professor S Leonard Syme, Chair

Self-rated health is an independent predictor of future health outcomes, including morbidity and mortality. Therefore, in a public health context, for high proportions of populations to report very good or excellent subjective health is in itself an important end point. To achieve this, public health interventions need to be informed by knowledge of the determinants of both illness and wellbeing in different groups. In this study, I used self-rated health as the outcome measure, and studied the characteristics of individuals in a population who rated their health as excellent or very good (classed as 'good' self rated health), versus those who rated it as only good, fair, or poor (classed as 'poor' self-rated health). A broad range of risk and protective multi-domain determinants of health were included in the analysis as predictor variables. The study data were drawn from the CARDIA study, a United States cohort started in 1985 to investigate the development of coronary artery disease risk factors in a young adult population.

In the first analysis, I utilized classification tree methods to segment the study sample of 3649 individuals, to identify subgroups with some shared characteristics and relatively homogenous self-rated health status. Lifestyle, social and community influences, and living and working conditions were all associated with self-rated health. Combinations of these factors differed by population subgroup. Physical activity rating emerged as the most important variable in the single tree classification, and the model suggested interaction of lifestyle and medical factors with socioeconomic factors, income and education.

In the second analysis, the study sample was first divided into subsets based on total family income. I investigated the characteristics associated with self-rated health within each income subset using classification trees. The findings suggested a social gradient for several health determinants. The proportion of good self-rated health increased with higher income category, and the proportion of poor self-rated health decreased. Within population subgroups stratified by income, the combinations of factors that were associated with self-rated health, and the predictor variable that ranked as most important relative to self-rated health differed. This is suggestive of potentially important differences in the factors that are responsible for self-rated

health and health inequalities among different income groups that have dissimilar social, cultural and economic contexts.

The third analysis extended the single classification tree analysis with the application of random forests. This method produced an ensemble of classification trees, which improved accuracy and produced more robust variable importance measures. Despite the inclusion of a wide range of predictor variables representing fixed factors, lifestyle and medical conditions, social and community influences, and living and working conditions, the model selected education and income as the highest-ranking variables associated with self-rated health in the study sample. This highlights the importance of addressing social determinants of health and inequities.

This dissertation contributes to the literature on the determinants of self-rated health, and adds a novel application of classification tree analysis and random forests methods to the study of self-rated health. Capturing the complex interplay of factors affecting health in populations can be difficult with parametric multivariate regression. These models may not capture the full array of variables influencing health. Recursive partitioning methods can serve as an initial tool to suggest population subgroups that might have homogenous risks of an outcome, and identify the relative importance of risk and protective factors in population subgroups for further inquiry. This knowledge is valuable in developing appropriate and targeted public health interventions that focus on specific needs.

This dissertation is dedicated to
my parents, for believing it a long time ago,
and to Sridhar, for making it real.

Table of Contents

Acknowledgements

I have been extremely fortunate to have Len Syme as my advisor. At the start, he gave me the freedom to explore and develop my ideas. He has inspired and encouraged me throughout my doctoral studies. I am most grateful for his guidance as a mentor, the impact of which will remain with me far beyond the end of this dissertation.

Meredith Minkler gave considered and insightful comments, which helped me to improve the quality of my work. Her interest and faith in my research, along with consistently enthusiastic support during the months of dissertation writing were greatly encouraging and much appreciated.

Alan Hubbard has provided advice and clarity on statistical techniques that have informed a core component of this work. I have benefited greatly from his expertise, which has helped both in evolving my analysis and further stimulating my interest in different statistical methods.

Steve Sidney shared his experience and knowledge as principal investigator of the CARDIA study at Oakland. This brought a much valued and important perspective to this work.

I am thankful to Bill Satariano for his advice and encouragement both as Chair of my oral qualifying examination, and as a member of my dissertation committee.

Nancy Adler and Catarina Kiefe assisted in the early development of my proposal and in forming a link with the CARDIA study. Linda Neuhauser's constructive feedback and thought-provoking questions on my research proposal, helped mature this dissertation.

I am also grateful to staff at the CARDIA Coordinating Center for their assistance, and to participants in the CARDIA study whose involvement has made this research possible.

UC Berkeley is a great place to be a student. I have gained so much from my time there, and learned a great deal from the faculty and students with whom I had contact.

Colleagues in public health in England, including Dr Gillian Maudsley, Dr Daniel Seddon and Professor Margaret Whitehead have encouraged my progress, and offered the flexibility and support required to complete my dissertation whilst in post as a clinical lecturer and specialty registrar in public health.

I would also like to thank the many other colleagues, friends, and family members (particularly A&A) in California and England who regularly inquired, *"how's it going?",* and played a role in supporting me and my doctoral studies. My parents have always provided a solid background of love and encouragement. They are a constant inspiration to me both personally and professionally. A special mention is due to our daughter for her essential contribution of fun, comedy and happy distraction. Finally, this would not have started or finished without the immense love, patience and wisdom of my husband, who has helped in countless ways to make this happen.

Chapter 1 Introduction

Self-rated health represents a summary measure of the multi-domain determinants that shape health in individuals. Though simply assessed, as a single measure item on a Likert scale, self-rated health has been hailed as providing an indication of global health status in a way that nothing else can (1). How a person evaluates and ranks his or her wellbeing has been shown to provide a valid and valuable assessment of objective or overall health status, and is acknowledged and used as a key indicator in health research.

Self-rated health is known to be an independent predictor of future health outcomes. Numerous studies have examined the relationship of self-rated health with mortality, and found that the association persists independently of several objective indicators including physician-rated health (1-4). These results signify the "*profound meaning of the answers evoked by this easily communicated and deceptively simple question"* (4)(p.103).

It is clear therefore, that in a public health context, for high proportions of populations to report very good or excellent subjective health is in itself an important end point. Furthermore, recognizing exactly what influences subjective health in populations is vital to developing effective interventions. An understanding of the common attributes within population groups that either confer susceptibility to illness, (or to poor self-rated health), or support wellbeing (and excellent self-rated health) is important if action is to be targeted at mitigating adverse conditions and promoting protective factors. This is the focus of my dissertation. In this study, I use self-rated health as the outcome measure, and study the characteristics of individuals in a population who rate their health as excellent or very good, versus those who rate it as only good, fair, or poor.

Health is a complex construct. The health implications of a single risk factor or exposure may not be universally identical. That is, they may depend on interaction with coexisting variables, so that different combinations of risk or protective factors produce different outcomes. My aim is to investigate the joint influence on health of a broad range of multi-domain factors. I will assess the relative importance of demographic, lifestyle, medical, and psychosocial factors, and indicators of living and working conditions on self-rated health. The results should be interpretable with a view to understanding differences in self-rated health status and so be meaningful in both a statistical and practical sense. To achieve this, I have applied recursive partitioning analytic methods, to a study population drawn from the CARDIA Study (Coronary Artery Risk Development in Young Adults). As discussed in more detail below, this prospective cohort study was initiated in 1985 to study the impact of lifestyle and other factors on the development of coronary artery disease in young adults. A total of 5115 men and women were recruited in four urban areas of the United States: Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota, and Oakland, California. The CARDIA study population at baseline was approximately balanced in terms of subgroups of race/ethnicity, gender, education and age (~18-24 and ~25-30). A majority of the group has been examined at each of the seven follow-up examinations, in years 2, 5, 7, 10, 15, 20 and 25.

Background

Self-rated health

One of the most frequently used methods to assess self-rated is a single global question on overall health, to the effect of, *"in general, would you say your health is excellent, very good, good, fair, or poor?"* Alternative phrasings of the question exist with differing response options (5, 6). The wording for frame of reference may also vary, and be classed as *'non comparative'* (whether health is rated as excellent, good, fair, or poor), *'age comparative'* (whether health is rated better, same or worse compared to other people of their age), and *'time comparative'* (health rated compared to how it was at a previous point in time) (7). There is evidence that though wording may differ, the questions appear to result in parallel assessments of subjective health so that the exact phrasing is unimportant (1, 4-7). However, a few studies conclude that there *is* sensitivity to reference points, so that differently phrased self-rated health questions are not comparable measures in older age groups, and may not have comparable relationships with mortality (8, 9). Bailis et al. (10) further interpreted self-rated health as reflecting both a *spontaneous assessment* of health, as well as *enduring self-concept* views. Using data from the National Population Health Survey in Canada, they found that within a 2-year period, self-rated health status changed for almost half the sample of adults, reflecting changes in self-reported physical and mental health and health practices (spontaneous assessment). Self-rated health was also found to be a significant independent predictor of self-rated health 2 years later, reflecting an enduring self-concept of health to some degree.

Assessments of self-rated health are simple to ascertain and provide broad measures of population health that go beyond just morbidity and mortality (11). A question on self-rated health has been included in some large-scale national level surveys in the US, including, the National Health Interview Survey (NHIS) (12), the National Health and Nutrition Examination Survey (NHANES) (13) and the Behavioral Risk Factor Surveillance System (BRFSS) (14). In the United Kingdom (UK), the 2001 census included a question on self-rated general health for the first time, and this was repeated in the 2011 Census (15, 16).

Self-rated health as a predictor of future health outcomes

The first clear empirical evidence of self-rated health as an important predictor of mortality is credited to the analysis of the Manitoba Longitudinal Study on Aging. Early (within first two years) and late (years 3 to 7) mortality for persons with poor self-rated health was 2.92 and 2.77 times higher than that of those with excellent self-rated health, controlling for age, sex, objective health status and residence location (2). In their 1997 review of 27 community studies on self-rated health and mortality, Idler and Benyamini discovered a dose-response pattern so that the probability of death was highest for the lowest rating of subjective health, and stronger effects were observed for men compared with women (1). The relationship between self-rated health and mortality is seen to persist independently of numerous objective indicators, including objective measures of health status, reported by individuals or medical staff. In a meta-analysis by DeSalvo et al. (17), there was a two-fold higher mortality risk for persons with poor self-rated health compared with persons with excellent self-rated health. Consequently, self-rated health has been identified as an independent predictor of a range of other health outcomes including functional ability (18, 19), morbidity (20) and health care utilization (21).

Determinants of self-rated health

Internationally, several studies have sought to investigate the correlates of self-rated health with a view to explaining this predictive capacity. Literature identifies various independent determinants of self-rated health in different populations, from demographic, lifestyle, medical, psychosocial and socioeconomic domains. Generally, lower health ratings are associated with increasing age (22-25), being female (7, 22, 26), being of black (26)or Hispanic ethnicity (25, 26) compared with white. Higher education and income are positively associated with higher self-rated health status (25-29). Behavioral factors associated with poorer self-rated health include diet, physical inactivity, smoking, alcohol consumption, and higher body weight (27, 30-32). Associations with self-rated health have also been observed for chronic medical morbidity and physical functioning, fatigue, lack of energy, number of medications, and negative affect (31, 33). Psychosocial variables include lack of social support, sense of community belonging (34), low perceived control over life, indicators of happiness, and working conditions (27, 28, 31). As described by Kaplan and Camacho (3), cross sectional studies have demonstrated higher rates of poor perceived health in people who also report higher levels of social isolation, negative life events, depression, job problems, unhappiness, life dissatisfaction and unemployment. Poor self-rated may be a common feature linking psychosocial factors to disease outcomes via a decrease in host resistance (3, 35).

Considering a number of the studies that have sought to explore the determinants or correlates of self-rated health, what conclusions can be drawn? There are two major issues to consider – the lack of consensus across studies, and the methods used.

Differences in the main determinants of self-rated health are evident across populations drawn from different age groups (24, 36-38), occupations (39-43), and geographies, such as USA (26), Canada (38, 44), Spain (36), Sweden (27), Greece (22, 45), Syria (46), China (47) , Japan (48), Pakistan (49), and Australia (23). By country, some determinants of self-rated health may be culturally shaped in that they reflect specific gender roles, social norms and expectations (46).

In the elderly, self-evaluations of health in community residents have reflected objective health status, such that those with better self-perceived health also report fewer illnesses, take fewer medications, and suffer from less disability (50). In a study by Benyamini et al., 487 elderly participants from the Rutgers Aging and Health Study were asked to rate health factors on which they formed their assessment of self-rated health status (51). Each of 42 health-related factors was perceived as at least somewhat relevant for some of the individuals when judging their health. Factors relating to overall physical functioning and 'vitality' were rated highly. People with higher ratings of self-rated health were more likely to rate risk factors and indicators of good physical and psychological health; individuals with poor or fair ratings were more likely to rate factors such as medications or current symptoms and illnesses as important factors when rating their own health (51).

In adolescent populations, behavioral factors, physical health, family structure and income, and psychological distress have all proven to be relevant to self-rated health (52). Tremblay et al explored factors related to self-perceived health in a sample of Canadian adolescents using a cross-sectional design, with data from the 2000-2001 Canadian Community Health Survey. The odds of reporting very good or excellent health were lower for teens living in houses with lower income or educational attainment. Smoking, episodic heavy drinking, physical inactivity during leisure time, being obese, and low fruit and vegetable consumption were independently related to lower health ratings (38). Even between the ages of 15 and 17 years, girls were less likely to report good/excellent health compared with boys, and also more likely to have a chronic condition and experience of depression in the previous year. Younger age groups also seem to take health-compromising behaviors into consideration, such as cannabis and hard drug use, which in turn may be reflected in their poorer self-rated health assessments (53).

A large study of two occupational cohorts, using data from the Whitehall study in England, and the Gazel cohort in France concluded that self-rated health was essentially a reflection of physical and mental health (42). In populations with chronic illness and disability, psychological resources, particularly a sense of high mastery and self esteem (44), and psychosocial factors (54) were shown to associate with better health. Occupational groups may be quite homogenous in terms of socioeconomic factors or working conditions, and therefore, these factors will not necessarily emerge as relevant to self-rated health. Similarly, in groups defined by presence of a medical condition that is uniform across the population, it is comprehensible that other domains of health determinants (psychological factors, social support and networks, or economic resources,) might drive disparities in subjective health.

Differences in determinants of self-rated health are hardly surprising when considering dissimilar populations. In fact, when attempting to unpack the concept of self-rated health in a particular population context, the value is in capturing the diversity, that is, the unique determinants of health that are most important in that context and subgroup, and in gaining insight into the particular needs that could be addressed through public health intervention.

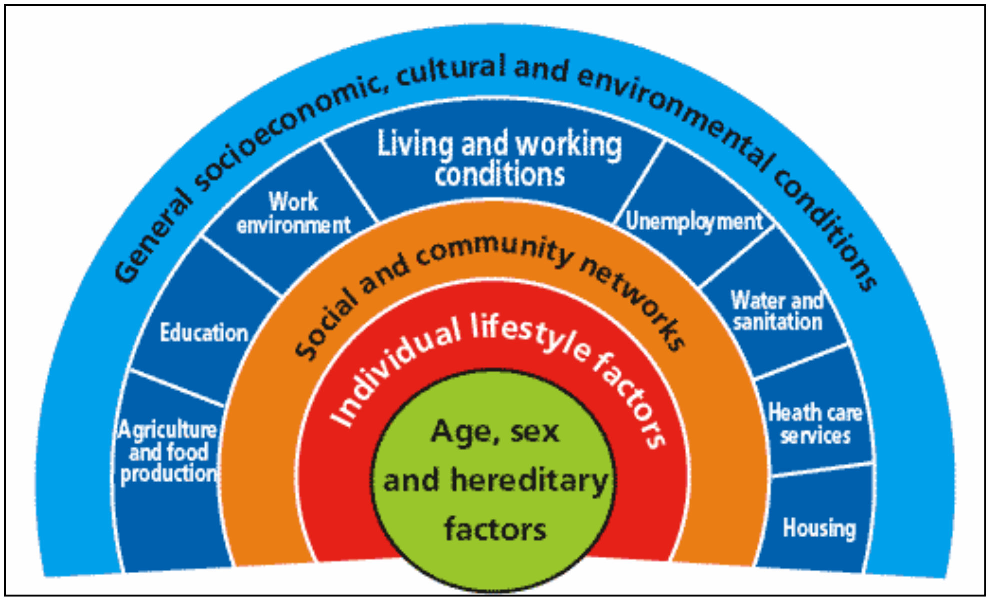
Previous quantitative studies on correlates or determinants of self-rated health, as discussed above, have used multivariate regression to model the relationship between multi-domain variables and self-rated health based in both cross-sectional and cohort studies (55). When considering the influence of such a broad spectrum of health determinants on an outcome, it is difficult however, to satisfy the requirements of traditional parametric statistical models in terms of the underlying data structure of the predictor variables, or to model a large number of variables and interactions. The results of traditional models may also be less meaningful when translated to a practical or clinical setting (56). In a review of a sample of fifty-six published studies on determinants of self-rated health, a number of problems were identified in relation to multivariate modeling. These included over-fitting, nonconformity to a linear gradient, and lack of reporting of tests for interactions (57). A further observation was that though self-rated health is a multidimensional measure, most studies do not cover its various components concomitantly.

Methods

Analytic approach

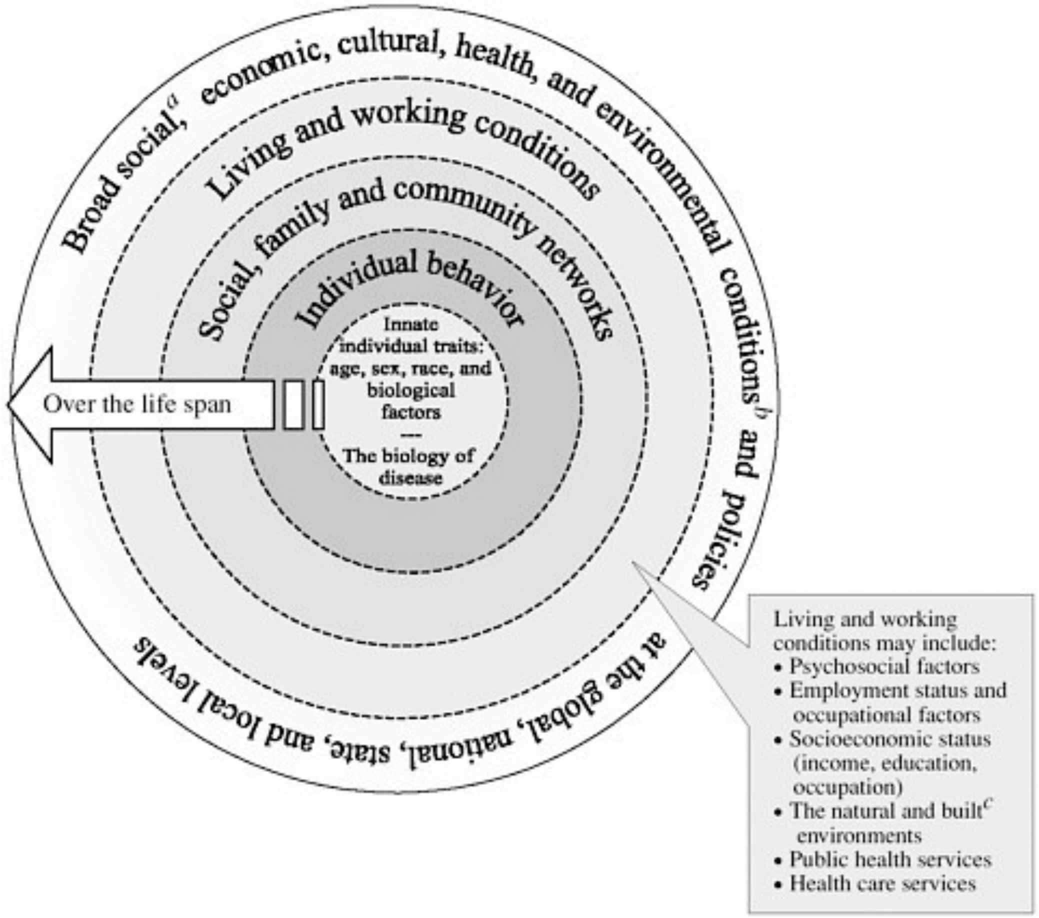
It is known that a broad range of factors is relevant to health. The ecological model of health emphasizes the relationships among these multiple determinants, and assumes that health is affected by interaction between these factors, including biology, behavior and the social and physical environment (58). Models that reflect the layered multi-domain influences on health include those by Bronfenbrenner (59), Dahlgren and Whitehead (60) (Figure 1) and Kaplan et al. (61). An ecological model of population health (Figure 2) adapted from Dahlgren and Whitehead is included in the 2003 Institute of Medicine report on *Educating Public Health Professionals for the 21st Century* (58).

Figure 1: The main determinants of health (Dahlgren and Whitehead, 1991) (60)


General socioeconomic, cultural and environmental conditions
Living and working conditions
Work environment
Education
Agriculture and food production
Unemployment
Water and sanitation
Heath care services
Housing
Social and community networks
Individual lifestyle factors
Age, sex and hereditary factors

Recognizing exactly what influences subjective health in populations is vital to developing effective interventions. An ecological model of health promotion considers both individual and social environmental factors as targets for public health action (62). Kaplan et al. also highlighted the importance of a public health approach that, "*does not exclusively privilege the proximal,* [*or focus on molecular explanations of disease*] *but seeks opportunities for understanding and intervention at both upstream and downstream vantage points*" (61)(p.42). The ecological perspective proposed by Bronfenbrenner considers multiple interacting layers of environmental influences as affecting individuals. Accordingly, ecological research seeks to "*control in as many theoretically relevant ecological contrasts as possible within the constraints of practical feasibility and rigorous experimental design*. This is in contrast to classical experiments that attempt to '*control out*' potential confounders whilst focusing on a single variable (59)(p.518).

Figure 2: The Ecological Model of Population Health (*The Future of the Public's Health*, IOM 2003) (58)


Broad social,[a] economic, cultural, health, and environmental conditions[b] and policies at the global, national, state, and local levels
Living and working conditions
Social, family and community networks
Individual behavior
Innate individual traits: age, sex, race, and biological factors
---
The biology of disease
Over the life span
Living and working conditions may include:
• Psychosocial factors
• Employment status and occupational factors
• Socioeconomic status (income, education, occupation)
• The natural and built[c] environments
• Public health services
• Health care services

Thus, for this dissertation, in seeking to explore factors associated with self-rated health, rather than consider only independent effects, or highlight single domains of importance and focus on these, I propose that in a real-life context, all domains and layers would be pertinent to self-rated health in some way, but the relative importance may vary in different populations. Furthermore, it is not necessary that determinants or correlates are uniform across populations. I aim to identify differences between population subgroups in terms of which factors or combinations of factors are significant to self-rated health.

For this study, I use a dataset drawn from the CARDIA study, and apply recursive-partitioning methods, namely classification tree analysis and random forests to understand which variables or interactions of variables drive the phenomenon of self-rated health status. As noted above, the CARDIA dataset is a prospective cohort study initiated in 1985 to study the impact of lifestyle and other factors on the development of coronary artery disease in young adults, based in four urban areas of the United States(63). The CARDIA Publications and Presentations Committee and the Institutional Review Board at the University of California, Berkeley, have approved the protocol for this study.

The CARDIA dataset will be systematically split into groups of maximum homogeneity in terms of the specified outcome, thus identifying subgroups that have relatively homogenous self-rated health status. At each split, the classifier is the predictor variable that splits the remaining population into the most homogenous groups. This process continues either up to a point where a predetermined number of individuals exist in each group, or to the point of saturation, where the largest possible tree is grown. Following refinement with pruning and cross-validation, in the final tree, the population is ultimately divided into a number of relatively homogenous groups, each categorized by a specific level of self-rated health. As a result, the tree structure created from the data displays how groups are generated, and identifies associations between specific self-rated health status and exposures selected by the tree model.

This type of analysis gives a simple characterization of the conditions in terms of risk and protective factors that determine when an individual is in one class of self-rated health rather than another. To the best of my knowledge, this is the first application of classification tree analysis to study the characteristics associated with self-rated health.

The CARDIA dataset

The CARDIA study (Coronary Artery Risk Development in Young Adults) was initiated in 1985. It is a prospective cohort study designed to study the impact (both favorable and unfavorable) of lifestyle and other factors on the development of coronary artery disease in young adults. A total of 5115 men and women were recruited in four urban areas of the United States: Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota, and Oakland, California. The recruiting age was 18-30 years. The CARDIA study population at baseline was approximately balanced in terms of subgroups of race/ethnicity, gender, education and age (~18-24 and ~25-30). Cohort examinations have been carried out at year 2 (1987-1988), year 5 (1990-1991), year 7 (1992-1993), year 10 (1995-1996), year 15 (2000-2001) year 20 (2005-2006), and year 25 (2010-2011). A majority of the group has been examined at each of the seven follow-up examinations: 91% (4624), 86% (4352), 81% (4086), 79% (3950), 74% (3672), 72% (3547), and 72% (3499), respectively.

A fuller description of the design, recruitment and characteristics of the CARDIA cohort is given in the paper by Friedman et al. (63). Further information on the CARDIA study is also available online (at www.cardia.dopm.uab.edu).

For this study, data were taken from the year-15 examination of the cohort, conducted in 2000-2001, and collected through interviewer and self-administered questionnaires (with the exception of information on race/ethnicity taken from the 1985-1986 data collection, and family history taken from the 1995 data collection). In year-15, data were collected on both self-rated health and a broad range of health-related factors of interest in this study. Though similar data were collected in later years, I was interested in an initial study of a relatively young adult population.

From an original total of 5115 participants at baseline, 3672 were examined in year 15. From the year 15-group, all participants who had a response for self-rated health, and were coded as male or female were included in the final study sample of 3649 participants. The predictor variables chosen for this study are based on domains considered to be fundamental determinants of health (Table 1). A number of different models exist; I have based the selection on the levels presented in the model proposed by Dahlgren and Whitehead (60) (Figure 1 above). In the CARDIA study, self-rated health was assessed in the CARDIA examinations as part of the SF-12 questionnaire, a validated and widely used tool to assess quality of life (64).

Table 1: Predictor variables based on the determinants of health selected from the CARDIA study dataset

Level 1: Age, sex & hereditary factors ▪ Age ▪ Sex ▪ Race / ethnicity ▪ Family history
Level 2: Individual lifestyle factors ▪ Medical history ▪ Diet ▪ Physical activity ▪ Smoking ▪ Alcohol ▪ Tobacco ▪ Illicit drug use
Level 3: Social & community influences ▪ Social support / network *("feeling that family friends really care")* ▪ Sense of close knit neighborhood, neighborhood cohesion
Level 4: Living & working conditions ▪ Education ▪ Income ▪ Housing - rent or own house ▪ Employment - working versus unemployed ▪ Medical insurance ▪ Access to health services ▪
Level 5: General socioeconomic, cultural and environmental conditions This level will not be directly reflected in the choice of variables. All the data are collected from study sites within the United States of America, during a known timeframe, and thus knowledge of these forces during the period of study collection may be useful background information in interpretation of results. Factors for level 5 will represent other relevant influences on health, as below.
Level 5: Other influences on health ▪ Control & adequacy of resources *("how hard is it to pay for basics")* ▪ Optimism for the future ▪ Experience of discrimination due to gender / race-ethnicity / socioeconomic status ▪ Limited function *("unable to do moderate activity"*) ▪ Some type of on-going chronic burden

Classification tree analysis

A brief description of the classification and regression tree model process follows concepts and theory introduced in the seminal book, *Classification and Regression Trees*, by Breiman, Friedman, Olshen and Stone (65). Their work is considered referenced throughout this section, and notation appears as in the text.

Classification and regression trees (CART) are a form of recursive partitioning, and may be used:
(i) to produce an accurate classifier, that is, to create a systematic way of predicting class membership given a set of measurements on a case or,
(ii) to uncover the predictive structure of a problem, that is, to try and gain an understanding of the conditions in terms of measurement / explanatory variables that determine which class an object is in.
Classification trees are used when the outcome is dichotomous, and regression trees when the outcome is continuous. This study uses classification tree analysis.

Tree construction begins with the root node, t_1, which comprises the entire dataset. A binary question is applied to split the parent node into 2 subsets. The proportion of cases in the node answering yes, P_L, goes to the left node, t_L. The proportion of cases in the node answering no, P_R, goes to the right node, t_R. The resulting nodes are more homogenous or 'purer' than the data in the root node. At every node, the split adopted is that which maximizes the goodness of split function, i.e. results in the largest decrease in impurity possible by a split of node *t*. In this analysis, the splitting rule uses the Gini Index of diversity as a measure of node impurity. For each explanatory variable in the data set, the best split is selected based on a particular threshold or cut off point for that variable. These best splits are all ranked according to reduction in impurity achieved. The variable and its split point that most reduces the impurity of the root or parent node is selected for the split. Splitting continues so that the largest possible tree is grown; at every node the data in each of the descendant subsets are purer than the parent subset, and until all the terminal nodes are small. Each terminal node is a subgroup of the dataset, and is designated by a class label. There may be 2 or more terminal subsets with the same label. The tree is then selectively pruned or recombined upward to produce a decreasing sequence of sub trees. Cross-validation or test sample estimates pick out the sub tree with the lowest estimated misclassification rate.

Cross-validation gives an internal estimate of the true misclassification rate of the tree model. In *v*-fold cross-validation, the cases in the data set are randomly divided into *v* subsets or folds of as equal size as possible. Ten fold cross-validation is a commonly used value and shown to be optimal. In this case, each fold in turn is withheld whilst the remaining 9 folds are used to build a test tree. The remaining fold is used as an independent test sample. The 10-fold cross-validation error estimate is calculated by averaging across all 10 trees. Cross-validation is considered parsimonious with data; every case is used in tree construction and exactly once in the test sample.

Use of CART has a number of advantages. It is a non-parametric technique, which is valuable when dealing with a broad complex set of predictor variables, and a large sample size. CART analysis handles high dimensionality, a mixture of data types and non-standard data structure well, whilst providing insight into the predictive structure of the data (65). Missing variables can be adjusted for within the analysis, and since the structure of the classification is apparent, the results in terms of population subsets are relatively easy to interpret.

CART analysis has been applied in clinical studies. Podgorelec et al. described tree-based methods as providing a *simple representation of gathered knowledge,* and a valuable tool for medical decision making due to their high classification accuracy (66). In partitioning a study population by clinical signs and symptoms, tree-based methods have been used to establish high-risk groups (67). Trees may also capture information not reflected in traditional parametric analysis. Nelson et al., in a study of disease risk groups, concluded that recursive partitioning uncovered interactions between variables that could be overlooked in traditional application of logistic regression to case control data unless they are modeled based on a priori knowledge (68).

The same positive features of CART are applicable also to public health research. CART is a useful segmentation technique to identify groups with the purpose of targeting public health action. Forthofer and Bryant (69) highlight a key benefit of CART analysis: it is particularly useful when the aim is to identify differences and appreciate specific features of subgroups; seeing the relative importance of different factors may also shed light on where action could be prioritized. In contrast, parametric multivariate regression essentially models the average effect of a predictor on the outcome in a population (69, 70). Whilst this is clearly important for many studies, it may not offer the best approach in terms of informing intervention development. Also, the same intervention is not necessarily appropriate across an entire population. BeLue et al. demonstrated this in a study applying CART analysis to a population of US adolescents aged 12-17 years. They identified subsets of the population with differing risk and protective factors related to obesity, demonstrating that in terms of public health intervention, often *"one size does not fit all'* (71).

In summary, self-rated health is established as a key and informative indicator in health research, and a recognized predictor of future health outcomes. For populations to report high fractions of excellent or good self-rated health is in itself a valuable outcome. To achieve this, public health interventions need to be informed by an understanding of the characteristics associated with good or poor self-rated health. Previous studies have examined the independent predictors or correlates of self-rated health. In this study, in the context of the ecological model of health, I seek to examine the joint impact of multi-domain and multilevel variables on health, and understand their relative importance in population subgroups. This is achieved through the application of recursive partitioning methods.

Structure of dissertation

The next three chapters are structured as follows:

In chapter 2, I investigate and compare the characteristics of individuals in the study population who appraise their health as excellent or very good, versus those who rate it as only good, fair, or poor. I utilize classification tree analysis to segment the CARDIA study sample of 3649 individuals, and identify subgroups with some shared characteristics and relatively homogenous self-rated health status. A range of predictor variables is used to represent a wide spectrum of multi-domain influences on health.

In chapter 3, the CARDIA study sample is divided into subsets based on total family income, to create a socioeconomic gradient. I then investigate the characteristics, in terms of risks and protective factors, associated with self-rated health within each income subset. Using classification tree analysis, my aim is to explore whether these factors differ by subset across the socioeconomic gradient.

In chapter 4, I extend the analysis in chapter 2, exploring characteristics associated with self-rated health, with the application of random forests. This method produces an ensemble of classification trees, which improves accuracy compared with a single tree, and more robust variable importance measures. Predictor variables are ranked in order of decrease in node impurity in the model. The use of recursive partitioning methods can effectively highlight subgroups with relatively homogenous risk for a self-rated health status, and important risk and protective factors for further inquiry. This is valuable in developing appropriate interventions that relate to the real-life mix of influences that impact health in different populations.

In chapter 5, I summarize the findings from chapters 2, 3 and 4, with further discussion and conclusions.

Chapter 2 Characteristics associated with self-rated health in the CARDIA Study: classification tree analysis

Introduction

Health is determined by multiple interacting influences. To reduce health inequalities, action should be informed by knowledge of the drivers of both illness and wellbeing in different groups. For many conditions, it is well known that even well studied biological risk factors fail to account for all the disease that occurs (72). Meanwhile, psychosocial factors and socioeconomic conditions are associated with multiple disease outcomes. Thus, to be effective, public health strategies must take into account upstream conditions, the 'causes of the causes' (73), so that interventions are designed to consider both mitigation of adverse conditions and promotion of protective factors. Furthermore, we need to improve our understanding of the common attributes within population groups that either confer susceptibility to illness, or support resilience and wellbeing.

Self-rated health represents a summary measure of the multi-domain determinants that shape health in individuals. Though simply assessed, as a single measure item on a Likert scale, self-rated health has been hailed as providing an indication of global health status in a way that nothing else can. How a person evaluates and ranks their wellbeing provides a valid and valuable assessment of objective or overall health status, and is acknowledged and used as a key indicator in health research. Of further significance is that self-rated health has a predictive value in terms of future health-related outcomes, which is well established. In many longitudinal studies, self-rated health has been shown to predict subsequent health, even after adjusting for potential confounders of known importance, and after taking account of objective physician-rated health.

Several studies have demonstrated the significant independent effect of self-rated health on mortality in different countries and age groups (1-4). The validity of self-rated health as a summary measure of health has been demonstrated by studies that show that it is an independent predictor of mortality. The probability of death has been shown to be highest for the lowest rating of subjective health, and stronger effects have been observed for men compared with women. In a meta-analysis by DeSalvo et al., there was a two-fold higher mortality risk for persons with poor self-rated health compared with persons with excellent self-rated health (17). This relationship has been observed for several racial/ethnic groups including blacks, whites and Hispanics (74, 75). Generally, lower health ratings are associated with increasing age (22, 23), being female (7, 22, 26), and being of black or Hispanic ethnicity compared with white (25, 26). Higher education and income are positively associated with higher self-rated health status (25-28).

In this paper, I investigate and compare the characteristics of individuals in a population who appraise their health as excellent or very good versus those who consider it to be just good, fair or poor, using self-rated health as the outcome measure. My aim is to segment the study population into subgroups, based on similar health status, and identify the influences that determine group membership. This process highlights factors suitable to an early intervention focus. The use of classification trees as a data analysis tool in this context results in both a specification of groups and a simple characterization in terms of variable importance. Classification trees are a form of recursive partitioning and are useful in uncovering predictive structure, that is, an understanding of the conditions in terms of measurement or explanatory variables that determine which class an individual is in (65).

I have utilized classification tree analysis to conceptualize profiles of factors, representing the main influences on health, and study their association with self-rated health status. Such audience segmentation approaches have been shown to be of value in a public health context (70) as they can serve as an initial tool to suggest population subgroups that might have homogenous risk of disease, or as in this study, similar self-rated health status. However, in contrast to the usual underlying cluster analysis, classification trees operate through supervised machine learning algorithms. As a result, rather than simply grouping based on like factors, predictors are modeled on the specified condition or outcome. This type of analysis offers several advantages by considering the joint impact of a broad range of health determinants. Beyond the independent effects of single predictors, and small groups of variables, recognizing the 'real-life' manner in which an array of influences can interact to shape health outcomes is essential. This type of knowledge is useful in guiding the design of public health programs so that they are focused and relevant to need, and hence more successful in improving health outcomes.

Methods

Data: The CARDIA Study

I utilized cross-sectional data collected during the CARDIA Study (Coronary Artery Risk Development in Young Adults). The CARDIA longitudinal study began in 1985 with a cohort of 5115 black and white men and women, aged between 18 and 30 at baseline, recruited in Birmingham, Alabama; Chicago, Illinois; Minneapolis, Minnesota; and Oakland, California. For this study, data were taken from the year 15 examination of the cohort, conducted in 2000-2001, through interviewer and self-administered questionnaires (with the exception of race/ethnicity information taken from the 1985-1986 data collection, and family history information taken from the 1995 data collection). From an original total of 5115 participants, 3672 were followed up in year 15. From the year 15 group, all participants who had a response for self-rated health, and were coded as male or female were included in the final study sample of 3649 participants.

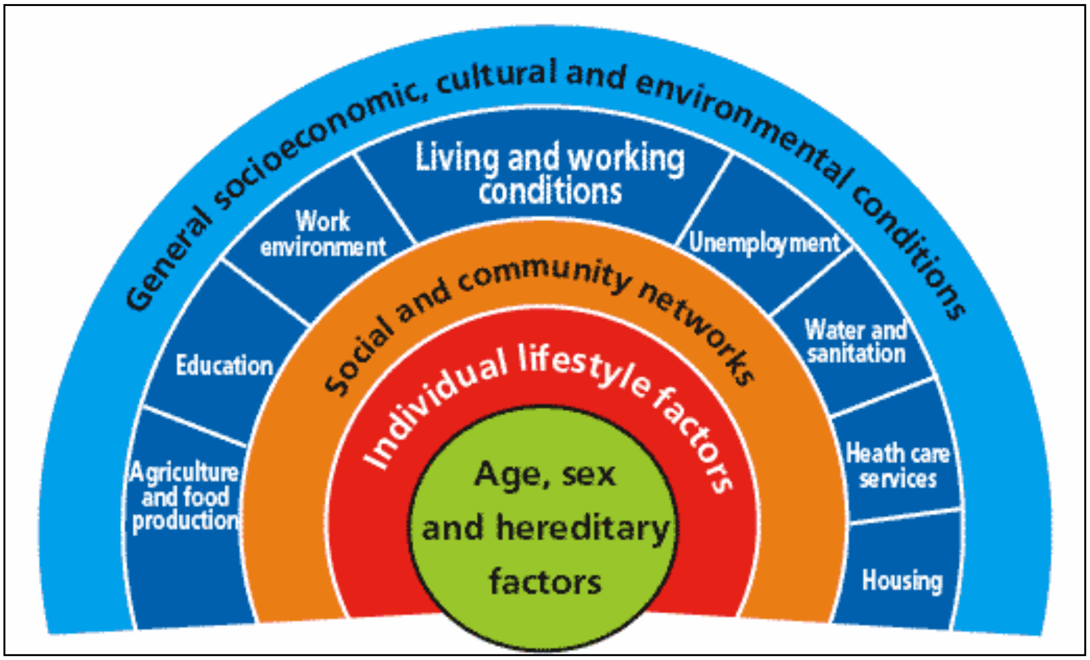
Outcome variable

Self-rated health was assessed on a five-point scale, by the question, *"In general would you say your health is excellent, very good, good, fair or poor?"* Responses were categorized by grouping together excellent or very good as 'good' self-rated health, and responses of good, fair or poor, as 'poor' self-rated health.

Predictor variables

A range of predictor variables was used in the analysis in order to represent a wide spectrum of multi-domain health determinants. These were based on the layers of influence on health from the Dahlgren and Whitehead model (60)(Figure 1): age, sex and hereditary factors; individual lifestyle factors and medical history; social and community influences; living and working conditions. The method of assessment in the CARDIA questionnaire and the format in which variables were included in this study analysis are outlined in Table 1.

Figure 1: The Main Determinants of Health (Dahlgren and Whitehead, 1991) (60)


General socioeconomic, cultural and environmental conditions
Living and working conditions
Work environment
Education
Agriculture and food production
Unemployment
Water and sanitation
Heath care services
Housing
Social and community networks
Individual lifestyle factors
Age, sex and hereditary factors

Age, sex and hereditary factors

Variables included age in years, sex, race/ethnicity (Hispanic, black, or white), and family history of maternal or paternal diabetes, high blood pressure, stroke, angina, or heart attack.

Individual lifestyle factors and medical history

For medical history, respondents were asked in the questionnaire if a doctor or nurse had ever said that they had any of a list of major diseases or health problems. Conditions included in the analysis were self-reported history of high blood pressure or hypertension, high blood cholesterol, heart disease, asthma, chronic bronchitis, emphysema, diabetes, thyroid disease, liver disease, kidney disease, cancer or malignant tumor, HIV, stroke or transient ischemic attack (TIA), multiple sclerosis, epilepsy, nervous/emotional or mental disorder, and depression.

Physical activity over the past year was reported on a 5-point scale from physically inactive to active, compared with others of the same age and sex.

For information on diet, specifically for frequency of consumption of fast food, respondents gave the number of times per week that they ate out at any restaurant from a list of fast food chains.

Smoking was included as having smoked previously, at least 5 cigarettes a day for at least 3 months; being a current smoker, and number of cigarettes smoked per day on average.

For alcohol intake, I used continuous variables relating to number of drinks per week of wine, beer or hard liquor.

Drug intake was recorded as history of ever using marijuana, crack, non-crack cocaine, amphetamines, or opiates.

Social and community influences

To represent social support, indicator variables were created representing how much family or friends were perceived to care, and whether they could be relied on to talk with about worries. For neighborhood cohesion, indicators were whether respondents felt people were willing to help their neighbors; whether they lived in a close-knit neighborhood; if people in the neighborhood could be trusted; whether people in the neighborhood generally get along with each other, and if they share the same values.

Living and working conditions

Variables measuring the highest grade of school completed, total combined family income over the previous 12 months, and housing and employment status were utilised, along with level of difficulty paying for basics and medical care, availability of health insurance, and access to health services. Additional variables were experience of discrimination due to gender, race/ethnicity or socioeconomic status, in 7 different settings, reported on-going chronic burden, and measures of optimism for the future and control over life events.

Though macro-level general socioeconomic, cultural and environmental conditions were not directly reflected through the choice of variables, all data were collected from urban study sites within the United States, during a known timeframe, and thus knowledge of these forces during the period of study collection is useful background information in interpretation of results.

Data Analysis

Data analysis was carried out using IBM SPSS Statistics v19.0.0. I produced descriptive summaries for the study sample. I used the chi-square test for independence (with Yates' continuity correction for 2x2 tables) to assess the bivariate relationship between individual categorical predictor variables and self-rated health, and Mann Whitney U Tests for continuous predictor variables and self-rated health, following tests for normality of distribution.

Classification trees were used to segment the population, and identify subgroups with similar characteristics and self-rated health status. Indicator variables were created for a number of predictors; in order to minimise restriction of the model a priori, not all variables for which participants were asked to choose a response from ordinal categories, were dichotomized (physical activity rating, education, income and chronic burden). Responses of 'don't know' to a question were grouped with and treated as missing data. The parameters of the tree were specified as: cross validation with 10 sample folds, minimum number of cases, 100 for parent node, and 50 for child node.

I note that interpreting the form of the tree, particularly the combination of variables that define nodes as causal is very dubious, particularly in finite samples (e.g., in the low 1000s). Small variations in the data can produce very large variations in the structure of the tree, even though the prediction accuracy might not vary much at all. This is due to the fact that this is not a well-defined statistical parameter, and as such, there is no rigorous theory for providing inference on the structure of the tree. Thus, the interpretation of the tree is entirely exploratory and the results are interpreted in that light.

Table 1: Predictor variables used to analyze factors associated with self-rated health in the CARDIA study

Variable	**Method of assessment in CARDIA questionnaire**	**Variable format for study analysis**
Age, sex and hereditary factors		
Age	In years	Age in years
Sex	Male or Female	Male=1, Female=2
Race/ethnicity	Hispanic, black (Not Hispanic), white (not Hispanic)	Hispanic, black (not Hispanic), white, not Hispanic
Family History	Did your natural mother / father ever have any of the following: diabetes, high blood pressure, stroke, angina, heart attack?	Indicator variable for history of maternal or paternal diabetes, high blood pressure, stroke, angina, heart attack

Individual lifestyle factors and medical history		
Medical history – presence of disease	Has a doctor or nurse ever said that you have: high blood pressure; high blood cholesterol; heart disease; asthma; chronic bronchitis; emphysema; diabetes; thyroid disease; liver disease; kidney disease (excluding nephritis or glomerulonephritis); cancer or malignant tumour; HIV; stroke or TIA (transient ischemic attack), multiple sclerosis, epilepsy (seizures), nervous / emotional or mental disorder, depression?	Indicator variables for history of disease for each condition.
Diet	Thinking about how often you eat out, how many times in a week do you eat breakfast, lunch, or dinner out in a place such as: McDonald's, Burger King, Wendy's, Arby's, Pizza Hut, or Kentucky Fried Chicken?	Continuous variables: Number of times per week that breakfast, lunch, or dinner eaten out in fast food restaurant.
Physical activity	Compared to other people your age and sex, what number would you choose for rating your physical activity during the past year?	5 point rating of physical activity (physically inactive to very active) 1=physically inactive 2=less than moderately active 3=moderately active 4=more than moderately active 5=very active
Smoking / Tobacco	Have you ever smoked cigarettes regularly for at least 3 months? By regularly we mean at least 5 cigarettes per week almost every week? Do you still smoke cigarettes regularly? By regularly we mean at least 5 cigarettes per week almost every week? How many cigarettes do you smoke per day on the average?	Indicator variables for history, for at least 3 months, of being regular: Cigarette smoker (at least 5 per week almost every week) Still smoke cigarettes regularly (at least 5 per week almost every week) Continuous variable: number of cigarettes smoked per day on average (1 pack=20 cigarettes)
Alcohol	Number of drinks per week of wine (about a 5oz. glass). Number of drinks per week of beer (1 beer is a 12 oz. glass, can, or bottle). Number of drinks per week of hard liquor (each shot of 1½oz. counted as 1 drink).	Continuous variables for number of drinks per week for wine, beer and hard liquor
Illicit drug use	History of <u>ever using:</u> marijuana / crack / other forms of cocaine that are not crack (including powder, free base, and coca paste) / amphetamines ("Speed" or "Uppers") / opiates for non-medical reasons (Heroin, Dilaudid, Morphine, Demerol)?	Categorical indicator variables for history of drug use for each drug type

Social and community influences		
Social support / network ("feeling that family friends really care")	How much do members of your family or friends really care about you? How much can you rely on them if you need to talk about your worries? 1=Not at all, 2=A little, 3=Some, 4=A lot	Responses of some + a lot grouped together, and not at all + a little grouped together forming indicator variables for : Family members or friends are perceived to care Can rely on family members or friends if need to talk about worries.
Sense of close knit neighborhood, neighborhood cohesion	In thinking about the neighborhood in which you live, how strongly do you agree or disagree that: People around here are willing to help their neighbors. This is a close-knit neighborhood. People in this neighborhood can be trusted. People in this neighborhood generally don't get along with each other. People in this neighborhood do not share the same values. 1=Strongly Agree 2=Agree, 3=Neutral, 4=Disagree, 5=Strongly Disagree	Strongly agree + agree grouped together; and neutral + disagree + strongly disagree grouped together, forming indicator variables for: People willing to help their neighbors. Live in close-knit neighborhood. People in the neighborhood can be trusted. Disagree + strongly disagree grouped together, and strongly agree + agree + neutral grouped together, forming indicator variables for: People in the neighborhood generally get along with each other. People in the neighborhood share the same values
Living and working conditions		
Education	What is the highest grade (or year) of regular school you have completed? 01-08=elementary school 09-12=High School 13-16=College 17-20+=Graduate School	Ordinal categories
Income	Which of these categories best describe your total combined family income for the past 12 months? This should include income (before taxes) from all sources, wages, veteran's benefits, help from relatives, rent from properties, and so on. 1 = Less than $5,000 2 = $5,000 - $11,999 3 = $12,000 - $15,999 4 = $16,000 - $24,999 5 = $25,000 - $34,999 6 = $35,000 - $49,999 7 = $50,000 - $74,999 10=$75,000 - $99,999 11=$100,000 and greater	Ordinal categories

Housing - rent or own house	Is the home where you live: 1 Owned or being bought by you (or someone in the household)? 2 Rented for money? 3 Occupied without payment of money or rent? 4 Other (specify)	Categorical indicator variable for own home versus rented, occupied or other
Employment - working versus unemployed	The following categories might best describe your current main daily activities and/or responsibilities: Are you working full-time? Are you working part-time? Are you unemployed or laid off? 1 No 2 Yes 9 No answer	Categorical indicator variable for unemployed status
Control & adequacy of resources ("how hard is it to pay for basics")	How hard is it for you (and your family) to pay for the very basics like food and heating? How hard is it for you (and your family) to pay for medical care? Would you say it is: 1 Very hard 2 Hard 3 Somewhat hard 4 Not very hard	1+2+3 grouped together versus 4 to create indicator variables for: Hard to pay for basics Hard to pay for medical care
Medical insurance	In the past two years, have you always had health insurance or other coverage for medical care? Health insurance coverage refers to health insurance (like Blue Cross/Blue Shield) or participation in an HMO. Other than government programs, health insurance can be obtained through an employer, union, or school. Are you covered by health insurance of this type? Are you self-insured? That is, do you or someone else pay totally for your health insurance? 1=yes, 0=no	Categorical indicator variables for: Always had health insurance or other coverage for medical care in the past two years. Covered by health insurance like Blue Cross/Blue Shield or participation in an HMO; health insurance obtained through an employer, union, or school. Self-insured

Access to health services	Was there anytime during the past two years when you did not seek medical care because it was too expensive or health insurance did not cover it? Do not include dental care. 1 No 2 Yes 8 Not sure/refused Overall, how hard has it been for you to get the health services you have needed? 1 Very hard 2 Fairly hard 3 Not too hard 4 Not hard at all 9 No answer	Categorical indicator variables: 0=no, 1=yes did not seek medical care in past 2 years due to cost Has been hard overall getting health services 1='hard' (grouped 1+2) 0='not hard' (grouped 3+4)
Experience of discrimination due to: Gender Race/ethnicity or colour Socioeconomic position or social class	Have you ever experienced discrimination, been prevented from doing something, or been hassled or made to feel inferior in any of the following seven situations because of gender / Race-ethnicity or color / Socioeconomic position or social class: At school Getting a job Getting housing At work At home Getting medical care On the street or in a public setting	Categorical indicator variables for experience of discrimination due to gender, Race-ethnicity or color, and socioeconomic position or social class for each setting
Some type of on-going chronic burden	Please indicate whether you have experienced any of these strains for longer than 6 months. Serious ongoing health problem (yourself). Serious ongoing health problem (someone close to you). Ongoing difficulties with your job or ability to work Ongoing financial strain Ongoing difficulties in a relationship with someone close to you 1=No 2=Yes, but not very stressful 3=Yes, moderately stressful 4=Yes, very stressful	Ordinal categories for each of 5 options

Optimism for the future	Please indicate how strongly you agree or disagree with each of the following statements. I have little control over the things that happen to me I often feel helpless in dealing with the problems of life I'm always optimistic about my future 1=Strongly Agree 2=Agree 3=Neutral 4=Disagree 5=Strongly Disagree	Dichotomous categorical indicator variables: Have no control over the things that happen (grouped 1+2 and 3+4+5) Feel helpless in dealing with the problems of life (grouped 1+2 and 3+4+5) Always optimistic about future (grouped 1+2 and 3+4+5)

Results

In the final study sample of 3649 participants, there were 2135 individuals (58.5%) in the good self-rated health category, and 1514 individuals (41.5%) in the poor self-rated health category. The mean age was 40.2 years. There were 2036 women (55.8%) and 1613 men (44.2%). In the sample, 52.6%, (1921) individuals were white, 47.1% (1717) were black, and 0.3% (11) were Hispanic; 18% of the sample had finished high school only, 25% had completed 4 years in college, and 5.6% had completed 4 years in Graduate School. The majority of individuals owned their home (2507, 68.7%) (Table 2).

Using chi-square tests for independence (with Yates' correction for continuity for 2x2 tables), there were significant associations between self-rated health status and sex, race/ethnicity, physical activity rating, cigarette smoking, perceived social support and neighborhood cohesion, total family income, home ownership, unemployment, health insurance, difficulty paying for basics, optimism for the future, and serious ongoing health problems, at significance level $p<0.05$ (Table 2). Most medical and family history variables were found to be significantly associated with self-rated health in bivariate analysis at the $p<0.05$ levels apart from paternal stroke, cancer and HIV. Due to small cell numbers, the result for multiple sclerosis ($p=0.234$) and emphysema ($p=0.008$) may not have been valid at $p<0.05$ level.

For the poor self-rated health category, a Mann Whitney U Test revealed significantly higher number of fast food meals per week, cigarettes smoked per day, and liquor drinks per week ($p<0.05$). The number of wine drinks per week was higher in the good self-rated health group ($p<0.05$) (Table 3).

Table 2: Relationship between selected predictors and self-rated health

Predictor variables		Self-rated health category				
		Poor		Good		Chi-square for independence
		Count	%	Count	%	
Sex	Male	626	38.8%	987	61.2%	p=0.004*
	Female	888	43.6%	1148	56.4%	
Race/ethnicity	Hispanic	5	45.5%	6	54.5%	p<0.05*
	Black, not Hispanic	887	51.7%	830	48.3%	
	White, not Hispanic	622	32.4%	1299	67.6%	
Physical activity rating **1= physically inactive** **5=very active**	1	170	72.6%	64	27.4%	p<0.05
	2	372	60.2%	246	39.8%	
	3	730	45.3%	881	54.7%	
	4	142	22.5%	490	77.5%	
	5	96	17.6%	449	82.4%	
Still smoke cigarettes regularly	No	264	38.9%	415	61.1%	p<0.05*
	Yes	452	56.4%	350	43.6%	
Social support **Family members perceived to care**	No	56	62.9%	33	37.1%	p<0.05*
	Yes	1457	40.9%	2102	59.1%	
Live in close-knit neighborhood	No	971	46.0%	1138	54.0%	p<0.05*
	Yes	543	35.3%	997	64.7%	
Total family income	Less than $5,000	67	77.0%	20	23.0%	p<0.05
	$5,000 - $11,999	97	69.3%	43	30.7%	
	$12,000 - $15,999	66	58.9%	46	41.1%	
	$16,000 - $24,999	131	54.8%	108	45.2%	
	$25,000 - $34,999	183	53.5%	159	46.5%	
	$35,000 - $49,999	268	47.1%	301	52.9%	
	$50,000 - $74,999	305	38.6%	486	61.4%	

	$75,000 - $99,999	193	36.6%	334	63.4%	
	≥ $100,000	183	22.9%	615	77.1%	
Own home	No	582	51.2%	554	48.8%	p<0.05*
	Yes	929	37.1%	1578	62.9%	
Unemployed	No	1324	40.1%	1981	59.9%	p<0.05*
	Yes	185	55.6%	148	44.4%	
Had health insurance past 2 years	No	241	52.4%	219	47.6%	p<0.05*
	Yes	1269	39.9%	1913	60.1%	
Difficulty paying for basics	No	1054	36.2%	1861	63.8%	p<0.05*
	Yes	452	62.6%	270	37.4%	
Optimistic for future	No	578	54.5%	482	45.5%	p<0.05*
	Yes	936	36.2%	1653	63.8%	
Serious personal ongoing health problems (self) > 6 months	1	1003	34.6%	1899	65.4%	p<0.05
	2	174	60.2%	115	39.8%	
	3	200	73.5%	72	26.5%	
	4	137	73.7%	49	26.3%	
	Total	1514	41.5%	2135	58.5%	
*Continuity correction						

Table 3: Mann Whitney U Tests

Test Statistics (grouping variable: self - rated health category)

	Age (years)	Fast food meals per week	Number of cigarettes per day
Mann-Whitney U	1582321.500	1152260.000	73845.500
Wilcoxon W	3862501.500	2975355.000	135973.500
Z	-1.084	-6.277	-2.014
Asymp. Sig. (2-tailed)	.278	.000	.044

	Wine drinks per week	Beer drinks per week	Liquor drinks per week
Mann-Whitney U	879174.500	985571.000	937558.000
Wilcoxon W	1527265.500	1634801.000	2457454.000
Z	-6.029	-.358	-3.212
Asymp. Sig. (2-tailed)	.000	.720	.001

Ranks

	Self-rated health category	N	Mean Rank	Sum of Ranks
Age (years)	0 poor SRH	1514	1847.37	2796923.50
	1 good SRH	2135	1809.13	3862501.50
	Total	3649		
Fast food meals per week	0 poor SRH	1381	1765.63	2438340.00
	1 good SRH	1909	1558.59	2975355.00
	Total	3290		
Number cigarettes per day	0 poor SRH	457	419.41	191671.50
	1 good SRH	352	386.29	135973.50
	Total	809		
Wine drinks per week	0 poor SRH	1138	1342.06	1527265.50
	1 good SRH	1743	1505.60	2624255.50
	Total	2881		
Beer drinks per week	0 poor SRH	1139	1435.29	1634801.00
	1 good SRH	1743	1445.55	2519602.00
	Total	2882		
Liquor drinks per week	0 poor SRH	1139	1489.86	1696949.00
	1 good SRH	1743	1409.90	2457454.00
	Total	2882		

Classification tree analysis

In the study sample (n=3649), 58.5% of individuals had good self-rated health, and 41.5% (n=1514) had poor self-rated health. The classification tree structure is shown in Figure 2 with further detail in Table 4. There were 15 terminal nodes in the classification tree model and an overall misclassification rate of 31% based on cross-validation. There were 8 nodes with predominantly good self-rated health, ranging from 54.9% of the subgroup to 92.9%. There were 7 nodes with predominantly poor self-rated health, ranging from 50.3% of the subgroup to 84.7%. The primary split of the sample was on physical activity rating. The other variables selected for classification in the model were total family income, serious on-going health problem, highest year of school completed, perception that neighbours can be trusted, difficulty paying for basics, history of high blood pressure, on-going difficulty with job or working ability, and number of cigarettes smoked per day.

The normalized importance of the independent variables is shown in Figure 3, and detail of values in Table 5. Physical activity rating was the highest. The next 5 were those splitting variables that appeared in the final classification tree model: chronic burden from serious on-going health problems, highest year school completed, total family income, difficulty paying for basics, and history of high blood pressure. Race/ethnicity appears next which was also a discriminating factor between the subgroups with the highest proportions of good and poor self-rated health.

The subgroup (node 1) that was more than moderately active or very active, had a higher proportion of good self-rated health of 79.8% (n=939) compared with the total study sample, and those who were moderately to physically inactive. The higher physical activity group in Node 1 was further split by total family income. Where this was less than $25,000 - $34,999, the proportion of good self-rated health was lower at 59.9% (n=163), and the group was not split further by other predictors (terminal node 3). In the subgroup with total family income greater than $25,000 - $34,999 (node 4), the proportion of good self-rated health was higher at 85.7% (n=776).

In the second pathway from node 2, in the presence of a serious on-going health problem, 75.8% (n=449) of the group were classed in the poor self-rated health category, a higher proportion than in the total population and in node 5. In the next split, whether respondents felt that they could trust their neighbours also had an impact on proportion of poor self-rated health in the subgroup. The subgroup (node 12) who reported that people in their neighbourhood could not be trusted, had a higher fraction of poor self-rated health at 84.7% (n=238) compared with those who felt they could, proportion of poor self-rated health, 67.8% (n=211).

Figure 2: Classification Tree Analysis (see Table 4 for further detail)

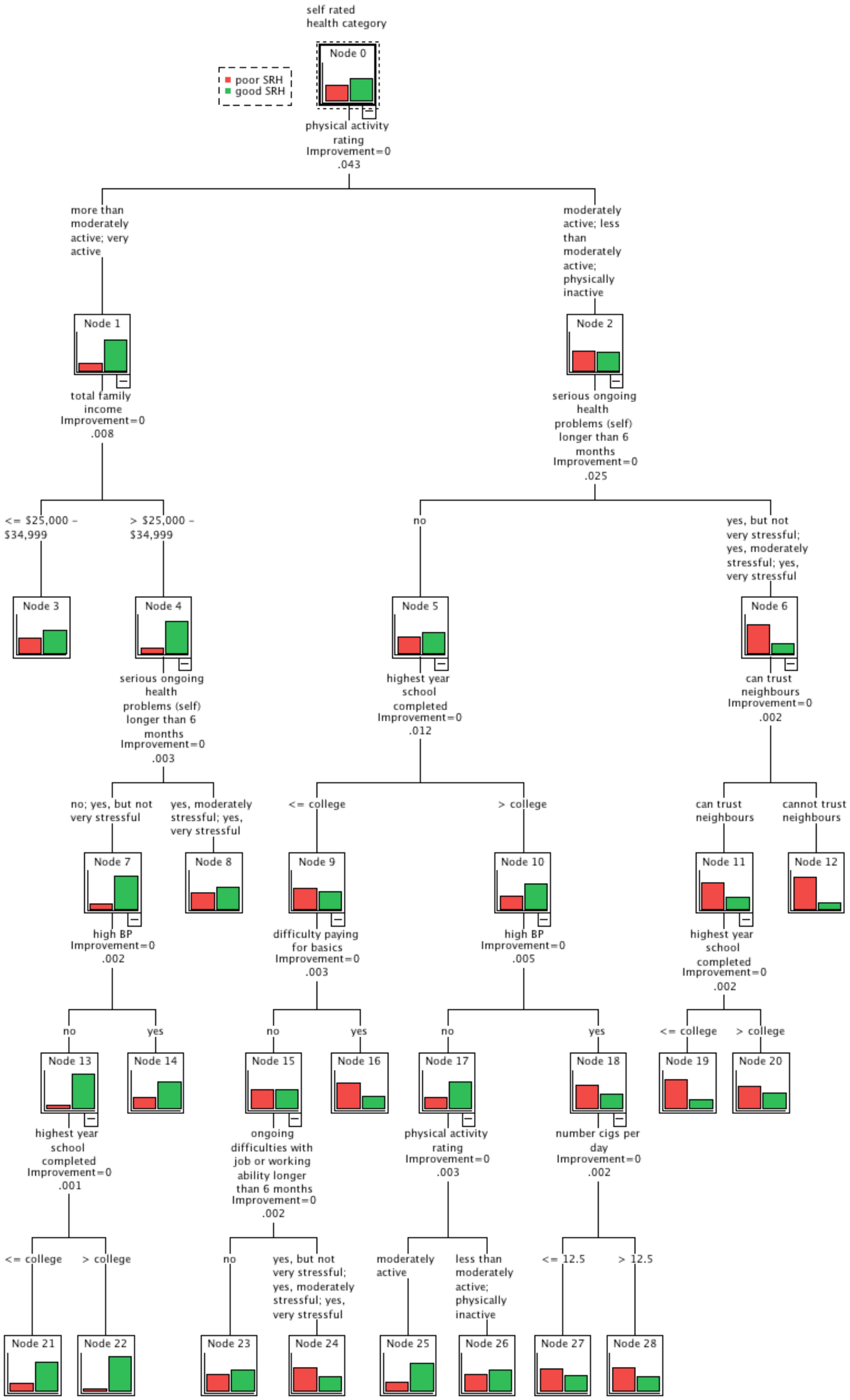


Table 4: Classification tree (Figure 1) nodes in table format

	0 poor SRH		1 good SRH		Total	
Node	**N**	**Percent**	**N**	**Percent**	**N**	**Percent**
0	1514	41.5%	2135	58.5%	3649	100.0%
1	238	20.2%	939	79.8%	1177	32.3%
2	1276	51.6%	1196	48.4%	2472	67.7%
3	109	40.1%	163	59.9%	272	7.5%
4	129	14.3%	776	85.7%	905	24.8%
5	827	44.0%	1053	56.0%	1880	51.5%
6	449	75.8%	143	24.2%	592	16.2%
7	102	12.1%	740	87.9%	842	23.1%
8	27	42.9%	36	57.1%	63	1.7%
9	497	55.4%	400	44.6%	897	24.6%
10	330	33.6%	653	66.4%	983	26.9%
11	211	67.8%	100	32.2%	311	8.5%
12	238	84.7%	43	15.3%	281	7.7%
13	75	10.0%	674	90.0%	749	20.5%
14	27	29.0%	66	71.0%	93	2.5%
15	325	50.3%	321	49.7%	646	17.7%
16	172	68.5%	79	31.5%	251	6.9%
17	263	30.2%	609	69.8%	872	23.9%
18	67	60.4%	44	39.6%	111	3.0%
19	122	77.2%	36	22.8%	158	4.3%
20	89	58.2%	64	41.8%	153	4.2%
21	32	22.1%	113	77.9%	145	4.0%
22	43	7.1%	561	92.9%	604	16.6%
23	205	45.1%	250	54.9%	455	12.5%
24	120	62.8%	71	37.2%	191	5.2%
25	151	24.6%	462	75.4%	613	16.8%
26	112	43.2%	147	56.8%	259	7.1%
27	31	58.5%	22	41.5%	53	1.5%
28	36	62.1%	22	37.9%	58	1.6%

Figure 3: Normalized importance (see table 5 for further detail)

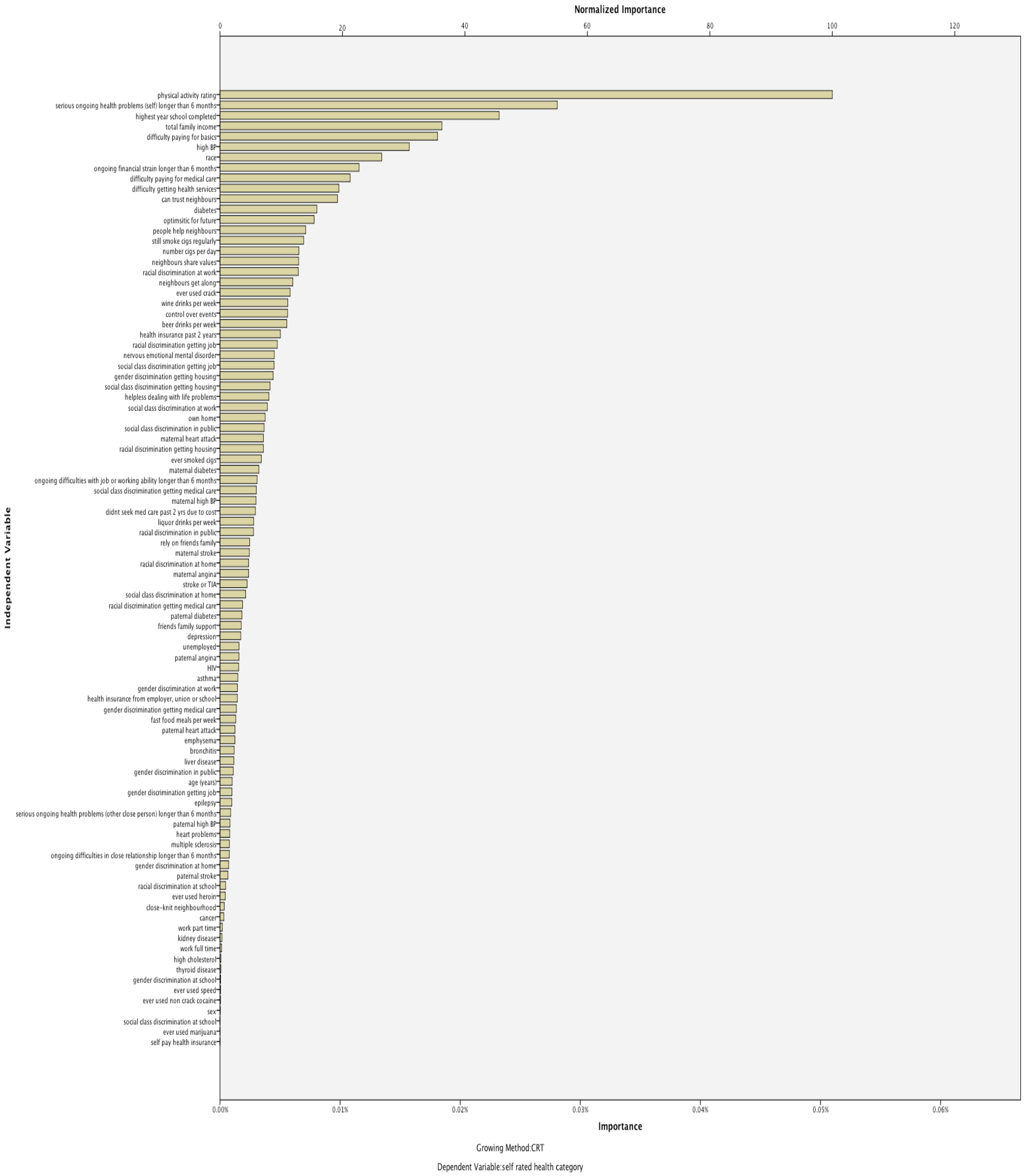


Table 5: Normalized importance (values >10%)

Independent Variable Importance		
Independent Variable	**Importance**	**Normalized Importance**
Physical activity rating	.051	100.0%
Serious personal ongoing health problems longer than 6 months	.028	55.1%
Highest year school completed	.023	45.6%
Total family income	.018	36.2%
Difficulty paying for basics	.018	35.5%
High blood pressure	.016	30.9%
Race / ethnicity	.013	26.4%
Ongoing financial strain longer than 6 months	.012	22.7%
Difficulty paying for medical care	.011	21.3%
Difficulty getting health services	.010	19.4%
Can trust neighbors	.010	19.2%
Diabetes	.008	15.8%
Optimistic for future	.008	15.4%
People help neighbors	.007	14.0%
Still smoke cigs regularly	.007	13.7%
Number cigs per day	.007	12.9%
Neighbors share values	.007	12.8%
Racial discrimination at work	.007	12.8%
Neighbors get along	.006	11.9%
Ever used crack	.006	11.5%
Wine drinks per week	.006	11.1%
Control over events	.006	11.0%
Beer drinks per week	.006	10.9%
Growing Method: CRT Dependent Variable: self rated health category		

I separately compared the characteristics of two subgroups identified in the classification tree model with the rest of the study population based on psychosocial and socioeconomic variables that did not appear as splitting variables in the final CART model:

(1) Node 22 which had the highest proportion of good self-rated health (92.9%, n=561) – characterized by higher physical activity rating; higher income bracket; no serious on-going health burden to self; or if present, not very stressful; no history of hypertension; highest year of school completed is graduate level.

(2) Node 12 which had the highest proportion of poor self-rated health (84.7%, n=238) – characterized by lower physical activity rating; serious on-going personal health problem; perception that people in neighbourhood cannot be trusted.

Comparing proportions with z tests, membership in node 22 was also associated with being white, owning a home, being employed, feeling that friends and family care and can be relied upon, perception of neighbours helping each other, getting along, sharing values, and the neighbourhood being close knit. Node 22 had a significantly higher proportion of respondents who felt that they had control over life events, were not helpless dealing with life problems, and were optimistic for the future (Table 6).

Membership in node 12 was associated with being black, not owning a home, being unemployed, not feeling that family and friends care, or can be relied upon for support, perception that neighbours don't help each other, the neighbourhood is not close knit, neighbours don't get along, don't share values. Respondents felt they had no control over life events, felt helpless dealing with life problems, and were not optimistic for the future (Table 7).

Table 6: Characteristics of node 22 subgroup (highest proportion good self-rated health) compared with remainder of study population [a]

		Membership in node 22	
		No (Remainder of study population)	**Yes (Node 22 subgroup)**
		(A)	(B)
Race/ethnicity	3 Hispanic		
	4 Black, not Hispanic	**B**	
	5 White, not Hispanic		**A**
Own home	0 no	**B**	
	1 yes		**A**

Unemployed	0 no		**A**
	1 yes	**B**	
Friends family support	0 don't feel friends/family care	**B**	
	1 feel friends/family care		**A**
Rely on friends family	0 cant rely on friends and family	**B**	
	1 can rely on friends/family		**A**
People help neighbors	0 neighbors don't help each other	**B**	
	1 neighbors help each other		**A**
Close-knit neighborhood	0 neighborhood not close knit	**B**	
	1 close knit neighborhood		**A**
Neighbors get along	0 neighbors don't get along	**B**	
	1 neighbors get along		**A**
Neighbors share values	0 neighbors don't share values	**B**	
	1 neighbors share values		**A**
Control over events	0 no control over events	**B**	
	1 control over events		**A**
Helpless dealing with life problems	0 not helpless dealing with life problems		**A**
	1 helpless dealing with life problems	**B**	
Optimistic for future	0 not optimistic about future	**B**	
	1 optimistic about future		**A**

Results are based on two-sided tests with significance level 0.05. For each significant pair, the key of the category with the smaller column proportion appears under the category with the larger column proportion.

a. Tests are adjusted for all pairwise comparisons within a row of each innermost sub table using the Bonferroni correction.

Table 7: Characteristics of node 12 subgroup (highest proportion poor self-rated health) compared with remainder of study population[a]

		Membership in node 12	
		No (Remainder of study population)	Yes (Node 12 subgroup)
		(A)	(B)
Race/ethnicity	3 Hispanic		
	4 Black, not Hispanic		**A**
	5 White, not Hispanic	**B**	
Own home	0 no		**A**
	1 yes	**B**	
Unemployed	0 no	**B**	
	1 yes		**A**
Friends family support	0 don't feel friends/family care		**A**
	1 feel friends/family care	**B**	
Rely on friends family	0 cant rely on friends and family		**A**
	1 can rely on friends/family	**B**	
People help neighbors	0 neighbors don't help each other		**A**
	1 neighbors help each other	**B**	
Close-knit neighborhood	0 neighborhood not close knit		**A**
	1 close knit neighborhood	**B**	
Neighbors get along	0 neighbors don't get along		**A**
	1 neighbors get along	**B**	
Neighbors share values	0 neighbors don't share values		**A**
	1 neighbors share values	**B**	
Control over events	0 no control over events		**A**
	1 control over events	**B**	
Helpless dealing with life problems	0 not helpless dealing with life problems	**B**	
	1 helpless dealing with life problems		**A**
Optimistic for future	0 not optimistic about future		**A**
	1 optimistic about future	**B**	

Results are based on two-sided tests with significance level 0.05. For each significant pair, the key of the category with the smaller column proportion appears under the category with the larger column proportion.

a. Tests are adjusted for all pairwise comparisons within a row of each innermost subtable using the Bonferroni correction.

Discussion

This study found that multi-domain health determinants - lifestyle (physical activity), social and community influences (neighbourhood cohesion), living and working conditions (education, income, and adequacy of resources) - are associated with self-rated health in the study sample. I also found that combinations of these factors differ by population subgroup. Physical activity rating emerged as the most important variable in classification with a higher level of physical activity associated with good self-rated health. Of interest is that in those with higher physical activity, there appears to be interaction of lifestyle and medical factors with socioeconomic factors, income and education. In the classification tree, with similar levels of physical activity or chronic burden related to a serious health problem, subgroups with higher income or education were also those with higher proportions of good self-rated health. Furthermore, the subgroup with even moderate activity (node 25, 75.4%, n=462) had a higher proportion of good self-rated health compared with less than moderate or no physical activity (node 26, 56.8%, n=147). The overall misclassification rate for the single tree of 31% is similar to that found in previous studies using classification tree analysis (71, 76).

Node 4 represented a subgroup with higher physical activity rating and higher income. When this was split on history of a serious personal on-going health problem, lack of associated stress was associated with a higher proportion of good self-rated health, (node 7, 87.9%, n=740) than when some degree of stress was present (node 8, 57.1%, n=36). Information on the nature of this stress was unavailable in our dataset but appears relevant to explore further as an element potentially amenable to intervention. Psychosocial factors, such as social support, positive social relationships, an optimistic outlook on life, perceived control over life outcomes, and a sense of purpose and direction in life, have been identified as health protective factors; psychosocial factors also influence positive health behaviors (77). Previous studies have shown that in populations with chronic illness and disability, psychological resources, particularly a sense of high mastery and self-esteem were shown to associate with better self-rated health (44, 54).

In the comparison of nodes 22 and 12, race/ethnicity, home ownership and employment status emerged as differences in profiles between the two groups. The subgroup with the highest proportion of good self-rated health (node 22, 92.9%) had a statistically significant higher proportion of respondents who were white, owned their home and were employed. This subgroup was also characterised by higher physical activity rating, income greater than $25,000, no serious on-going health problem, or one not associated with stress, and highest year of education as beyond college (i.e. Graduate School). The subgroup with the highest proportion of poor self-rated health (node 12, 84.7%) had a statistically significant higher proportion of respondents who were black, did not own their home and were unemployed. This subgroup was also characterised by lower physical activity rating, severe on-going health problems and lack of trust in their neighbors.

I dichotomized the outcome variable, self-rated health, into ‘good’ and ‘poor’ categories. This meant that original responses of good were grouped with poor or fair. I chose to separate the very good and excellent responses into one group, as they were more definite positive statements of better health. Respondents may have regarded a response of good, being the center of a 5-point scale, as a neutral or ‘average’ value. This also resulted in more equal group sizes. Fair and good self-ratings of health have been associated with higher mortality; risk is not associated solely with the poor group but there is a gradient (4). A limitation of this study is the use of cross-sectional data so that I report only associations rather than determinants of self-rated health. Though the predictor variables were selected from an existing strong dataset to represent multiple layers of influences on health, a few may not optimally represent the characteristic of interest. For example, diet is included only by way of fast food intake. Nevertheless, the CARDIA dataset has allowed me to consider several multi-domain variables relating to the ecological model of health (58, 60), and thus conduct a valuable exploratory study.

In the bivariate analysis, using the chi-square test, a few predictor variables were not significantly associated with self-rated health at the p<0.05 level, or the test was invalid due to small cell numbers. However, I chose to include these in the classification model. My interest was in exploring the way in which multilevel factors could potentially interact in different population subgroups leading to different health outcomes, even if alone they would not necessarily be independent predictors. Effect modification would be relevant in the context of public health action as there may be additional risk or protection with regard to health outcomes depending on how variables interact. In addition, for some factors the test results may have been invalid simply due to small cell numbers.

Classification tree modelling has a number of advantages. It is a non-parametric technique, which is valuable when dealing with such a broad set of predictor variables, and a large sample size. Missing variables can be adjusted for within the analysis, and since the structure of the classification is apparent, the results in terms of population subsets are relatively easy to interpret. Responses that were classed as ‘don’t know’ in the original CARDIA data collection were labeled in this study as ‘missing’. It is possible that this could introduce some bias if people who responded to certain questions with ‘don’t know’ were more likely to have a particular self-rated health status. A further limitation is that this is a non-probabilistic method. There is no rigorous theory for providing inference on the structure of the tree; my interpretation of the tree is entirely exploratory and the results are interpreted in that light. As noted above, single tree structures are unstable, and small variations in the data can produce very large variations in the structure of the tree, even though the prediction accuracy might not vary much at all. To address this, I extend the analysis in chapter 4 to include random forests.

Earlier studies have used parametric multivariate regression to identify correlates of self-rated health, and independent determinants, and establish which domain of health determinants is most important for predicting self-rated health. As the setting and study population have differed by age, geography and occupation, therefore unsurprisingly, the determinants of health have differed across the population groups. In other populations, researchers have concluded that self-rated health is largely determined by physical and mental health as opposed to non-health circumstances too. This study also suggests earlier findings that have demonstrated an association between lower self-reported health, and lower income, less education and being black. However, rather than seeking to reach a consensus on the determinants of self-rated health across entire populations, this study adds that there is value in determining the different influences among population subgroups, and acknowledging and addressing these when designing interventions to improve health outcomes.

Knowledge of the relationship between single predictors and outcomes is clearly essential. However, the strength of conceptualising the potential determinants of self-rated health or other health outcomes using this type of analytic approach is that classification trees build upon individual factor-outcome relationships, and add detail on interactions and multilevel contributions. This may be a better reflection of how multiple influences on health interact in practice, where relationships between health determinants are not necessarily simple or represented by linear models. Single elements of the broad range of health determinants reflect only some aspect of health but without consideration of cofactors, are incomplete predictors of overall health status (78). Studies have shown that biomedical risk factors account for only a fraction of ill health. The results of the Whitehall study, for example, highlighted that traditional 'medical' risk factors (such as cholesterol level, smoking, systolic blood pressure, and diabetes) explained only a third of the observed socioeconomic gradient in health (79). Thus upstream factors, or at least the impact of the 'causes of the causes' (73) have to be considered in tandem with other foci of health promotion programs, even if they may be more difficult to remedy.

The classification tree model also divides the population into population subgroups that are relatively homogenous in terms of outcome. This offers greater insight into needs of distinct subgroups rather than modelling an average relationship across the population (69, 70). This is particularly of relevance when the goal is to inform interventions. For example, BeLue et al. (71) demonstrated this in a study applying classification tree analysis to a population of US adolescents aged 12-17 years. Youth obesity is caused by multiple factors. The findings that obesity-related risk and protective factors emerged differently among distinct socio-demographic subgroups of the adolescent study population indicate that interventions that work for one subgroup may not be optimal for another. Intervention would need to be tailored to meet the needs of different subgroups. Friel et al. (76) applied multivariate classification tree analysis to study the profiles (by way of socio-demographic, socioeconomic factors and health-related lifestyle behaviors) of adults who comply with fruit and vegetable dietary recommendations. This was based on the rationale that food choice is complex and influenced by economic social and environmental context. Social characteristics were not the same for males and females in terms of relative importance in predicting fruit and vegetable consumption. This has implications for setting dietary strategies and policy.

I am not aware of previous work that has applied a classification model to self-rated health in order to consider population subgroups and interacting multilevel factors in this way. The results of this initial exploratory study raise issues that are relevant to public health practice. Lifestyle choices are rooted in our socioeconomic context. Thus, even if we are interested in intervention only on behavioral factors, such as physical exercise, smoking or alcohol consumption, there is value in understanding the concurrent upstream setting of living and working conditions, and how other factors might be associated with, and influence or restrict, these choices. Contextualizing risk factors is important if action is to be taken on the *"fundamental factors that put people at risk of risks"* (80)(p.85). Second, multivariate logistic regression models demonstrate the average relationship between predictor and outcome over the population. Classification tree analysis can serve as an initial tool to suggest population subgroups that might have homogenous risks of disease or an outcome. Intervention strategies can be developed and tailored to address these unique characteristics and needs within subgroups. A better understanding of need-based on multilevel health determinants would allow typically limited resources to be targeted. Third, a central message of the Marmot Report, *Fair Society, Healthy Lives* (81), was the need for *proportionate universalism*. In order to reduce the steepness of the gradient in health, we need to engage in a population approach but target those most in need. Health is shaped in a complex manner. The impact upon health of a single exposure or risk factor may vary across the population depending on position on the gradient, and therefore through coexisting factors and specific socioeconomic context. This concept is analogous to the framework used by infectious disease epidemiologists, in which there is consideration of the agent, the host, and susceptibility, or the environment. A better understanding of how these multi-domain factors come together in population subgroups is essential to designing public health interventions that are most suitable for the individuals they target.

Chapter 3 Characteristics associated with self-rated health in the CARDIA Study: a study of population subgroups along a socioeconomic gradient

Introduction

To reduce health inequalities and the steepness of the socioeconomic gradient in health, the Marmot report recommends proportionate universalism (81). This requires a population approach to public health action that also includes targeting those most in need. A key first step in this process is to understand the influences on health outcomes amongst different population subgroups. This is essential to designing public health interventions that are most appropriate for the individuals they target.

In this study, I investigate the distribution of known health determinants across socioeconomic groups, in a population drawn from the CARDIA study (Coronary Artery Risk Development in Young Adults), and explore the factors associated with self-rated health within different socioeconomic groups.

Self-rated health is a key parameter gathered in many health surveys (12-14, 16). It is known to be an independent predictor of a range of future health outcomes, including functional ability (18, 19), morbidity (20) and health care utilization (21). Several other studies have examined the relationship of self-rated health with mortality, and found that the association persists independently of numerous objective indicators (1). Therefore, in a public health context, for high proportions of populations to report good subjective health is in itself an important end point.

Ecological models reflect the multiple wider influences on health and the interactions between them that determine health; in addition to addressing biological risk factors, public health action should regard the impact on health of these upstream environmental, social and behavioral factors, and develop appropriate policy (58). Given the complexity of the socioeconomic status and health gradient, Adler et al. (82) suggested that a number of conceptual and methodological issues have constrained earlier research. It is difficult to unravel in full the mechanisms that might contribute to the gradient using only analysis techniques such as parametric multivariate regression. These methods are limited in their ability to capture a large number of multi-domain interrelated variables (78). In this paper, I aim to capture the joint impact and relative importance of multi-domain health determinants in relation to the chosen outcome measure, self-rated health, by applying classification tree analysis. Classification tree analysis is a form of recursive partitioning. This method is useful in uncovering the predictive structure of a problem, that is, to try and gain an understanding of the conditions in terms of explanatory variables that determine which class (of outcome) an individual is in (65). Classification tree methods may capture information such as particular interactions not reflected in traditional parametric analysis (68). The method is particularly useful when the aim is to identify differences, and appreciate specific features, of subgroups; seeing the relative importance of different factors may also shed light on where action could be prioritized (69).

Lower social position is related to greater exposure to hazardous behavioral, psychosocial and material risk factors (83, 84). Resources available to those in higher socioeconomic positions afford a better chance of protection from disease risk and the consequences of illness (80). This supports the notion that public health intervention across the gradient needs to be proportionally weighted depending on need. However, it is unclear whether the relative importance of factors, in terms of impact on health outcomes, remains the same between different socioeconomic groups. Dahlgren and Whitehead recommend that reduction in health inequities will be aided by an understanding of which risk factors are important for different socioeconomic groups, and whether these differ from those for the overall population (83). I investigate this further in this study, and hypothesize that the combinations of factors associated with health outcomes differ between population subgroups, based on lifestyle choices, psychosocial factors, and living and working conditions. This would imply that an intervention that is appropriate for one subgroup would not be optimal for another due to differing priorities and need (71). If this is the case, such knowledge has value in the context of public health action.

In this study, I divide the data sample from the CARDIA study into groups based on total family income to represent socioeconomic status. I then investigate the health-related factors associated with self-rated health within the different population subgroups. Using classification tree analysis, my goal is to understand whether these differ across the socioeconomic gradient.

Methods

I utilized cross-sectional data collected as part of the CARDIA study, a United States cohort study started in 1985 to investigate the development of coronary artery disease risk factors in a young adult population. For this study, data were taken from the year 15 examination of the cohort, conducted in 2000-2001, through interviewer and self-administered questionnaires (with the exception of race/ethnicity information taken from the 1985-1986 data collection, and family history information taken from the 1995 data collection). In year-15, data were collected on both self-rated health and a broad range of health-related factors of interest in this study. Though similar data were collected in later years, I was interested in an initial study of a relatively young adult population. From an original total of 5115 participants, 3672 were followed up in year 15. From the year 15 group, all participants who had a response for self-rated health, and were coded as male or female, were included in the final study sample of 3649 participants.

For each individual in the study sample, data were available for self-rated health, and a wide range of multi-domain health determinants (Table 1). I selected the variables based on the Dahlgren and Whitehead model on the determinants of health (60) to include age, sex and hereditary factors; individual lifestyle factors; social and community influences, and living and working conditions.

In the CARDIA study, self-rated health was assessed on a five-point scale, by the question, *"In general would you say your health is excellent, very good, good, fair or poor?"* Responses were categorized by grouping together excellent or very good as 'good' self-rated health, and responses of good, fair or poor, as 'poor' self-rated health. I chose to separate the very good and excellent responses into one group, as they were more definite positive statements of better health. Respondents may have regarded a response of good, being the center of a 5-point scale, as a neutral or 'average' value. This also resulted in more equal group sizes. Indicator variables were created for a number of variables as outlined previously in chapter 2 (page 18, Table 1).

Data Analysis

Data analysis was carried out using IBM SPSS Statistics v19.0.0. In the CARDIA study, respondents were asked to report their total family income over the previous 12 months. I split the CARDIA study sample into 5 groups based on income category. I used the Mantel-Haenszel chi-square test for trend to assess the relationship between the multi-domain predictor variables and ordinal income categories.

For each income category subgroup, I ran a classification tree analysis using the predictor variables (see chapter 2, page 18, Table 1 for more detail), excluding total family income. Classification trees were used to segment the population, and identify the variables selected by the model as associated with self-rated health. With the smaller samples for each tree analysis, I proportionately reduced the tree growing criteria to a minimum parent node size of 20 and child node size of 10 (compared with the full sample classification tree analysis in chapter 2). Again, as emphasized in chapter 2, one has to consider this as exploratory data analysis as such an estimate of variable importance is not a robust measure (since small changes in the data can result in large changes in the structure of the tree, as well as variables selected). However, it serves as an interesting tool for highlighting potential risk groups in future studies.

Table 1: Multi-domain health determinants drawn from the CARDIA study

Level 1: Age, sex & hereditary factors

- Age
- Sex
- Race / Ethnicity
- Family history of medical conditions

Level 2: Individual lifestyle factors

- Medical history
- Diet
- Physical activity
- Smoking
- Alcohol
- Tobacco
- Illicit drug use

Level 3: Social & community influences

- Social support / network
  - Feeling that friends and family care
  - Can rely on friends and family
- Sense of close-knit neighborhood, neighborhood cohesion
  - Neighbors help each other
  - Live in close-knit neighborhood
  - Neighbors can be trusted
  - Neighbors generally get along with each other
  - Neighbors share the same values

Level 4: Living & working conditions

- Education
- Income
- Housing - rent or own house
- Employment - working versus unemployed
- Control and adequacy of resources
  - Difficulty paying for basics or medical care
- Medical insurance
- Access to health services
- Experience of discrimination due to gender, race/ethnicity or colour, socioeconomic position, or social class in 7 settings (at school, getting a job, getting housing, at work, at home, getting medical care, on the street, or in a public setting)

Results

The income distribution for the CARDIA study sample is shown in Table 2. Data on total family income was available for 3605 respondents. As noted above, the study sample was divided into 5 subsets based on income. The categories formed, and the size of the study sample population in each, are shown in Table 3. The proportions of the study sample within each category were: 16% under $25,000; 25.3% $25,000-$50,000; 21.9% $50,000 -$75,000; 14.6% $75,000 - $100,000; and 22.1% $100,000 and over.

Table 4 shows the distribution of the predictor variables by income category. Using the Mantel-Haenszel chi-square test for trend, there was a significant association (p<0.05) between income category and self-rated health status. There was also a social gradient for a number of predictor variables related to lifestyle, social influences, and living and working conditions. For the indicators of chronic burden, there were significant associations between 3 of the indicators and income category using the chi-square test for independence (p<0.05). The specific indicators for which there were associations were experience of strain for longer than 6 months due to serious on-going personal health problem; on-going financial strain; or on-going difficulties in a relationship with someone close.

Table 2: Income distribution in the CARDIA study sample

		Value	Count	Percent
Values	1	**< $5,000**	**87**	**2.4%**
	2	**$5,000 - $11,999**	**140**	**3.8%**
	3	**$12,000 - $15,999**	**112**	**3.1%**
	4	**$16,000 - $24,999**	**239**	**6.5%**
	5	**$25,000 - $34,999**	**342**	**9.4%**
	6	**$35,000 - $49,999**	**569**	**15.6%**
	7	**$50,000 - $74,999**	**791**	**21.7%**
	10	**$75,000 - $99,999**	**527**	**14.4%**
	11	**≥$100,000 and greater**	**798**	**21.9%**
Total			**3605**	**1.2%**

Table 3: Income categories based on total family income

		Frequency	Percent	Valid Percent	Cumulative Percent
Income category	1 under $25,000	578	15.8	16.0	16.0
	2 $25,000-$50,000	911	25.0	25.3	41.3
	3 $50,000-$75,000	791	21.7	21.9	63.2
	4 $75,000-$100,000	527	14.4	14.6	77.9
	5 $100,000 plus	798	21.9	22.1	100.0
	Total	3605	98.8	100.0	
Missing		4	1.2		
Total		3649	100.0		

Table 4: Distribution of selected predictor variables by income category and Mantel Haenszel chi-square test for trend

Variable	**Income categories N (% within income category)**						
	<$25,000	**$25,000-$50,000**	**$50,000-$75,000**	**$75,000-$100,000**	**$100,000 plus**	**Total**	**p value for chi - square test for trend: linear-by-linear association**
Self-rated health							
Poor	361 (62.5)	451 (49.5)	305 (38.6)	193 (36.6)	183 (22.9)	1493	<0.05
Good	217 (37.5)	460 (50.5)	486 (61.4)	334 (63.4)	615 (77.1)	2112	
Sex, race/ethnicity and hereditary factors							
Male	215 (37.2)	392 (43.0)	327 (41.3)	254 (48.0)	404 (50.6)	1591	<0.05
Female	363(62.8)	519 (57.0)	464 (58.7)	274 (52.0)	394 (49.4)	2014	
Hispanic	4 (0.7)	3 (0.3)	2 (0.3)	0 (0)	2 (0.3)	11	<0.05[a]
Black	431 (74.6)	529 (58.1)	344 (43.5)	210 39.8)	167 (20.9)	1681	
White	143 (24.7)	379 (41.6)	445 (56.3)	317 (60.2)	629 (72.8)	1913	
Maternal diabetes	97 (20.6)	120 (15.2)	109 (15.4)	53 (11.4)	54 (7.6)	433	<0.05
Maternal high blood pressure (BP)	246 (55.0)	348 (47.5)	294 (45.2)	190 (45)	226 (34.1)	1304	<0.05
Maternal stroke	52 (10.8)	49 (6.1)	54 (7.6)	21 (4.4)	27 (3.8)	203	<0.05
Maternal angina	45 (10.0)	72 (9.5)	52 (7.7)	47 (10.5)	36 (5.2)	252	<0.05

Maternal heart attack	59 (12.4)	64 (8.0)	40 (5.6)	29 (6.2)	24 (3.4)	216	<0.05
Paternal diabetes	58 (15.6)	108 (16.1)	78 (12.5)	57 (13.1)	83 (12.3)	384	<0.05
Paternal high BP	163 (48.4)	265 (45.4)	250 (43.8)	186 (48.3)	272 (44.7)	1136	>0.05
Paternal stroke	46 (12.2)	80 (11.6)	64 (9.9)	44 (10.0)	43 (6.3)	277	<0.05
Paternal angina	58 (16.7)	93 (15.3)	77 (13.3)	65 (16.3)	119 (18.8)	412	>0.05
Paternal heart attack	78 (20.9)	149 (21.5)	108 (16.6)	102 (23.1)	143 (21.2)	580	>0.05
Lifestyle factors and medical history							
High BP	153 (26.7)	162 (18)	113 (14.4)	90 (17.2)	78 (9.8)	596	<0.05
High Cholesterol	80 (14.4)	151 (17.1)	142 (18.3)	93 (18.1)	163 (20.7)	629	<0.05
Heart disease	70 (12.3)	97 (10.8)	90 (11.5)	52 (10.1)	94 (11.9)	403	>0.05
Diabetes	42 (7.4)	61 (6.8)	39 (4.9)	31 (5.9)	33 (4.1)	206	>0.05
Transient Ischemic Attack / stroke	10 (1.7)	4 (0.4)	10 (1.3)	4(0.8)	1 (0.1)	29	<0.05
Asthma	84 (14.7)	114 (12.6)	94 (11.9)	61 (11.6)	86 (10.8)	439	<0.05
Chronic bronchitis	36 (6.3)	59 (6.5)	31 (3.9)	16 (3.0)	17 (2.1)	159	<0.05
Liver Disease	27 (4.7)	18 (2.0)	16 (2.0)	10 (1.9)	17 (2.1)	88	<0.05
Epilepsy	22 (3.8)	9 (1.0)	8 (1.0)	7 (1.3)	6 (0.8)	52	<0.05
Nervous emotional or mental disorder	81 (14.1)	66 (7.3)	52 (6.6)	34 (6.5)	33 (4.1)	266	<0.05
Depression	121 (21.1)	152 (16.9)	122 (15.5)	74 (14.1)	97 (12.2)	566	<0.05
Physical activity 1 inactive	61 (10.6)	69 (7.6)	46 (5.8)	28 (5.3)	28 (3.5)	232	<0.05*
2	89 (15.5)	148 (16.3)	163 (20.7)	99 (18.8)	113 (14.2)	612	
3	262 (45.5)	418 (45.9)	364 (46.1)	220 (41.7)	331 (41.5)	1595	
4	78 (13.5)	140 (15.4)	115 (14.6)	107 (20.3)	187 (23.5)	627	
5 very active	86 (14.9%)	135 (14.8)	101 (12.8)	73 (13.9)	138 (17.3)	533	
Current smoker	248 (76.1)	244 (63.0)	136 (43.7)	70 (39.3)	87(34.0)	785	<0.05
Social and community influences							
Friends and family care	533 (92.2)	883 (96.9)	786 (99.4)	522 (99.2)	791 (99.1)	3515	<0.05
Can rely on friends and family	444 (76.8)	768 (84.3)	706 (89.3)	485 (92.2)	740 (92.7)	3143	<0.05
People help neighbors	352 (60.9)	609 (66.8)	605 (76.5)	406 (77.0)	658 (82.5)	2630	<0.05
Close knit neighborhood	214 (37.0)	337 (37.0)	324 (41.0)	234 (44.4)	416 (52.1)	1525	<0.05
Can trust neighbors	213 (36.9)	453 (49.7)	477 (60.3)	346 (65.7)	607 (76.2)	2096	<0.05
Neighbors get along	307 (53.1)	602 (66.1)	612 (77.4)	414 (78.6)	670 (84.0)	2605	<0.05

Neighbors share values	191 (33.0)	367 (40.3)	389 (49.2)	298 (56.5)	507 (63.5)	1752	<0.05
Living and working conditions							
Own home	183 (31.7)	510 (56.0)	610 (77.1)	452 (85.8)	730 (91.5)	2485	<0.05
Unemployed	158 (27.4)	63 (6.9)	36 (4.6)	28 (5.3)	38 (4.8)	323	<0.05
Hard to pay for basics	295 (51.1)	245 (27.0)	111 (14.1)	38 (7.2)	21 (2.6)	710	<0.05
Hard to pay for medical care	300 (52.0)	243 (26.8)	93 (11.8)	32 (6.1)	33 (4.1)	701	<0.05
Health insurance coverage past 2 years	371 (64.3)	771 (84.6)	735 (92.9)	507 (96.2)	775 (97.1)	3159	<0.05
Did not seek medical care last 2 years due to cost	134 (23.2)	112 (12.3)	50 (6.3)	22 (4.2)	24 (3.0)	342	<0.05
Control over events	417 (72.1)	746 (81.9)	687 (86.9)	462 (87.8)	736 (92.2)	3048	<0.05
Helpless in dealing with life problems	105 (18.2)	98 (10.8)	58 (7.3)	26 (4.9)	28 (3.5)	315	<0.05
Optimistic for future	359 (62.1)	619 (67.9)	563 (71.2)	403 (76.5)	616 (77.2)	2560	<0.05

[a] 5 cells (33.3%) have expected count less than 5. The minimum expected count is 1.61.

*chi-square test reported

Self-rated health

The proportion of 'good' (excellent to very good self-rated health) increased with higher income category, and the proportion of 'poor' (good, fair or poor) self-rated health decreased.

Sex, race/ethnicity and hereditary factors

There was an increasing proportion of males and whites with higher income. There was an inverse social gradient for the proportion of respondents with family history of maternal diabetes, maternal blood pressure, maternal stroke, maternal heart attack, and paternal diabetes and paternal stroke. Higher income groups had higher age respondents (Table 5).

Lifestyle factors and medical history

Medical conditions showing a significant inverse social gradient included high blood pressure, diabetes, stroke, asthma, chronic bronchitis, liver disease, nervous/ emotional or mental disorder, and depression. There was no significant trend for heart disease in this study sample, and the proportion of respondents with high cholesterol increased with higher income. The chi-square test was significant (p<0.05) for the relationship between physical activity rating and income category. The lowest income category (< $25,000) had a higher proportion of respondents who rated themselves physically inactive compared to the highest income category ($100,000 plus). Wine drinks per week increased with income category, and fast food consumption, beer and hard liquor (bourbon, vodka, etc.) drinks per week decreased with income category (Table 5).

Social and community influences

There were significant social gradients for social and community influences. Larger fractions of respondents in higher income categories reported good social support and sense of neighbourhood cohesion (Table 4).

Living and working conditions

The proportion of home ownership increased with higher income category, and unemployment decreased. There was an inverse social gradient for proportion of respondents reporting difficulty in paying for basics and medical care, and not seeking medical care due to cost. The proportion of respondents with health insurance over the previous two years increased with higher income category. Respondents in higher income categories were also more likely to be optimistic for the future, have a feeling of control over life events and be less likely to report feeling helpless in dealing with life problems (Table 4).

Table 5: Kruskall-Wallis test for continuous variables

	Income categories	N	Mean Rank	p value
Age (years)	1 under $25k	**578**	**1721.02**	**p<0.05**
	2 $25k to $50k	**911**	**1713.82**	
	3 $50k to $75k	**791**	**1781.14**	
	4 $75k to $100k	**527**	**1830.34**	
	5 $100k+	**798**	**1967.8**	
	Total	**3605**		
Fast food meals per week	1 under $25k	**527**	**1652.03**	**p<0.05**
	2 $25k to $50k	**843**	**1744.65**	
	3 $50k to $75k	**712**	**1669.79**	
	4 $75k to $100k	**489**	**1568.58**	
	5 $100k+	**683**	**1462.08**	
	Total	**3254**		
Wine drinks per week	1 under $25k	**392**	**1272.1**	**p<0.05**
	2 $25k to $50k	**675**	**1314.13**	
	3 $50k to $75k	**626**	**1351.49**	
	4 $75k to $100k	**427**	**1425.96**	
	5 $100k+	**731**	**1675.66**	
	Total	**2851**		
Beer drinks per week	1 under $25k	**393**	**1640.29**	**p<0.05**
	2 $25k to $50k	**675**	**1433.56**	
	3 $50k to $75k	**626**	**1337.31**	
	4 $75k to $100k	**427**	**1366.21**	
	5 $100k+	**731**	**1416.64**	
	Total	**2852**		
Liquor drinks per week	1 under $25k	**393**	**1545.7**	**p<0.05**
	2 $25k to $50k	**675**	**1442.39**	
	3 $50k to $75k	**626**	**1355.81**	
	4 $75k to $100k	**427**	**1430.93**	
	5 $100k+	**731**	**1405.69**	
	Total	**2852**		

Classification tree analysis

Classification tree analysis was carried out on each income category and used self-rated health as the outcome measure:

Under $25,000 subgroup:

The proportion of the subset with poor self-rated health was 62.5% (n=361), and with good self-rated health was 37.5% (n=217). The classification tree is shown in Figure 1. There were 3 terminal nodes with predominantly good self-rated health ranging from 55.6% to 87.5%. The overall misclassification rate based on cross-validation was 38%. There were 6 terminal nodes with predominantly poor self-rated health ranging from 58.1% to 95.7%. The primary split of the sample was on presence of a serious ongoing health problem. The other variables selected in the model were family history of paternal angina, history of high blood pressure, ability to rely on friends and family for support, consumption of fast food meals per week, number of cigarettes smoked per day, ongoing difficulties in a close relationship, and ongoing financial strain. Factors associated with a relatively higher proportion of good self-rated health are low number of cigarettes per week, relying on friends and family and not eating fast food meals. In terms of normalized importance (Table 6), the highest ranking variables also included diabetes, nervous, emotional or mental disorder, depression, serious ongoing health problems in another close person, and family history of paternal heart attack.

$25,000 to $50,000 subgroup:

The proportion of the subset with fair to poor self-rated health was 49.5% (n=451), and with good self-rated health was 50.5% (n=460). The classification tree is shown in Figure 2. There were 18 terminal nodes - 12 with predominantly poor self-rated health (highest 96.7%) and 6 with predominantly good self-rated health (highest 92.9%). The overall misclassification rate based on cross-validation was 37%. Factors selected by the model included: physical activity, nervous/emotional or mental disorder, serious ongoing health problem, current smoker, cigarettes per day, neighbors help each other, social class discrimination getting housing, highest year of school completed, ongoing financial strain, number of cigarettes per day, paternal diabetes, ongoing difficulties in a close relationship. Physical activity was the most important variable by normalized importance (Table 6).

Figure 1: Classification tree for income category under $25,000

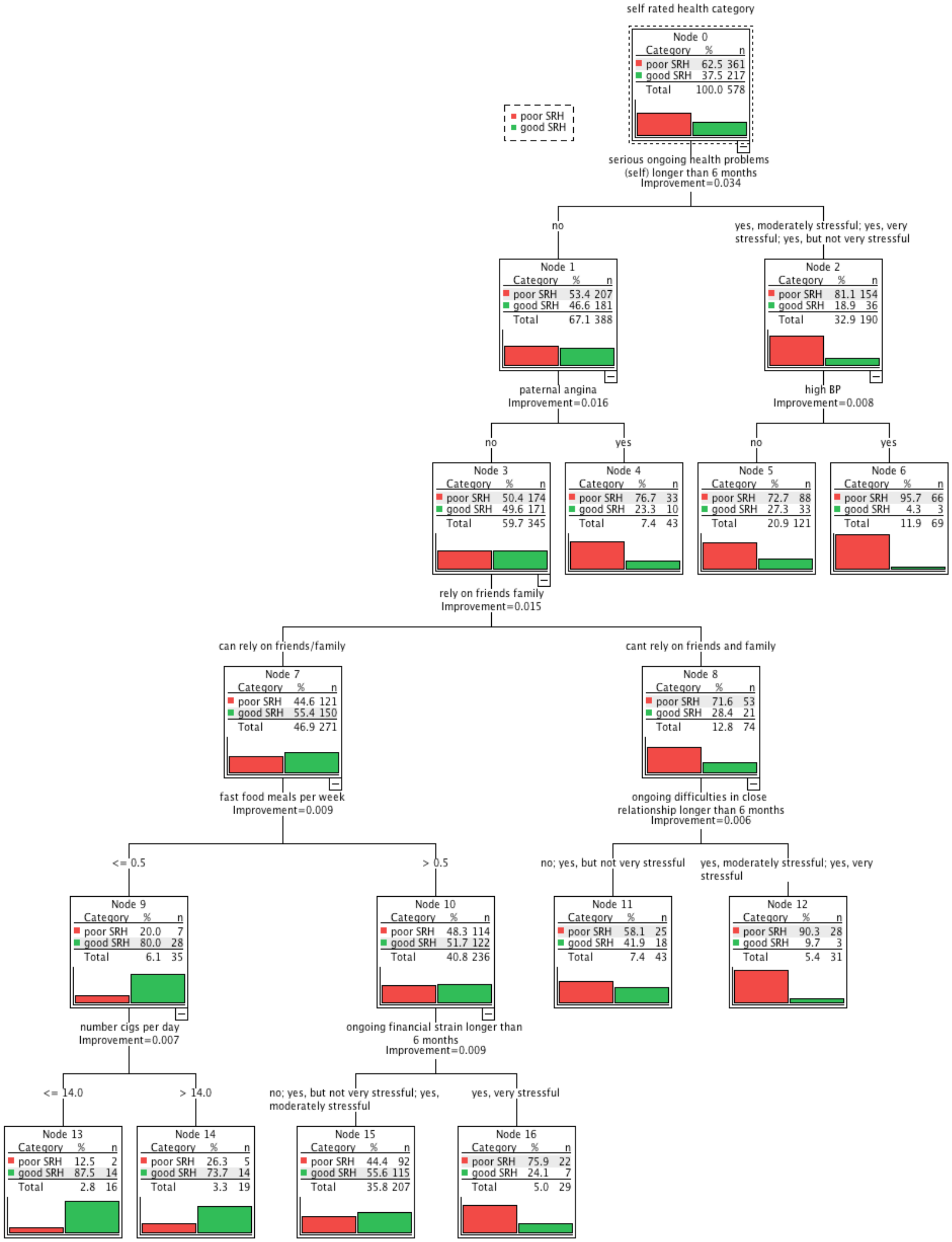
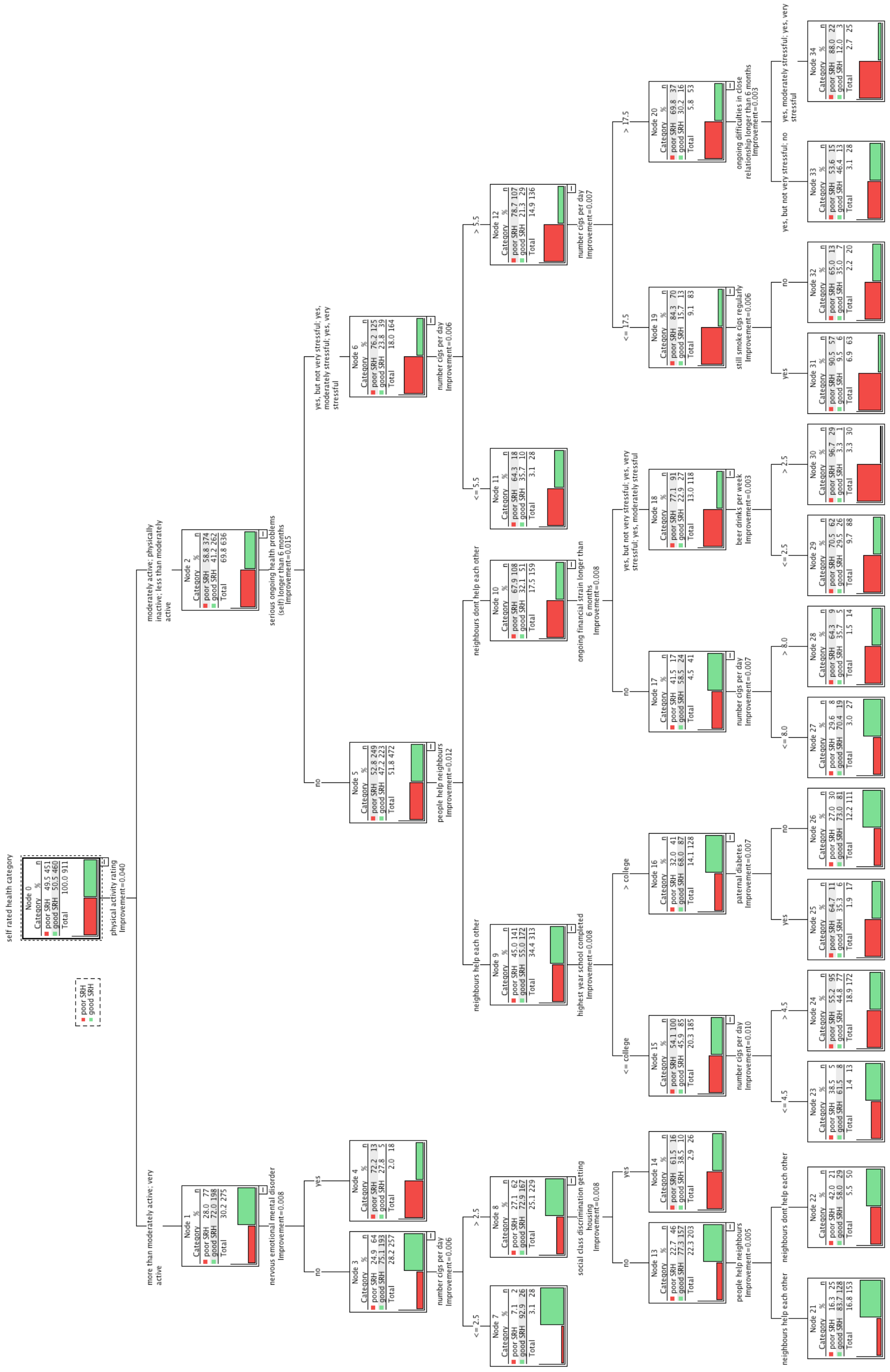


Figure 2: Classification tree for income category $25,000-$50,000



$50,000 to $75,000 subgroup:

The proportion of the subset with good self-rated health was 61.4% (n=486), and with poor self-rated health was 38.6% (n=305). The classification tree is shown in Figure 3. There were 16 terminal nodes. The overall misclassification rate based on cross-validation was 30%. Six terminal nodes had predominantly good self-rated health (highest 95.9%), and 10 had predominantly poor self-rated health (highest 94.4%). Variables selected in the model were physical activity, serious ongoing personal health problem, smoking (number cigarettes per day), optimism for the future, alcohol consumption (wine drinks per week), and highest year school completed. Physical activity was ranked the highest by normalized importance (Table 6).

$75,000 to $100,000 subgroup:

The proportion of the subset with good self-rated health was 63.4% (n=334) and with poor self-rated health was 36.6% (n=193) (Figure 4). There were 8 terminal nodes. The overall misclassification rate based on cross-validation was 34%. Four terminal nodes had predominantly poor self-rated health (highest 85.7%), and four had predominantly good self-rated health (highest 82.1%). The highest split was on number of cigarettes per day. Other variables selected by the model included physical activity rating, serious ongoing personal health problem. Cigarettes per day was the highest ranked variable by normalized importance (Table 6).

$100,000 plus subgroup:

The proportion of the subset with good self-rated health was 77.1% (n=615) and with poor self-rated health was 22.9% (n=183) (Figure 5). There were 5 terminal nodes, 2 with predominantly poor self-rated health (highest 70.4%) and 3 with predominantly good self-rated health (highest 93.7%). The overall misclassification rate based on cross-validation was 23%. Variables selected by the model were physical activity, high blood pressure and serious ongoing health problems. In the normalized importance ranking, the variables appearing in the classification tree were the 3 highest ranked; highest year of education completed was the fourth ranked variable (Table 6).

Figure 3: Classification tree for income category $50,000-$75,000

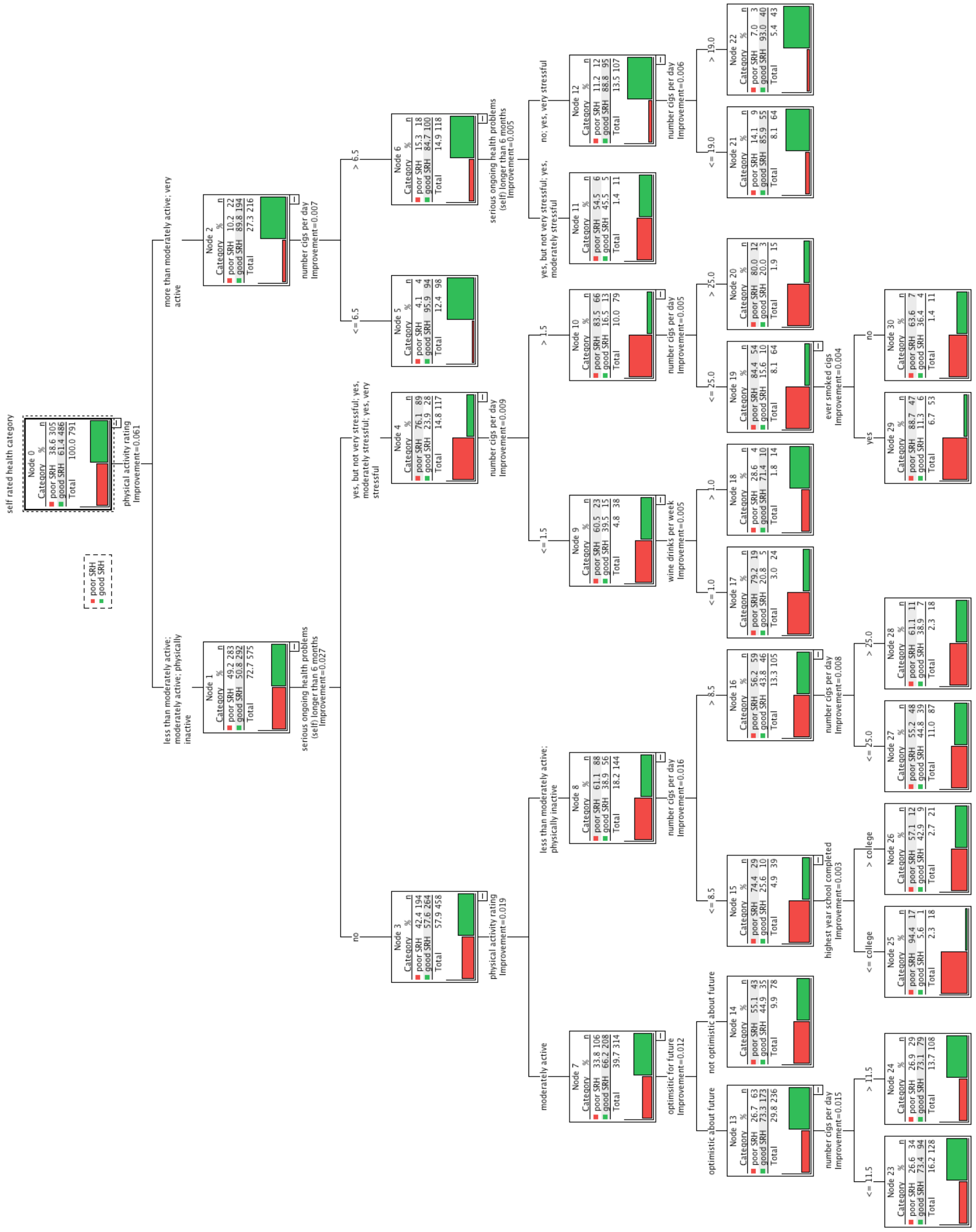


Figure 4: Classification tree for income category $75,000-$100,000

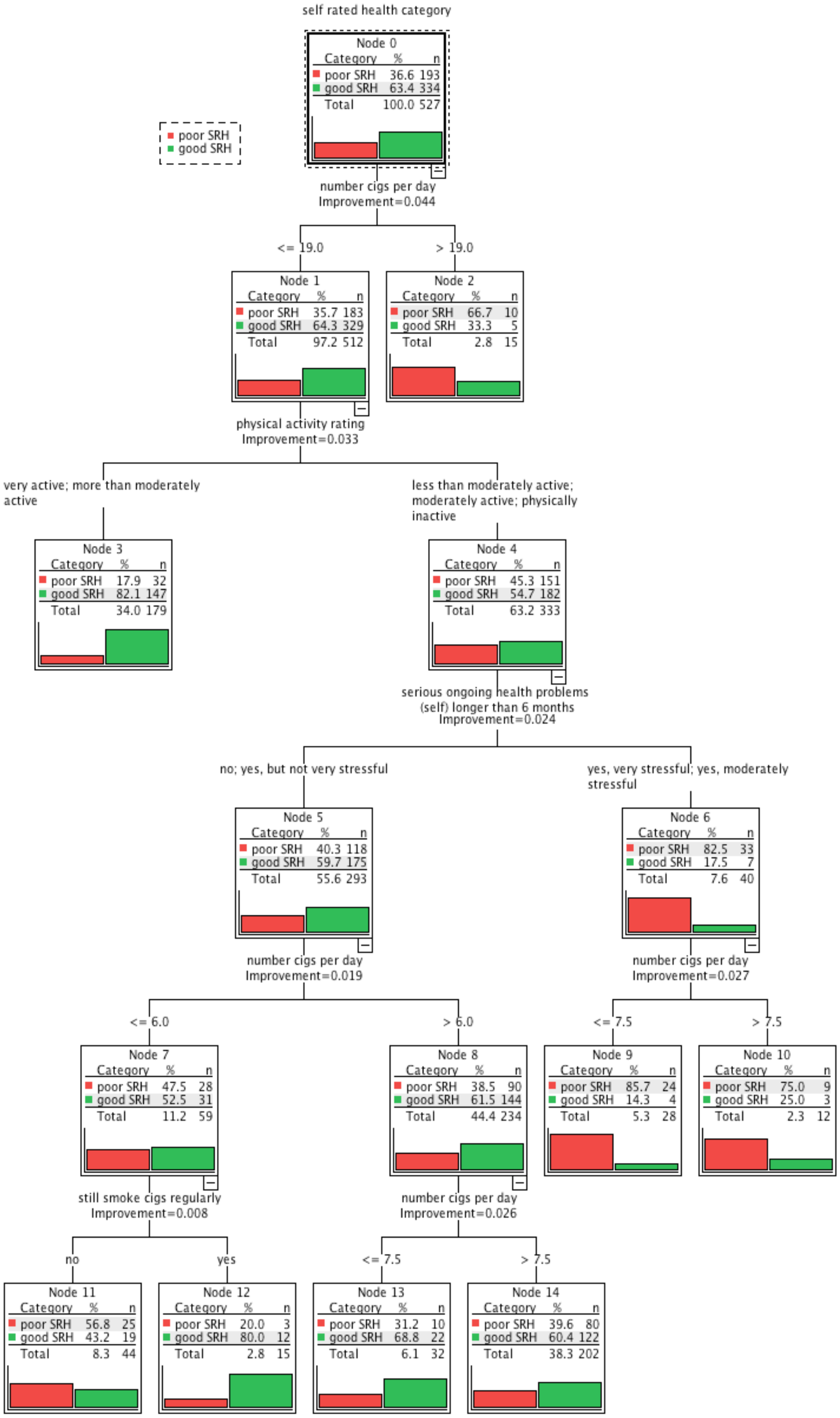


Figure 5: Classification tree for income category $100,000 plus

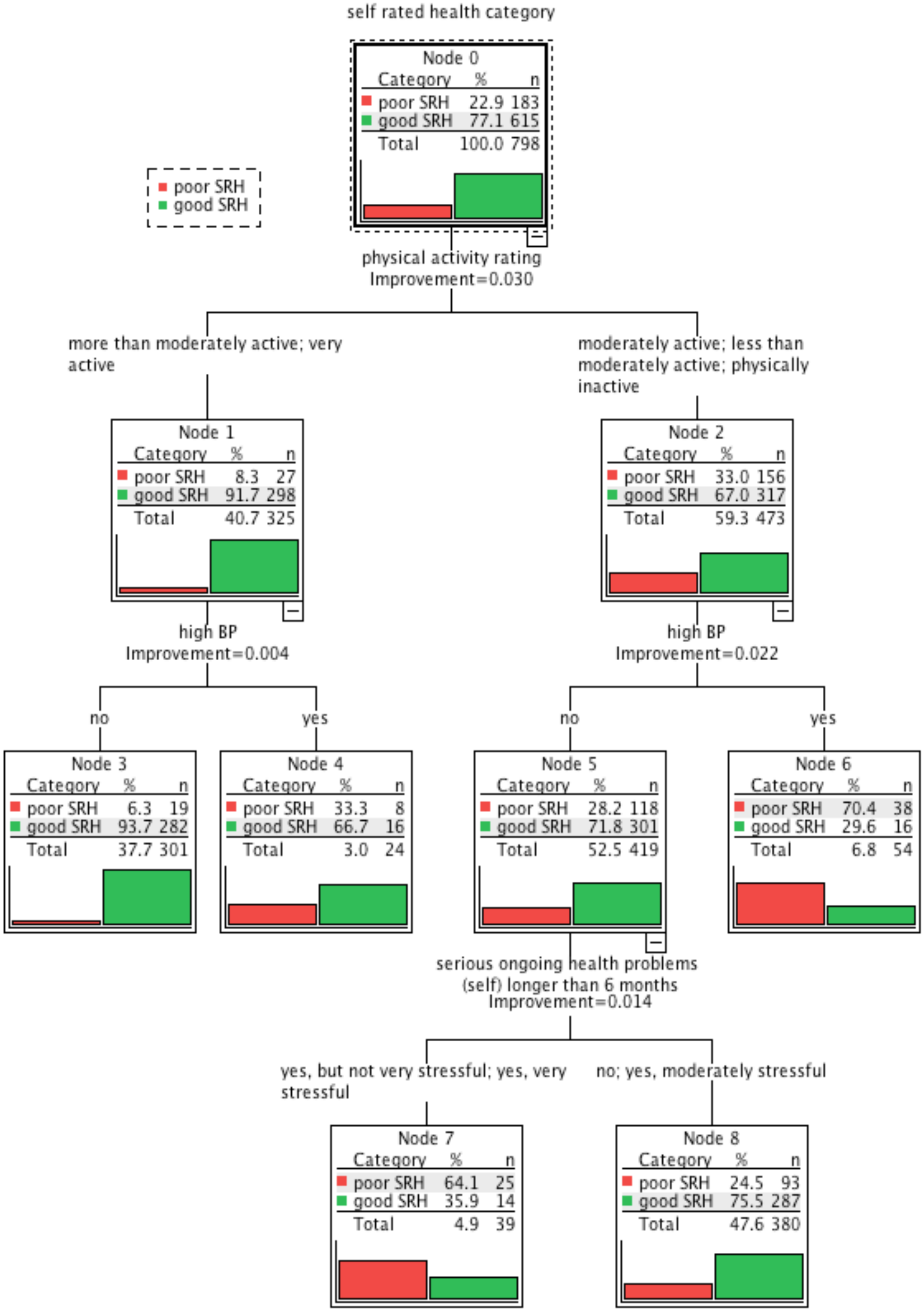


Table 6 shows the factors associated with self-rated health in each income category subgroup of the study population. The shaded boxes show the normalized importance of all the variables appearing in the classification trees; the remaining values represent the other variables that appeared between the highest and lowest ranked variables in the tree model, according to normalized importance values. The combinations of factors associated with health are different by subgroup with no variables, other than a serious ongoing health problem to self and physical activity, appearing in all categories.

Table 6: Combinations of factors chosen by classification tree model for subgroups in different income categories with normalized importance values

		Normalized independent variable importance in classification tree model (%)				
Variable		**Under $25,000**	**$25,000-$50,000**	**$50,000-$75,000**	**$75,000-$100,000**	**$100,000 plus**
Race / Ethnicity			22.6			
Family history	Paternal angina	46.1				
	Paternal diabetes		19.4	5.9		
	Paternal stroke				10.9	
Medical history	High BP	28.2				86.9
	Diabetes	43.2	28.8			
	Liver disease			7.5		
	Chronic bronchitis			7.2		
	Mental nervous emotion	39.4	24.0	6.3		
	Depression	38.2		5.9		
Lifestyle factors	Fast food meals per week	34.0	19.2	12.0		
	Physical activity	38.4	100	100	32.2	100
	Current smoker		29.7		6.5	
	Cigarettes per day	47.1	96.2	92.8	100	
	Ever smoked			5.6		
	Alcohol wine drinks per week		24.2	10.7	7.8	
	Alcohol beer drinks per week		14.7		9.4	
	Liquor drinks per week		18.2	6.4		
	Drug use ever use speed		18.1			
Living and working conditions	Rely on friends and family	43.2	15.5			
	Neighbors help each other		43.3			
	Close-knit neighborhood		19.7			
	Trust neighbors		36.2			
	Neighbors get along		26.0		6.9	
	Neighbors share values				7.0	

	Education		31.1	8.1		(9.3)
	Employment					
	Hard to pay for basics		19.9		6.5	
	Social class discrimination - housing		24.9			
	Racial discrimination at work				6.8	
	Optimism			15.2		
	Control				7.7	
	Health problem - self	100.00	48.3	42.9	21.7	47.2
	Health problem - close person	31.9	15.2			
	Close relationship	20.3	22.9	5.7	8.1	
	Financial strain	87.2	36.5	7.4		
	Job or working ability	26.2				

Shaded boxes show the normalized importance of the variables appearing in the classification trees

Discussion

This study suggests a social gradient for several health determinants relating to lifestyle factors, social and community influences, and living and working conditions. Previous studies have reported similar observations (84), and a social gradient in self-rated health (24). However, this study adds an additional element. The analysis suggested that within population subgroups, stratified by income, the combinations of factors that are associated with self-rated health are different. The predictor variables that are ranked most important in relation to self-rated health differ in sub groups depending on income category. Again, though I cannot assert this ranking is robust (and no inference exists for the difference of the variable importance ranking between income sub-groups), it is suggestive of potentially important differences in the factors that are responsible for self-rated health among income groups.

With successively higher income categories, the proportion of respondents who were male, white, and with good self-rated health increased. Respondents in higher income categories were more likely to report support from, and ability to rely on, friends and family, neighborhood trust, mutual help and cohesion, and home ownership. They were also more likely to have health insurance, and report a sense of control over life events and optimism for the future. There was an inverse social gradient with decreasing fraction of family history of maternal diabetes, hypertension, stroke and heart disease, and paternal diabetes or stroke. Most medical conditions (diabetes, stroke, chronic bronchitis, asthma, liver disease, epilepsy, nervous/emotional or mental condition, depression) showed an inverse social gradient, apart from high cholesterol that increased with income, and heart disease, which showed no significant trend in this sample. Respondents in lower income categories were more likely to have failed to seek health care due to cost in the previous 2 years, be unemployed, have difficulty paying for basics and medical care, and report feeling helpless in dealing with life problems. The results suggest the potential contribution of these factors in the socioeconomic gradient in health; their impact on health depends on interaction with co-variables, and these change with the gradient.

Adler and Stewart's review (85) discussed the eras though which research on socioeconomic status and health has progressed. The first model of a threshold effect between poverty and health was later refined following evidence of a graded association. Subsequent eras studied mechanisms linking socioeconomic status and health, and considered multilevel influences, and most recently, interactions among factors. My primary exploratory study proposes an alternative approach to investigating unique and specific needs within population subgroups along the socioeconomic gradient, and adds to the work on the joint impact of health determinants.

Also, in this study, though several medical conditions were included as predictor variables, it appears that it may not be the absolute presence of disease that is most important. Instead, it is the chronic burden due to a serious health-related condition that was important across all income categories. For the lower income subgroups, many of the measures of chronic burden were more likely. These indicators also represent associated stress and not just the presence of the problem per se. In the presence of a chronic disease, in addition to the medical condition, strain could be due to medical costs, access to services, lack of support in dealing with illness, impact on daily life, and work. Social and community influences, resources in terms of paying for basics, medical care, and health insurance were greater in higher income categories, suggesting a possible protective or buffering role from the associated stress. According to a Census Bureau report, it is estimated that in 2012, 16.3 percent of the US population, or 49.9 million people, were without health insurance (86).

Similar levels of exposure to a risk factor may lead to differential impacts on different socioeconomic groups, due to differences in social, cultural and economic environments; these differential impacts may also occur as low income groups are more likely to be exposed to several risk factors simultaneously (83). Morello-Frosch et al. also highlight the cumulative impact of exposures and vulnerabilities in low-income communities; these contribute to poor health outcomes and inequalities through a convergence of multiple environmental hazards and social stressors. Previous work has also explored the link between neighborhood socioeconomic status and self-rated health. Some of the variance in the association is explained through psychosocial explanatory factors, such as loneliness, perceived stress and optimism (87). Thus, mitigating cumulative effects requires a multi-level and multi-dimensional approach to policy and intervention (87, 88). The WHO Task Force on Research Priorities for Equity in Health called for research studying the *"interrelationships between individual factors and social context that increase or decrease the likelihood of achieving and maintaining good health"* (p950); a focus on single proximal risk factors in isolation, fails to take into account the impact of social context and position, and the interaction of the multiple mechanisms underlying health inequity (89).

Through use of the CARDIA dataset I was able to study a wide range of multilevel determinants, though, due to the use of a cross sectional study sample, I can report only associations with self-rated health. Additional limitations to the study include the self-reported nature of certain factors, and also the potential of health selection in our use of income as the marker of socioeconomic status. A single classification tree has a degree of instability, and further research is needed to refine the analysis, with bootstrap methods and random forests, for example. As choice of socioeconomic indicator depends on the postulated mechanisms by which it affects health (90), I justify the use of income as I am interested in socioeconomic status in terms of disparities in material resources, and access to resources that affect living and working conditions, lifestyle choices, and the neighborhoods in which people live.

Classification tree analysis is particularly well suited to dealing with a large number of multilevel variables and identifying combinations of predictor variables. The results are easily interpretable in terms of associations, interactions, and identifying potentially valuable intervention points. Within subgroups of the population based on income, the combinations of factors that are associated with self-rated health differ, and the relative importance of factors is also different. The results of this study suggest therefore, that any one intervention may not be optimal or even appropriate across the whole population. The most effective action for improving health should be based on an understanding of differing patterns of risk and protective factors in smaller population subgroups. This is more likely to result in implementation of interventions that are of most value in terms of health impact and appropriateness to need. This is analogous to an audience segmentation approach that has been used to guide social marketing in public health in recent years (70). Though social marketing has been generally applied with a view to modifying health behaviors, our work indicates that the principles can be extended to include intervention targeting broader social determinants.

Chapter 4 Characteristics associated with self-rated health in the CARDIA Study: random forests analysis

Introduction

Recursive partitioning methods, such as classification and regression trees (CART), can capture the impact of complex interactions within a set of multi-domain health determinants. They also can suggest population subgroups that share similar health outcomes. I utilized classification tree analysis in my previous study of the characteristics associated with self-rated health, using a dataset drawn from the CARDIA study (Coronary Artery Risk Development in Young Adults). In this chapter, I extend the analysis with the application of random forests, acknowledging the lack of robustness inherent in over-interpreting a single classification tree fit to a finite data set. The random forests method produces an ensemble of classification trees, resulting in significant improvements in accuracy compared with a single tree, and more robust variable importance measures (91).

Classification tree analysis systematically splits the study population into subgroups of maximum homogeneity (for the outcome of interest), with some shared characteristics, that is, common predictor variables. CART has a number of advantages. It is a non-parametric method that is valuable when dealing with a large set of predictor variables and a large sample size, and can deal with missing variables. Simple tree structures produced by the analysis display population subgroups, and provide an understanding of the conditions, in terms of predictor variables, that determine which outcome category an individual is in (65). Trees reflect interactions between multilevel health determinants, and suggest where interventions might optimally be focused.

Classification tree analysis has been applied in clinical epidemiology settings. First, as a diagnostic tool, the output of these methods have been referred to as analogous to clinical decision making, identifying clusters of signs and symptoms, and producing a simple representation of gathered knowledge that is easily interpretable (66). Second, for prediction of outcomes or complications based on prognostic indicators (92). Third, to identify groups with varying risk in clinical settings - classification trees have been found to uncover interactions between variables that may be overlooked in the traditional application of logistic regression methods to case-control data (67, 68).

There are, however, a number of limitations to classification tree analysis (93, 94) and naturally, the choice of analysis method needs to be based on the features that are most suitable for the research question (95). Tree structures are prone to instability and even minor changes to the dataset can result in significant changes to the tree structure. With high dimensional data, there may be a large number of variables and a relatively small number of observations. This is referred to as the small n, large p problem (96) and occurs, for example, in genetics research when studying the relationship between thousands of genes and a disease outcome. Inference based on a single classification tree is not meaningful. Forests are a method of addressing these limitations. Classification accuracy is improved by growing an ensemble of trees and using the generated trees to determine the most popular class (91).

Random forests are constructed using the following algorithm (91, 93): a bootstrap sample is drawn - *n* observations are sampled with replacement from the original sample. Recursive partitioning is applied to the bootstrap sample. At each node, from the original complete set of predictor variables, a random subset of variables are selected, and the tree splits are restricted based on these; this reduces correlation between trees. Trees are generated without pruning; bagging (bootstrap aggregation) decreases the variance created by the lack of pruning (97) . Splitting of the data continues until no further splits are possible; this may be when the node is homogenous (all of one class), or there are no more predictor variables on which to split (97). These steps are repeated a predetermined number of times to form a forest of trees. The forest-based classification is made by majority vote from all trees. In the single classification tree, cross-validation is used to estimate tree accuracy (65). In random forests, there is an integral internal error rate produced as a result of the bagging process; each tree created in the random forest ensemble is produced using a different bootstrap sample from the study dataset, whilst approximately one third of the cases are unselected. This is the out of bag sample, which is then put into the tree to get a classification. All the class predictions are compared to the true classes, to produce the out of bag error rate. This is the best indicator of stability (96-98). Typical fitting of random forests to data does not seek to find the "smallest" model, and so it tends to over-fit. This means that though random forests is a very good predictor, it may in fact systematically underestimate the importance of variables, given a phenomenon equivalent to over-adjustment in more standard epidemiological parlance (99). However, it is a robust and valuable tool, particularly used in conjunction with CART.

In this chapter, I apply random forests to refine the classification tree analysis on characteristics associated with self-rated health with a view to producing more robust variable importance estimates. Combined with the results from the CART analysis (chapter 2), these can be used as a more reliable basis on which to consider potential public health action, so that intervention strategies can be developed and prioritized to address factors associated with good or poor self-rated health.

Methods

I utilized cross-sectional data collected as part of the CARDIA study, a United States cohort study started in 1985 to investigate development of coronary artery disease risk factors in a young adult population. For this study, data were taken from the year 15 examination of the cohort, conducted in 2000-2001, through interviewer and self-administered questionnaires (with the exception of race/ethnicity information taken from the 1985-1986 data collection, and family history information taken from the 1995 data collection). From an original total of 5115 participants, 3672 were followed up in year 15. From the year-15 group, all participants who had a response for self-rated health, and were coded as male or female, were included in the final study sample of 3649 participants.

For each individual in the study sample, data were available for self-rated health, and a wide range of multi-domain health determinants (Table 1). We selected the variables based on the Dahlgren and Whitehead model (60) to include age, sex and hereditary factors, and to represent individual lifestyle factors, social and community influences, and living and working conditions.

In the CARDIA study, self-rated health was assessed on a five-point scale, by the question, *"In general would you say your health is excellent, very good, good, fair or poor?"* Responses were categorised by grouping together excellent or very good as 'good' self-rated health, and responses of good, fair or poor, as 'poor' self-rated health. Indicator variables were created for a number of variables as outlined previously in chapter 2 (page 18, Table 1).

I used the Random Forests package in R, through the R integration package RanFor (R version 2.12.2 Copyright © 2011 The R Foundation for Statistical Computing [http://www.r-project.org]) in IBM SPSS Statistics v19.0.0. I specified 1000 trees in the random forest model to generate variable importance measures, and used the default value for number of predictor variables sampled at each node. For classification trees, this is the square root of the number of predictors. I set the parameters to impute missing values in scale variables as the variable's median value, and for categorical variables, as the modal value (100).

Table 1: Multi domain health determinants drawn from the CARDIA study

Level 1: Age, sex & hereditary factors

- Age
- Sex
- Race / Ethnicity
- Family history of medical conditions

Level 2: Individual lifestyle factors

- Medical history
- Diet
- Physical activity
- Smoking
- Alcohol
- Tobacco
- Illicit drug use

Level 3: Social & community influences

- Social support / network
  - Feeling that friends and family care
  - Can rely on friends and family
- Sense of close-knit neighborhood, neighborhood cohesion
  - Neighbors help each other
  - Live in close-knit neighborhood
  - Neighbors can be trusted
  - Neighbors generally get along with each other
  - Neighbors share the same values

Level 4: Living & working conditions

- Education
- Income
- Housing - rent or own house
- Employment - working versus unemployed
- Control and adequacy of resources
  - Difficulty paying for basics or medical care
- Medical insurance
- Access to health services
- Experience of discrimination due to gender, race/ethnicity or colour, socioeconomic position, or social class in 7 settings (at school, getting a job, getting housing, at work, at home, getting medical care, on the street, or in a public setting)

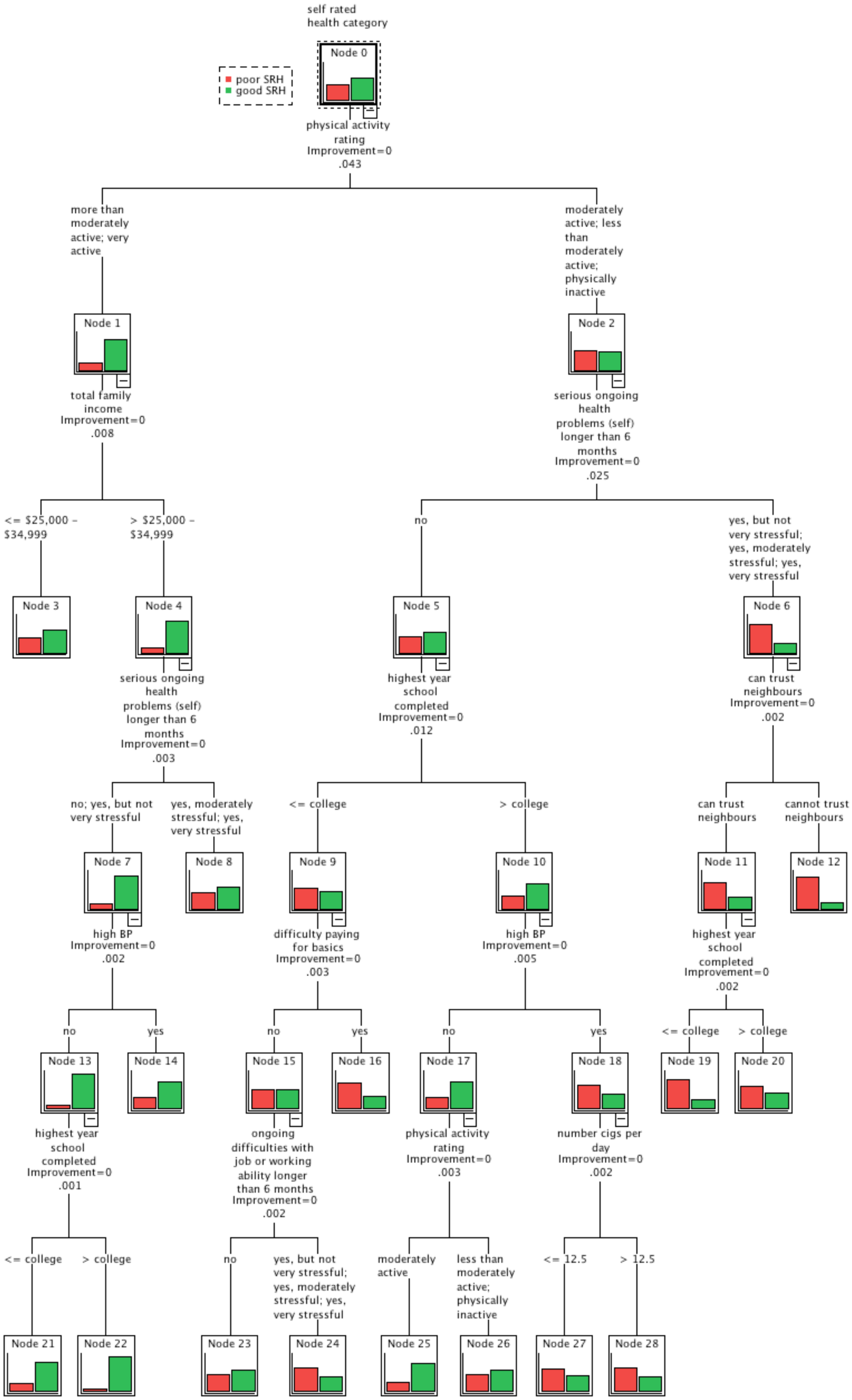
Results

The results of the single classification tree analysis (as in chapter 2) are shown in Figure 1. We found that multi domain health determinants – lifestyle, social and community influences, and living and working conditions are associated with self-rated health in the study sample, and combinations of factors differed by population subgroup. The normalised importance of the highest ranked variables (20% or greater) is shown in Table 2. Physical activity rating emerged as the most important variable in the single classification tree with a higher level of physical activity associated with good self-rated health. There appeared to be interaction of lifestyle and medical factors with socioeconomic factors, income and education. The next 5 variables in the ranking were those splitting variables that appeared in the final classification model: chronic burden from serious on-going health problems, highest year school completed, total family income, difficulty paying for basics, and history of hypertension. Race / ethnicity, ranked next, was also a discriminating factor between the subgroups with the highest proportions of good and poor self-rated health.

Table 2: Highest ranked predictor variables by normalized importance (> 20%) for single classification tree analysis

Independent Variable	Normalized Importance
Physical activity rating	100.00%
Serious on-going personal health problem (self) longer than 6 months	55.10%
Highest year school completed	45.60%
Total family income	36.20%
Difficulty paying for basics	35.50%
High Blood Pressure	30.90%
Race / Ethnicity	26.40%
On-going financial strain longer than 6 months	22.70%
Difficulty paying for medical care	21.30%

Figure 1: Single Classification Tree Analysis


self rated health category
poor SRH
good SRH
Node 0
physical activity rating Improvement=0 .043
more than moderately active; very active
moderately active; less than moderately active; physically inactive
Node 1
total family income Improvement=0 .008
Node 2
serious ongoing health problems (self) longer than 6 months Improvement=0 .025
<= $25,000 - $34,999
> $25,000 - $34,999
no
yes, but not very stressful; yes, moderately stressful; yes, very stressful
Node 3
Node 4
serious ongoing health problems (self) longer than 6 months Improvement=0 .003
Node 5
highest year school completed Improvement=0 .012
Node 6
can trust neighbours Improvement=0 .002
no; yes, but not very stressful
yes, moderately stressful; yes, very stressful
<= college
> college
can trust neighbours
cannot trust neighbours
Node 7
high BP Improvement=0 .002
Node 8
Node 9
difficulty paying for basics Improvement=0 .003
Node 10
high BP Improvement=0 .005
Node 11
highest year school completed Improvement=0 .002
Node 12
no
yes
no
yes
no
yes
<= college
> college
Node 13
highest year school completed Improvement=0 .001
Node 14
Node 15
ongoing difficulties with job or working ability longer than 6 months Improvement=0 .002
Node 16
Node 17
physical activity rating Improvement=0 .003
Node 18
number cigs per day Improvement=0 .002
Node 19
Node 20
<= college
> college
no
yes, but not very stressful; yes, moderately stressful; yes, very stressful
moderately active
less than moderately active; physically inactive
<= 12.5
> 12.5
Node 21
Node 22
Node 23
Node 24
Node 25
Node 26
Node 27
Node 28

Random Forests

The variable importance output from random forests analysis is shown in Table 3, color-coded by domain of predictors (see Key, Table 3). Education and income were associated with the greatest decreases in node impurity, (179.63 and 178.81, respectively), and reflected the highest variable importance. Physical activity rating was ranked 3rd at 151.00, chronic burden due to serious health problem (101.50), and chronic burden due to financial difficulty (97.50) ranked 4th and 5th. Four out of the five highest ranked variables were the same in the single classification tree normalized importance (chapter 2, page 31, Table 5) and the random forests ranking (Table 3 below). The fifth variable in the single tree ranking was difficulty paying for basics and in random forests, it was on-going financial strain; these can be considered related concepts.

Table 3: Random forests: predictor variables and decrease in node impurity

Key: predictor variables color-coded by domain of health determinant

⬜ Age, sex and hereditary factors (family history)
🟩 Individual lifestyle factors
🟥 Medical conditions
🟦 Social and community influences
🟧 Living and working conditions

Predictor Variable	*Decrease in Node Impurity
🟧 **Highest year school completed**	179.63
🟧 **Total family income**	178.813
🟩 **Physical activity rating**	151.002
🟧 **Chronic burden – personal health problem (self)**	101.496
🟧 **Chronic burden – on going financial strain**	97.501
⬜ **Age**	77.056
🟩 **Fast food meals per week**	64.264
🟧 **Chronic burden – job or working ability**	55.196
🟧 **Chronic burden – relationship difficulties**	52.747
🟦 **Neighbors trust each other**	50.073
🟧 **Chronic burden – serious health problem (in other)**	46.847
🟥 **High blood pressure**	43.969
🟩 **Beer drinks per week**	39.318
🟧 **Optimistic for future**	38.652
🟧 **Hard to pay for basics**	37.019

Cigarettes per day	32.801
Wine drinks per week	32.419
Liquor drinks per week	28.892
Neighbors help each other	25.117
Neighbors get along	23.768
Own home	21.692
Discrimination due to race-ethnicity/colour at work	21.376
Race /ethnicity	20.369
High cholesterol	20.338
Discrimination due to race-ethnicity/colour in public	19.924
Neighbors share values	19.758
Control over life events	18.89
Hard to pay for medical care	17.485
Sex	16.926
Close knit neighborhood	16.882
Maternal high blood pressure	16.316
History marijuana use	16.065
Discrimination due to race-ethnicity/colour getting a job	15.862
Paternal high blood pressure	15.39
Discrimination due to gender in public	15.029
Discrimination due to gender at work	15.019
History non crack cocaine use	14.343
Depression	14.152
Asthma	13.382
Work part time	13.238
Work full time	13.131
Still smoke cigarettes regularly	12.544
Ever smoked cigarettes at least 3 months	12.506
Can rely on friends and family	12.467
Discrimination due to socioeconomic position in public	12.433
History of amphetamine use	12.41
Helpless in dealing with life problems	12.41
Health insurance from Employer/Union/School	12.262
Maternal diabetes	11.552
Discrimination due to gender getting a job	11.452
Paternal heart attack	11.405
Paternal diabetes	11.384

Discrimination due to race-ethnicity/colour at school	11.281
Diabetes	10.95
Heart disease	10.402
Difficulty getting health services	10.268
Discrimination due to socioeconomic position at work	10.228
History crack use	9.791

Discrimination due to gender at school	9.741
Discrimination due to socioeconomic position getting housing	9.593
Maternal angina	9.532
Discrimination due to socioeconomic position getting a job	9.015
Discrimination due to gender getting housing	8.926
Paternal angina	8.909
Self-paid health insurance	8.898
Always had health insurance past 2 years	8.781
Nervous, emotional or mental disorder	8.689
Paternal stroke	8.68
Discrimination due to socioeconomic position at school	8.537
Discrimination due to race-ethnicity or colour getting housing	8.399
Unemployed	8.2
Maternal heart attack	8.13
Did not seek medical care last 2 years due to cost	7.55
Discrimination due to gender getting medical care	7.272
Discrimination due to gender at home	7.125
Chronic bronchitis	7.016
Kidney disease	6.799
Thyroid disease	6.544
Maternal stroke	6.412
History opiate use (non-medical)	6.401
Discrimination due to socioeconomic position getting medical care	6.167
Discrimination due to race-ethnicity or colour getting medical care	4.704
Cancer	4.518
Liver disease	3.759
Discrimination due to socioeconomic position at home	2.983

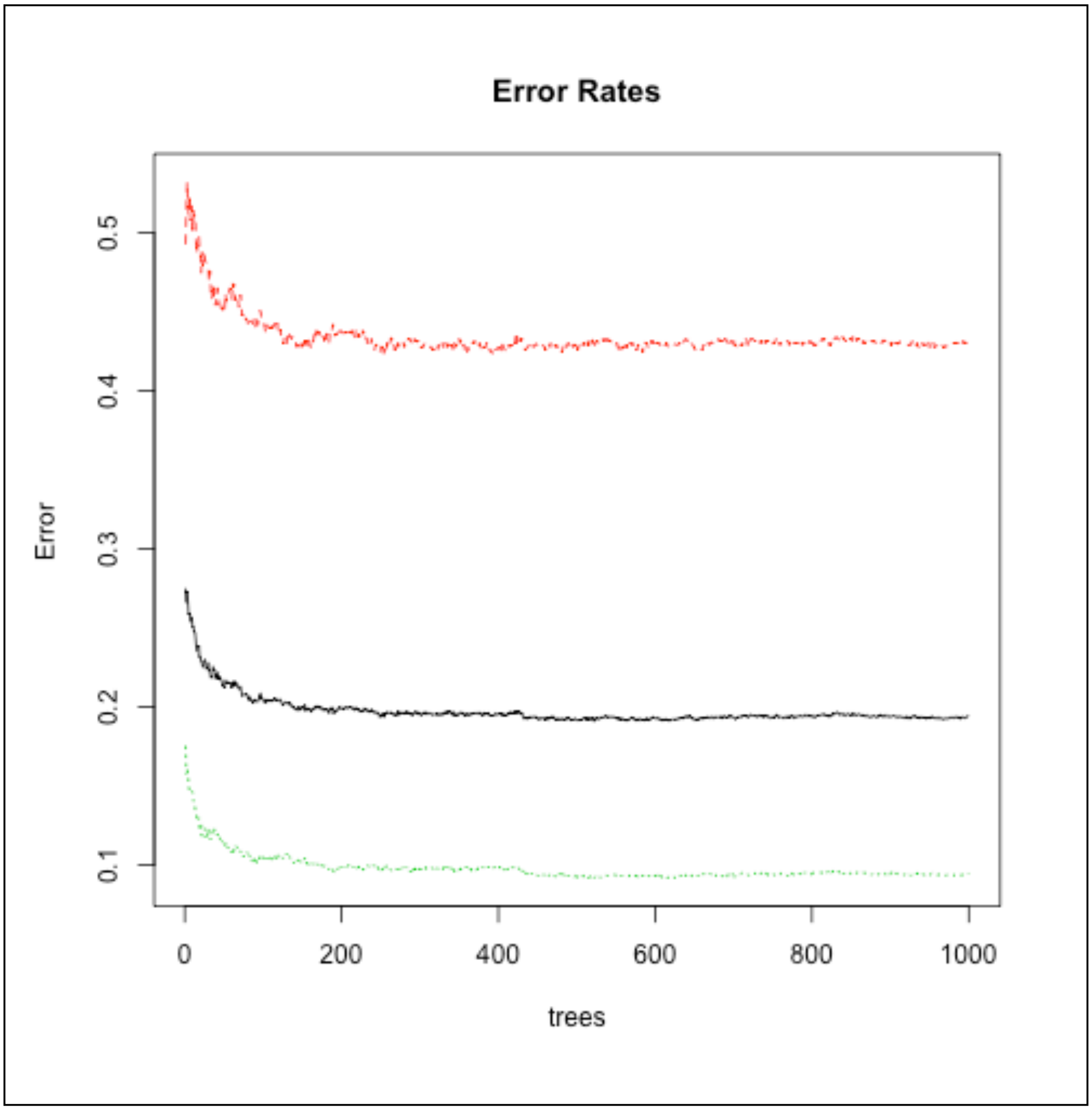
Friends and family care	2.682
Discrimination due to race-ethnicity or colour at home	2.245
Epilepsy	2.137
HIV	1.414
Stroke or TIA	0.87
Multiple Sclerosis	0.467
Emphysema	0.122

*Decrease in node impurity is the total decrease in node impurities from splitting the variable averaged over all trees measured by the Gini index.

The highest ranked variables in the random forests model represent living and working conditions, lifestyle, and social and community influences. The presence of a serious health condition appears high in the ranking. This represents chronic burden associated with a condition rather than just presence of disease itself. Most of the predictor variables indicating history of a particular medical condition were ranked low, apart from high blood pressure and high cholesterol.

The out of bag error rates varied depending on the number of trees grown in the forest (Figure 2). There was no major decrease in error above approximately 200 trees. The overall estimated out of bag error rate was 19%. Compared with the cross-validated error estimate of 31% for the single classification tree, the error is improved through random forests,

Figure 2: Out of bag error rates plotted against number of trees in random forests analysis



Confusion Matrix of Predictions

	0	1	Class Error
0 (red)	**860.000**	**655.000**	**.432**
1(green)	**341.000**	**3259.000**	**.095**

Overall estimated out of bag error rate is 0.195 (black)

Rows are actuals; columns are predicted

0=poor/1=good self-rated health

Discussion

In this study, I found that education and income were the most important variables identified in association with the outcome of self-rated health category. Random forests analysis is a more robust estimate of variable importance compared with the output of the single classification tree. The comparison of the out of bag error rate with the cross-validated error rates reflects the reduction of error through random forests. Despite the inclusion of a wide range of predictor variables representing fixed factors, lifestyle and medical conditions, social and community influences, and living and working conditions, the analysis selected the socio-economic variables as most important based on the mean decrease in Gini index. The random forests analysis was generally consistent with our findings from the single tree, reflecting the association of education and income with self-rated-health status. The ranking of variables by random forest suggest that self-rated health is not merely a reflection of objective health. Based on the single classification tree, the combinations of multi-domain risk and protective factors were different by population subgroup.

Though we included several disease variables, the indicators for presence of disease per se are not highly ranked in the random forests model, though chronic burden associated with serious health condition was ranked 4th highest. This might imply that it is elements of stress, and lack of social or financial resources that could be key issues in whether illnesses become a stressful burden or not. Previous studies have shown that patients with chronic conditions are known to also have good self-rated health if they have high levels of mastery (44). The previous analysis in chapter 3 reflects that individuals across the socioeconomic gradient are susceptible to illness. Prevalence of diseases is lower, and prevalence of excellent to very good self-rated health is higher, with higher income groups. Higher income subgroups also have higher proportions of protective buffering factors such as more financial resources, social support, and community cohesion. This was shown in the trend analysis of predictor variables by increasing income categories in the CARDIA study sample. When considering public health action for different populations, an understanding of the broader social determinants associated with good and poor self-rated health is important in deciding the best intervention pathway to improve health. Even lifestyle choices are made within a specific socioeconomic context constraining or enabling choices. Recursive partitioning methods are suited to this type of analysis, and to uncovering these relationships.

Limitations of classification and regression trees relate to the instability with regard to inferences in the importance of variables provided by a single tree. Relatively newer recursive partitioning techniques offer methods to address these. Bagging trees consists of multiple trees grown out of bootstrap samples that are combined by averaging (for regression) or by simple vote for classification (101). Random forests add the further dimension of a random sampling of the predictor variables. Random predictor selection controls the bias (102). There is some loss of interpretability with random forests, as trees are not graphically represented due to large numbers, and limitations in interpretation due to an inherent over-fitting. However, random forests also produce more robust measures of variable importance. Recent developments in semi-parametric methods add further value to these methods. Using a counterfactual approach to generate variable importance, the distribution of the outcome of interest can be compared with its theoretical distribution if the variable of interest is set to the lowest risk (103, 104). This is especially valuable in producing parameters that translate well to public health practice. Variable importance analysis by fitting multiple Population Intervention Models (PIMs) produces a parameter that is analogous to attributable risk. Under certain assumptions, the parameter can be considered as an actual causal effect of the exposure variable on the outcome, or as measuring the hypothetical effect of an intervention in which everyone in the population is made to be like the members of the target group (104, 105). The advantage of these methods is that they can use the power of methods such as random forests that are very good at flexibly fitting the data, while still providing interpretable and robust estimates of variable importance.

Social marketing has been proposed as a method of achieving behavior change with lifestyle modification through targeted health promotion programs (106). Although social marketing's target audience is usually made up of consumers, it is used also to influence policy makers who can address the broader social and environmental determinants of health (107). In social marketing, audience segmentation information shapes behavior change strategies, but the principles can be extended to include intervention targeting broader social determinants and used to conceptualize a profile of population subgroups in terms of protective and risk factors. Firstly, from commercial marketing, deep knowledge of the customer can be translated into knowledge of the communities and population subgroups that are to benefit from public health interventions. A key component of social marketing is audience segmentation, in which information on distinct population subgroups, and knowledge of the social and cultural environments in which the people act on behavioral decisions is gathered. Lemon highlighted the utility of audience segmentation strategies in public health as they identify relatively homogenous subsets of the population as a basis for targeted interventions appropriate to unique needs (70). Recursive partitioning methods are a means of achieving this efficiently, and offer added value in identifying special needs of population subgroups. Audience segmentation is typically done through cluster analysis methods; classification tree methods, however, model these factors on an outcome, which adds a valuable dimension. Epidemiological studies utilizing parametric multivariate regression techniques typically model the average relationship in the population between an exposure and an outcome. This can result in intervention being aimed only at the 'average' participant (69, 70).

The selection of education and income as the highest-ranking variables association with self-rated health in the CARDIA study sample highlights the importance of addressing social determinants of health and inequities. Capturing the complex interplay of factors affecting health in subgroups can be difficult using parametric multivariate logistic regression. These models may not capture the full array of variables influencing health. The use of recursive partitioning methods and semi-parametric variable importance methods with public health survey data can effectively highlight important risk and protective factors amongst population subgroups for further inquiry. This is useful in developing appropriate interventions that relate to the real-life mix of circumstances affecting communities.

Chapter 5 Discussion

This dissertation contributes to the literature on the determinants of self-rated health and adds a novel application of classification and regression tree and random forests methods to the study of self-rated health status.

In chapter 2, I segmented the CARDIA study sample using classification tree analysis. Multi-domain health determinants – lifestyle, social and community influences (neighborhood cohesion), living and working conditions (education, income and adequacy of resources) - were associated with self-rated health in the study sample, and combinations of factors varied by subgroup. Physical activity rating emerged in the single tree model as the most important variable, with a higher level of physical activity associated with better self-rated health. For the subgroup with higher physical activity, the tree model suggested interaction of lifestyle and medical factors with socioeconomic factors, income and education. The subgroup with the highest proportion of good self-rated health (92.9%, n=561), compared with the remainder of the study population, was characterized by the following factors: higher physical activity rating; higher income bracket; no serious on-going personal health burden; or if present, not very stressful; no history of hypertension; highest year of school completed is graduate level. The subgroup with the highest proportion of poor self-rated health (84.7%, n=238), compared with the remainder of the study population, shared different characteristics: lower physical activity rating; serious on-going personal health problem; perception that people in the neighborhood cannot be trusted.

In chapter 3, I divided the study sample by total family income to create 5 subsets representing a socioeconomic gradient. There was a significant association between income category and self-rated health status; the proportion of excellent or very good ('good') self-rated health increased with income, and the proportion of good, fair or poor ('poor') self-rated health decreased. A social gradient was observed for several health determinants related to lifestyle, social influences, and living and working conditions. This study also found that the factors selected by the classification tree model as associated with self-rated health differed across each income-based subgroup; only the variables indicating a serious ongoing personal health problem, and physical activity appeared in all income subgroups. This implies that the constellation of health determinants associated with self-rated health varies in different socioeconomic groups; the impact upon health outcome of particular factors depends on interaction with co-variables, and these combinations change with the gradient.

In chapter 4, random forests methods were applied to extend the analysis in chapter 2. Education and total family income were the most important variables identified in association with self-rated health category. Despite the inclusion of a wide range of predictor variables representing fixed and lifestyle factors, medical conditions, social and community influences, and living and working conditions, the analysis model selected these socioeconomic variables as most important based on the mean decrease in Gini Index.

There are some limitations to this study. Though the predictor variables were selected from an existing strong dataset to represent multiple layers of influences on health, a few may not optimally represent the characteristic of interest. For example, diet is included only by way of fast food intake. Responses that were classed as 'don't know' in the original CARDIA data collection were labeled in this study as 'missing'. It is possible that this could introduce some bias if people who responded to certain questions with 'don't know' were more likely to have a particular self-rated health status. Tree-based methods are prone to instability, so that small perturbations in the data can produce large variations in tree structure even though prediction accuracy might not vary at all, though this may be less problematic since the focus here is on understanding specific influences within one population. I have attempted to address this by the application of random forests in chapter 4; the ensemble of trees improves predictive accuracy and provides more robust variable importance measures. Even so, there is no rigorous theory for providing inference on the structure of a single tree, and the output of random forests too, in the absence of a Type 1 error rate, is best considered a rank ordering of key variables worthy of further investigation.

This study has a number of strengths, and results that have relevance to public health practice. First, classification and regression tree techniques are particularly suited to uncovering associations between several multi-domain characteristics and self-rated health. The health of both individuals and social groups results from interaction between structural, material, social, cultural, and behavioral factors. Therefore, whilst understanding the individual effects of health determinants is important, in this study, I also aimed to look at combinations or the joint impact of factors, and the relative importance of multiple layers of influence. Single elements of the broad range of health determinants reflect only some aspect of health but without consideration of cofactors, are incomplete predictors of overall health status (78). Findings from chapters 3 and 4 suggest that self-rated health does not simply reflect medical health. There appear to be co-factors that enable some individuals to rate health as good despite presence of disease. Patients with chronic conditions are known to also have good self-rated health if they have high levels of psychosocial resources (44). In chapter 3, presence of chronic burden associated with serious health condition appears across the social gradient. This could imply that it is elements of stress, and lack of social or financial resources that could be key issues in whether illnesses become a stressful burden or not. People across the socioeconomic gradient develop disease. Why do some people maintain a perception of good self-rated health and others do not? Across the study population, the proportion of good self-rated health increased with higher income. This may be because individuals in these higher income subgroups also have buffering factors, such as more resources, social support, and community cohesion.

The tree models in this study selected education and income as ‘important’ associations of self-rated health despite the inclusion of several other health-related variables. Even lifestyle indicators are the result of a specific socioeconomic context constraining or enabling choices, and thus intervention aimed at, say reducing smoking or alcohol consumption should regard the broader context of living conditions (80). Studies have shown that biomedical risk factors account for only a fraction of ill health. The results of the Whitehall study, for example, highlighted that traditional ‘medical’ risk factors (such as cholesterol level, smoking, systolic blood pressure, and diabetes) explained only a third of the observed socioeconomic gradient in health (79). Thus upstream factors, or at least the impact of the ’causes of the causes’ have to be considered in tandem with other foci of health promotion programs, even if they may be more difficult to remedy.

A second strength of the approach taken in this study is that classification and regression trees produce relatively homogenous population subgroups, in terms of outcome, that are defined by selected common characteristics (70). This study demonstrated an alternative methodological approach to viewing the potential determinants of differences in health status across the socioeconomic gradient, and identified the relative importance of factors associated with self-rated health in income-based subsets. Adler et al. proposed that given the complexity of the socioeconomic gradient, statistical procedures other than traditional regression might better examine combinations of outcomes and assess complex interrelated variables (82). This study has applied such a creative approach even though this is an exploratory study of associations, and the results are interpreted in this light. Results in chapter 3 reflected a positive association between higher income and better self-rated health but also that combinations of health determinants associated with self-rated health differed across the gradient. Strategies to reduce levels of poor self-rated health are better developed with an understanding of the unique characteristics and needs of specific subgroups across the socioeconomic gradient. In the context of a proportionate universalism approach to reducing health inequalities, the findings imply that as well as differences in the degree or intensity of public health action required across the gradient, differences in the *type* of action are also likely to be important (as health determinants appear in different combinations, and with different relative importance in different income groups).

Thirdly, the non-parametric approach used in this study is not dependent on the data following a particular distribution. This is pertinent given the aim of simultaneously considering categorical, ordinal, and continuous variables from several health-related domains. Breiman describes statistical modelling as having two cultures: data modelling assumes a stochastic data model; algorithmic modelling treats the data mechanism as unknown. In the first of these models, with complex high dimensionality datasets, including different types of variables, there is a risk of making incorrect assumptions on the structure of the underlying data being multivariate normal. Breiman argues, *“If the model is a poor emulation of nature, the conclusions may be wrong”* (108)(p.202). Alternate non-parametric approaches have different drawbacks. Dimension reduction with principle components or factor analysis, results in the original predictor variables being transformed into a reduced set of components. However, their individual effect is no longer clearly identifiable (96). Portrait et al. discussed the difficulties of processing the rich set of indicators needed to capture the concept of health, in their work applying Grade of Membership analysis to form a typology of elderly individuals' health status (78). This approach recognized the multidimensionality of the data, and conceptualized health status or outcome as graded participation into several aspects of health. The results are generated as a number of hypothetical pure types or groups, along with numerical weightings of the affinity of individuals with pure types. For some research questions, this type of output is not easily translatable to be of value in public health practice. Sudat et al. also emphasize that in high dimensional datasets, traditional regression approaches can also produce model parameters with little real world interpretability (103). The results of classification tree analysis are easier to translate. In this study, the use of cross-sectional data prevents causal inference. Nevertheless, the different patterns of risk and protective factors discovered in smaller distinct population subgroups point to where further study and development of interventions could best be focused. Recursive partitioning techniques resemble audience segmentation used in commercial settings, which have also been adopted by public health for use with social marketing. Principles of social marketing techniques, generally applied with a view to modifying health behaviors, could be extended to include action targeting broader social determinants. Intervention may be more successful if tailored to meet needs of smaller population subgroups rather than a uniform approach across large communities.

Conclusion

This primary study raises a number of areas for further investigation. Further work is needed to study the causal nature of the associations identified, and understand the mediating factors between poor socioeconomic status and self-rated health that are amenable to change. Application of newer semi-parametric methods can estimate the potential positive public health impact of intervention through measures analogous to attributable risk. Further work and creative methods are needed to unravel causal relationships between risk and protective factors and self-rated health, and more importantly, seek solutions to improve health status. It would be of interest to combine qualitative and quantitative work, and further link multidimensional characteristics and self-rated health to biochemical markers of health status.

This study reflects well how combinations of health-related factors are associated with differing self-rated health outcomes in population subgroups, depending on co-variables and socioeconomic context. This concept is somewhat analogous to the framework used by infectious disease epidemiologists, in which there is consideration of the agent, host, and susceptibility or the environment. The predictive importance of current self-rated health status is well established. Therefore, in addition to preventing specific disease, in a public health context, aiming for populations to have high proportions of good subjective health is itself an important end point. Developing knowledge of how multi-domain factors come together in population subgroups is essential to designing appropriate and effective public health strategies that can achieve this.

References

1. Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. J Health Soc Behav. 1997 Mar;38(1):21-37.

2. Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. Am J Public Health. 1982 Aug;72(8):800-8.

3. Kaplan GA, Camacho T. Perceived health and mortality: a nine-year follow-up of the human population laboratory cohort. Am J Epidemiol. 1983 Mar;117(3):292-304.

4. Idler EL, Kasl SV, Lemke JH. Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982-1986. Am J Epidemiol. 1990 Jan;131(1):91-103.

5. Bjorner J, Fayers P, Idler E. Measures for clinical trials. In: Fayers PM, editor. Assessing quality of life in clinical trials: methods and practice. Second ed. Oxford: Oxford University Press 2005. p. 309-25.

6. Fayers PM, Sprangers MA. Understanding self-rated health. Lancet. 2002 Jan 19;359(9302):187-8.

7. Eriksson I, Unden AL, Elofsson S. Self-rated health. Comparisons between three different measures. Results from a population study. Int J Epidemiol. 2001 Apr;30(2):326-33.

8. Sargent-Cox KA, Anstey KJ, Luszcz MA. The choice of self-rated health measures matter when predicting mortality: evidence from 10 years follow-up of the Australian longitudinal study of ageing. BMC Geriatr. 2010;10:18.

9. Manderbacka K, Kareholt I, Martikainen P, Lundberg O. The effect of point of reference on the association between self-rated health and mortality. Social science & medicine. 2003 Apr;56(7):1447-52.

10. Bailis DS, Segall A, Chipperfield JG. Two views of self-rated general health status. Social science & medicine. 2003 Jan;56(2):203-17.

11. Zack MM, Moriarty DG, Stroup DF, Ford ES, Mokdad AH. Worsening trends in adult health-related quality of life and self-rated health-United States, 1993-2001. Public Health Rep. 2004 Sep-Oct;119(5):493-505.

12. CDC. Centers for Disease Control and Prevention. National Health Interview Survey. [June 2012]; Available from: http://www.cdc.gov/nchs/nhis.htm.

13. CDC. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey. [June 2012]; Available from: http://www.cdc.gov/nchs/nhanes.htm.

14. CDC. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System. [June 2012]; Available from: http://www.cdc.gov/brfss/.

15. Office for National Statistics. 2001 Census for England and Wales. Newport 2012 [30th July 2012]; Available from: http://www.ons.gov.uk/ons/guide-method/census/census-2001/index.html.

16. Office for National Statistics. 2011 Census for England and Wales. Newport 2012 [30 July 2012]; Available from: http://www.ons.gov.uk/ons/guide-method/census/2011/index.html.

17. DeSalvo KB, Bloser N, Reynolds K, He J, Muntner P. Mortality prediction with a single general self-rated health question. A meta-analysis. J Gen Intern Med. 2006 Mar;21(3):267-75.

18. Idler EL, Kasl SV. Self-ratings of health: do they also predict change in functional ability? The journals of gerontology Series B, Psychological sciences and social sciences. 1995 Nov;50(6):S344-53.

19. Lee Y. The predictive value of self assessed general, physical, and mental health on functional decline and mortality in older adults. Journal of epidemiology and community health. 2000 Feb;54(2):123-9.

20. Moller L, Kristensen TS, Hollnagel H. Self rated health as a predictor of coronary heart disease in Copenhagen, Denmark. Journal of epidemiology and community health. 1996 Aug;50(4):423-8.

21. Wolinsky FD, Culler SD, Callahan CM, Johnson RJ. Hospital resource consumption among older adults: a prospective analysis of episodes, length of stay, and charges over a seven-year period. J Gerontol. 1994 Sep;49(5):S240-52.

22. Daniilidou NV, Gregory S, Kyriopoulos JH, Zavras DJ. Factors associated with self-rated health in Greece: a population-based postal survey. Eur J Public Health. 2004 Jun;14(2):209-11.

23. Shadbolt B. Some correlates of self-rated health for Australian women. American Journal of Public Health. 1997;87:951-6.

24. McFadden E, Luben R, Bingham S, Wareham N, Kinmonth AL, Khaw KT. Social inequalities in self-rated health by age: cross-sectional study of 22,457 middle-aged men and women. BMC Public Health. 2008;8:230.

25. Pleis J, Ward B, Lucas J. Summary health statistics for U.S. adults: National Health Interview Survey, 2009. National Center for Health Statistics. . Vital Health Stat 2010;10(249).

26. Franks P, Gold MR, Fiscella K. Sociodemographics, self-rated health, and mortality in the US. Soc Sci Med. 2003 Jun;56(12):2505-14.

27. Molarius A, Berglund K, Eriksson C, Lambe M, Nordstrom E, Eriksson HG, et al. Socioeconomic conditions, lifestyle factors, and self-rated health among men and women in Sweden. European journal of public health. 2007 Apr;17(2):125-33.

28. Bobak M, Pikhart H, Hertzman C, Rose R, Marmot M. Socioeconomic factors, perceived control and self-reported health in Russia. A cross-sectional survey. Social science & medicine. 1998 Jul;47(2):269-79.

29. Shields M, Shooshtari S. Determinants of self-perceived health. Health reports / Statistics Canada, Canadian Centre for Health Information 2001 Dec;13(1):35-52.

30. Manderbacka K, Lundberg O, Martikainen P. Do risk factors and health behaviours contribute to self-ratings of health? Soc Sci Med. 1999;48(12):1713-20.

31. Benyamini Y, Leventhal H. Self assessments of health: What do people know that predicts their motality? Research on Aging. 1999;21:477-500.

32. Ferraro KF, Yu Y. Body weight and self-ratings of health. Journal of health and social behavior. 1995 Sep;36(3):274-84.

33. Kempen GI, Miedema I, van den Bos GA, Ormel J. Relationship of domain-specific measures of health to perceived overall health among older subjects. J Clin Epidemiol. 1998 Jan;51(1):11-8.

34. Shields M. Community belonging and self-perceived health. Health reports / Statistics Canada, Canadian Centre for Health Information. 2008 Jun;19(2):51-60.

35. Syme SL, Berkman LF. Social class, susceptibility and sickness. Am J Epidemiol. [Review]. 1976 Jul;104(1):1-8.

36. Giron P. Determinants of self-rated health in Spain: differences by age groups for adults. European journal of public health. 2012 Feb;22(1):36-40.

37. Verropoulou G. Key elements composing self-rated health in older adults: a comparative study of 11 European countries. European Journal of Ageing. 2009;6(3):213-26.

38. Tremblay S, Dahinten S, Kohen D. Factors related to adolescents' self-perceived health. Health Rep. 2003;14 Suppl:7-16.

39. Mikolajczyk RT, Brzoska P, Maier C, Ottova V, Meier S, Dudziak U, et al. Factors associated with self-rated health status in university students: a cross-sectional study in three European countries. BMC Public Health. 2008;8:215.

40. Vaez M, Laflamme L. First-year university students' health status and socio-demographic determinants of their self-rated health. Work. 2002;19(1):71-80.

41. Vingilis E, Wade TJ, Adlaf E. What factors predict student self-rated physical health? J Adolesc. 1998 Feb;21(1):83-97.

42. Singh-Manoux A, Martikainen P, Ferrie J, Zins M, Marmot M, Goldberg M. What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies. Journal of epidemiology and community health. 2006 Apr;60(4):364-72.

43. Haddock CK, Poston WS, Pyle SA, Klesges RC, Vander Weg MW, Peterson A, et al. The validity of self-rated health as a measure of health status among young military personnel: evidence from a cross-sectional survey. Health Qual Life Outcomes. 2006;4:57.

44. Cott CA, Gignac MA, Badley EM. Determinants of self rated health for Canadians with chronic disease and disability. J Epidemiol Community Health. 1999 Nov;53(11):731-6.

45. Darviri C, Fouka G, Gnardellis C, Artemiadis AK, Tigani X, Alexopoulos EC. Determinants of self-rated health in a representative sample of a rural population: a cross-sectional study in Greece. Int J Environ Res Public Health. 2012 Mar;9(3):943-54.

46. Asfar T, Ahmad B, Rastam S, Mulloli TP, Ward KD, Maziak W. Self-rated health and its determinants among adults in Syria: a model from the Middle East. BMC Public Health. 2007;7:177.

47. Xu J, Zhang J, Feng L, Qiu J. Self-rated health of population in Southern China: association with socio-demographic characteristics measured with multiple-item self-rated health measurement scale. BMC Public Health. 2010;10:393.

48. Sun W, Watanabe M, Tanimoto Y, Shibutani T, Kono R, Saito M, et al. Factors associated with good self-rated health of non-disabled elderly living alone in Japan: a cross-sectional study. BMC Public Health. 2007;7:297.

49. Ahmad K, Jafar TH, Chaturvedi N. Self-rated health in Pakistan: results of a national health survey. BMC Public Health. 2005 May 19;5:51.

50. Fillenbaum GG. Social context and self-assessments of health among the elderly. J Health Soc Behav. 1979 Mar;20(1):45-51.

51. Benyamini Y, Leventhal EA, Leventhal H. Elderly people's ratings of the importance of health-related factors to their self-assessments of health. Soc Sci Med. 2003 Apr;56(8):1661-7.

52. Vingilis ER, Wade TJ, Seeley JS. Predictors of adolescent self-rated health. Analysis of the National Population Health Survey. Can J Public Health. 2002 May-Jun;93(3):193-7.

53. Wade TJ, Vingilis E. The development of self-rated health during adolescence: an exploration of inter- and intra-cohort effects. Can J Public Health. 1999 Mar-Apr;90(2):90-4.

54. Bosworth HB, Siegler IC, Brummett BH, Barefoot JC, Williams RB, Vitaliano PP, et al. The relationship between self-rated health and health status among coronary artery patients. J Aging Health. 1999 Nov;11(4):565-84.

55. Froom P, Melamed S, Triber I, Ratson NZ, Hermoni D. Predicting self-reported health: the CORDIS study. Prev Med. 2004 Aug;39(2):419-23.

56. Lewis R. An introduction to Classification and Regression Tree (CART) Analysis 2000 Annual Meeting of the Society for Academic Emergency Medicine; San Francisco, California2000.

57. Mantzavinis GD, Pappas N, Dimoliatis ID, Ioannidis JP. Multivariate models of self-reported health often neglected essential candidate determinants and methodological issues. J Clin Epidemiol. 2005 May;58(5):436-43.

58. Gebbie K, Rosenstock L, Hernandez L. Who Will Keep the Public Healthy? Educating Public Health Professionals for the 21st Century. Washington, DC: The National Academies Press; 2003.

59. Bronfenbrenner U. Toward an Experimental Ecology of Human Development. American Psychologist. 1977;32:513-31.

60. Dahlgren G WM. Policies and strategies to promote social equity in health. . Stockholm: Institute for Future Studies1991.

61. Kaplan G, Everson S, Lynch J. The contribution of social and behavioral research to an understanding of the distribution of disease: a multilevel approach. . In: Smedley B, Syme S, editors. Promoting Health: Intervention Strategies from Social and Behavioral Research Washington, DC: National Academy Press. ; 2000. p. 37-80.

62. McLeroy KR, Bibeau D, Steckler A, Glanz K. An ecological perspective on health promotion programs. Health Educ Q. [Review]. 1988 Winter;15(4):351-77.

63. Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR, Jr., et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. Journal of clinical epidemiology. 1988;41(11):1105-16.

64. Ware J, Jr., Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996 Mar;34(3):220-33.

65. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Boca Raton: Chapman & Hall / CRC; 1984.

66. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. Journal of Medical Systems. 2002;26(5):445-63.

67. Kershaw TS, Lewis J, Westdahl C, Wang YF, Rising SS, Massey Z, et al. Using clinical classification trees to identify individuals at risk of STDs during pregnancy. Perspect Sex Reprod Health. 2007 Sep;39(3):141-8.

68. Nelson LM, Bloch DA, Longstreth WT, Jr., Shi H. Recursive partitioning for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. Journal of clinical epidemiology. 1998 Mar;51(3):199-209.

69. Forthofer M, Bryant C. Using Audience-Segmentation Techniques to Tailor Health Behaviour Change Strategies. Am J Health Behav. 2000;24(1):36-43.

70. Lemon S, Roy J, Clark M, Friedmann P, Rakowski W. Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. Annals of Behavioural Medicine. 2003;26(3):172-81.

71. BeLue R, Francis LA, Rollins B, Colaco B. One size does not fit all: identifying risk profiles for overweight in adolescent population subsets. J Adolesc Health. 2009 Nov;45(5):517-24.

72. Syme SL. Social determinants of health: the community as an empowered partner. Prev Chronic Dis. 2004 Jan;1(1):A02.

73. Marmot M. Achieving health equity: from root causes to fair outcomes. Lancet. 2007 Sep 29;370(9593):1153-63.

74. McGee DL, Liao Y, Cao G, Cooper RS. Self-reported health status and mortality in a multiethnic US cohort. Am J Epidemiol. 1999 Jan 1;149(1):41-6.

75. Chandola T, Jenkinson C. Validating self-rated health in different ethnic groups. Ethn Health. 2000 May;5(2):151-9.

76. Friel S, Newell J, Kelleher C. Who eats four or more servings of fruit and vegetables per day? Multivariate classification tree analysis of data from the 1998

Survey of Lifestyle, Attitudes and Nutrition in the Republic of Ireland. Public Health Nutr. 2005 Apr;8(2):159-69.

77. WHO. World health report 2002. Reducing risks, promoting healthy life. Geneva: World Health Organization2002.

78. Portrait F, Lindeboom M, Deeg D. Health and mortality of the elderly: the grade of membership method, classification and determination. Health Econ. 1999 Aug;8(5):441-57.

79. van Rossum CT, Shipley MJ, van de Mheen H, Grobbee DE, Marmot MG. Employment grade differences in cause specific mortality. A 25 year follow up of civil servants from the first Whitehall study. J Epidemiol Community Health. 2000 Mar;54(3):178-84.

80. Link BG, Phelan J. Social conditions as fundamental causes of disease. Journal of health and social behavior. [Review]. 1995;Spec No:80-94.

81. Marmot M. Fair society, healthy lives: strategic review of health inequalities in England post-2010. The Marmot Review. London2010.

82. Adler N, Boyce T, Chesney M, Cohen S, Folkman S, Kahn R, et al. Socioeconomic Status and Health: The Challenge of the Gradient. American Psychologist. 1994;49(1):15-24.

83. Dahlgren G, Whitehead M. European Strategies for tackling social inequities in health: levelling up, Part 2. . Copenhagen: WHO Regional Office for Europe2007.

84. Marmot MG, Fuhrer R, Ettner SL, Marks NF, Bumpass LL, Ryff CD. Contribution of psychosocial factors to socioeconomic differences in health. Milbank Q. 1998;76(3):403-48, 305.

85. Adler NE, Stewart J. Health disparities across the lifespan: meaning, methods, and mechanisms. Annals of the New York Academy of Sciences. 2010 Feb;1186:5-23.

86. DeNavas-Walt C, Proctor BD, Smith JC. Income, Poverty, and Health Insurance Coverage in the United States: 2010. P60-239. Washington, DC: U.S. Census Bureau2011.

87. Wen M, Hawkley LC, Cacioppo JT. Objective and perceived neighborhood environment, individual SES and psychosocial factors, and self-rated health: an analysis of older adults in Cook County, Illinois. Social science & medicine. 2006 Nov;63(10):2575-90.

88. Morello-Frosch R, Zuk M, Jerrett M, Shamasunder B, Kyle AD. Understanding the cumulative impacts of inequalities in environmental health: implications for policy. Health Aff (Millwood). 2011 May;30(5):879-87.

89. WHO Task Force on Research Priorities for Equity in Health and the WHO Equity Team. Priorities for research to take forward the health equity policy agenda.2005.

90. Lynch J, Kaplan G. Socioeconomic Position. In: Berkman L, Kawachi I, editors. Social Epidemiology. New York: Oxford University Press; 2000. p. 13-35.

91. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

92. Goldman L, Cook EF, Johnson PA, Brand DA, Rouan GW, Lee TH. Prediction of the need for intensive care in patients who come to the emergency departments with acute chest pain. The New England journal of medicine. 1996 Jun 6;334(23):1498-504.

93. Zhang H, Singer BH. Recursive Partitioning and Applications. 2nd Edition ed. New York: Springer Science+Business Media; 2010.

94. Marshall RJ. The use of classification and regression trees in clinical epidemiology. Journal of clinical epidemiology. 2001 Jun;54(6):603-9.

95. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J Chronic Dis. 1984;37(9-10):721-31.

96. Strobl C, Malley J, Tutz G. An Introduction to Recursive Partitioning. Technical Report Number 55: Department of Statistics, University of Munich2009.

97. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet. 2010;11:49.

98. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9:307.

99. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009 Jul;20(4):488-95.

100. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2(3):18-22.

101. Sutton CD. Classification and regression trees, bagging, and boosting. In: Rao CRea, editor. Handbook of Statistics: Data Mining and Data Visualization. Amsterdam, the Netherlands: Elsevier Publishing; 2005. p. 303-29.

102. Prasad AM, Iverson LR, Liaw A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems. 2006;9:181-99.

103. Sudat SE, Carlton EJ, Seto EY, Spear RC, Hubbard AE. Using variable importance measures from causal inference to rank risk factors of schistosomiasis infection in a rural setting in China. Epidemiol Perspect Innov. 2010;7:3.

104. Hubbard AE, Laan MJ. Population intervention models in causal inference. Biometrika. 2008;95(1):35-47.

105. Ritter SJ, Jewell N, Hubbard AE. Variable Importance Analysis with the multiPIM R Package. UC Berkeley Division of Biostatistics Working Paper Series 2011.

106. Choosing Health: Making healthy choices easier. Public Health White Paper. In: (UK) DoH, editor. London: The Stationery Office; 2004.

107. Grier S, Bryant CA. Social marketing in public health. Annu Rev Public Health. 2005;26:319-39.

108. Breiman L. Statistical Modeling: The Two Cultures. Statistical Science. 2001;16(3):199-231.