# SCORING AND RELATIVE RISK ANALYSIS IN NUTRITION AND PHYSICAL ACTIVITY

A Dissertation

by

ELI SAMUEL KRAVITZ

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Raymond J. Carroll |
| Committee Members, | Tanya P. Garcia |
| | N. Sivakumar |
| | Suojin Wang |
| Head of Department, | Jianhua Huang |

August 2019

Major Subject: Statistics

# ABSTRACT

This work presents three analyses of the NIH-AARP Study of Diet and Health. Each analysis develops recommendations for nutritional intake or physical behaviors, or alters existing recommendations. New statistical methodology for nonlinear and nonparametric regression is introduced. Each methodology results in consistent estimation of relative risk of disease (cancer, mortality, etc.). Technical details and proofs are collected in separate appendices for each chapter.

First, a major collaborative project to create a composite scoring system for physical activity is presented. A scoring system allows quick assessment of physical activity levels which can then be used to estimate disease risk. This score, denoted Physical Behavior Score (PBS), is verified to predict mortality using a subset of the NIH-AARP Study of Diet and Health withheld for validation. The Physical Behavior Score is highly predictive of mortality. Women in the highest quintile of scores had a 54% reduction in all-cause mortality risk, and men in the highest quintile had a 45% reduction in all-cause mortality risk.

Next, the Healthy Eating Index is used as a case study to provide a general method for reevaluating composite scores. The Healthy Eating Index breaks nutritional intake into 12 components. A method is presented that can be used to reassess the relative importance of these components using a weighted logistic regression model applied across many populations and diseases. Variable selection is performed by taking an asymptotic approximation and adding an adaptive Lasso penalty. This approximation simplifies variable selection into a simple least squares minimization. Oracle properties of this variable selection technique are established, which is different from the usual one population and one disease context.

Finally, the problem of the first chapter in which a physical activity score is created then applied to analysis of disease or mortality is revisited. Sample splitting is used to partition the sample into two disjoint subsets, using the first subset to build the score and using the remaining data to estimate relative risk of this score. For parametric models, the limiting distribution of risk estimates is derived. An obvious question is what happens if multiple sample splits are performed.

It is shown that as the number of sample splits increases, the combination of multiple sample splits is effectively equivalent to performing no sample splits. This suggests there is no clear benefit to performing multiple splits.

# ACKNOWLEDGMENTS

This work would not be possible without my advisor Dr. Raymond Carroll. I am lucky to have him as a mentor and friend. I have grown as a statistician, and as a person, thanks to his guidance and encouragement. I will miss having a resident statistical expert down the hall from me, but I will miss my friend and mentor even more.

I worked with a wonderful group of collaborators on material in Chapter 2. In particular, I would like to thank Dr. Sarah Kozey-Keadle, who I worked with closely from the inception of the project. Dr. Keadle taught me so much about her field and how to present complicated statistical analyses to a general audience. Of everything I worked on during my PhD, I am most proud of this paper.

Dr. David Ruppert was helpful in developing the material in Chapter 4. Working with Dr. Ruppert was one of my favorite experiences in graduate school. Dr. Ruppert taught me how to make an educated guess, test it with computer simulations, and then refine my guess. This is one of the most useful skill I learned during my studies.

I would like to thank my committee members: Dr. Tanya Garica, Dr. Suojin Wang, and Dr. Sivakumar. Dr. Garica was a constant source of encouragement. Dr. Wang was one of the first faculty member I met when I moved here, and I'm glad we have remained close through my PhD. Dr. Sivakumar managed to push some mathematics into my brain.

I am thankful for the many friends I have made at Texas A&M. Some of them are still here, and many of them have moved away. My life is better for having known them. I have shared an office with Yabo Niu for years, and it will be strange not seeing him every day.

Kim Ritchie has funded me as a Technology Teaching Assistant (TTA) for my entire time at Texas A&M. I appreciate having a supervisor I could talk and joke with, and I appreciate how invested she is in all the TTA's.

Dr. Longnecker provided me constant guidance and support, especially in the early stages of my PhD. I made more trips to his office than I can possibly count. My time in graduate school would not have been the same without him.

I am thankful for my family who have supported me during this time of my life. In particular I would like to thank my parents, my sister Hannah, and my aunt Paula. Hannah, I hope we are both Dr. Kravitz soon.

I am thankful to my new family in Texas: Brittney and Franklin the dog. With you rural Texas feel like New York City. I look forward to the next step in our lives.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supported by a dissertation committee consisting of Professor Raymond Carroll as committee chair, Professors Tanya Garcia and Suojin Wang of the Department of Statistic, and Professor N. Sivakumar of the Department of Mathematics

All data analyzed in this dissertation was provided by Dr. Raymond Carroll. Dr. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030). The work depicted in Chapter 2 was published in 2019 in an article listed in the footnote of Chapter 2. The work in Chapter 4 was done under the joint supervision of Professor David Ruppert of the Department of Statistical Science at Cornell University and Raymond Carroll.

All other work conducted for the dissertation was completed independently.

## Funding Sources

Graduate study was supported by a fellowship from Texas A&M University and from employment as a Technology Teaching Assistant in the Department of Statistics.

## TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES vi

TABLE OF CONTENTS vii

LIST OF FIGURES x

LIST OF TABLES xi

1. INTRODUCTION 1

2. DEVELOPMENT AND TESTING OF AN INTEGRATED SCORE FOR PHYSICAL BEHAVIORS 5

  2.1 Introduction 5
  2.2 Methods 6
  2.3 End Point Ascertainment and Covariate Assessment 8
  2.4 Statistical Analyses 8
  2.5 Covariates 9
  2.6 Development of the PBS 10
  2.7 Example Physical Behavior Score 11
  2.8 Predictive Validity of PBS 12
  2.9 Results 12
  2.10 Discussion 14
  2.11 Conclusions 18
  2.12 Technical Description of Score Development 19

3. RE-EVALUATING COMPOSITE SCORES: ADAPTIVE LASSO VARIABLE SELECTION FOR NONLINEAR MODELS 27

  3.1 Introduction 27
  3.2 A Nonlinear Model Across Populations 28
  3.3 Variable Selection 31
    3.3.1 Least Squares Approximation and Adaptive Lasso 31
    3.3.2 Model Selection and Oracle Properties 32
  3.4 Data Analysis 35
    3.4.1 Background 35

3.4.2 Results ........ 37
3.5 Simulations ........ 38
3.5.1 Subset Analysis ........ 38
3.5.2 Variable Selection and Coverage ........ 38
3.6 Discussion ........ 41

4. SAMPLE SPLITTING AS AN M-ESTIMATOR WITH APPLICATION TO PHYSICAL ACTIVITY SCORING ........ 42

4.1 Introduction ........ 42
4.2 Motivating Example: Physical Activity and Survival ........ 43
4.3 Sample Splitting Methodology ........ 44
4.3.1 Basic Formulation ........ 44
4.3.2 A Single Split ........ 45
4.3.3 Many Splits ........ 46
4.4 Asymptotic Theory ........ 47
4.4.1 Single Split Asymptotics ........ 47
4.4.2 Multiple Split Asymptotics ........ 48
4.4.3 Increasing Number of Splits Asymptotics ........ 48
4.5 Simulations ........ 49
4.6 Data Analysis ........ 50
4.6.1 Step 1: Developing the Score ........ 51
4.6.2 Step 2: Rescaling the Score ........ 53
4.6.3 Step 3: Risk Prediction Based on the Score ........ 54
4.7 Discussion ........ 56

5. SUMMARY ........ 57

REFERENCES ........ 58

APPENDIX A. SUPPLEMENTARY MATERIAL FOR CHAPTER 2 ........ 65

A.1 R Code to Generate Physical Behavior Score ........ 65
A.2 Testing Proportional Hazards Assumption ........ 65
A.3 Physical Activity Questionnaire and Met Conversion ........ 65

APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER 3 ........ 70

B.1 R Code and Data ........ 70
B.2 Proof of Theorem 1 ........ 70
B.3 Proof of Theorem 2 ........ 71
B.4 Proof of Theorem 3 ........ 72
B.5 Variance Calculation of Rescaled Coefficients ........ 72

APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER 4 ........ 74

C.1 R Code and Data ........ 74

C.1.1 Assumptions .......... 74
C.1.2 Taylor Expansions from Sections 4.4.1 and 4.4.2 .......... 75
C.1.3 Theory for Infinite Splits from Section 4.4.3 .......... 76
C.1.4 Asymptotic Equivalence to Stacked Estimating Equations .......... 77

# LIST OF FIGURES

FIGURE Page

2.1 Shape constrained additive regression (SCAR) returns 8 sets of predicted values, one for each function that is fit to describe the relationship between physical activity and mortality. .......... 22

2.2 The relative hazard of all-cause mortality plotted from the 5th percentile, 53.5, to the maximum score of 100, separately for men and women. .......... 23

2.3 Stratified analysis of Physical Behavior Score in relation to all-cause mortality .......... 24

2.4 Unscaled shape constrained functions .......... 25

2.5 Distribution of physical behavior scores in the NIH-AARP Diet and Health Study Cohort, 2004-2006 .......... 26

4.1 Three of the eight marginal model plots from the binary regression model. .......... 52

4.2 Log odds ratio of the physical activity score as the number of sample splits increases 55

A.1 Schoenfeld residuals plotted against time. A straight line indicates that the coefficients fit with Cox regression do not change with time. .......... 66

A.2 Log Hazard, equivalent to log of negative log of Survival time, against time. .......... 67

A.3 NIH-AARP Study of Diet and Health questionnaire for physical activity. .......... 68

A.4 NIH-AARP Study of Diet and Health questionnaire for sedentary behaviors .......... 69

# LIST OF TABLES

TABLE Page

2.1 Description and pairwise correlations for each of the physical behavior score components in the NIH-AARP Diet and Health Study Cohort, 2004-2006. Rows A-E are measured in MET-hr/wk, and Rows F-H are measured in hr/day. Reprinted with permission from Wolters Kluwer Health Inc ...... 9

2.2 Reported physical behaviors and corresponding Physical Behavior Score for maximum, median, and two hypothetical people in the NIH-AARP Diet and Health Study Cohort, 2004-2011 ...... 11

2.3 Baseline Participant Characteristics and Physical Behavior Components by quintile of physical behavior score in the NIH-AARP Diet and Health Study Cohort, 2004-2006. ...... 14

2.4 Comparison of physical behavior score to groups components for all-cause and cause-specific mortality. ...... 15

3.1 Description of the Healthy Eating Index 2005 scoring system ...... 29

3.2 Summary of the NIH-AARP data for cancer occurrence. ...... 35

3.3 Summary of the NIH-AARP data for mortality. Cardiovascular disease has been abbreviated as CVD. ...... 36

3.4 Results from Section 3.4.2 when the outcome of interest is cancer occurrence, various types of mortality, and aggregated all-cause mortality ...... 37

3.5 Results of the subset analysis in Section 3.5.1 when all-cause mortality is the outcome of interest. ...... 39

3.6 Simulation results from Section 3.5.2 showing coverage and the percent of time the true model was selected. ...... 40

4.1 Results of the simulation in Section 4.5 using the linear model. ...... 49

4.2 Results of the simulation in Section 4.5 using the logistc model. ...... 50

4.3 Each of the 8 physical activity variables with their expected relationship and marginal model when predicting survival. ...... 51

4.4 Example physical activity score developed using half of the data ...... 54

A.1 Physical activity variables are expressed as MET-hrs per week, which is the MET-value for each activity multiplied by the reported hours per week. . . . . . . . . . . . . . . . . . . . . . 69

# 1. INTRODUCTION

The work in this thesis presents several analyses of the NIH-AARP Study of Diet and Health Schatzkin et al. (2001) and the new methodology developed to perform these analyses. Each chapter develops a *composite scores* or alters an existing composite score. A composite score compares a person's health behavior to an idealized standard and assigns them a number, usually between 0 and 100, to indicate their compliance to this health behavior. A composite score of 100 indicates perfect compliance while a score of 0 indicates poor compliance. This measure of health behavior is applied across populations and health outcomes (colon cancer, breast cancer, etc.) to create a single interpretable score. This single measure of health is used to estimate disease risk, rather than using the health behavior directly. Composite scores simplify statistical analysis by converting a complicated health behavior into a single continuous covariate. Composite scores are used extensively in nutritional epidemiology (Guenther et al., 2008a, 2013a; Panagiotakos et al., 2006; McCullough et al., 2002).

The new methodology in this dissertation focuses on regression models of the form

$$g\{E(Y_i|X_i,Z_i)\} = \beta_0(X_i^{\mathrm{T}}\theta) + Z_i^{\mathrm{T}}\beta, \tag{1.1}$$

where $g(\cdot)$ is a known link function, $Y_i$ is a response of interest which is generally binary, $(X_i, Z_i)$ are two types of covariates, and $(\theta, \beta)$ are unknown parameters. The $L^2$ norm of $\theta$ is constrained so $\beta$ and $\theta$ are both identifiable, though other constraints are possible. Model 1.1 is a very general nonlinear model which includes a range of statistical models as specific cases. $\mathbf{X}$ and $\theta$ define a shape-constrained spline model in Chapter 2, a weighted logistic regression model in Chapter 3, and a combination of 4-parameter logistic functions, linear functions, and quadratic functions in Chapter 4. In each application, $\beta_0$ estimates the relative risk of disease. Consistent estimation of $\beta_0$ and asymptotically consistent variance estimates are emphasized.

Chapter 2 presents collaborative work to develop a composite score for physical activity. Using the NIH-AARP Study of Diet and Health, physical activity is discretized into eight components. A Generalized Additive Model (Hastie and Tibshirani, 1986; Wood, 2017) is proposed to described the relationship between physical activity and survival. That is,

$$\mathrm{pr}(Y_i = 1, X_i, Z_i) = H\{\textstyle\sum_{j=1}^{J} f_j(X_{ij}) + Z_i^{\mathrm{T}}\beta\}, \tag{1.2}$$

where $H(\cdot)$ is the logistic distribution function, $X_{ij}$ are physical activity measurement, $f_j$ is an unknown smooth function, $Z_i$ is a vector of covariates, and $\beta$ is an unknown parameter. We have additional information that can be incorporated into our model: the relationship between the eight components and survival have well established functional forms in the physical activity literature. For example, kinesiologists have shown the relationship between vigorous activity and survival is relatively concave and non-decreasing. That is, vigorous activity is beneficial to overall health but the positive benefits eventually plateau. Using the methods of Chen and Samworth (2016), we use this information to enforce shape constraints on each of the smooth functions in the Generalized Additive Model so the fitted functions, $\widehat{f}_j(\cdot)$, are consistent with existing literature. The functions are not allowed to vary freely, but instead must satisfy a specified shape constraint.

After the shape constrained additive model is fit, the fitted values are transformed to be between 0 and 100 to put these fitted values on a comparable scale to other composite scores. People with a high probability of survival are assigned a score close to 100 and people with a low probability of survival are assigned a score close to 0. This new composite score is verified to predict mortality in an independent population using a Cox proportional hazards model.

Chapter 3 gives a general methodology for reevaluating composite scores through weighted logistic regression and variable selection. The 2005-Health Eating Index Guenther et al. (2008a) is used as an example to motivate the methods. A previous analysis by Reedy et al. (2008) for predicting colorectal cancer uses the model

$$\mathrm{pr}(Y_i = 1 | X_i, Z_i) = H(\beta_0 \textstyle\sum_{j=1}^{J} X_{ij} + Z_i^{\mathrm{T}}\beta), \tag{1.3}$$

where $Y_i = 1$ denotes occurrence of colorectal cancer, $Z_i$ is a vector of covariates, $X_{ij}$ is the Healthy Eating Index score for a particular dietary components, $\sum_{j=1}^{J} X_{ij}$ is the combined Healthy Eating Index score, and $(\beta_0, \beta)$ is a vector of unknown parameters. $\beta_0$ estimates the relative risk of colorectal cancers as a function of Healthy Eating Index scores.

Model 1.3 is extended by adding parameters which allow the effect of diet to vary by population and disease. Model 1.3 becomes,

$$\text{pr}(Y_{ik\ell} = 1 | X_{ikj}, Z_{ik\ell}) = H(\beta_{k\ell} \textstyle\sum_{j=1}^{J} X_{ijk\ell} + Z_{ik\ell}^{\mathrm{T}} \theta_{k\ell}), \tag{1.4}$$

where $k$ denotes population and $\ell$ denotes disease. Here $\beta_{k\ell}$ gives the relative risk of disease $\ell$ in population $k$ as a function of a person's Healthy Eating Index score. Next, a vector of weights, $\theta$, is introduced to (1.4) to allow the relative importance of each nutritional component to vary. Model 1.4 becomes

$$\text{pr}(Y_{ik\ell} = 1 | X_{ikj}, Z_{ik\ell}) = H(\beta_{k\ell} \textstyle\sum_{j=1}^{J} X_{ijk\ell} \theta_j + Z_{ik\ell}^{\mathrm{T}} \theta_{k\ell}), \tag{1.5}$$

If $\theta_j > 1$ the relative importance of the $j^{th}$ dietary component is increased, and if $\theta_j < 1$ the relative importance is decreased. Finally, we add an adaptive Lasso penalty (Zou, 2006) to the likelihood associated with (1.5). The penalty is only on $\theta$. This is used to perform variable selection to see if any nutritional components are unnecessary in predicting disease risk.

Chapter 3 considers verifying a composite score is predictive of a disease without an independent population for validation. To alleviate concerns of overfitting, Model (1.1) is fit by sample splitting: first estimating $\theta$ on a portion of sample then estimating $\beta$ using $\widehat{\theta}$ and the remaining sample. The analysis from Chapter 2 is used as a case study, with the shape constrained model replaced with a parametric model. Sample splitting is used to first build the score on half the dataset, and then separately evaluate its use in disease or mortality risk.

Sample splitting is formalized with estimating equations and M-estimator theory (Huber, 1964, 1967; Stefanski and Boos, 2002). Assume two responses $(\mathcal{W}_i, \mathbf{Y}_i)$ are observed and two covariate vectors are observed, $(\mathbf{X}_i, \mathbf{Z}_i)$. There are two parameters, $\theta$ and $\beta$. The response $\mathcal{W}$ and $(\mathbf{X}, \mathbf{Z}, \theta)$ are related by the estimating equation $\Psi(\mathcal{W}, \mathbf{X}, \mathbf{Z}, \theta)$. The response $\mathbf{Y}$ and $(\mathbf{X}, \mathbf{Z}, \beta, \theta)$ are related $\mathcal{K}(\mathbf{Y}, f(\mathbf{X}; \theta), \mathbf{Z}, \beta)$, where $f(\cdot;\cdot)$ is a known function. Assume $\delta = (\delta_1, ..., \delta_n)$ is a vector of independent and identically distributed Bernoulli$(\pi)$. $\delta$ is used as an indicator variable to denote which portion of the sample is used to estimate $\theta$ and which portion is used to estimate $\beta$. The parameter $\theta$ is estimated by solving

$$0 \;\;=\;\; \textstyle\sum_{i=1}^{n} \delta_i \Psi(\mathcal{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \theta),$$

and $\beta$ is estimated by solving

$$0 \;\;=\;\; \textstyle\sum_{i=1}^{n} (1-\delta_i)\mathcal{K}(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \beta, \widehat{\theta}).$$

This is extended to the case of many splits. The estimates of $\theta$ and $\beta$ are combined with the sample mean. The limiting distribution of $(\widehat{\theta}, \widehat{\beta})$ is derived. From this confidence intervals and hypothesis tests can be constructed.

# 2. DEVELOPMENT AND TESTING OF AN INTEGRATED SCORE FOR PHYSICAL BEHAVIORS [*]

## 2.1 Introduction

Accumulating evidence indicates that a variety of physical behaviors affect health in positive and negative ways. A physically active lifestyle, including time spent in moderate and vigorous physical activities and strength training, improves longevity and quality of life and decreases risk of many chronic diseases. (Office of Disease Prevention and Health Promotion, 2019) Mortality risk is also inversely associated with light intensity activity (Katzmarzyk, 2014) positively associated with time spent in sedentary behaviors (The Obesity Society et al., 2016; Grøntved and Hu, 2011) and with both too little and too much sleep (Xiao et al., 2014; Ma et al., 2016).

Analyzing the independent vs. combined effects of several different physical behaviors on health outcomes is challenging. Studies have historically focused on isolating an independent behavior while adjusting for other behaviors – for example, does the association between sedentary time and mortality risk persist when adjusting for time spent in moderate-vigorous intensity physical activity (Ekelund et al., 2016)? More recently, recognizing time trade-offs among sleep, sedentary behavior, and physical activity, researchers have examined the effect of substituting time spent in one behavior with another using isotemporal substitution or compositional analyses (Mekary et al., 2013; Chastin et al., 2015) – for example, does replacing one hour of sitting with one hour of light-intensity activity result in mortality benefits (Mekary et al., 2013; Whitaker et al., 2017; Stamatakis et al., 2015; Matthews et al., 2015, 2016)? Both of these approaches provide insight into the behavior-disease relationship, but neither quantifies the joint impact of different physical behaviors on health and longevity.

Studies in nutritional epidemiology demonstrate the utility of indices to characterize overall

[*] Reprinted with permission from Wolters Kluwer Health Inc: Medicine & Science in Sports & Exercise, "Development and Testing of an Integrated Score for Physical Behaviors", Advance online publication, 2019, by Keadle, Sarah Kozey, Kravitz, Eli S., Matthews, Charles, Tseng, Marilyn, and Carroll, Raymond J. In accordance Wolters Kluwer's copyright no modifications have been made except formatting.

patterns of behavior (Guenther et al., 2013b, 2008b,a; Hu et al., 2000), but to date there have been few attempts to integrate multiple physical behaviors into a single index. Investigators for the European Investigation into Cancer and Nutrition study developed an index that included physical activity at work, sport, cycling, television viewing and computer use by selecting the self-reported activity variables that were the most strongly associated with accelerometer counts Cust et al. (2008). Another study used National Health and Nutrition Examination Survey data to examine the cross-sectional relationship of a physical index that included combined moderate-vigorous physical activity, sitting time, grip strength and estimated fitness with various health-outcomes Keadle et al. (2017). Both approaches relied on simple scoring systems that do not account for established non-linear relationships of health with aerobic activity and sedentary time Arem et al. (2015); Matthews et al. (2016), and neither method considered sleep duration despite its established association with health Yin et al. (2017). Moreover, to our knowledge, no previous studies have examined the prospective association of such an index with mortality.

There is a need for new approaches to analyze the joint effects of distinct but inter-related physical behaviors with health outcomes. We developed a physical behavior score (PBS) (ranging from 0-100) that integrated different types and intensities of physical activity, sitting and sleep behaviors. To demonstrate the predictive validity of the new score, we estimated the relationship between it and mortality from any-cause and sub-types including cardiovascular disease, cancer and other mortality in a large prospective cohort of American adults.

## 2.2 Methods

The National Institutes of Health-AARP Diet and Health Study cohort was established in 1995-1996, when 566,398 AARP members (50–71 yr) in six states and two metropolitan areas responded to a questionnaire about their medical history, diet, and demographics Schatzkin et al. (2001). Between 2004-2006, 313,835 participants completed a follow-up questionnaire that asked detailed questions about active and sedentary behaviors, medical history, and risk factors. Those eligible for this analysis (N =163,016) personally responded to both questionnaires, were free of major diseases at the start of follow-up (2004–2006), and had sufficiently complete exposure data. Specific reasons for exclusions are as follows: questionnaire was completed by proxy respondents (N=18,600), and at the time of follow-up questionnaire, self-rated poor health (N=38,550), pre-existing degenerative or chronic diseases (e.g., Parkinson's, end stage renal disease) (N= 22,475), or missing primary exposure data (i.e., physical activity, sleep duration or sedentary time) (N= 71,194). Questionnaire completion was considered to imply informed consent and the U.S. National Cancer Institute's Special Studies institutional review board approved the study.

The physical activity questionnaire asked how much time per week was spent in 16 activities during the past 12 months (Figure A.3). Activities were classified as exercise and sports (8 questions) or as non-exercise activities (8 questions), which included household chores and lawn and garden activities. For each of the physical activity questions, response options were: none, 5 min, 15 min, 30 min, 1 hour, 1.5, 2-3, 4-6, 7-10, >10 hours/wk. The energy cost of each activity was assigned using standard methods, and physical activity energy expenditure was calculated (MET-hrs/d). Three sitting questions asked about the number of hours spent "in a typical 24-hour period during the past 12 months", with eight possible response options: None, $<3$, 3-4, 5-6, 7-8, 9-10, 11-12, or $\geq 12$ hours/day (Figure A.4). The exercise items have been validated against physical activity diaries, (r=0.62 and 0.65) (Chasan-Taber et al., 1996; Wolf et al., 1994). Estimates of physical activity from the survey have been correlated with total energy expenditure as assessed by doubly labeled water (r=0.33) and estimates of sitting time were significantly, although weakly, correlated with activPAL accelerometer (r=0.16) Matthews et al. (2018)

For this analysis, survey responses were classified into one of eight physical behaviors (including five types of physical activity, two types of sedentary behavior, and sleep duration) as follows:

1. Light household activity (MET-hrs/wk; 1 question): cooking, cleaning, laundry, dusting.

2. Moderate-vigorous household activity (MET-hrs/wk; 6 questions): household chores (e.g., vacuuming), moderate outdoor chores (e.g., weeding), vigorous outdoor chores (e.g., carrying lumber), home repairs (e.g., painting), caring for children, caring for another adult.

3. Moderate Exercise (MET-hrs/wk; 3 questions): walking for exercise, walking for other daily

activities, playing golf.

4. Vigorous Exercise (MET-hrs/wk; 5 questions): tennis, swimming laps, bicycling, jogging, other aerobic exercise.

5. Weight training (MET-hrs/wk; 1 question): weight training using free weights and machines.

6. Sitting watching television, video or DVD (hrs/day; 1 question).

7. Other sitting (hrs/day; 1 question): reading, knitting, using a computer.

8. Sleeping at night or napping during the day (hrs/day; 1 question).

For each category, extreme values ($>$ 95th percentile) were truncated to the 95th percentile, plus random error.

### 2.3 End Point Ascertainment and Covariate Assessment

Vital status was determined through linkage with the Social Security Administration Death Master File and the National Death Index. The primary end points for our analysis were mortality from all causes, and cause-specific mortality. Cause specific mortality was assigned using International Classification of Diseases, 10th revisions (ICD-10) codes. We categorized cancer mortality (C00-C44, C45.0, C45.1, C45.7, C45.9, C48-C97, and D12-D48). Cardiovascular disease mortality included ICD-10 coded I00-I09, I10-I13, I20-I51, I60-I69 and I70-I78. Remaining causes of death were categorized as other-causes. Mortality follow-up was through December 31, 2011. Demographic characteristics (sex, race/ethnicity, education) were assessed on the baseline questionnaire and other covariates (age, smoking, body mass index [BMI)] based on self-reported height and weight, health status and disease history) were based on values reported on the follow-up questionnaire.

### 2.4 Statistical Analyses

Overall approach to development and testing. First, we analyzed descriptive statistics for each of the eight physical behaviors and determined Spearman correlations among them (Table 2.1).

Overall, the correlations between components were weak, but statistically significant, with the exception of vigorous exercise and weight training (R=0.43). The next highest correlations were for moderate-vigorous household activity and moderate exercise (R=0.28) and light-intensity household activity (R=0.28). To develop the PBS, we took a data-driven approach by using generalized additive models Hastie and Tibshirani (1986), adjusting for covariates, to quantify the relationship between each physical behavior and survival. We then rescaled the fitted probabilities to produce an overall score ranging from 0-100. Finally, we examined predictive validity of the composite PBS. Specific details of the analyses are described below.

| Physical Behavior | Median | (25th, 75th) | Range | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A. Moderate exercise | 13.7 | (7.2, 28.8) | (0.0, 68.8) | 1 | | | | | | | |
| B. Vigorous exercise | 0.00 | (0.0, 12.5) | (0.0, 50.0) | 0.15 | 1 | | | | | | |
| C. Light-intensity Household | 7.0 | (3.3, 21.3) | (0.0, 43.2) | 0.20 | 0.04 | 1 | | | | | |
| D. Moderate-vigorous Household | 13.3 | (5.4, 28.8) | (0.0, 83.0) | 0.28 | 0.08 | 0.28 | 1 | | | | |
| E. Weight Training | 0.0 | (0.0, 0.0) | (0.0, 9.7) | 0.14 | 0.43 | -0.01 | 0.03 | 1 | | | |
| F. Sitting other than TV | 3.0 | (3.0, 5.0) | (0.0,12.0) | 0.03 | 0.02 | 0.06 | -0.03 | 0.01 | 1 | | |
| G. Television sitting | 3.5 | (1.5, 3.5) | (0.0, 8.5) | -0.05 | -0.10 | 0.00 | -0.06 | -0.09 | 0.04 | 1 | |
| H. Sleep | 7.5 | (5.5, 7.5) | (0.0, 10.5) | 0.01 | 0.01 | -0.01 | 0.00 | 0.02 | 0.01 | 0.02 | 1 |

Table 2.1: Description and pairwise correlations for each of the physical behavior score components in the NIH-AARP Diet and Health Study Cohort, 2004-2006. Rows A-E are measured in MET-hr/wk, and Rows F-H are measured in hr/day. Reprinted with permission from Wolters Kluwer Health Inc

### 2.5 Covariates

Covariates were selected to be consistent with previous analyses of mortality and physical activity and sedentary time in this dataset Matthews et al. (2015). The fully adjusted models for both the score development and predictive validity were adjusted for the following covariates: age (yr), sex, education (<12 y, high school graduate, some college, college graduate, unknown), smoking history (never, stopped 10+ yr, stopped 5–9 yr, stopped 1–4 yr, stopped <1 yr, current smoker, unknown), race/ethnicity (Non-Hispanic White, Non-Hispanic Black, Other, unknown), overall health (excellent, very good, good, fair, unknown), body mass index (<25, 25–29.9, 30+ kg/$m^2$, unknown), physician diagnosed depression (yes, no, or missing), physician diagnosed heart disease (yes, no, missing).

## 2.6 Development of the PBS

A technical description of the score development is available in Section 2.12. We randomly selected half of the sample to develop the PBS. To obtain the scores for each of the eight physical behavior variables, we fit a logistic regression with survival (non-mortality) as the outcome, using a non-parametric Generalized Additive Model (Hastie and Tibshirani, 1986) using a randomly selected half of the sample. Based on previous research, the shape of the relationship between each physical behavior and mortality was incorporated into the model using the Shape Constrained Additive Regression(SCAR) method of Chen and Samworth (2016)(SCAR package; https://cran.r-project.org/web/packages/scar/index.html). Specifically, the dose-response relationship between aerobic activity and survival is concave and increasing (Arem et al., 2015), sedentary time is concave and decreasing, and hours of sleep is concave (Grøntved and Hu, 2011; Yin et al., 2017; Prince et al., 2014). Models were adjusted for the covariates described above.

Maximum logit scores from each of the eight behavior-survival functions were summed to produce a maximum total score, such that components more strongly associated with survival contributed more towards the total score (Figure 2.4). This total was rescaled to range from 0-100, with 0 being highest risk and 100 being lowest risk, to be consistent with composite scores from other fields, such as the Healthy Eating Index (Guenther et al., 2008a) Sedentary behavior values were reverse coded because more sedentary time decreases survival probability. Because of its inverted U-shaped relationship with survival, scores for sleep duration were scaled such that that long-sleepers (>10 hours/d) and those whose reported no hours of sleep received a 0 score. Figure 2.1 shows plots for the rescaled values for each of the eight components, with y-axes indicating the relative importance of each physical behavior to the overall PBS. Table 2.2 shows the final scoring system, and Figure 2.5 shows the histogram of the total score across the data set.

|  | Maximum Score |  | Median Values |  |
|---|---|---|---|---|
| **Physical Behavior** | Criteria for Max | Max PBS | Reported behavior | Assigned PBS |
| Moderate Exercise | >50 MET-hrs/wk | 32 | 13.7 MET/hrs | 27.25 |
| Vigorous Exercise | >20 MET-hrs/wk | 10 | 0 MET/hrs | 0 |
| Light Household Activity | >3 MET-hrs/wk | 3 | 7 MET/hrs | 2.47 |
| MVPA Household Activity | >20 MET-hrs/wk | 25 | 13.25 MET/hrs | 22.9 |
| Weight Training | >2 MET-hrs/wk | 3 | 0 MET/hr | 0 |
| Sitting Other than TV | 0 hrs/day | 5 | 3 hrs/day | 4.66 |
| Hours of TV Sitting | 0 hrs/day | 14 | 3.5 hrs/day | 10.32 |
| Hours of Sleep | 7.5 hrs/day | 8 | 7.5 hrs/day | 7.54 |
| Total Score (quintile) |  | 100 (Q5) |  | 75.14 (Q3) |

|  | Person A |  | Person B |  |
|---|---|---|---|---|
| **Physical Behavior** | Reported behavior | Assigned PBS | Reported behavior | Assigned PBS |
| Moderate Exercise | 8 MET/hrs | 24.36 | 8 MET/hrs | 24.36 |
| Vigorous Exercise | 0 MET/hrs | 0 | 0 MET/hrs | 0 |
| Light Household Activity | 3.5 MET/hrs | 2.42 | 3.5 MET/hrs | 2.42 |
| MVPA Household Activity | 4 MET/hrs | 17.9 | 4 MET/hrs | 17.9 |
| Weight Training | 0 MET/hr | 0 | 0 MET/hr | 0 |
| Sitting Other than TV | 3 hrs/day | 4.66 | 5 hrs/day | 4.11 |
| Hours of TV Sitting | 3.5 hrs/day | 10.32 | 7 hrs/day | 3.1 |
| Hours of Sleep | 7 hrs/day | 7.54 | 6 hrs/day | 6.5 |
| Total Score (quintile) |  | 69.94 (Q2) |  | 61.12 (Q1) |

Table 2.2: Reported physical behaviors and corresponding Physical Behavior Score for maximum, median, and two hypothetical people in the NIH-AARP Diet and Health Study Cohort, 2004-2011. Person A meets guidelines NIH guidelines with average time spent sitting and Person B meets NIH guidelines with high reported sedentary behavior.
**Note:** PBS is physical behavior score. MET-hrs/wk is the Metabolic equivalent (MET)-value for each activity multiplied by the reported hours per week; MVPA is moderate-vigorous intensity physical activity. Moderate $\geq$ 3 METs; Vigorous $\geq$ 6 METs. PBS score for quintiles are Q1 (0 to 66.47); Q2 (>66.47 to 73.63); Q3 (>73.63 to 79.03); Q4 (>79.03 to 84.85); Q5 (>85.85). Reprinted with permission from Wolters Kluwer Health Inc

## 2.7 Example Physical Behavior Score

In Table 2.2 we show a breakdown of how much each behavior contributes to the overall score. Moderate exercise, vigorous exercise and moderate-vigorous household activity accounted for the majority of the score (57/100 points). We also provide the PBS for a hypothetical person who is at the median for all reported behaviors, resulting in a score of 75.14 (Q3). (13.7 MET-hrs/wk of moderate activity, 0 MET-hrs/wk of vigorous activity per week, 7 MET-hrs/wk of light household activity, 13.25 MET-hrs/wk of MVPA household activity, 0 MET-hrs/wk of weight training, 3.0 hrs/day sitting non watching television, 3.5 hrs/day of television sitting, and 7.5 hrs of sleep). Table 2.2 also shows two hypothetical people who achieve physical activity recommendations (8 MET-hrs/wk of moderate exercise and 4 MET-hrs/week of MVPA household activity) but vary by levels of sitting time and sleep, which results in different PBS scores and classifies them in different quintiles. The provided R code in Section A.1 includes a function to calculate the PBS either for an individual or for a dataset with information on the same eight physical behaviors.

### 2.8 Predictive Validity of PBS

To model the relationship between the PBS and mortality risk, we used Cox Proportional Hazards Regression to examine mortality risk across PBS quintiles in the second half of the sample. We tested the proportional hazards assumption using Schoenfeld residuals Schoenfeld (1982) and found this assumption was not violated. See Figures A.1 and A.2 We also determined the mortality risk by quintiles of aerobic activity and sedentary time, and sleep quartiles (Q1: <5hr/day; Q2 is 5-7 hr/day; Q3: 7-8 hr/day [referent], Q4 is >9hr/day) in multivariate adjusted models as a comparator for the PBS. Separate models were fit for all-cause and cause-specific mortality (cardiovascular disease, cancer and other causes). We repeated the categorical (quintile or quartile) analysis separately for men and women, and also conducted a sex-stratified Cox proportional hazards regression analysis using PBS as a continuous variable. For the continuous analysis, the referent point was set at the 5th percentile (PBS of 53.5). We also conducted stratified analyses by sex, age group (median split), BMI categories (normal weight BMI <25 kg/$m^2$; overweight BMI 25-29.9 kg/$m^2$; obese $\geq$30 kg/$m^2$) and self-reported health status (fair, good, very good and excellent). All subgroup analyses were adjusted for covariates. All models adjusted for the same confounders, and all analyses were done in R (version 3.4.3), with an alpha-level of 0.05.

### 2.9 Results

Table 2.3 shows the baseline participant characteristics by PBS quintile. There tended to be more variation in physical activity compared to sedentary time and sleep across quintiles. The quintile cut-offs were Q1 (0 to 66.47); Q2 (>66.47 to 73.63); Q3 (>73.63 to 79.03); Q4 (>79.03 to 84.85); Q5 (>85.85). Participants in the first quintile (Q1) tended to be more obese, have lower education levels, were less likely to report health status as excellent compared to those in Q5. Over an average of 6.6y of follow-up, there were 8,732 deaths (3,503 Cancer, 2732 CVD, 2,497 other cause). There was a strong graded decline in risk of all-cause mortality across quintiles of PBS (Table 3). Compared to the first quintile of PBS, HRs (95% CI) were 0.72 (0.68, 0.77), 0.64 (0.60, 0.69), 0.58 (0.555, 0.62), and 0.53 (0.49, 0.57) for quintiles 2-5 respectively. Results were similar but stronger for cardiovascular disease mortality (Q5 vs Q1= 0.42 [0.37, 0.48]) and other mortality (Q5 vs Q1 = 0.42 [0.36, 0.48]). The PBS was also associated with graded decreased risk of cancer mortality (Q5 vs Q1 = 0.75 [0.68, 0.85]).

We then compared the magnitude of the mortality associations using the combined PBS to the individual behavioral components in isolation. For all-cause mortality, when comparing Q5 to Q1, the HR for PBS 0.53 (0.49, 0.57) was a stronger association than those observed for the individual score components (i.e., aerobic activity 0.63 (0.57, 0.69), sedentary time 0.74 (0.67, 0.82) and sleep duration 1.24 (1.17, 1.30), (7-8 hrs vs 9+h/day)). Aerobic activity consistently had stronger associations than sedentary time or sleep for all-cause and cause-specific mortality. For cause-specific mortality, the relationship with PBS was consistently stronger (7-13% lower HR) than for that of aerobic activity alone (Table 2.4).

The relationship between physical behaviors and mortality was stronger for women than for men, although there was still a strong, dose-response relationship between PBS score for both men and women (Figure 2.2). In sex-specific quintiles for all-cause mortality, women in the highest quintile had 54% reduction in all-cause mortality (Q5 vs Q1= 0.46 [0.41, 0.52)), while men in the highest quintile had a 45% reduction in mortality risk (Q5 vs Q1= 0.55 [0.50, 0.60)). This observed sex-difference was consistent across cause-specific mortality (Supplemental Digital Content Table 2). To investigate potential confounding and reverse causation, stratified analyses were conducted for all-cause mortality by sub-groups. We found the relationship between PBS and mortality was associated with decreased mortality risk in both younger and older groups, across category of self-

|  |  | **Q1** | **Q2** | **Q3** | **Q4** | **Q5** |
|---|---|---|---|---|---|---|
| Age (yrs, mean [SD]) |  | 71.2 (5.4) | 70.7 (5.4) | 70.4 (5.4) | 70.1 (5.3) | 69.7 (69.7) |
| Sex (% Female) |  | 54.7 | 57.2 | 59.6 | 60.7 | 60.6 |
| White (%) |  | 90.8 | 92.6 | 93.4 | 93.8 | 93.9 |
| BMI category (%) | Normal | 24.9 | 29.7 | 34.2 | 37.5 | 42.7 |
|  | Overweight | 36.2 | 40.6 | 40.7 | 40.9 | 40.3 |
|  | Obese | 33.1 | 24.5 | 20.5 | 16.9 | 12.6 |
| Smoking Status (%) | Never | 35.4 | 37.8 | 38.9 | 40.4 | 41.2 |
|  | Current | 7.4 | 6 | 5.2 | 4.3 | 2.8 |
| Education | High school | 36.5 | 34.6 | 31.8 | 28.5 | 23.3 |
|  | College graduate | 37.3 | 39.9 | 43.1 | 46.8 | 52.5 |
| Depression (%) |  | 16.6 | 12.8 | 11 | 9.8 | 8.8 |
| Heart Disease (%) |  | 26 | 21.1 | 19.7 | 18.8 | 17.4 |
| Health status (%) | Excellent | 7.7 | 10.7 | 13.6 | 16.9 | 23.3 |
|  | Very Good | 28 | 36.7 | 40.4 | 43.7 | 46 |
|  | Good | 42.2 | 40.4 | 37.1 | 32.9 | 26.6 |
|  | Fair | 22.2 | 12.2 | 8.8 | 6.4 | 4.1 |
| Physical Behavior Components (median [ 25th, 75th ]) |  |  |  |  |  |  |
| Moderate exercise |  | 4.4 (2.5 9.4) | 9.4 (5.1 18.0) | 13.7 (7.3 25.9) | 20.7 (11.6 36.0) | 28.8 (15.1 51.1) |
| Vigorous exercise |  | 0.0 (0.0 0.0) | 0.0 (0.0 1.9) | 0.0 (0.0 7.3) | 3.8 (0.0 17.5) | 18.3 (8.8 36.5) |
| Light household activity |  | 5.3 (1.5 12.5) | 6.3 (2.5 21.3) | 7.8 (3.8 21.3) | 12.5 (3.8 22.8) | 12.5 (6.3 30.0) |
| Household activity |  | 3.5 (0.9 8.4) | 9.1 (4.5 18.1) | 14.4 (7.5 28.3) | 19.3 (10.4 36.0) | 26.4 (16.0 47.0) |
| Weight training |  | 0.0 (0.0 0.0) | 0.0 (0.0 0.0) | 0.0 (0.0 0.0) | 0.0 (0.0 0.9) | 0.9 (0.0 5.3) |
| TV sitting |  | 3.5 (3.5 5.5) | 3.5 (1.5 5.5) | 3.5 (1.5 3.5) | 1.5 (1.5 3.5) | 1.5 (1.5 3.5) |
| Non-TV sitting |  | 5.0 (3.0 7.0) | 5.0 (3.0 7.0) | 3.0 (3.0 5.0) | 3.0 (3.0 5.0) | 3.0 (3.0 5.0) |
| Sleep |  | 7.5 (5.5 7.5) | 7.5 (5.5 7.5) | 7.5 (5.5 7.5) | 7.5 (7.5 7.5) | 7.5 (7.5 7.5) |

Table 2.3: Baseline Participant Characteristics and Physical Behavior Components by quintile of physical behavior score in the NIH-AARP Diet and Health Study Cohort, 2004-2006. Note: Demographic characteristics (education, sex, race/ethnicity) were assessed on the baseline questionnaire (1995) and other variables were assessed on follow-up questionnaires (2004-2006) (i.e., age, physical behaviors smoking, body mass index [BMI)] based on self-reported height and weight, health status and disease history) were based on values reported on the follow-up questionnaire. Exercise and sport activities are measured in METhr/wk and sedentary behaviors are measured in hr/day. MET-hrs/wk is the Metabolic equivalent (MET)-value for each activity multiplied by the reported hours per week Moderate ≥ 3 METs; Vigorous ≥ 6 METs.Reprinted with permission from Wolters Kluwer Health Inc

reported health status and BMI categories. (Figure 2.3).

## 2.10 Discussion

This paper presents a new method to combine a number of distinct physical activities, sedentary behaviors and sleep into an integrated score. In a large sample of US adults, we showed this score

|  |  | **Q1** | **Q2** | **Q3** | **Q4** | **Q5** |
|---|---|---|---|---|---|---|
| All-cause | Physical Behavior Score | ref | 0.72 (0.68, 0.77) | 0.64 (0.60, 0.69) | 0.58 (0.55, 0.62) | 0.53 (0.49, 0.57) |
|  | Aerobic activity | ref | 0.74 (0.68, 0.80) | 0.63 (0.57, 0.69) | 0.65 (0.58, 0.72) | 0.63 (0.57, 0.69) |
|  | Sedentary time | ref | 0.88 (0.80, 0.96) | 0.84 (0.78, 0.91) | 0.77 (0.71, 0.84) | 0.74 (0.67, 0.82) |
|  | Sleep | - | 1.15 (1.07, 1.23) | 1.02 (0.98, 1.09) | Ref | 1.24 (1.17, 1.30) |
| CVD | Physical Behavior Score | ref | 0.63 (0.56, 0.71) | 0.58 (0.52, 0.65) | 0.52 (0.47, 0.58) | 0.42 (0.37, 0.48) |
|  | Aerobic activity | ref | 0.69 (0.58, 0.79) | 0.53 (0.41, 0.64) | 0.54 (0.42, 0.66) | 0.55 (0.43, 0.67) |
|  | Sedentary time | ref | 0.88 (0.74, 1.01) | 0.78 (0.65, 0.89) | 0.68 (0.56, 0.80) | 0.65 (0.52, 0.78) |
|  | Sleep | - | 1.20 (1.07, 1.34) | 1.01 (0.92, 1.11) | Ref | 1.24 (1.14, 1.36) |
| Cancer | Physical Behavior Score | ref | 0.91 (0.83, 1.01) | 0.81 (0.73, 0.90) | 0.79 (0.72, 0.88) | 0.75 (0.68, 0.85) |
|  | Aerobic activity | ref | 0.92 (0.82, 1.01) | 0.84 (0.73, 0.94) | 0.83 (0.73, 0.94) | 0.82 (0.72, 0.87) |
|  | Sedentary time | ref | 0.91 (0.79, 1.02) | 0.91 (0.79, 1.01) | 0.92 (0.81, 1.02) | 0.88 (0.80, 0.96) |
|  | Sleep | - | 1.12 (1.02, 1.21) | 0.97 (0.88, 1.05) | ref | 1.11 (1.02, 1.21) |
| Other | Physical Behavior Score | ref | 0.64 (0.57, 0.72) | 0.53 (0.47, 0.60) | 0.44 (0.38, 0.50) | 0.42 (0.36, 0.48) |
|  | Aerobic activity | ref | 0.63 (0.52, 0.75) | 0.58 (0.46, 0.70) | 0.55 (0.43, 0.67) | 0.51 (0.38, 0.64) |
|  | Sedentary time | ref | 0.87 (0.74, 1.01) | 0.77 (0.65, 0.90) | 0.68 (0.56, 0.80) | 0.65 (0.52, 0.78) |
|  | Sleep | - | 1.24 (1.09, 1.39) | 1.15 (1.06, 1.26) | ref | 1.40 (1.29, 1.51) |

Table 2.4: Comparison of physical behavior score to groups components for all-cause and cause-specific mortality. Multivariate models adjusted for age (yr), sex, education (< 12 y, high school graduate, some college, college graduate, unknown), smoking history (never, stopped 10+ yr, stopped 5–9 yr, stopped 1–4 yr, stopped <1 yr, current smoker, unknown), race/ethnicity (Non-Hispanic White, Non-Hispanic Black, Other, unknown), overall health (excellent, very good, good, fair, unknown), body mass index (<25, 25–29.9, 30+ kg/$m^2$, unknown), physician diagnosed depression (yes, no, or missing), physician diagnosed heart disease (yes, no, missing). For aerobic activity Q1 is low physical activity (0-27.8 MET-hr/wk), for sedentary time Q1 is high sitting (10.5+ hr/day) and for sleep referent (Q3) is 7-8 hr/day, Q1 is <5hr/day, Q2 is 5-7 hr/day and Q4 is >9hr/day. Aerobic activity, sedentary time and sleep were mutually adjusted in the same models. Reprinted with permission from Wolters Kluwer Health Inc

has strong predictive validity for both men and women. It also showed a strong-dose-response relationship with mortality risk in both men and women, and it was more strongly associated with mortality risk than its individual components, indicated by non-overlapping confidence intervals (Table 2.4). In addition to being novel from a statistical perspective, the physical behavior score has important practical applications that may advance the field of physical behavior epidemiology.

A primary application of the PBS is in its potential to estimate the overall association between a comprehensive range of physical behaviors, morbidity and mortality. In 2012, Lee and colleagues estimated the global mortality burden due to lack of leisure-time physical activity was 5.3 million deaths annually, but this estimate was based on inactivity alone Lee et al. (2012). Given that the observed risk estimates using the PBS were stronger than for aerobic activity alone, our study suggests that the burden due to the combination of low activity, high sitting time and sleep duration may even higher. Moreover, physical behaviors are inherently interrelated and synergistic, and this type of indexed score accounts for this relationship. An important next step will be to determine if the scoring system we empirically developed in this cohort can be translated to other studies that assess physical behaviors in similar questionnaires (Mekary et al., 2013; Chasan-Taber et al., 1996; Wolf et al., 1994; Patel et al., 2017). To facilitate this, we have provided R-code to calculate the PBS in Section A.1.

A second application of this approach is in epidemiologic analyses using the composite score reflecting multiple physical behaviors as a single covariate. Given the growing list of physical behaviors that have been linked with health-related outcomes, it may be advantageous to use an overall index that represents multiple domains and types of physical behavior. The UK Biobank has elected a similar approach for accelerometer-measured activity, by including a single summary variable representing overall movement as the primary covariate for activity, rather than categorizing sedentary, light and moderate-vigorous physical activity (Doherty et al., 2017). Our approach may require additional questions to ensure the range of physical behaviors are covered, but it allows researchers to adjust for multiple dimensions of physical behavior without having to enter separate variables for each behavior into their model. This application of the PBS will have particular importance in studies where physical behavior is a covariate rather than the primary exposure of interest, when investigating the link between physical behaviors and rare disease outcomes or in small samples where statistical efficiency is a primary concern.

The statistical approach we used is innovative. While shape constrained regression has been used in a variety of fields(see Chen and Samworth (2016)) this is, to our knowledge, the first time it has been applied in the analysis of physical activity, sedentary behavior, and sleep. This approach is a unique way of letting previous knowledge guide the choice of statistical model. Additionally, when the relationship between two variables is known to satisfy a predefined relationship, shape constrained regression has been shown to give results that are, on average, closer to the truth than comparable methods without shape constraints (Chen and Samworth, 2016; Pya and Wood, 2015) This may help to explain the strong, statistically significant results we obtained.

There is a large body of evidence in nutritional epidemiology that has demonstrated the conceptual and etiologic value in assessing dietary patterns (i.e., combinations of foods and nutrients) in relationship to health (Guenther et al., 2008a,b; Hu et al., 2000; Guenther et al., 2013c). Dietary pattern analyses recognize that, in contrast to isolating a single component of food (e.g., carbohydrates), the foods people eat are likely to be correlated and synergistic. Dietary pattern indices have been developed a posterori (using a data driven approach like principal component analyses) or a priori, using score-based approaches (e.g., the Healthy Eating Indices derived from federal dietary guidelines). Our approach borrows strengths from each of these, a priori we identified key physical behaviors to be included in the PBS, and then used a data-driven approach to weight each of the individual components of the total score in relation to the outcome of interest – in this case, survival.

The PBS recognizes that individuals can achieve health benefits through different combinations of behavior, for example engaging in high volume of moderate exercise, or engaging in a lot of household activity, limiting sedentary time and adequate sleep. The strongest contributors to the PBS score were moderate exercise and television viewing, which is consistent with previous research and the 2018 Physical Activity Guidelines, which state that adults who sit less and do any amount of moderate-to-vigorous physical activity gain health benefits (Ekelund et al., 2016; Piercy et al., 2018). This may have implications for developing and evaluating interventions that target multiple behaviors, which is an important area for future research.

There are important limitations to note. Our sample is fairly well educated, predominantly white older adults (59-80 years). Although this is an important demographic given the aging US population, we do not know if these results generalize to younger samples or those with different racial/ethnic compositions. Physical behaviors were self-reported, which is subject to recall and measurement error. However, the use of a questionnaire enables details about activity domain that are not differentiated well using an accelerometer (e.g., household vs. leisure) and activity types (e.g., biking and weight lifting) that are not accurately captured by activity monitors. In contrast to a previous day recall or activity monitoring, the survey was not designed to capture a complete 24-hr cycle of daily activity, sitting and sleep. Activities were reported as duration per week and then converted to MET-hrs/week to account for differing energy cost of different activities. Future research is needed to determine the utility of this approach using instruments like previous day recalls or activity monitors that are designed to assess a complete 24-hr period (Matthews et al., 2018). Sleep duration is strongly associated with health (Yin et al., 2017), but there are other aspects of sleep quality that are associated with health outcomes that we did not assess (Kabat et al., 2018).

This study has important strengths including a large sample with considerable statistical power. The questionnaire included a wide range of activities, which enabled the multi-dimensional evaluation of physical behaviors. The statistical approach is a novel application that incorporates both the strength and shape of associations into an easily interpretable overall score (0-100). The method for developing a PBS score presented in this paper would generalize to a different sample and/or a different instrument to assess physical behavior, but may result in different weighting of the components. An additional strength is the dissemination of R-code for individuals to calculate their PBS score based on these data, and for researchers to apply to full datasets to estimate PBS in independent samples.

### 2.11 Conclusions

This paper presents a statistical method to generate a composite physical behavior that has high predictive validity for mortality outcomes. Although widespread in other areas of epidemiology (Guenther et al., 2008a, 2013a, 2008b) this is one of the first attempts to characterize integrate multiple distinct physical behaviors into a single physical composite score. This score can be applied to quantify the overall disease burden of physical behaviors rather than looking at different types of physical activity in isolation, and it can be used as a parsimonious covariate to adjust for physical behaviors. Future research is needed to test this approach in an independent sample and with different health outcomes.

## 2.12 Technical Description of Score Development

Using the NIH-AARP Study of Diet and Health (Schatzkin et al., 2001), we break physical activity into 8 discrete components. The relationship between physical activity and survival for each activity component was first modeled using a Generalized Additive Model (Hastie and Tibshirani, 1986; Wood, 2017). Suppose $Y_i$ is a binary indicator of survival of the $i^{th}$ person until the end of the study, $X_i = (X_{i1}, \ldots, X_{i8})$ is a vector containing the $i^{th}$ person's 8 physical activity measurements, and $Z_i$ is the vector containing the additional covariates like age, sex, education, and other demographic information. We model the probability of survival as

$$\mathrm{pr}(Y_i = 1 | X_i, Z_i) = H\{\textstyle\sum_{j=1}^{8} f_j(X_{ij}) + Z_i^{\mathrm{T}}\theta\}, \tag{2.1}$$

where $H(\cdot)$ is the logistic distribution function, each $f_j$ is an unknown smooth function, and $\theta$ is a vector of unknown parameters.

The expected relationship between each physical behavior and mortality - the functions, $f_j$, in (2.1) - can be restricted to be consistent with previous studies. In the statistics literature, these restrictions are referred to as shape constraints. The dose-response relationship between aerobic activity and survival has been established as somewhat concave and non-decreasing (Arem et al., 2015). That is, aerobic activity has positive effect but levels off after a certain point. Sedentary time is known to have a negative effect on overall health (Grøntved and Hu, 2011; Prince et al., 2014). Sleep is known to be beneficial in reasonable doses but too much or too little sleep is indicative of poor health (Yin et al., 2017), suggesting a concave, parabolic relationship with overall health. To incorporate this information into Model (2.1) we used the Shape Constrained Additive Regression (SCAR) method of Chen and Samworth (2016) for fitting Generalized Additive Models with shape constraints.

The fitted values, as they are reported by SCAR, represent the logits of the effect of each physical behavior on survival. This is not a scale that is desirable to work with. The regression results should be scaled to be between 0 and 100 to put it on the same scale as scores from other fields, such as the Healthy Eating Index (Guenther et al. 2008). Fitted values from (2.1) indicating a low probability of survival are translated to values near 0, and fitted values indicating a high probability of survival are translated to a value near 100. This creates the scoring system for physical activity.

Time spent sitting adversely impacts survival and fitted values for sitting variables are negative. Since one of the aims of this analysis is to create a score which ranges from 0 to 100, a number lower than 0 is undesirable. In the scoring system we develop, a person with a score of 0 should be getting no positive benefits from physical activity, and a score below 0 has no meaning. Fitted values were rescaled to force the functions to take on only positive values. Denote the scaled function by $f^*(\cdot)$ and the index of the function corresponding to TV sitting as $j = A$ and the index for non-TV sitting as $j = B$ Both function, $f_A(\cdot)$ and $f_B(\cdot)$, are rescaled by adding the absolute value of the minimum of both functions. That is,

$$f_j^*(X_{ij}) = f_j(X_{ij}) + |\min_j f_j(X_{ij})| \tag{2.2}$$

for $j = A, B$. The smallest value fitted value of the functions $f_A^*(\cdot)$ and $f_B^*(\cdot)$ is 0 after applying (2.2).

Additionally, the estimated function for hours of sleep becomes negative for values greater than about 9.5 hours of sleep. Instead of shifting the function by a constant (such as $|\underset{j}{min}|$ for the sitting functions) negative values are set to 0, so that,

$$f^*(X_{ij}) = f_j(X_{ij}) * \mathcal{I}\{f(X_{ij}) > 0\}, \tag{2.3}$$

where $\mathcal{I}\{f(X_{ij}) > 0\}$is an indicator function. This function takes on value 1 if $f(X_{ij})$ is positive and 0 if $f(X_{ij})$ is negative.

For simple notation, all shape constrained functions are referred to by $f_j^*$ even if they have not been modified as in (2.2) and (2.3). As a final step, the maximum values of each $f_j^*$ is calculated and then summed. Denote this quantity as T. Formally,

$$T = \textstyle\sum_{j=1}^{J} \max_{j} f_j^*(X_{ij}). \tag{2.4}$$

The fitted values for every subject in the dataset are then divided by T. This is the Physical Behavior Score reported in the paper. The result is denoted by $\mathcal{S}_i$, i.e.,

$$\mathcal{S}_i = 100/T \textstyle\sum_{j=1}^{J} f_j^*(X_{ij}) \tag{2.5}$$

At this stage, $\mathcal{S}$ is between 0 and 100 for every individual in the NIH-AARP dataset. A plot of the original function is given in Figure 2.4 and the distribution of physical activity scores is given in Figure 2.5.

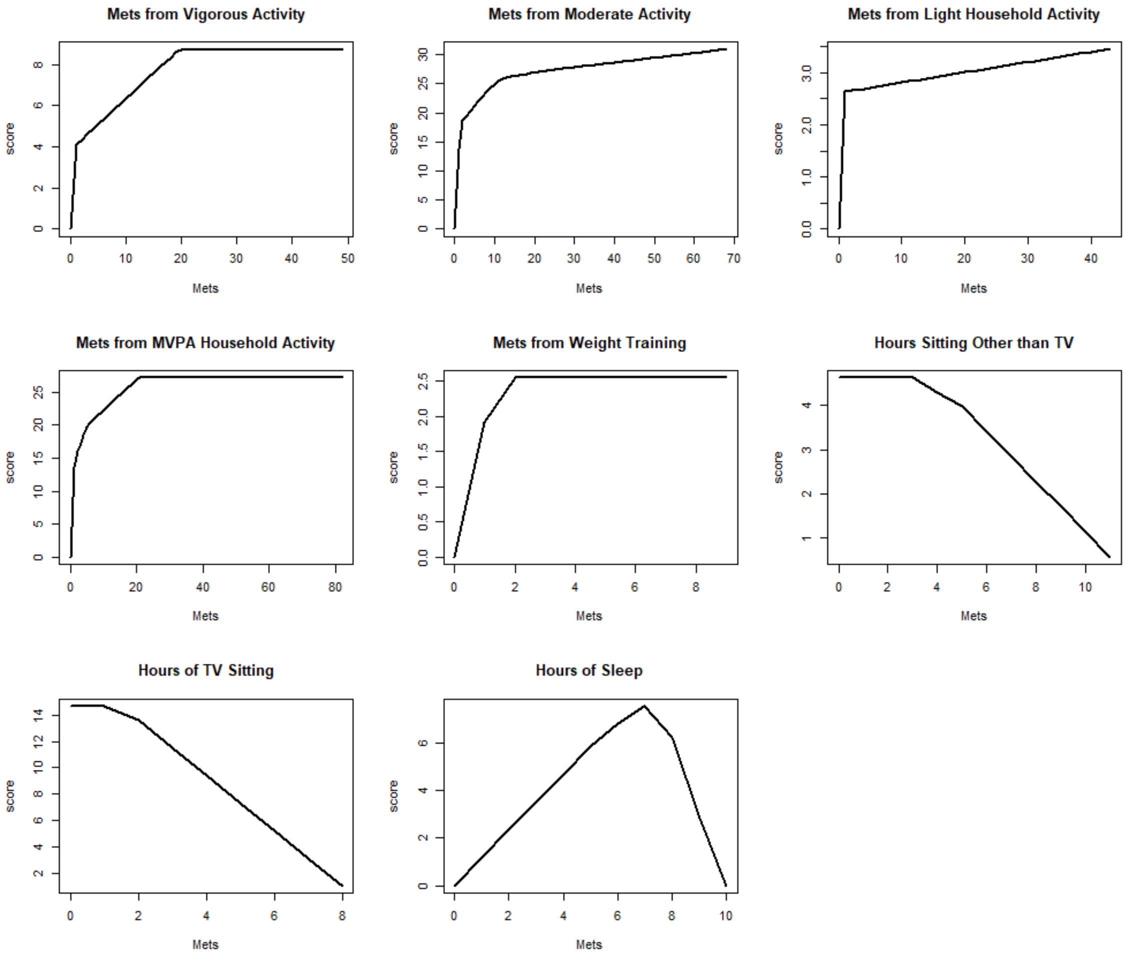


Figure 2.1: Shape constrained additive regression (SCAR) returns 8 sets of predicted values, one for each function that is fit to describe the relationship between physical activity and mortality. The values on the x-axis are MET's expended during a particular physical behavior or hours spent in a particular physical behavior. The values on the y-axis show the relative importance of each physical behavior on the total Physical Behavior Score. Reprinted with permission from Wolters Kluwer Health Inc

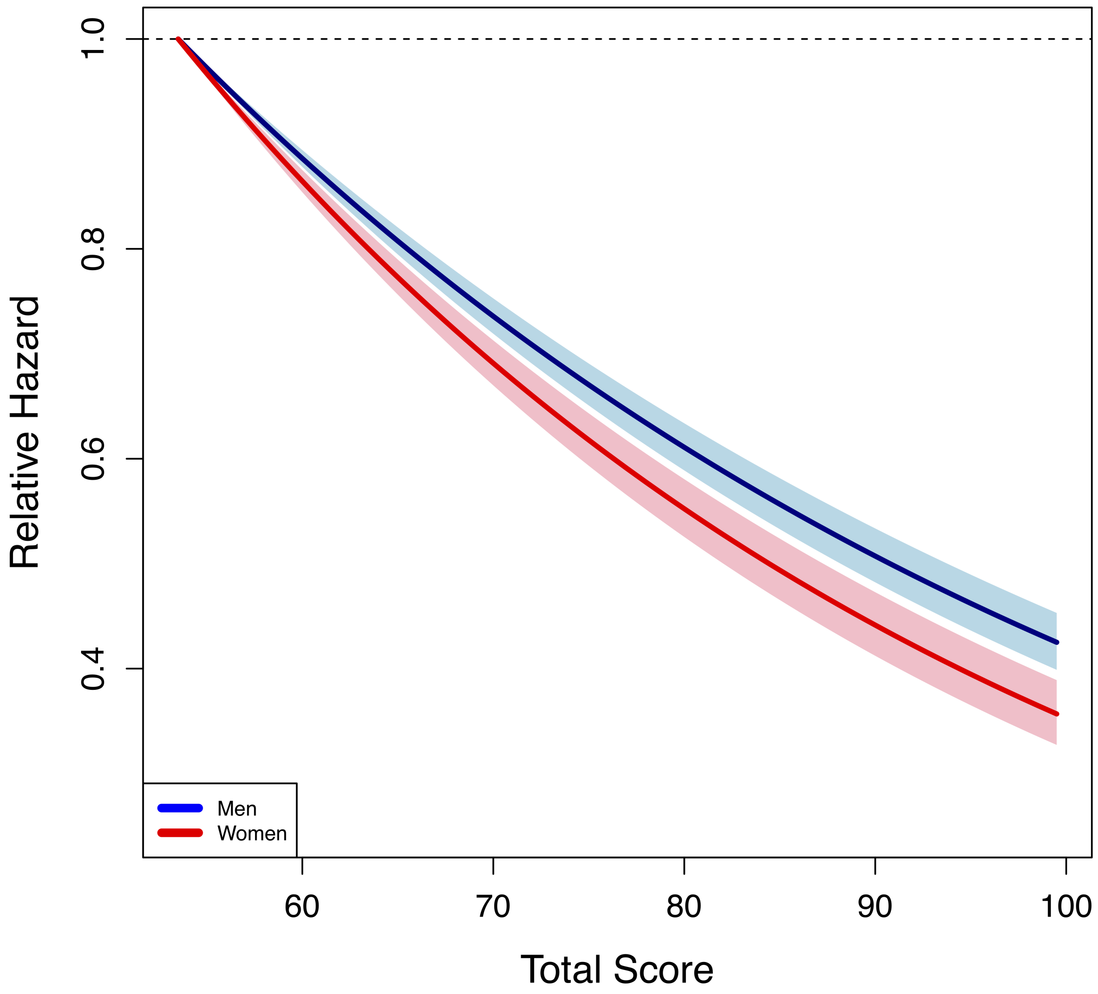


Figure 2.2: The relative hazard of all-cause mortality plotted from the 5th percentile, 53.5, to the maximum score of 100, separately for men and women. Model adjusted for age (yr), sex, education (< 12 y, high school graduate, some college, college graduate, unknown), smoking history (never, stopped 10+ yr, stopped 5–9 yr, stopped 1–4 yr, stopped <1 yr, current smoker, unknown), race/ethnicity (Non-Hispanic White, Non- Hispanic Black, Other, unknown), overall health (excellent, very good, good, fair, unknown), body mass index (<25, 25–29.9, 30+ kg/$m^2$, unknown), physician diagnosed depression (yes, no, or missing), physician diagnosed heart disease (yes, no, missing). Shading indicates 95% confidence intervals. Reprinted with permission from Wolters Kluwer Health Inc

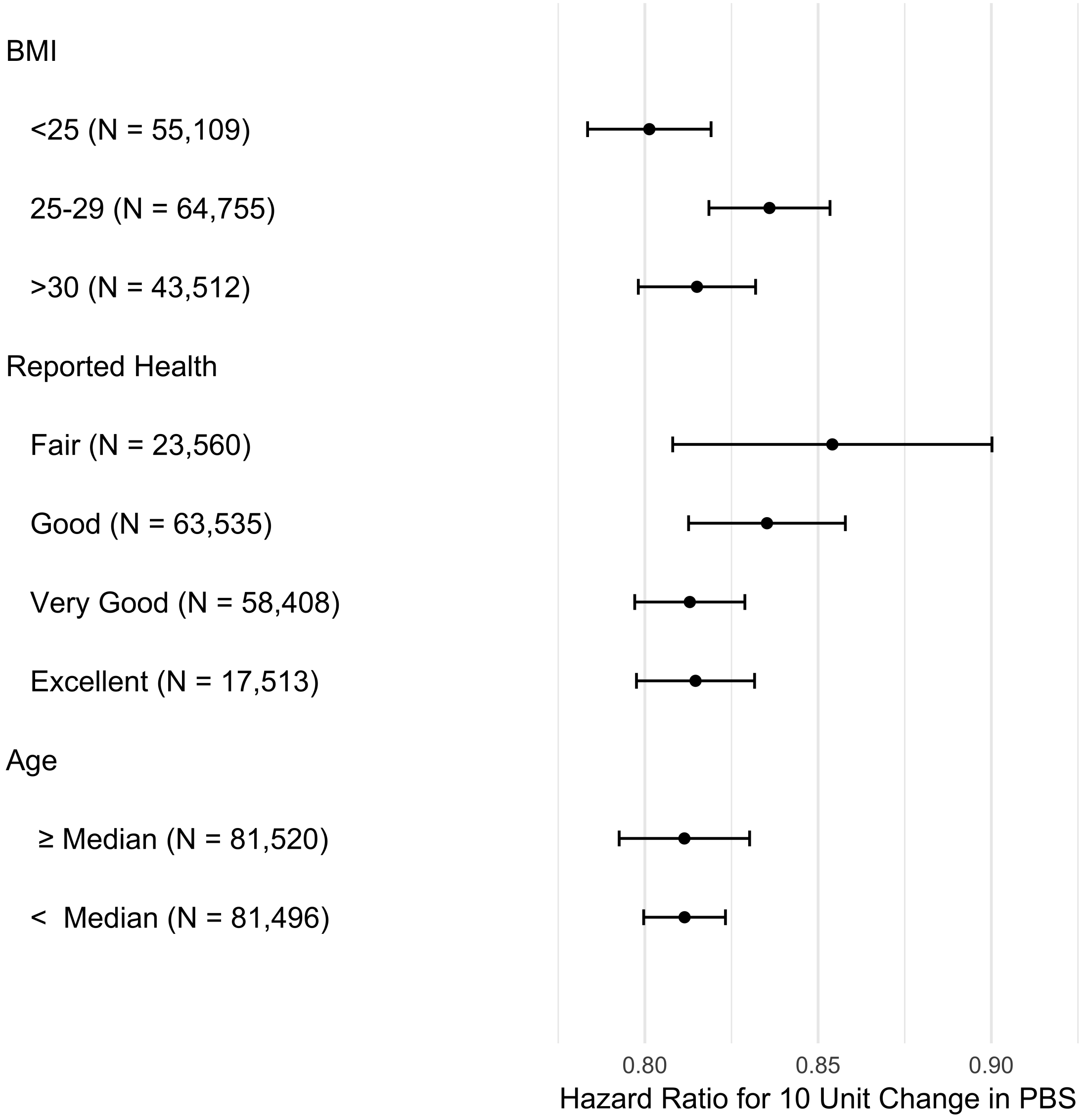


Figure 2.3: Values are Hazard Ratio and confidence interval per 10 unit increase in Physical Behavior Score. All models were adjusted for age (yr), sex, education (<12 y, high school graduate, some college, college graduate, unknown), smoking history (never, stopped 10+ yr, stopped 5–9 yr, stopped 1–4 yr, stopped <1 yr, current smoker, unknown), race/ethnicity (Non- Hispanic White, Non-Hispanic Black, Other, unknown), overall health (excellent, very good, good, fair, unknown), body mass index (<25, 25–29.9, 30+ kg/$m^2$, unknown), physician diagnosed depression (yes, no, or missing), physician diagnosed heart disease (yes, no, missing) Reprinted with permission from Wolters Kluwer Health Inc

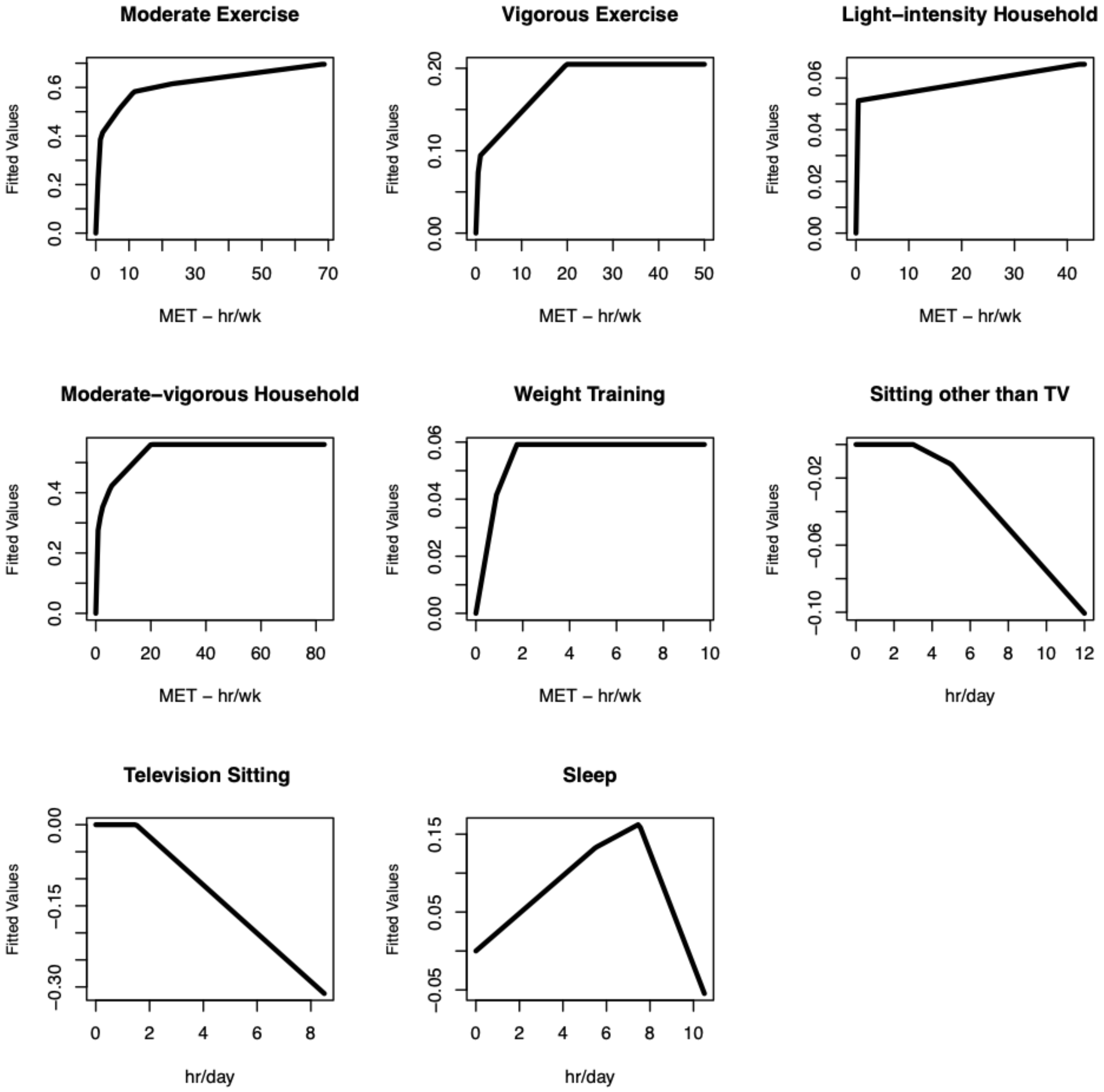


Figure 2.4: Shape constrained additive regression (SCAR) returns 8 sets of predicted values, one for each function that is fit to describe the relationship between physical activity and mortality. The values on the x-axis are MET's expended during a particular physical behavior or hours spent in a particular physical behavior. The values on the y-axis show the additive effect of each physical behavior on the logits (log-odds) of survival. These are the fitted values described in Section 2.12.Reprinted with permission from Wolters Kluwer Health Inc

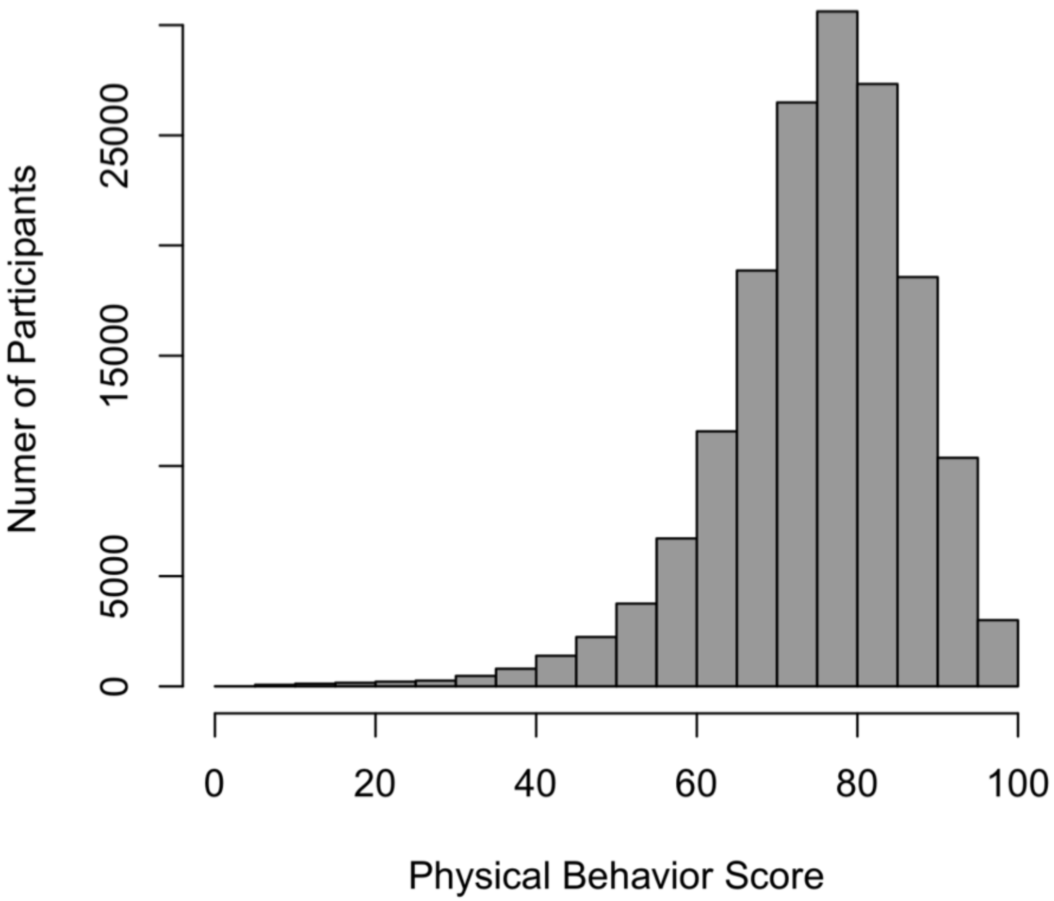


Figure 2.5: Distribution of physical behavior scores in the NIH-AARP Diet and Health Study Cohort, 2004-2006. Reprinted with permission from Wolters Kluwer Health Inc

# 3. RE-EVALUATING COMPOSITE SCORES: ADAPTIVE LASSO VARIABLE SELECTION FOR NONLINEAR MODELS

## 3.1 Introduction

In epidemiology and other public health fields there is a need to reduce complicated behavioral patterns into simple, interpretable terms. A composite score, also referred to as an index, is commonly used to achieve this end, so as to be applied to different populations and different health outcomes. Composite scoring systems compare an individual's health behavior to an idealized standard. Based on compliance to a set of health behaviors, an individual is assigned a score between 0 and 100. A score of 0 indicates poor compliance, and 100 is a theoretical perfect behavior.

We look at one particular challenge of working with composite scoring systems derived to be applied to multiple populations and diseases: Given an existing scoring system, are all components in the system necessary? We follow the ideas put forward by Ma et al. (2017) and fit a logistic regression model using people from many populations who suffer from many diseases to develop a score. We include nonlinear terms in our regression model which capture the effect of the score on the risk of particular disease in a population of interest. Ma et al. (2017) perform a similar analysis but include a nonparametric component in their model that we do not. The authors found that the flexibility of their semiparametric model was not needed in their real-world data analysis.

We include an adaptive lasso penalty (Zou, 2006) to perform variable selection. The literature for the Lasso is well developed but does not apply to our particular model because of identifiability issues discussed in Section 3.2. Additionally, typical software packages for fitting Lasso problems such as *glmnet* (Friedman et al., 2010) or Least Angle Regression (Efron et al., 2004) are not able to handle these nonlinear models. To remedy this, we use the Least Squares Approximation of Wang and Leng (2007). The Least Squares Approximation allows us to translate our estimation problem into a simpler asymptotically equivalent least squares minimization. We establish that our variable selection technique chooses, asymptotically, the correct subset of components and has the optimal convergence rate. That is, it has oracle properties.

While our methods are general, we apply them to the 2005 Healthy Eating Index and use the 2005 Healthy Eating Index to motivate our methods. We will refer to the 2005 Healthy Eating Index as the Healthy Eating Index, omitting the year. The Healthy Eating Index is based on the key recommendations of the 2005 Dietary Guidelines for Americans (`http://www.health.gov/dietaryguidelines/dga2005/document/default.htm`). The index includes ratios of interrelated dietary components to energy (caloric) intakes. The 2005 Healthy Eating Index comprises 12 distinct component scores and a total summary score. See Table 3.1 for a list of these components and the standards for scoring, and see Guenther et al. (2008b,a) for details on how the Healthy Eating Index was developed and validated.

Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy intake, assessed, and ascribed a score. There are other, competing measures of diet such as the 2010 Healthy Eating Index (Guenther et al., 2013a), the Modified Mediterranean Diet Score (Trichopoulou et al., 2005) and the MedDietScore (Panagiotakos et al., 2006), all of which are associated with lowered mortality risk and better overall health. Our aims are (a) to suggest improvements to the dietary guidelines of the Healthy Eating Index; and (b) to use model selection techniques to evaluate the relative importance of the 12 components. We find the unexpected fact that empty calories (SoFAAS) are not predictive of increased mortality risk.

## 3.2 A Nonlinear Model Across Populations

We will to use the term *disease* in a generic way, until our data analysis. The term should be understood to mean a collection of health outcomes, which, for example, could be various combinations overall mortality, mortality from various diseases, different chronic conditions, or the development of different cancers.

Denote $j = 1, ..., J$ as the index of the Healthy Eating Index components. There are $k = 1, ...K$ populations and $\ell = 1, ...L_K$ diseases in each population. There are $i = 1, ...n_{k\ell}$ individuals be evaluated for disease $\ell$ in population $k$. The data observed are

| Component | Units | Healthy Eating Index 2005 score calculation |
|---|---|---|
| Total Fruit | cups | $\min(5, 5 \times (\text{density}/.8))$ |
| Whole Fruit | cups | $\min(5, 5 \times (\text{density}/.4))$ |
| Total Vegetables | cups | $\min(5, 5 \times (\text{density}/1.1))$ |
| DOL | cups | $\min(5, 5 \times (\text{density}/.4))$ |
| Total Grains | ounces | $\min(5, 5 \times (\text{density}/3))$ |
| Whole Grains | ounces | $\min(5, 5 \times (\text{density}/1.5))$ |
| Milk | cups | $\min(10, 10 \times (\text{density}/1.3))$ |
| Meat and Beans | ounces | $\min(10, 10 \times (\text{density}/2.5))$ |
| Oil | grams | $\min(10, 10 \times (\text{density}/12))$ |
| Saturated Fat | % of energy | if density $\geq 15$ score = 0<br>else if density $\leq 7$ score = 10<br>else if density $> 10$ score $= 8 - (8 \times (\text{density} - 10)/5)$<br>else, score = $10 - (2 \times (\text{density} - 7)/3)$ |
| Sodium | milligrams | if density $\geq 2000$ score=0<br>else if density $\leq 700$ score=10<br>else if density $\geq 1100$<br>score $= 8 - \{8 \times (\text{density} - 1100)/(2000 - 1100)\}$<br>else score $= 10 - \{2 \times (\text{density} - 700)/(1100 - 700)\}$ |
| SoFAAS | % of energy | if density $\geq 50$ score = 0<br>else if density $\leq 20$ score=20<br>else score $= 20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$ |

Table 3.1: Description of the Healthy Eating Index 2005 scoring system. Here, "SoFAAS" are calories from solid fats, alcoholic beverages and added sugars, while "DOL" are dark green and orange vegetables and legumes. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1000 and dividing by usual intake of kilo-calories. For saturated fat, density is $9 \times 100$ usual saturated fat (grams) divided by usual calories, i.e., the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, i.e., the division of usual intake of SoFAAS by usual intake of calories. The total Healthy Eating Index-2005 score is the sum of the individual component scores.

- $Y_{ik\ell}$ is a binary indicator of disease $\ell$ for the $i^{\text{th}}$ person in population $k$.
- $(X_{i1}, ..., X_{iJ})$ is the Healthy Eating Index score for person $i$ with components $j = 1, ..., J$. In the 2005 Healthy Eating Index, $J = 12$.
- For each population and disease, there may be different covariates which include terms such as age, ethnicity, education, body mass index, smoking, physical activity, etc. These covari-

ates are denoted as $Z_{ik\ell}$.

To better capture the risk of a particular disease, we introduce a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_J)^{\mathrm{T}}$ which allows flexible rescaling of the Healthy Eating Index components. We model the probability of a subject $i$ of population $\ell$ having disease $k$ as

$$\mathrm{pr}(Y_{ik\ell} = 1 | X_{ikj}, Z_{ik\ell}) = H(\beta_{k\ell} \textstyle\sum_{j=1}^{J} X_{ijk\ell}\alpha_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}) = p_{ik\ell}, \tag{3.1}$$

where $H(\cdot)$ is the logistic distribution function. The parameter $\boldsymbol{\alpha}$ reweighs each person's original Healthy Eating Index score, $\sum_{j=1}^{J} X_{ijk\ell}$, to create a new modified score, $\sum_J X_{ijk\ell}\alpha_j$. If a particular $\alpha_j$ is greater than 1, this indicates that a particular HEI component should be given more relative importance in the 0-to-100 score than in the Healthy Eating Index while a $\alpha_j < 1$ indicates it should be given less importance. The term $\beta_{k\ell}$ allows the effect of the modified score to vary with population and disease. There is novelty in this modeling approach. We are able to provide a *single* measure of diet for every population and disease. To emphasize this, and for notational convenience, we omit the subscripts over $k$ and $\ell$ in $\mathbf{X}$ This modeling approach is beneficial to public health practitioners as a single predictor $\sum X_{ij}\alpha_J$ can be used for any disease/population of interest, and the effect of this predictor, $\beta_{l\ell}$, can be reassessed as needed.

Model (3.1) results in the composite loglikelihood function

$$L_n(\alpha, \beta, \theta) = \textstyle\sum_{\ell}\sum_{k}\sum_{i}\{Y_{ik\ell}\mathrm{log}(p_{ik\ell}) + (1 - Y_{ik\ell})\mathrm{log}(1 - p_{ik\ell})\}.$$

The multiplication of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in model (3.1) means that they are not identifiable, and some kind of constraints are needed. A natural constraint would be to enforce that the maximum value of the new score, $\sum_J X_{ij}\alpha_j$, is 100. We do eventually use this constraint, but not initially. That particular constraint makes (3.1) difficult to fit from a computational perspective. Instead we begin by setting $\beta_{11} = -1$. The remaining parameters, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\theta$ are estimated in an iterative profiling procedure, first fixing $\boldsymbol{\alpha}$ and maximizing $L_n$ with respect to $\boldsymbol{\beta}$ and $\theta$, then fixing $\boldsymbol{\beta}$ and $\theta$ and maximizing over $\boldsymbol{\alpha}$. This processes is repeated until convergence. Ma et al. (2017)

provide guarantees that this procedure will converge to the correct value of the parameters.

Once the estimates have converged, we rescale the $\alpha$ coefficients so the new score is between 0 and 100. Define $\mathbf{c}_{max} = (c_{max,1}, \ldots, c_{max,J})^{\mathrm{T}}$as the maximum value that the original Healthy Eating Index assigns to a particular dietary component. Each element of $\boldsymbol{\alpha}$ is set to $\alpha_j^* = \alpha_j / \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{c}_{max}$. This puts the newly assigned score on a scale from 0 to 100, and the constraint on $\boldsymbol{\alpha}$ allows $\beta_{11}$ to be estimated by refitting (3.1) with $\boldsymbol{\alpha}^*$ in place of $\boldsymbol{\alpha}$.

## 3.3 Variable Selection

### 3.3.1 Least Squares Approximation and Adaptive Lasso

In our context of multiple diseases and populations, we next establish which Healthy Eating Index components have no effect on health status. To test this we add an adaptive lasso penalty (Zou, 2006) to our likelihood (3.2). Like all Lasso style penalties, the adaptive Lasso can induce perform variable selection by forcing some coefficients to take on a value of 0. The adaptive Lasso has a number of desirable properties that are explored in Section 3.3.2.

In our problem, we focus only on penalization of the $\boldsymbol{\alpha}$ parameters. In principle then, we would minimize

$$L_n(\beta, \alpha, \theta) + \lambda \textstyle\sum_{j=1}^{J} |\widehat{\alpha}_{full,j}|^{-\gamma} |\alpha_j|, \tag{3.2}$$

with respect to $\boldsymbol{\beta}, \boldsymbol{\alpha}$, and $\theta$, where $\lambda$ is the tuning parameter, $\gamma$ is a prespecified positive number, and $\widehat{\alpha}_{full,j}$ is an estimate of $\alpha_j$ which has not been subject to any penalization, found by maximizing (3.2). If a component in $\boldsymbol{\alpha}$ is shrunk to 0 by the adaptive lasso method, we take this as an indication that a particular Healthy Eating Index component is unnecessary in predicting outcomes of interest.

However, in practice, there is a computational problem. Typical tools for fitting Lasso problems such as *glmnet* (Friedman et al., 2010) or Least Angle Regression (Efron et al., 2004) are designed for standard linear and generalized linear models and do not handle the term $\sum_j X_{ij}\alpha_j$ correctly, because they cannot penalize the $\boldsymbol{\alpha}$ coefficient without also penalizing the $\boldsymbol{\beta}$ coefficient. For a conceptually simple and computationally fast solution, Wang and Leng (2007) proposed a Least

Squares Approximation for unifying computation of all Lasso models. Consider a problem with parameters $\Psi = (\psi_1, \ldots, \psi_P)$ and a loss function $L_n(\Psi)$. Let $\widetilde{\Psi}$ be the minimizer of $L_n(\cdot)$. The authors show that any reasonable loss function, in parameters denoted as $\Psi$,

$$L_n(\Psi) + \textstyle\sum_{j=1}^{d} \lambda_j |\psi_i|,$$

can be expressed as an asymptotically equivalent least squares problem

$$Q(\Psi) = (\widetilde{\Psi} - \Psi)^{\mathrm{T}} \widehat{\Sigma}^{-1} (\widetilde{\Psi} - \Psi) + \textstyle\sum_{j=1}^{d} |\psi_j|,$$

where $\widetilde{\Psi}$ is the vector than minimizes $L_n(\cdot)$ and $\widehat{\Sigma}$ is an asymptotically consistent estimate of the covariance matrix of $\widetilde{\Psi}$.

We approximate our loglikelihood as

$$L_n(\Theta) + \lambda \textstyle\sum_{j=1}^{J} |\widehat{\alpha}_{full,j}|^{-\gamma} |\alpha_j| \approx (\widetilde{\Theta} - \Theta)^{\mathrm{T}} \widehat{\Sigma}^{-1} (\widetilde{\Theta} - \Theta) + \lambda \textstyle\sum_{j=1}^{J} |\widehat{\alpha}_{full,j}|^{-\gamma} |\alpha_j|, \qquad (3.3)$$

where $\Theta = (\beta, \alpha, \theta)$.

The Least Squares Approximation in (3.3) can be solved with standard optimization software or fit as a Gaussian family adaptive lasso model using the *glmnet* R package. Denote $\widehat{\Theta}_{LSA}(\lambda)$ as the value which minimizes the right hand side of (3.3) as a function of $\lambda$. In our computation we take $\gamma = 1$, typical choice, though it can be set to any value satisfying Lemma 1 and Lemma 2 in Section 3.3.2.

### 3.3.2 Model Selection and Oracle Properties

Variable selection procedures ideally should possess the *oracle* property. Fan and Li (2001) give an overview for what it means for a selection procedure to have the oracle property. Denote $A = \{j : \Theta_j \neq 0\}$ and $\widehat{A}(\lambda) = \{j : \widehat{\Theta}(\lambda)_{LSA,j} \neq 0\}$ The procedure should have

- **Selection Consistency**: $\mathrm{pr}\{\widehat{A}(\lambda) = A\} \to 1$.

- **Optimal Estimation Rate**: $\sqrt{n}(\widehat{\Theta}_{\delta,\widehat{A}_\delta} - \Theta_A) \to N(0, \Sigma_A)$ in distribution, where $\Theta_A$ are the nonzero components of $\Theta$ and $\Sigma_A$ is the covariance matrix knowing the true subset of predictors

We derive the selection consistency and optimal estimation rate of the Least Squares Approximation in our problem in a similar manner as Zhang et al. (2015). The following result provides the selection consistency for $\widehat{\theta}_{LSA}(\lambda)$ .The proof of this theorem, as well as all the proofs in Section 3.3.2, is provided in Section B.

**Lemma 1.** *As $n \to \infty$, if $n^{1/2}\lambda \to 0$, and $n^{(1+\gamma)/2}\lambda \to \infty$, then*

$$pr\{\widehat{A}(\lambda) = A\} \to 1.$$

Wang and Leng's proof of oracle properties relies extensively on they they call the *covariance assumption*. The covariance assumption specifies a strict relationship between the asymptotic covariance matrix of the full model and the asymptotic covariance matrix of an over-fitted model. The exact assumption is stated as follows: Let $\Sigma$ denote the variance of the limiting distribution of the parameters of the full model. Denote $\Omega = \Sigma^{-1}$, and $\Omega^{(\mathcal{S})}$ as the sub-matrix of $\Omega$ corresponding to the sub-model $\mathcal{S}$. Let $\Sigma_{\mathcal{S}}$ denote the variance of the limiting distribution of model $\mathcal{S}$, and $\Omega_{\mathcal{S}} = \Sigma_{\mathcal{S}}^{-1}$. The covariance assumption states that $\Omega^{(\mathcal{S})} = \Omega_{\mathcal{S}}$ for any over-fitted $\mathcal{S}$.

The variance-covariance matrix of model (3.1) is fit using a sandwich estimator. A sandwich estimator has the form $\Sigma = J^{-1}HJ^{-T}$ where $J = J_n(\Theta) = \nabla L_n(\Theta)$, $H = H_n(\Theta) = \nabla^2 L_n(\Theta)$, and $J^{-T} = (J^{-1})^{\mathrm{T}}$. See Carroll et al. (2006) Section A.3.1 for a detailed treatment of sandwich estimators. In general, $\Sigma^{(S)} = (J^{-1}HJ^{-T})^{(S)} \neq J_S^{-1}H_S J_S^{-T} = \Sigma_S$, and therefore covariance matrices derived from sandwich estimators will not satisfy the covariance assumption of Wang and Leng.

The selection consistency of Theorem 1 does not rely on the Wang and Leng's covariance assumption, but the optimal estimation rate does. Therefore, Wang and Leng's theory will not guarantee asymptotically consistent parameter or variance estimates. However, we can get parameter and variance estimates by fitting the model (3.1) using only the selected components. This is explained in the following theorem.

**Lemma 2.** *Let A denote the set of nonzero covariates, $\theta_A$ denote these nonzero covariates, $\Sigma_A$ denote the covariance matrix of the nonzero covariates, and $\widehat{\theta}(A)$ denote the estimates of $\theta_A$ found by fitting the logistic model from Section 3.2. As n $\to \infty$, if $n^{1/2}\lambda \to 0$ and $n^{(1+\gamma)/2}\lambda \to \infty$, then*

$$\sqrt{n}\{\widehat{\Theta}(A) - \Theta_A\} \to N(0, \Sigma_A).$$

The results of Lemma 1 and Lemma 2 rely the proper choice of $\lambda$. Like all Lasso methods, the Least Squares Approximation provides a solution for any $\lambda$, however the optimal value of $\lambda$ must be selected. For finding the best fitting penalized model, Wang and Leng propose a BIC style criterion, namely

$$BIC(\lambda) = \{\widehat{\Theta}_{LSA}(\lambda) - \widetilde{\Theta}_{full}\}^{\mathrm{T}}\widehat{\Sigma}^{-1}\{\widehat{\Theta}_{LSA}(\lambda) - \widetilde{\Theta}_{full}\} + g_n/\{n\text{log}(n)\},$$

where $g_n$ is the number of nonzero coefficients in $\widehat{\Theta}_{LSA}(\lambda)$. Define $A_{true}$ as the set of nonzero coefficients. The interval $(0, \infty)$ can be partitioned into three disjoint sets depending on whether $\widehat{A}(\lambda)$ is over-fit, under-fit, or equal to the true model:

$$\begin{aligned}\mathbb{R}_- &= \{\lambda \in (0,\infty) : \widehat{A}(\lambda) \subset A_{true}\},\\ \mathbb{R}_0 &= \{\lambda \in (0,\infty) : \widehat{A}(\lambda) = A_{true}\},\\ \mathbb{R}_+ &= \{\lambda \in (0,\infty) : \widehat{A}(\lambda) \supset A_{true}, \widehat{A}(\lambda) \neq A_{true}\}.\end{aligned}$$

Letting $\lambda^* \propto n^{-2/3}$ which satisfies Theorem 1, then we have,

$$\text{pr}\{\widehat{A}(\lambda^*) = A\} \to 1.$$

Additionally, we have the following result for any $\lambda \in R_-$ and $\lambda \in R_+$

**Lemma 3.** *As $n \to \infty$, if $\widehat{\Sigma}$ is a consistent estimate of the variance-covariance matrix of the limiting distribution of the full model, then*

$$pr\{\inf_{\lambda \in \mathbb{R}_- \cup \mathbb{R}_+} BIC(\lambda) > BIC(\lambda^*)\} \to 1 \tag{3.4}$$

Theorem 3 tell us that any $\lambda$ which produces the incorrect model, that is $\lambda \in R_-$ and $\lambda \in R_+$, will not be selected by the BIC criterion as the optimal tuning parameter. Wang and Leng's BIC criterion is consistent in selecting the optimal tuning parameter.

## 3.4 Data Analysis

### 3.4.1 Background

Of particular interest to nutritionists and epidemiologists is the relationship between diet and cancer as well as diet and mortality. We conduct our analysis on the 2005 NIH-AARP Study of Diet and Health. This longitudinal study tracks incidence of lung, colorectal, prostate, breast, and ovarian cancer in adults between the ages of 51-75, as well as cause of death for those who died while the study was conducted. Table 3.2 lists the number of adults surveyed as well as the breakdown of cancer by men and women, and Table 3.3 lists mortality. The study follows mortality caused by cancer, cardiovascular disease, as well as all other causes of mortality for both men and women.

| | Men | | Women | |
|---|---|---|---|---|
| Description | # Cases | Percentages | # Cases | Percentages |
| Sample size | 294,673 | | 199,285 | |
| Breast cancer | | | 7,736 | 3.88% |
| Ovarian cancer | | | 759 | 0.38% |
| Prostate cancer | 23,477 | 7.97% | | |
| Colorectal cancer | 4,693 | 1.59% | 2,291 | 1.15% |
| Lung cancer | 6,135 | 2.08% | 3,630 | 1.82% |

Table 3.2: Summary of the NIH-AARP data for cancer occurrence.

| Description | Men<br># Cases | Women<br># Cases |
|---|---:|---:|
| Sample size | 219,612 | 169,480 |
| CVD mortality | 8,112 | 4,028 |
| Cancer mortality | 12,247 | 7,344 |
| Other cause mortality | 10,821 | 6,547 |

Table 3.3: Summary of the NIH-AARP data for mortality. Cardiovascular disease has been abbreviated as CVD.

We consider three events of interest: cancer occurrence, morality, and all-cause mortality. Cancer occurrence is defined as diagnosis of any of the five types of cancer in Table 3.2, mortality is defined as mutually exclusive outcome of one the three cause in Table 3.3, and all-cause mortality is the aggregation of *any* type of mortality. We consider these outcomes separately and fit separate models for each outcome. For each outcome the analysis is as follows.

- Model (3.1) is fit using all the components of the 2005 Healthy Eating Index.
- The Least Squares Approximation with an Adaptive Lasso penalty is used to identify the relevant subset of Healthy Eating Index components.
- Model (3.1) is refit using only components selected by the Least Squares Approximation.

This results in three sets of selected components and three sets of parameter estimates

The $\boldsymbol{\alpha}^*$ coefficients, which correspond to the rescaled Healthy Eating Index described in Section 3.2, are provided for cancer occurrence, morality, and all-cause mortality in Table 3.4. The variance of the unscaled $\boldsymbol{\alpha}$ coefficients is calculated with sandwich estimator, and variance of the re-scaled $\boldsymbol{\alpha}^*$ is calculated using the Delta Method. This derivation is provided in Appendix B.5. We do not provide confidence intervals for components in $\boldsymbol{\alpha}^*$ which are set to 0 by the Least Squares Approximation.

| | Cancer | Mortality | All-Cause Mortality |
|---|---|---|---|
| Whole Grains | 3.98 (3.18, 4.78) | 5.59 (5.04, 6.13) | 5.61 (5.03, 6.18) |
| Total Grains | 1.34 (0.59, 2.09) | 0.87 (0.41, 1.33) | 0.93 (0.47, 1.40) |
| Whole Fruit | 1.56 (0.74, 2.37) | 0.22 (-0.12, 0.57) | 0.39 (0.02, 0.08) |
| Total Fruit | 2.71 (1.87, 3.55) | **0** | **0** |
| Total Veg. | 1.37 (0.47, 2.26) | 2.13 (1.57, 2.68) | 2.04 (1.46, 2.62) |
| DOL Veg. | 1.52 (0.65, 1.14) | 0.81 (0.39, 0.68) | 0.73 (0.30, 1.17) |
| Dairy | 0.90 (0.65, 1.14) | 0.53 (0.39, 0.68) | 0.58 (0.43, 0.74) |
| Meat & Beans | **0** | 1.06 (0.81, 1.31) | 0.98 (0.71, 1.25) |
| Oils | 0.58 (0.30, 0.86) | 0.59 (0.42, 0.77) | 0.63 (0.44, 0.81) |
| Sodium | 1.28 (0.95, 1.61) | 1.96 (1.78, 2.13) | 1.85 (1.66, 2.04) |
| Saturated Fats | 1.00 (0.72, 1.28) | 1.04 (0.88, 1.22) | 1.09 (0.91, 1.27) |
| Empty Calories | **0** | **0** | **0** |

Table 3.4: Results from Section 3.4.2 when the outcome of interest is cancer occurrence, various types of mortality, and aggregated all-cause mortality. Provided estimates are found by fitting the logistic regression model from Section 3.2 using only the subset of components chosen by the Least Squares Approximation. Parentheses are 95% confidence intervals. Bold 0's indicate components which are set to 0 by the Least Squares Approximation.

### 3.4.2 Results

The Healthy Eating Index puts a large penalty on diets high in empty calories. Empty calories, refereed to as SoFAAS in Table 3.1, and and made up of solid fats, alcohol and added sugars, make up 20 points of the Healthy Eating Index score. This means that someone with a diet high in empty calories will always be assigned a score below 80 regardless of their other nutritional intake. This is the largest contribution by a single component. This is in stark contrast to our results. In each analysis, the Least Squares Approximation forces Empty Calories to take a value of 0. That is, we find that empty calories are not very predictive of mortality.

Total grains appear to be undervalued by the Healthy Eating Index. For example, a person receiving a perfect score of 5 for whole grains in the original Healthy Eating Index would be reassigned a score of $5.61 \times 5 = 28.05$ if all-cause mortality was of interest. Similarly, our assessment gives total vegetables over twice its original weight when predicting mortality. It is also apparent that for any kind of mortality, the 2005 Healthy Eating Index may overstate the importance of fruit in general.

To some, the 2005 Healthy Eating Index may appear more appropriate for predicting cancer than it does for predicting mortality. This possible concern may have prompted the development of the Alternative Healthy Eating Index by McCullough et al. (2002).

## 3.5 Simulations

We justify the numeric results from Section 3.4.2 and the theory from Section 3.3.1 with two sets of simulations. The first simulation examines the stability of the estimation procedure as the sample size changes. In the second simulation we generate data from a similar model to the data example in Section 4.6.

### 3.5.1 Subset Analysis

To ensure that the results presented in Section 3.4.2 are robust and not an artifact of the large sample size, we split the data by a factor of $1/4$ and $1/8$ and rerun the analysis. We present the results when all-cause mortality is the outcome of interest. We want to ensure that the estimation procedure and the variable selection procedure perform are roughly similar for each subset. These results are given in Table 3.5. The results are fairly stable.

### 3.5.2 Variable Selection and Coverage

We simulate from the model

$$\mathrm{pr}(Y_{ik\ell}|X_{ikj}, Z_{ik\ell}) = H(\beta_{k\ell}\textstyle\sum_{j=1}^{J} X_{ijk\ell}\alpha_j + Z_{ik\ell}^{\mathrm{T}}\theta_{k\ell}) = p_{ik\ell}, \tag{3.5}$$

where there are $k = 2$ populations, $L_1$ = 3 diseases in the first population, and $L_2$ = 4 diseases in the second population. We set $\beta_1 = (-1, -0.08, -0.04)^{\mathrm{T}}$ and $\beta_2 = (-0.09, -0.06, -0.03, -0.01)^{\mathrm{T}}$. The Healthy Eating Index Measurements, $X_{ikj}$, and covariates, $Z_{ik\ell}$ are sampled without replaced from the NIH-AARP Study of Diet and Health. The Healthy Eating Index measurements for total fruit and whole fruit, as well as total grains and whole grains are summed together to create two components representing fruit and grains. This is done because the measurements are highly

|  | 1/8 (N= 48636) | 1/4 (N= 97273) | 1 (N = 389092) |
|---|---|---|---|
| Total Grains | 1.37 (-0.05, 2.8) | 1.39( 0.76, 2.03) | 0.93 (0.47, 1.40) |
| Total Fruit | **0** | **0** | **0** |
| Whole Fruit | 0.29 (-0.83, 1.41) | 0.47 (0.03 0.97) | 0.39 (0.46, 1.40) |
| Whole Grains | 5.26 (3.56, 6.96) | 5.11 (4.34, 5.87) | 5.60 (5.02, 6.18) |
| Total Veg. | 1.64 (0.13, 3.4) | 2.08 (1.29, 2.87) | 2.04 (1.46, 2.62) |
| DOL Veg. | 0.87 (-0.43, 2.17) | 0.60 (0.01, 1.19) | 0.73 (0.30, 1.17) |
| Dairy | 0.6 (0.14, 1.06) | 0.56 (0.35, 0.77) | 0.58 (0.43, 0.74) |
| Meat & Beans | 0.89 (0.10,1.68) | 0.97 (0.61, 1.33) | 0.98 (0.71, 1.25) |
| Oils | 0.81 (0.26, 1.36) | 0.65 (0.40 0.90) | 0.63 (0.44, 0.81) |
| Sodium | 1.9 (1.33, 2.48) | 2.04 (1.79, 2.30) | 1.85 (1.66, 2.04) |
| Saturated Fats | 1.08 (0.53, 1.63) | 0.95 (0.71,1.19) | 1.09 (0.91, 1.27) |
| Empty Calories | **0** | **0** | **0** |

Table 3.5: Results of the subset analysis in Section 3.5.1 when all-cause mortality is the outcome of interest. The fraction refers to the proportion of the original data set used to fit the model. It is followed in parenthesis by the sample size used in the analysis. All results refer to refitting the stratified model of Section 3.2 using the subset of components identified by the Least Squares Approximation. Point estimates are given and 95% confidence intervals follow in parenthesis. Bold 0's indicates parameters set to 0 by the Least Squares Approximation. The results are stable through across all subsets.

correlated. This means the dimension of $\mathbf{X}$ is 10, instead of 12 as in the real data set. We set $\boldsymbol{\alpha}$ to be a vector of length $10$ with $5$ nonzero components. The nonzero components of $\boldsymbol{\alpha}$ are set to $3, 3, 2.5, 2.5$ and $2$. Two continuous covariates are selected from $\mathbf{Z}$ and the parameters $\theta_{kl}$ are drawn from Uniform$[-2, 2]$. distribution We simulate $N = 1000$ data sets with a range of sample sizes. The $\gamma$ tuning parameter is set to 2.

The theory in Section 3.3.1 makes two guarantees: the probability of selecting the correct subset of predictors approaches $1$ and the asymptotic covariance matrix of the nonzero parameters, $\Sigma_T$ is correctly estimated. We demonstrate these claims by simulation. We test for variable selection with

$$N^{-1}\textstyle\sum_{i=1}^{N}\mathbb{I}(\widehat{\alpha}_{LSA,i} = \alpha_T),$$

where $\widehat{\boldsymbol{\alpha}}_{LSA,i}$ is the subset of $\boldsymbol{\alpha}$ chosen by the Least Squares Approximation on the $i^{th}$ simulation,

$\alpha_T$ is the set of true nonzero predictors, and "$=$" denotes set equality.

We demonstrate the asymptotic consistency of $\widehat{\Sigma}_T$ by showing that we can construct confidence intervals for the nonzero components which have proper coverage. Coverage is tested separately for each component of $\boldsymbol{\alpha}$ with

$$N^{-1}\textstyle\sum_{i=1}^{N}\mathbb{I}\{\alpha_j \in (\widehat{\alpha}_j - z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2}, \widehat{\alpha}_j + z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2})\},$$

where $\alpha_j$ is the $j^{th}$ nonzero component of $\boldsymbol{\alpha}$, $\Sigma_{jj}^{1/2}$ is the standard error of $\widehat{\alpha}_j$, and $z_{\alpha/2}$ is the upper $\alpha$ percentile of the standard normal distribution. The estimates $\widehat{\alpha}$ and $\widehat{\Sigma}$ refer to the estimates obtained after refitting (3.5) with only the variables selected by the Least Squares Approximation.

The results are given in Table 3.6. We point out that while the coverage probabilities are close to nominal at $n = 1000$, consistent variable selection requires larger sample sizes. Proper variable selection is seen at $n = 10000$, though acceptable results can be seen at smaller sample sizes. A large sample size is likely required because we use two asymptotic approximations: a sandwich estimator for the covariance matrix and the least squares approximation for variable selection. Regardless of the sample size requirements, the NIH-AARP Study of Diet and Health is more than large enough for consistent variable selection and proper confidence interval coverage.

| Sample size | % Selected | V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|---|---|
| 10000 | 97.9 | 94.6 | 95.6 | 96.0 | 94.7 | 96.8 |
| 5000 | 94.6 | 94.6 | 94.5 | 95.0 | 95.3 | 96.4 |
| 2500 | 94.4 | 94.4 | 94.9 | 94.6 | 94.6 | 96.0 |
| 1000 | 93.6 | 93.6 | 93.6 | 93.0 | 92.3 | 93.2 |

Table 3.6: SSimulation results from Section 3.5.2 showing coverage and the percent of time the true model was selected. The second column gives the proportion of the 1000 simulations which identified the correct 5 predictor subset. The remaining columns, V1 through V5, give the approximate coverage of 95% confidence intervals for the 5 nonzero predictors.

### 3.6 Discussion

Using nonlinear models and adaptive Lasso penalization, we have introduced a method to continually reassess composite score techniques. Our method produces parameter estimates and covariance estimates which are asymptotically consistent. Asymptotically, the penalization method chooses the correct subset of coefficients. Our empirical results are similar to Ma et al. (2017), though they included a nonparametric component in their model that we do not. The authors found that the flexibility of their semiparametric model was not needed in their actual data analysis. While highly technical, it is possible to expand our analysis into their framework.

If a researcher suspects that a particular composite score does not apply well to their population of interest, they use the methods outlined in this paper to reweigh the relative importance of each score component and see if each component is necessary. Analyzing composite scores in this way can lead to important new finding. Our analysis of the Healthy Eating Index suggests that the negative effects of empty calories may be overstated. Similarly, the relative importance of fruit and whole grains can change dramatically when considering mortality risk instead of cancer risk.

There is considerable scope for future work. Empirically, future work should address the correlation between many of the dietary components. It is, for example, impossible to consume Total Grains without also consuming Whole Grains. The components selected and parameter estimates may change if the collinearity is addressed. The methodology in this work may be may be extended to variable selection while using a nonparametric or semiparameteric model for diet. Diet may be modeled with a single index model (Carroll et al., 1997), letting the weighted sum of Healthy Index Scores vary freely as in Ma et al. (2017), or with a Generalized Additive Model (Wood, 2017), modeling each dietary component separately with a smooth function.

# 4. SAMPLE SPLITTING AS AN M-ESTIMATOR WITH APPLICATION TO PHYSICAL ACTIVITY SCORING

## 4.1 Introduction

In many applications it is useful to think of performing a statistical procedure in two stages: model building and validation. When building a classification model, it is commonplace to reserve a portion of the data set for testing predictive accuracy. Similarly, many nonparametric regression methods depend on an unknown tuning parameter, $\lambda$, which may be chosen using cross-validation. Here, model fit is determined using data which was not used to build the model. In both examples, reusing data for model fitting and validation without withholding a portion will result in overfitting.

This two stage procedure can be seen in real-world problems. Consider providing dietary recommendations in the form of *composite score* or *index*. One particular example is the Healthy Eating Index. The Healthy Eating Index was designed by the United States Department of Agriculture (USDA) to provide dietary recommendations to Americans (U.S. Department of Health and Human Services and U.S. Department of Agriculture, 2005). The Healthy Eating Index was later validated and confirmed to accurately measure diet and predict risk of disease with a 24-hour food recall study (Guenther et al., 2008c).

In Section 4.2, we introduce a scenario of building a composite score and providing risk estimates without a second independent population. We may choose to treat this problem like the nonparametric or classification problems and split the data set, $\mathcal{D}$, into two pieces, $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$. The partition, $\mathcal{D}_{in}$, serves as a "training" data set for model fitting while $\mathcal{D}_{out}$ serves as a "validation" data set for checking model fit and performing hypothesis tests. In this way we can mimic the building-to-validation of having two independent populations.

It is natural to extend this process: If one split is good, *many splits* must be better. A single split has variability; a fortuitous split may give parameter estimates very close to their true value while another split may yield poor estimates. We follow the ideas of Meinshausen et al. (2009), though in a different context than in their variable selection paper. The authors create $B$ copies of the data and partition each of these $b = 1, \dots, B$ copies to create, $\mathcal{D}_{in}^{(b)}$ and $\mathcal{D}_{out}^{(b)}$. Model fitting is done on each $\mathcal{D}_{in}^{(b)}$ and hypothesis tests are performed on $\mathcal{D}_{out}^{(b)}$.

We explore multiple splitting for parametric models in Section 4.3. We fit a parametric model on $\mathcal{D}_{in}^{(b)}$and then fit a model on $\mathcal{D}_{out}^{(b)}$using a function of the parameter estimates as a predictor. We find the surprising result that as the number of splits and the sample size approach $\infty$, split sample estimators become equivalent to fitting both models simultaneously using *stacked estimating equations* (Carroll et al., 2006, Appendix A.6.6).

Our method is similar to *divide-and-conquer algorithms* (Li et al., 2013; Battey et al., 2015), particularly the average mixture algorithm (Chang et al., 2017; Zhang et al., 2012). These algorithms reduce the computational cost of statistical inference on large datasets by dividing the dataset into many smaller subsets which can be quickly analyzed with traditional software (R, SAS, etc). The results from these subsets are then combined together, often with the sample mean. None of this work, to our knowledge, incorporates a dependent parameter and fits a second model to estimate this parameter.

This paper is organized as follows: In Section 4.2, we introduce a motivating example. In Section 4.3 we describe sample splitting using estimating equations and M-estimators. Section 4.4 shows theoretical results for a single split, a finite number of splits, and when the number of splits approach infinity. Section 4.5 has simulations that illustrate the asymptotic theory. Section 4.6 uses sample splitting to develop a physical activity index. Section 4.7 has a discussion. All technical details are collected in Section C.

### 4.2 Motivating Example: Physical Activity and Survival

This work is partly motivated by the creation and analysis of a physical behavior score to predict mortality. Researchers may be concerned about using the same data to create the score and use that score to estimate risk ratios. We have found sample splitting appealing to practitioners because the relative risks are estimated with different data from that which builds the score.

We build the physical behavior score using the NIH-AARP Study of Diet (Schatzkin et al., 2001). Participants self-reported physical behaviors which are then characterized into 8 discrete components. We apriori specify the expected relationship between these physical activities and survival to be consistent with the kinesiology literature. Using the training data $\mathcal{D}_{in}$ for a single sample split, we fit a binary regression model to survival that satisfies these relationships. The 8 components and their marginal models are listed in Table 4.3. The expected relationships are justified in Section 4.6. We rescale the fitted from this model so people with high levels of beneficial activity are assigned a score near 100 and people with low levels of beneficial activity are assigned a score close to 0. We denote the rescaled fitted values as $f(\mathbf{X};\widehat{\theta})$

We now ask: Is our composite score predictive of mortality? If so, how strong is the effect? We use $f(\mathbf{X};\widehat{\theta})$ as a predictor in a logistic regression, along with covariates, $\mathbf{Z}$. In the test data $\mathcal{D}_{out}$, we the model the probability of mortality as

$$\mathrm{pr}(Y_i = 1|X_i, Z_i) = H\{\beta_0 f(X_i;\widehat{\theta}) + Z_i^{\mathrm{T}}\beta\}, \tag{4.1}$$

where $H(\cdot)$ is the logistic distribution function. Our primary interest is in $\beta_0$, which describes the relative risks of survival as a function of the composite score.

## 4.3 Sample Splitting Methodology

Denote the total number of sample splits as $B$. We consider two cases: a single sample split ($B = 1$) and many splits ($1 < B < \infty$). We describe the algorithm for sample splitting using estimating equations. In Section 4.4, we provide an asymptotic theory for both of these cases, as well when $B \to \infty$.

### 4.3.1 Basic Formulation

In what follows, we use the term *estimating equation* to also include a *M-estimator equation*. We observe two types of responses $(\mathcal{W}_i, \mathbf{Y}_i)$ that are independent and identically distributed and two types of covariates, $(\mathbf{X}_i, \mathbf{Z}_i)$. There are two parameters, $\theta$ and $\beta$. The relationship between $\mathcal{W}$ and $(\mathbf{X}, \mathbf{Z}, \theta)$ is described by a model which has an estimating equation $\Psi(\mathcal{W}, \mathbf{X}, \mathbf{Z}, \theta)$. The relationship of $\mathbf{Y}$ and $(\mathbf{X}, \mathbf{Z}, \beta, \theta)$ is described a model that has an estimating equation $\mathcal{K}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \beta, \theta)$. We define $\theta$ and $\beta$ as the solutions to $\Psi(\cdot)$ and $\mathcal{K}(\cdot)$ respectively.

For a single data set, consistent estimation of $(\beta, \theta)$ can be done routinely by solving the stacked estimating equation

$$0 = \textstyle\sum_{i=1}^{n}\{\Psi^{\mathrm{T}}(\mathcal{W}_i, \theta), \mathcal{K}^{\mathrm{T}}(\mathbf{Y}_i, \beta, \theta)\}. \tag{4.2}$$

Asymptotic theory for such estimators is well-known (Huber, 1964, 1967; Stefanski and Boos, 2002).

In the physical activity problem of Section 4.2 $\mathcal{K}(\cdot)$ does not depend directly on $\theta$ but rather on a function of $\theta$ and $\mathbf{X}$, denoted with $f(\mathbf{X}; \theta)$. That is,

$$\mathcal{K}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \beta, \theta) = \mathcal{K}(\mathbf{Y}, f(\mathbf{X}; \theta), \mathbf{Z}, \beta).$$

Stacked estimating equations and the methods in this paper apply to this situation, but to simplify notation we do not include $f(\mathbf{X}; \theta)$ when writing $\mathcal{K}(\cdot)$. The proofs in the Appendix consider this case and make explicit the dependence of $\mathcal{K}(\cdot)$ on $f(\mathbf{X}; \theta)$, rather than just $\theta$.

### 4.3.2 A Single Split

In our experience, there is concern among practitioners that solving (4.2) on the complete data is somehow "cheating", because all the data are used to build the score while simultaneously estimating risk. The concern is that risk estimators in such a scenario will overstate the usefulness of the score. One way to alleviate this concern is to split the data into two disjoint groups, $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, and use these to estimate $\theta$ and $\beta$ respectively. Let $(\delta_1, ..., \delta_n)$ be independent and identically distributed Bernoulli$(\pi)$ random variables. In our applications we set $\pi = 1/2$ so that the splits are approximately equal size. We denote the first partition, $\mathcal{D}_{in}$, by $\delta = 1$, which estimates $\theta$ by solving

$$0 = \textstyle\sum_{i=1}^{n}\delta_i\Psi(\mathcal{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \theta). \tag{4.3}$$

In the second partition, $\mathcal{D}_{out}$, denoted by $\delta = 0$, we estimate $\beta$ by solving

$$0 \quad = \quad \textstyle\sum_{i=1}^{n}(1-\delta_i)\mathcal{K}(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \beta, \widehat{\theta}). \tag{4.4}$$

Since the two groups are independent, solving (4.3)-(4.4) alleviates concerns about overfitting. This is an analogue to the variable selection procedure of Wasserman and Roeder (2009), who also propose a single split, with variable selection conditioned on the data with $\delta = 1$.

### 4.3.3 Many Splits

Section 4.3.2 give a $\widehat{\beta}$ that depend on the particular random sample split. Meinshausen et al. (2009) and Dezeure et al. (2015) criticize this and call it a "p-value lottery" as the p-value can vary from 0 to 1 depending on the split. Instead, in their context, they suggest using multiple data splits to eliminate simulation variability from choosing only a single split. Such an approach would randomly split the data into two parts $b = 1, ..., B$ times to create $B$ partitions of the data.

Define $B$ to be the number of sample splits. We define an indicator vector for each sample split, $b$. For $b = 1, .., B$ and $i = 1, ..., n$, let $(\delta_{1b}, ..., \delta_{nb})_{b=1}^{B}$ be independent and identically distributed Bernoulli$(\pi)$ random variables. Set $\delta_{ib} = 1$ if the $i^{th}$ person is selected into the $b^{th}$ training set $\mathcal{D}_{in}^{(b)}$. Then solve

$$0 \quad = \quad \textstyle\sum_{i=1}^{n}\delta_{ib}\Psi(\mathcal{W}_i, \mathbf{X}_i, \mathbf{Z}_i, \theta).$$

to get an estimate $\widehat{\theta}_b$. The subscript denotes the dependence on the parameter estimate of the $b^{th}$ sample split.

Now, set $\delta_{ib} = 0$ if the $i^{th}$ person is selected into the test set, $\mathcal{D}_{out}^{(b)}$. We then get an estimate $\widehat{\beta}_b$ by solving

$$0 \quad = \quad \textstyle\sum_{i=1}^{n}(1-\delta_{ib})\mathcal{K}(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \beta, \widehat{\theta}).$$

This gives $B$ estimates of $\theta$ and $\beta$. We combine them with the sample mean to get $\widehat{\theta} =$

$B^{-1}\sum_{b=1}^{B}\widehat{\theta}_b$ and $\widehat{\beta} = B^{-1}\sum_{b=1}^{B}\widehat{\beta}_b$.

### 4.4 Asymptotic Theory

In this section we provide asymptotic theory for three cases of sample splitting: $B = 1, 1 < B < \infty$, and $B \to \infty$. We provide asymptotic expansions from which asymptotic normality and the asymptotic covariance matrix can be derived. Hypothesis tests can be performed using Wald test statistics and Normal-based confidence intervals.

#### 4.4.1 Single Split Asymptotics

Suppose there is only a single split, $b = 1$. Define

$$\begin{aligned}
\Omega_{nb}(\theta) &= n^{-1}\textstyle\sum_{i=1}^{n}\delta_{ib}E\{\partial\Psi(W,\theta)/\partial\theta^{\mathrm{T}}\};\\
\Lambda_{nb}(\beta,\theta) &= n^{-1}\textstyle\sum_{i=1}^{n}(1-\delta_{ib})E\{\partial\mathcal{K}(Y,\beta,\theta)/\partial\beta^{\mathrm{T}}\};\\
\Delta_{nb}(\beta,\theta) &= n^{-1}\textstyle\sum_{i=1}^{n}(1-\delta_{ib})E\{\partial\mathcal{K}(Y,\beta,\theta)/\partial\theta^{\mathrm{T}}\}.
\end{aligned}$$

Using standard M-estimation theory (Stefanski and Boos, 2002), and suppressing the dependence of $\{\Omega_{nb}(\theta), \Lambda_{nb}(\beta,\theta), \Delta_{nb}(\beta,\theta)\}$ on $(\beta,\theta)$, we find that

$$\begin{aligned}
n^{1/2}(\widehat{\theta}_b - \theta) &= -\Omega_{nb}^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\delta_{ib}\Psi(W_i,\theta) + o_P(1). &(4.5)\\
n^{1/2}(\widehat{\beta}_b - \beta) &= -\Lambda_{nb}^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}(1-\delta_{ib})\mathcal{K}(Y_i,\beta,\theta)\\
&\quad +\Lambda_{nb}^{-1}\Delta_{nb}\Omega_{nb}^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\delta_{ib}\Psi(W_i,\theta) + o_P(1). &(4.6)
\end{aligned}$$

Define

$$\mathcal{A}_{ib} = \begin{bmatrix} -\Omega_{nb}^{-1}\delta_{ib}\Psi(W_i,\theta) \\ -\Lambda_{nb}^{-1}(1-\delta_{ib})\mathcal{K}(Y_i,\beta,\theta) + \Lambda_{nb}^{-1}\Delta_{nb}\Omega_{nb}^{-1}\delta_{ib}\Psi(W_i,\theta) \end{bmatrix},$$

and define $\widehat{\mathcal{A}}_{ib}$ as $\mathcal{A}_{ib}$ with all unknown quantities replaced with their empirical estimates. Then the joint asymptotic covariance matrix of $(\widehat{\theta}_b, \widehat{\beta}_b)$ can be estimated by $n^{-1}S(\widehat{\mathcal{A}}_{ib})$, where $S(\cdot)$ is the sample covariance.

### 4.4.2 Multiple Split Asymptotics

Fix a finite number of splits, $1 < B < \infty$. The final estimates of $\theta$ and $\boldsymbol{\beta}$ are the average of the single split estimates from Section 4.4.1. Thus, the expansions of $\theta_b$ and $\beta_b$ in (4.5) and (4.6) can be replaced with their averages over $B$

$$\begin{aligned} n^{1/2}(\widehat{\theta}_b - \theta) &= -B^{-1}\textstyle\sum_{b=1}^{B} n^{-1/2}\Omega_{nb}^{-1}\sum_{i=1}^{n}\delta_{ib}\Psi(W_i,\theta) + o_P(1), \\ n^{1/2}(\widehat{\beta}_b - \beta) &= -B^{-1}\textstyle\sum_{b=1}^{B}\{\Lambda_{nb}^{-1}n^{-1/2}\sum_{i=1}^{n}(1-\delta_{ib})\mathcal{K}(Y_i,\beta,\theta) \\ &\qquad -\Lambda_{nb}^{-1}\Delta_{nb}\Omega_{nb}^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\delta_{ib}\Psi(W_i,\theta)\} + o_P(1). \end{aligned}$$

Define $\mathcal{A}_i = B^{-1}\sum_{b=1}^{B}\mathcal{A}_{ib}$, where $\mathcal{A}_{ib}$ is the same as in Section 4.4.1. Define $\widehat{\mathcal{A}}_i = \sum_{b=1}^{B}\widehat{\mathcal{A}}_{ib}$. The joint asymptotic covariance matrix of $(\widehat{\theta},\widehat{\beta})$ can be estimated consistently by $n^{-1}S(\widehat{\mathcal{A}}_i)$.

### 4.4.3 Increasing Number of Splits Asymptotics

We assume that $B$ increases at a slower rate than $n$, $B = B_n = o_P(n^{-1/2+a})$ for any $a > 0$. Since the user chooses $B$, this is not a restrictive assumption. Since the $\delta_{ib}$ are independent of the rest of the data, $B^{-1}\sum_{b=1}^{B}\delta_{ib} \to \pi$ and $B^{-1}\sum_{b=1}^{B}(1-\delta_{ib}) \to (1-\pi)$. The asymptotic expansions for $\widehat{\theta}$ and $\widehat{\beta}$ simplify to

$$\begin{aligned} n^{1/2}(\widehat{\theta} - \theta) &= -\Omega^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\pi\Psi(W_i,\theta) + o_P(1); \qquad (4.7) \\ n^{1/2}(\widehat{\beta} - \beta) &= -\Lambda^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}(1-\pi)\mathcal{K}(Y_i,\beta,\widehat{\theta}_b) + \\ &\qquad \Lambda^{-1}\Delta\Omega^{-1}\textstyle\sum_{i=1}^{n}\pi\Psi(W_i,\theta) + o_P(1). \qquad (4.8) \end{aligned}$$

The asymptotic expansions in (4.7) and (4.8) are also the expansion of $\widehat{\beta}$ and $\widehat{\theta}$ using stacked estimating equations with all the entire sample. This is justified in Appendix C.1.4. This shows that when $\pi = 1/2$, the estimates from sample splitting become asymptotically equivalent to estimates from the entire data set and not using any sample splitting.

## 4.5 Simulations

We simulate from two configurations. We consider a single split (B = 1) and B = 25 sample splits and set $\pi = 1/2$, so each $\mathcal{D}_{in}^{(b)}$ and $\mathcal{D}_{out}^{(b)}$ are approximately the same size. We express both configurations as regression models to simplify interpretation but can be expressed as estimating equations.

Both simulations are related to the problem of estimating the ratio of two normal random variables (Fieller, 1932, 1954), see Wang et al. (2015) for a recent review and some alternative methods. It is known that there is no confidence interval with guaranteed coverage and that has finite length with probability one. It is also known that asymptotic standard errors based on estimating equations and the delta method generally have somewhat less than nominal coverage probabilities except for larger sample sizes. We will see the Fieller phenomenon occur in our simulations.

The first configuration is a set of two linear regression models,

$$W_i = X_i^{\mathrm{T}}\theta + \epsilon_i; \qquad Y_i = \beta_0(X_i^{\mathrm{T}}\theta) + \epsilon_i. \tag{4.9}$$

We set $\theta = c(1,1,1)/\sqrt{3}$, $\beta_0 = 1/\sqrt{3}$, $\epsilon_i \sim N(0, 0.5)$, and vary $n = 25, 50, 100, 250, 500$. Results are in Table 4.1. Confidence intervals have approximately 95% coverage for $n > 250$.

| | Linear | |
|---|---|---|
| | B = 1 | B = 25 |
| n = 50 | 92.60 | 90.45 |
| n = 100 | 93.70 | 92.85 |
| n = 250 | 94.90 | 94.15 |
| n = 500 | 94.70 | 95.15 |
| n = 1000 | 95.10 | 94.95 |

Table 4.1: Results of the simulation in Section 4.5 using the linear model. Coverage of asymptotic 95% confidence intervals is provided for $B = 1$ and $B = 25$ sample splits.

The second configuration is two binary regression models, given as

$$\mathrm{pr}(W_i = 1|X_i) = H(X_i^{\mathrm{T}}\theta); \qquad \mathrm{pr}(Y_i = 1|X_i) \;=\; H\{\beta_0 f(X_i;\theta)\}, \tag{4.10}$$

where $H(\cdot)$ is the logistic distribution function. We set $f(X_i, \theta) = 2 + 1.5(X_i^{\mathrm{T}}\theta)^2$. The specific $f(\mathbf{X}, \theta)$ is set to mimic the case study in Section 4.2. This transformation also rescales the fitted values, although not to the same 0 to 100 range, and contains a nonlinear transformation of the parameters and data. The new scale is arbitrary; we choose the values 2 and 1.5, so there are are a reasonable number of both 0's and 1's in $\mathbf{Y}$. Again we set $\theta = (1, 1, 1)^{\mathrm{T}}/\sqrt{3}$ and $\beta_0 = 1/\sqrt{3}$ and vary $n = 50, 100, 250, 500$. Results for $B = 1$ and $B = 25$ are in Table 4.2, and are much the same as in Table 4.1, with the Fieller phenomenon being observed. Logistic regression naturally requires larger samples than linear regression to achieve nominal coverage.

| | Logistic | |
|---|---|---|
| | B = 1 | B = 25 |
| n = 50 | · | · |
| n = 100 | 88.95 | 85.40 |
| n = 250 | 93.85 | 90.40 |
| n = 500 | 93.40 | 94.00 |
| n = 1000 | 94.20 | 94.55 |

Table 4.2: Results of the simulation in Section 4.5 using the logistc model. Coverage of asymptotic 95% confidence intervals is provided for $B = 1$ and $B = 25$ sample splits. Results from the logistic model are unstable at $n = 50$ and are omitted.

### 4.6 Data Analysis

Participants in the NIH AARP Study of Diet and Health were ask to complete a questionnaire to measure physical behaviors, medical history, and risk factors for disease. Around a fifth of the total participants (N = 163,106) responded and met criteria for inclusion. Survey responses were translated to time or energy spent in five aerobic activities, two types of sitting activities, and sleep.

We model each of the physical activity components to be consistent with the existing kinesiology literature. The dose-response relationship between aerobic activity and survival has been established as somewhat concave and non-decreasing (Arem et al., 2015), with benefits for overall health which level off with increasing activity. Sedentary time is known to have a negative effect on overall health (Grøntved and Hu, 2011; Prince et al., 2014). Sleep is known to be beneficial in reasonable doses but too much or too little sleep is indicative of poor health (Yin et al., 2017), suggesting a concave, parabolic relationship with overall health. Table 4.3 lists the expected relationships and the marginal models we enforce to describe these relationships.

| **Activity** | **Expected Relationship** | **Marginal Model** |
|---|---|---|
| Vigorous Activity | Concave Increasing | 3-parameter Logistic |
| Moderate Activity | Concave Increasing | 3-parameter Logistic |
| Light Household Activity | Concave Increasing | 3-parameter Logistic |
| MVPA Household Activity | Concave Increasing | 3-parameter Logistic |
| Weight Training | Concave Increasing | 3-parameter Logistic |
| Hours Sitting Other than TV | Decreasing | Linear |
| Hours of TV Sitting | Decreasing | Linear |
| Hours of Sleep | Concave | Quadratic |

Table 4.3: Each of the 8 physical activity variables with their expected relationship and marginal model when predicting survival. The expected relationship is derived from physical activity literature.

### 4.6.1 Step 1: Developing the Score

On each $\mathcal{D}_{in}^{(b)}$we fit the model

$$\begin{aligned}\text{pr}(Y_i = 1 | X_i, Z_i) \quad &= \quad H\Bigg[\textstyle\sum_{j=1}^{5}\Big\{d_j - \dfrac{d_j}{1 + (X_{\text{aerob}_j}/c_j)^{b_j}}\Big\} + \theta_{\text{TV}} X_{\text{TV}} \\ &+ \quad \theta_{\text{Sit}} X_{\text{Sit}} + \theta_{\text{Sleep},1} X_{\text{Sleep}} + \theta_{\text{Sleep},2} X_{\text{Sleep}}^2 + Z_i^{\text{T}} \theta\Bigg], \qquad (4.11)\end{aligned}$$

where $Y_i = 1$ indicates survival until the end of the study, $X_i$ is a vector containing the 8 physical activity levels of the $i^{th}$ person, $Z_i$ is a vector of covariates including sex, race, education status, and an intercept, and $H(\cdot)$ is the logistic distribution function.

The first row of Figure 4.1, shows the fitted marginal models for three types of activity: moderate physical activity, sleep, and television sitting. The curves match their intended functional form: moderate physical activity is concave and increasing, sleep is concave, and television sitting is decreasing.

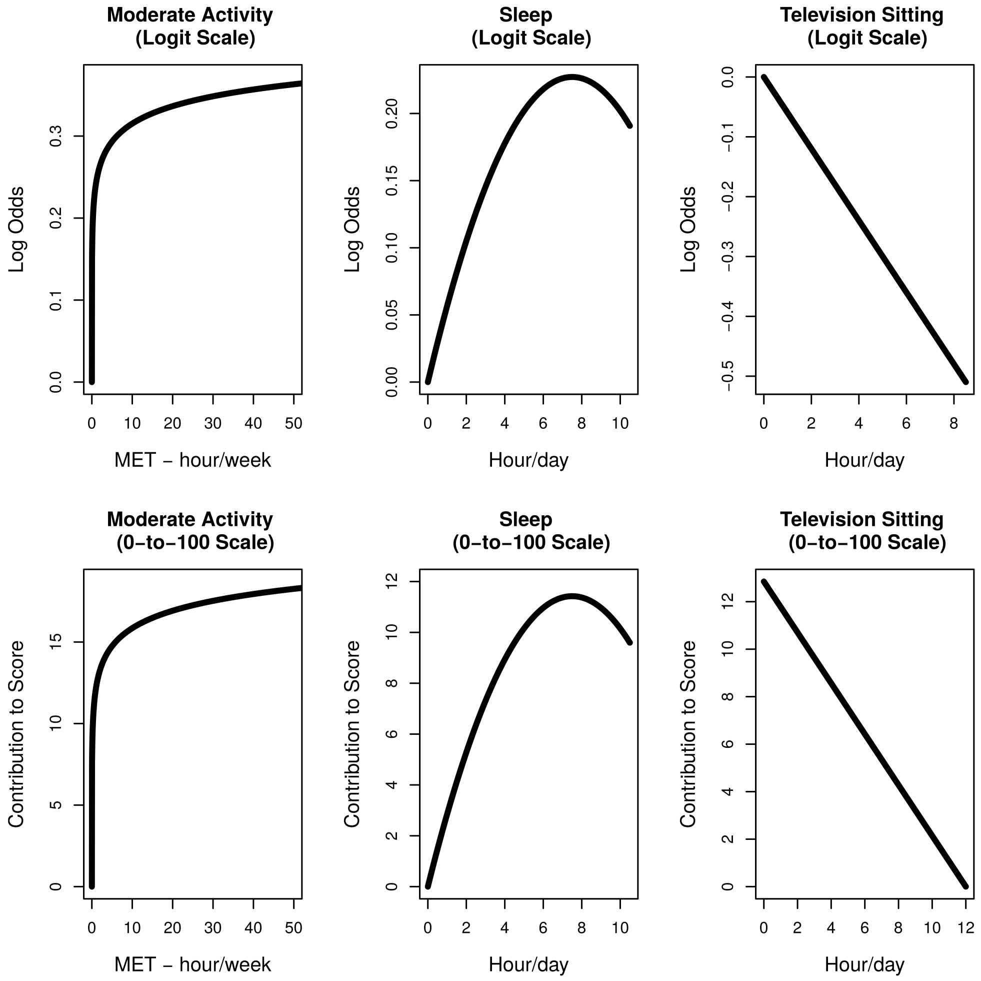


Figure 4.1: Three of the 8 marginal model plots from binary regression model. The first row shows the marginal models in the original scale while the second rows shows them in a 0-to-100 scale. Moderate activity is modeled to be concave and increase, sleep is modeled to be concave, and television sitting is modeled to be decreasing.

### 4.6.2 Step 2: Rescaling the Score

The fitted values from (4.11) are the logits of the effect of each type of physical activity on survival. We want to rescale the logits from the physical activity covariates so that their sum is between 0 and 100 with 0 being highest risk and 100 being lowest risk. While we fit model (4.11) with the additional covariates $Z$ to prevent confounding due to demographic information, we do not include the fitted values $Z_i^{\mathrm{T}}\,\widehat{\theta}$ when developing the score.

To rescale the logits, we first force all the physical activity marginal models to be positive by adding the absolute value of the minimum fitted value. For example, in the top row of Figure 4.1 we see that the marginal model for television sitting is negative for any amount of television sitting greater than 0. The function has a minimum of -0.5 at around 8 hours per day of sleep. By adding -0.5 to the fitted values, we can force this function to always be positive.

Next, we sum the maximum value obtained by each of the, now positive, marginal models and denote this with $T$. We can transform the fitted values with

$$\frac{T}{100}\Bigg[\textstyle\sum_{j=1}^{5}\Big\{d_j - \dfrac{d_j}{1+(X_{\mathrm{aerob}_j}/c_j)^{b_j}}\Big\} + \theta_{\mathrm{TV}}X_{\mathrm{TV}} \\ +\theta_{\mathrm{Sit}}X_{\mathrm{Sit}} + \theta_{\mathrm{Sleep,1}}X_{\mathrm{Sleep}} + \theta_{\mathrm{Sleep,2}}X_{\mathrm{Sleep}}^2\Bigg] \tag{4.12}$$

which puts the fitted values from physical activity on a scale from 0 to 100. We refer to the rescaled marginal models as the contributions to the total score. Three examples of these rescaled marginal models are shown in the bottom row of Figure 4.1. The contribution to the total score is given on the y-axis. Moderate activity, for example, accounts for up to 15 points of the total score of 100. Table 4.4 has an example of the physical activity score using the results from (4.12) on one particular split of the data. Table 4.4 lists the 8 physical behaviors with their relative contribution to a score of 100 and the criteria for receiving a perfect score in each criteria.

| Component | Contribution to Total | Criteria for Maximum |
|---|---|---|
| Vigorous Activity | 10 | >20 MET-hrs/wk |
| Moderate Activity | 30 | >50 MET-hrs/wk |
| Light Household Activity | 3 | >3 MET-hrs/wk |
| MVPA Household Activity | 25 | >20 MET-hrs/wk |
| Weight Training | 2 | > 2 MET-hrs/wk |
| Sitting Other than TV | 6 | < 3.5 hours |
| Hours of TV Sitting | 14 | < 2 hours |
| Hours of Sleep | 10 | 7.5 hours |
| Total | 100 | |

Table 4.4: Example physical activity score developed using half of the data. We fit the binary regression model from Section 4.2 and rescale the fitted values of the physical behaviors to be between 0 and 100. The middle column gives the proportion of the total score of 100 that each component can contribute. The third column gives the criteria for receiving the maximum score for each component.

### 4.6.3 Step 3: Risk Prediction Based on the Score

We denote the physical activity score created in the previous section with $f(\mathbf{X}, \theta)$. We then estimate the relationship between the physical behavior score and mortality using a Logistic Regression model on $\mathcal{D}_{out}^{(b)}$

$$\text{pr}(Y_i = 1|X_i, Z_i) \;=\; H\{\beta_0 f(X_i;\theta) + Z_i^{\mathrm{T}}\beta\}. \tag{4.13}$$

A plot of the estimates can be seen in Figure 4.2. The solid black line shows the value of $\widehat{\beta}_0$ with an increasing sample splits. For a particular number of sample splits, k, the solid blue shows $k^{-1}\sum_{b=1}^{k}\widehat{\beta}^{(b)}$. The dashed red lines are a 95% confidence interval. The overall mean of all these estimates is $\widehat{\beta} = -0.026$ and is calculated with $\widehat{\beta} = 50^{-1}\sum_{b=1}^{50}\widehat{\beta}^{(b)}$. It is denoted with a dashed black line.

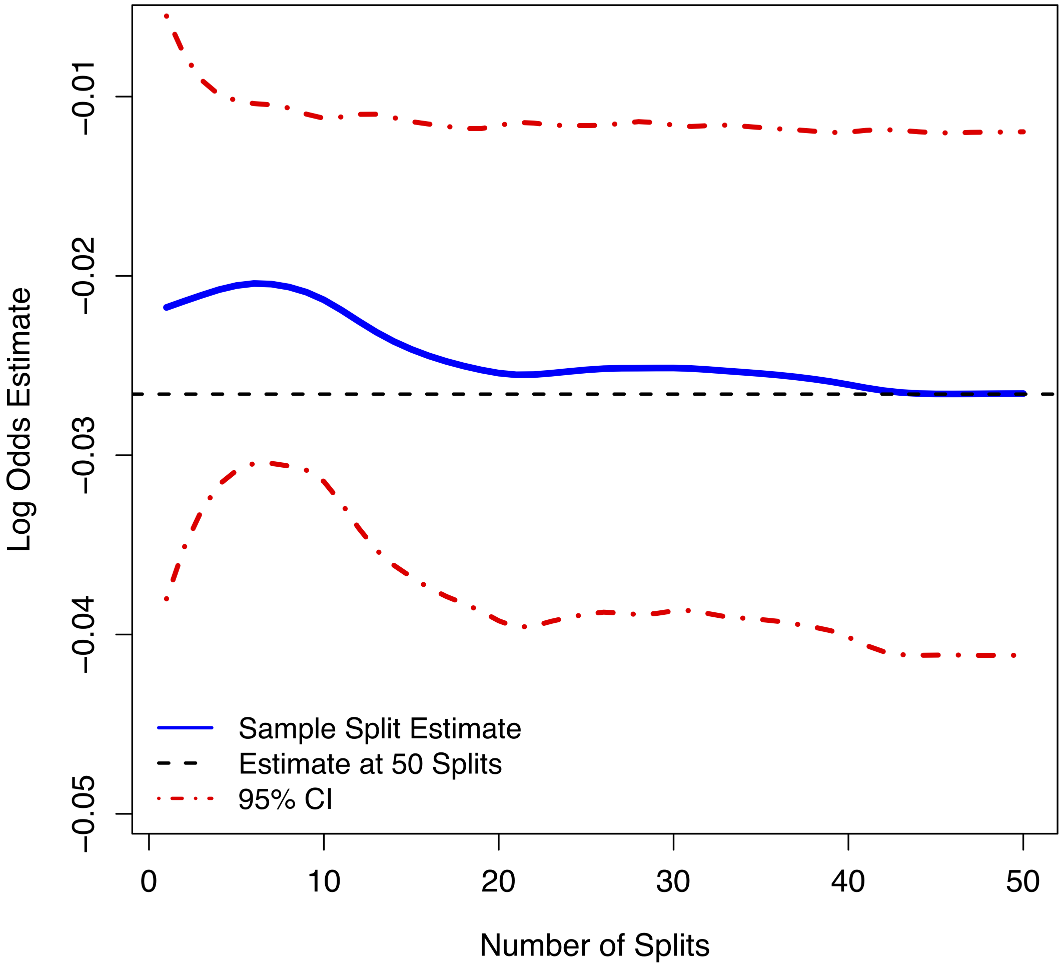


Figure 4.2: Log odds ratio of the physical activity score as the number of sample splits increases. The solid blue line denotes the estimate of $\beta$, the log odds ratio of the physical activity score as the number of sample splits increases. A 95% confidence interval is shown with the red dash-dotted lines. The final estimate of $\beta$, averaged over all 50 splits, is shown with a dashed black line at -0.026. The confidence intervals and parameter estimates are plotted using a smoothing spline.

There is variability when using a small number of number of splits. This is visible in Figure 4.2 for small values of $b$. Despite this $\widehat{\beta}_0$ changes little. The difference between the estimates of $\beta_0$ when $B = 1$ and $B = 50$ is only 0.004 and the confidence intervals are almost entirely overlapping.

### 4.7 Discussion

We explored sample splitting to perform validation without a second data set using estimating equations. We provided asymptotic expansions for a fixed sample split and a mean-aggregated split which can be used for hypothesis testing and asymptotic confidence intervals. In particular we are able to provide confidence intervals for a single split of the data. This scenario is of interest to practitioners who want a formal distinction between the data used for model building and the data used for validation (e.g. hypothesis testing).

In practice, there are two ways to evaluate a marginal score. One may treat the newly created score as a fixed quantity and use it as a predictor in a regression model. This ignores the variability due to estimating this score but results in smaller variance estimates. Instead, we consider the score as a function of unknown parameters which are estimated with uncertainty. This results in consistent parameter estimate which have higher variability than had the score been considered fixed.

We find that as the number of splits increases, results become equivalent to doing no splitting. Stated formally, the sample splitting estimator converges in probability to the stacked estimating equations estimator as the sample size and number of splits ($B$) approach infinity. It is tempting to split the data and average the results *many* times in an attempt to eliminate the simulation variability of a single split. This is, however, unnecessary. A single split uses two independent sample for training and then validation, mimicking the process described in Section 4.1 of developing a score on one population and validating it on a separate population. As one repeats sample splitting, there is longer independence between the training data and the validation data. In the extreme case with an "infinite" number of splits, results are equivalent to using all the entire dataset for model filling and validation, thus nullifying the original intention of splitting the dataset.

# 5. SUMMARY

In Chapter 2 we introduced a composite score to provide physical behavior recommendations. This is, to our knowledge, the first composite scoring system for physical activity. This score was developed using shape constrained regression modeling. Shape constraints restrict function estimates to be consistent with previous literature in physical activity.

In Chapter 3 we use the Healthy Eating Index to introduce a method of altering an existing composite score. This is done using nonlinear regression models combined with adaptive Lasso variable selection. We establish oracle properties of our estimators. We find the surprising result that empty calories do not seem to predict mortality.

Finally in Chapter 4 we looked at risk estimation when building a composite score using sample splitting. We developed theory for sample splitting in parametric models using estimating equation theory. We provided asymptotic expansion which can be used for hypothesis tests.

# REFERENCES

Arem, H., Moore, S. C., Patel, A., Hartge, P., De Gonzalez, A. B., Visvanathan, K., Campbell, P. T., Freedman, M., Weiderpass, E., Adami, H. O., et al. (2015). Leisure time physical activity and mortality: a detailed pooled analysis of the dose-response relationship. *JAMA Internal Medicine*, 175(6):959–967.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2015). Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91:242–250.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., and Stefanski, L. A. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC.

Chang, X., Lin, S.-B., Wang, Y., et al. (2017). Divide and conquer local average regression. *Electronic Journal of Statistics*, 11(1):1326–1350.

Chasan-Taber, S., Rimm, E. B., Stampfer, M. J., Spiegelman, D., Colditz, G. A., Giovannucci, E., Ascherio, A., and Willett, W. C. (1996). Reproducibility and validity of a self-administered physical activity questionnaire for male health professionals. *Epidemiology*, pages 81–86.

Chastin, S. F., Palarea-Albaladejo, J., Dontje, M. L., and Skelton, D. A. (2015). Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PloS one*, 10(10):e0139984.

Chen, Y. and Samworth, R. J. (2016). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):729–754.

Cust, A. E., Smith, B. J., Chau, J., van der Ploeg, H. P., Friedenreich, C. M., Armstrong, B. K., and

Bauman, A. (2008). Validity and repeatability of the epic physical activity questionnaire: a validation study using accelerometers as an objective measure. *International Journal of Behavioral Nutrition and Physical Activity*, 5(1):33.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, $p$-values and r-software hdi. *Statistical Science*, 30(4):533–558.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Van Hees, V. T., Trenell, M. I., Owen, C. G., et al. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one*, 12(2):e0169649.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Ekelund, U., Steene-Johannessen, J., Brown, W. J., Fagerland, M. W., Owen, N., Powell, K. E., Bauman, A., Lee, I.-M., Series, L. P. A., and Group, L. S. B. W. (2016). Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? a harmonised meta-analysis of data from more than 1 million men and women. *The Lancet*, 388(10051):1302–1310.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24:428–440.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B*, 16:175–185.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Grøntved, A. and Hu, F. B. (2011). Television viewing and risk of type 2 diabetes, cardiovascular disease, and all-cause mortality: a meta-analysis. *Journal of the American Medical Association*, 305(23):2448–2455.

Guenther, P. M., Casavale, K. O., Reedy, J., Kirkpatrick, S. I., Hiza, H. A., Kuczynski, K. J., Kahle,

L. L., and Krebs-Smith, S. M. (2013a). Update of the healthy eating index: Hei-2010. *Journal of the Academy of Nutrition and Dietetics*, 113(4):569–580.

Guenther, P. M., Kirkpatrick, S. I., Krebs-Smith, S. M., Reedy, J., Buckman, D. W., Dodd, K. W., and Carroll, R. J. (2013b). Evaluation of the healthy eating index-2010 (hei-2010).

Guenther, P. M., Kirkpatrick, S. I., Reedy, J., Krebs-Smith, S. M., Buckman, D. W., Dodd, K. W., Casavale, K. O., and Carroll, R. J. (2013c). The healthy eating index-2010 is a valid and reliable measure of diet quality according to the 2010 dietary guidelines for americans. *The Journal of nutrition*, 144(3):399–407.

Guenther, P. M., Reedy, J., and Krebs-Smith, S. M. (2008a). Development of the Healthy Eating Index-2005. *Journal of the American Dietetic Association*, 108:1896–1901.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M., and Reeve, B. B. (2008b). Evaluation of the healthy eating index-2005. *Journal of the American Dietetic Association*, 108:1854–1864.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M., and Reeve, B. B. (2008c). Evaluation of the healthy eating index-2005. *Journal of the American Dietetic Association*, 108(11):1854–1864.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Hu, F. B., Rimm, E. B., Stampfer, M. J., Ascherio, A., Spiegelman, D., and Willett, W. C. (2000). Prospective study of major dietary patterns and risk of coronary heart disease in men. *The American journal of Clinical Nutrition*, 72(4):912–921.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press.

Kabat, G. C., Xue, X., Kamensky, V., Zaslavsky, O., Stone, K. L., Johnson, K. C., Wassertheil-Smoller, S., Shadyab, A. H., Luo, J., Hale, L., et al. (2018). The association of sleep duration and quality with all-cause and cause-specific mortality in the women's health initiative. *Sleep* *medicine*, 50:48–54.

Katzmarzyk, P. T. (2014). Standing and mortality in a prospective cohort of canadian adults. *Medicine & Science in Sports & Exercise*, 46(5):940–946.

Keadle, S. K., Bluethmann, S., Matthews, C. E., Graubard, B. I., and Perna, F. M. (2017). Combining activity-related behaviors and attributes improves prediction of health status in nhanes. *Journal of Physical Activity and Health*, 14(8):626–635.

Lee, I.-M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., Group, L. P. A. S. W., et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The lancet*, 380(9838):219–229.

Li, R., Lin, D. K., and Li, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409.

Ma, Q.-Q., Yao, Q., Lin, L., Chen, G.-C., and Yu, J.-B. (2016). Sleep duration and total cancer mortality: a meta-analysis of prospective studies. *Sleep Medicine*, 27:39–44.

Ma, S., Ma, Y., Wang, Y., Kravitz, E. S., and Carroll, R. J. (2017). A semiparametric single-index risk score across populations. *Journal of the American Statistical Association*, 112(520):1648–1662.

Matthews, C. E., Keadle, S. K., Moore, S. C., Schoeller, D. S., Carroll, R. J., Troiano, R. P., and Sampson, J. N. (2018). Measurement of active and sedentary behavior in context of large epidemiologic studies. *Medicine and Science in Sports and Exercise*, 50(2):266.

Matthews, C. E., Keadle, S. K., Troiano, R. P., Kahle, L., Koster, A., Brychta, R., Van Domelen, D., Caserotti, P., Chen, K. Y., Harris, T. B., et al. (2016). Accelerometer-measured dose-response for physical activity, sedentary time, and mortality in us adults. *The American Journal of Clinical Nutrition*, 104(5):1424–1432.

Matthews, C. E., Moore, S. C., Sampson, J., Blair, A., Xiao, Q., Keadle, S. K., Hollenbeck, A., and Park, Y. (2015). Mortality benefits for replacing sitting time with different physical activities. *Medicine and Science in Sports and Exercise*, 47(9):1833.

McCullough, M. L., Feskanich, D., Stampfer, M. J., Giovannucci, E. L., Rimm, E. B., Hu, F. B., Spiegelman, D., Hunter, D. J., Colditz, G. A., and Willett, W. C. (2002). Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. *American Journal of Clinical Nutrition*, 76(6):1261–1271.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Mekary, R. A., Lucas, M., Pan, A., Okereke, O. I., Willett, W. C., Hu, F. B., and Ding, E. L. (2013). Isotemporal substitution analysis for physical activity, television watching, and risk of depression. *American Journal of Epidemiology*, 178(3):474–483.

Office of Disease Prevention and Health Promotion (2019). Physical PAGCR. `https://health.gov/paguidelines/`. Accessed: 2019-03-28.

Panagiotakos, D. B., Pitsavos, C., and Stefanadis, C. (2006). Dietary patterns: a Mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutrition, Metabolism and Cardiovascular Diseases*, 16(8):559–568.

Patel, A. V., Jacobs, E. J., Dudas, D. M., Briggs, P. J., Lichtman, C. J., Bain, E. B., Stevens, V. L., McCullough, M. L., Teras, L. R., Campbell, P. T., et al. (2017). The american cancer society's cancer prevention study 3 (cps-3): Recruitment, study design, and baseline characteristics. *Cancer*, 123(11):2014–2024.

Piercy, K. L., Troiano, R. P., Ballard, R. M., Carlson, S. A., Fulton, J. E., Galuska, D. A., George, S. M., and Olson, R. D. (2018). The physical activity guidelines for americans. *Jama*, 320(19):2020–2028.

Prince, S., Saunders, T., Gresty, K., and Reid, R. (2014). A comparison of the effectiveness of physical activity and sedentary behaviour interventions in reducing sedentary time in adults: a systematic review and meta-analysis of controlled trials. *Obesity Reviews*, 15(11):905–919.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, 25(3):543–559.

Reedy, J., Mitrou, P., Krebs-Smith, S., Wirfält, E., Flood, A., Kipnis, V., Leitzmann, M., Mouw, T., Hollenbeck, A., and Schatzkin, A. (2008). Index-based dietary patterns and risk of colorectal cancer: the NIH-AARP diet and health study. *American Journal of Epidemiology*, 168(1):38–48.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., et al. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the national institutes of health–american association of retired persons diet and health study. *American Journal of Epidemiology*, 154(12):1119–1125.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

Stamatakis, E., Rogers, K., Ding, D., Berrigan, D., Chau, J., Hamer, M., and Bauman, A. (2015). All-cause mortality effects of replacing sedentary time with physical activity and sleeping using an isotemporal substitution model: a prospective study of 201,129 mid-aged and older adults. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):121.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

The Obesity Society, Young, D. R., Hivert, M.-F., Alhassan, S., Camhi, S. M., Ferguson, J. F., Katzmarzyk, P. T., Lewis, C. E., Owen, N., and Perry, C. K. (2016). Sedentary behavior and cardiovascular morbidity and mortality: a science advisory from the american heart association. *Circulation*, 134(13):e262–e279.

Trichopoulou, A., Orfanos, P., Norat, T., Bueno-de Mesquita, B., Ocké, M. C., Peeters, P. H., van der Schouw, Y. T., Boeing, H., Hoffmann, K., Boffetta, P., et al. (2005). Modified Mediterranean diet and survival: EPIC-Elderly Prospective Cohort Study. *BMJ*, 330(7498):991.

U.S. Department of Health and Human Services and U.S. Department of Agriculture (2005). Dietary guidelines for Americans. Technical report.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Wang, Y., Wang, S., and Carroll, R. J. (2015). The direct integral method for confidence intervals for the ratio of two location parameters. *Biometrics*, 71(3):704–713.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37:2178.

Whitaker, K. M., Buman, M. P., Odegaard, A. O., Carpenter, K. C., Jacobs Jr, D. R., Sidney, S., and Pereira, M. A. (2017). Sedentary behaviors and cardiometabolic risk: an isotemporal substitution analysis. *American Journal of Epidemiology*, 187(2):181–189.

Wolf, A. M., Hunter, D. J., Colditz, G. A., Manson, J. E., Stampfer, M. J., Corsano, K. A., Rosner, B., Kriska, A., and Willett, W. C. (1994). Reproducibility and validity of a self-administered physical activity questionnaire. *International Journal of Epidemiology*, 23(5):991–999.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Xiao, Q., Keadle, S. K., Hollenbeck, A. R., and Matthews, C. E. (2014). Sleep duration and total and cause-specific mortality in a large us cohort: interrelationships with physical activity, sedentary behavior, and body mass index. *American Journal of Epidemiology*, 180(10):997–1006.

Yin, J., Jin, X., Shan, Z., Li, S., Huang, H., Li, P., Peng, X., Peng, Z., Yu, K., Bao, W., et al. (2017). Relationship of sleep duration with all-cause mortality and cardiovascular events: A systematic review and dose-response meta-analysis of prospective cohort studies. *Journal of the American Heart Association*, 6(9):e005947.

Zhang, X., Cao, J., and Carroll, R. J. (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics*, 71(1):131–138.

Zhang, Y., Wainwright, M. J., and Duchi, J. C. (2012). Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

# APPENDIX A

# SUPPLEMENTARY MATERIAL FOR CHAPTER 2

## A.1 R Code to Generate Physical Behavior Score

The R code to generate the Physical Behavior Score is available by request or online at `https://healthresearch.calpoly.edu/physical-behavior-score` and `https://github.com/kravitzel/Physical-Behavior-Score`.

## A.2 Testing Proportional Hazards Assumption

We justify the proportional hazard assumption in two ways. First, we plot the weighted Schoenfeld residuals against time. See Figure A.1 The relatively straight line in the plot indicates the proportional hazards assumption is valid. To see that quintile-based analysis is correct, we look at the log hazard, or equivalently the log of the negative log of survival time, stratified by the 5 quintiles of the PBS. Nonparallel or crossing curves indicate that the proportional hazards assumption is not valid. For very small values of time with few observations, the curves cross. However, as log time increases to reasonable values the curves become parallel and do not cross each other. See Figure A.1

## A.3 Physical Activity Questionnaire and Met Conversion

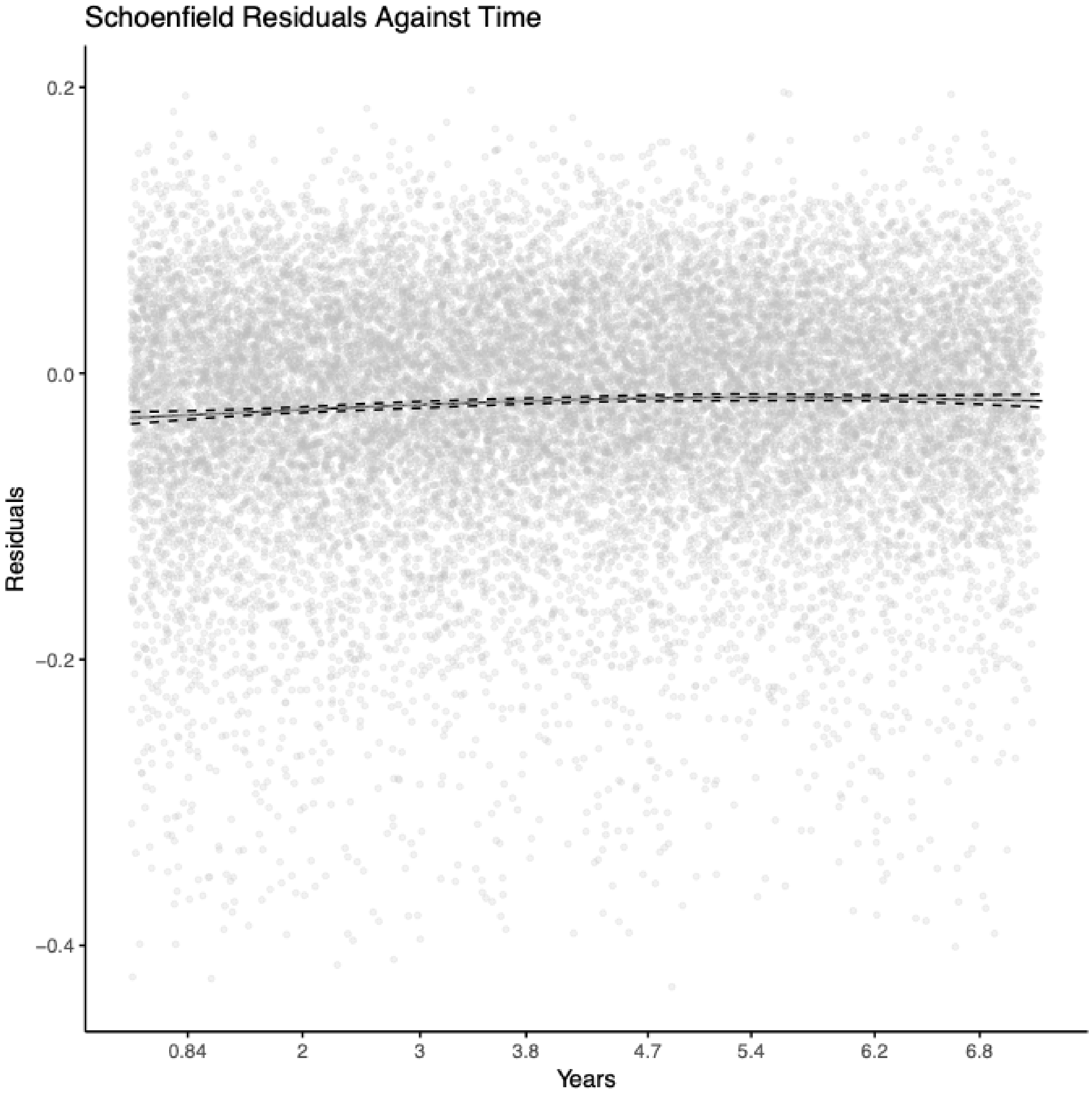


Figure A.1: Schoenfeld residuals plotted against time. A straight line indicates that the coefficients fit with Cox regression do not change with time. Reprinted with permission from Wolters Kluwer Health Inc

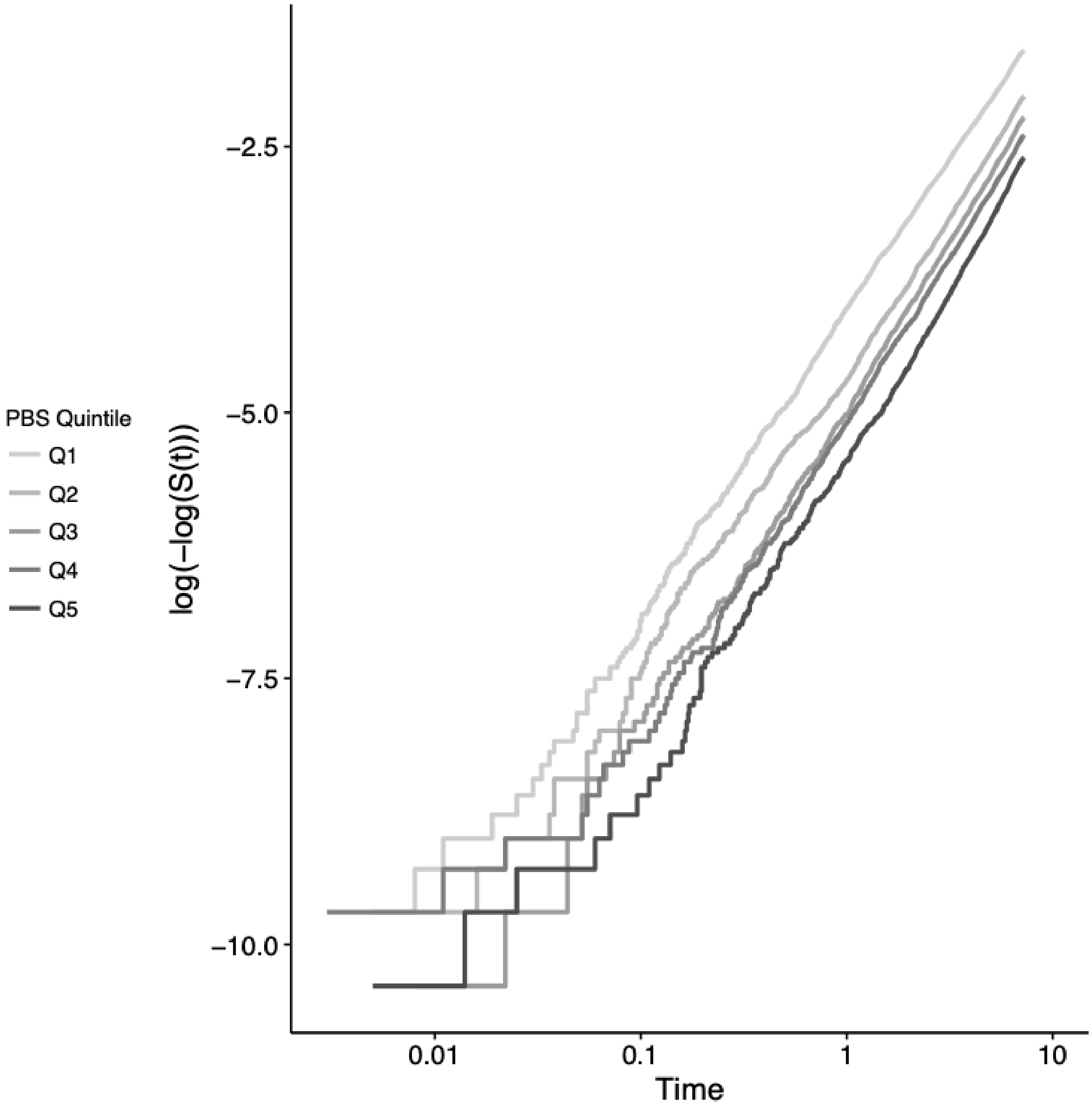
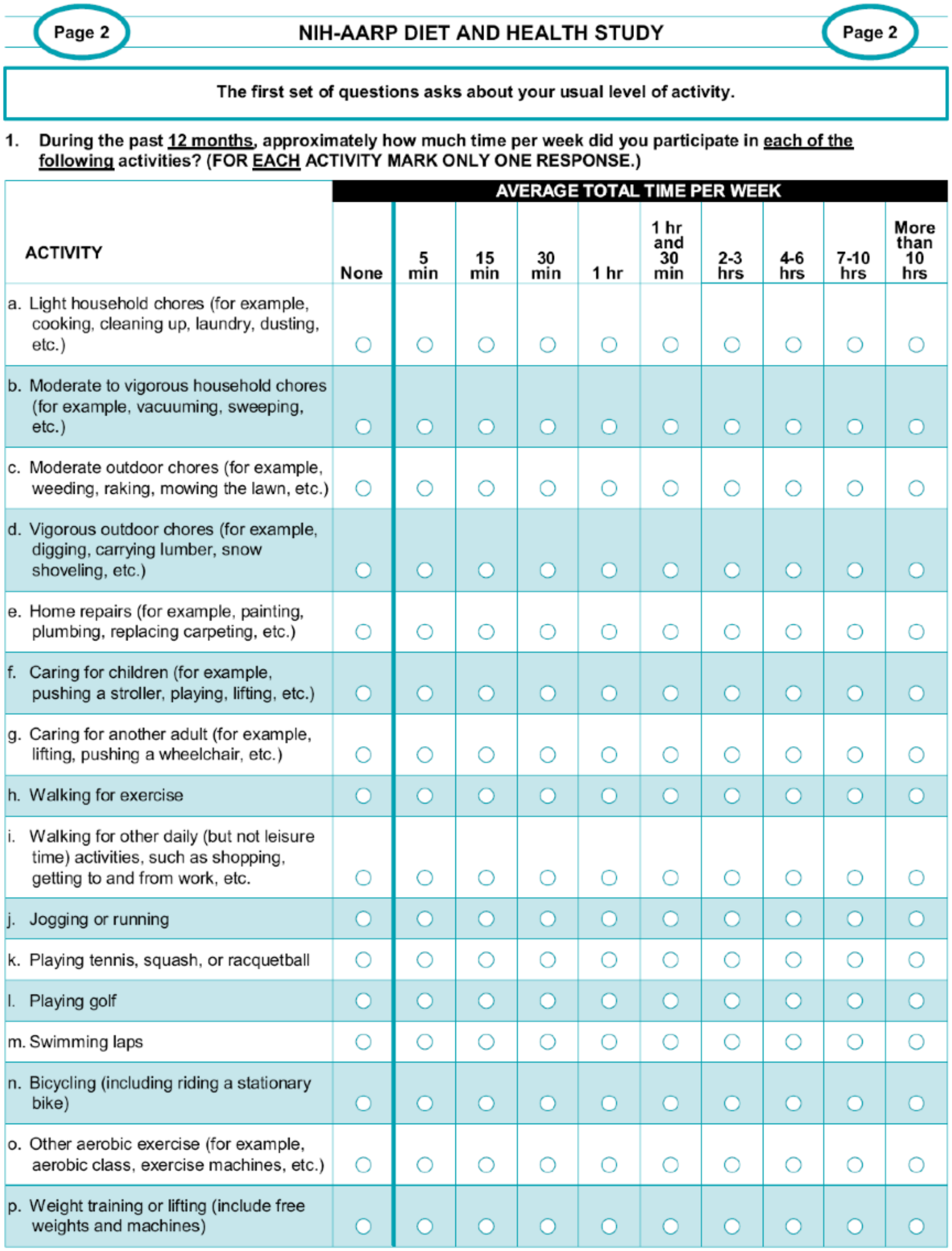


Figure A.2: Log Hazard, equivalent to log of negative log of Survival time, against time. Parallel lines indicate the proportional hazards assumption Is justified. Reprinted with permission from Wolters Kluwer Health Inc



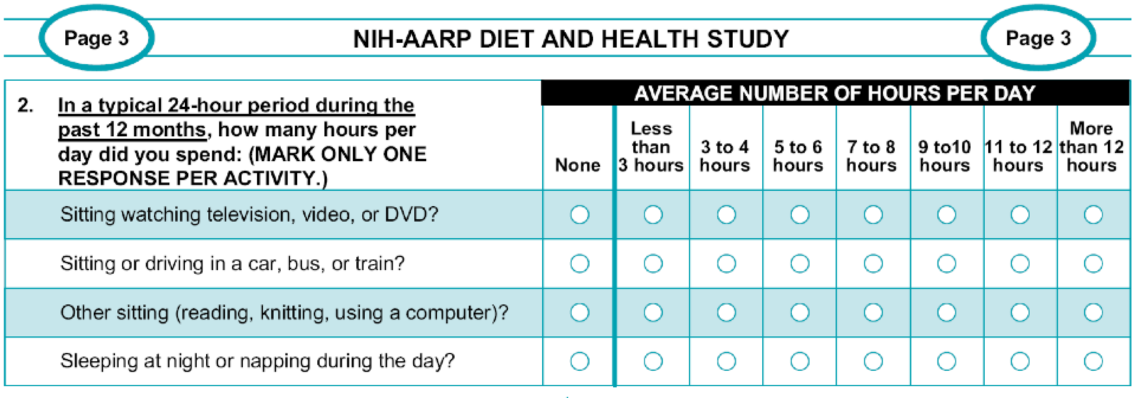
Page 2 NIH-AARP DIET AND HEALTH STUDY Page 2

**The first set of questions asks about your usual level of activity.**

**1. During the past 12 months, approximately how much time per week did you participate in each of the following activities? (FOR EACH ACTIVITY MARK ONLY ONE RESPONSE.)**

| | AVERAGE TOTAL TIME PER WEEK | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ACTIVITY | None | 5 min | 15 min | 30 min | 1 hr | 1 hr and 30 min | 2-3 hrs | 4-6 hrs | 7-10 hrs | More than 10 hrs |
| a. Light household chores (for example, cooking, cleaning up, laundry, dusting, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| b. Moderate to vigorous household chores (for example, vacuuming, sweeping, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| c. Moderate outdoor chores (for example, weeding, raking, mowing the lawn, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| d. Vigorous outdoor chores (for example, digging, carrying lumber, snow shoveling, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| e. Home repairs (for example, painting, plumbing, replacing carpeting, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| f. Caring for children (for example, pushing a stroller, playing, lifting, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| g. Caring for another adult (for example, lifting, pushing a wheelchair, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| h. Walking for exercise | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| i. Walking for other daily (but not leisure time) activities, such as shopping, getting to and from work, etc. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| j. Jogging or running | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| k. Playing tennis, squash, or racquetball | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| l. Playing golf | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| m. Swimming laps | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| n. Bicycling (including riding a stationary bike) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| o. Other aerobic exercise (for example, aerobic class, exercise machines, etc.) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| p. Weight training or lifting (include free weights and machines) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

Figure A.3: NIH-AARP Study of Diet and Health questionnaire for physical activity. Reprinted with permission from Wolters Kluwer Health Inc



Page 3 NIH-AARP DIET AND HEALTH STUDY Page 3

| 2. In a typical 24-hour period during the past 12 months, how many hours per day did you spend: (MARK ONLY ONE RESPONSE PER ACTIVITY.) | AVERAGE NUMBER OF HOURS PER DAY | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | None | Less than 3 hours | 3 to 4 hours | 5 to 6 hours | 7 to 8 hours | 9 to10 hours | 11 to 12 hours | More than 12 hours |
| Sitting watching television, video, or DVD? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Sitting or driving in a car, bus, or train? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Other sitting (reading, knitting, using a computer)? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Sleeping at night or napping during the day? | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

Figure A.4: NIH-AARP Study of Diet and Health questionnaire for sedentary behaviors. Reprinted with permission from Wolters Kluwer Health Inc

| **Activities** | **MET** (Compendium Code, 2001) |
|---|---|
| Light household chores | 2.5 (05040) |
| Moderate-to-vigorous indoor household chores | 3.5 (05026) |
| Moderate outdoor chores | 4.0 (06127) |
| Vigorous outdoor chores | 6.0 (08262) |
| Home repairs | 4.5 (08261) |
| Caring for children | 3.0 (05186) |
| Caring for adults. | 4.0 (05200) |
| Walking for exercise, | 4.3 (17200) |
| Walking for daily activities. | 2.9 (avg METs for codes 17161, 17270, 05060) |
| Jogging or running | 7.0 (12020) |
| Tennis, squash, racquetball | 7.3 (15675) |
| Playing golf | 4.8 (15255) |
| Swimming laps | 8.3 (18290) |
| Bicycling or stationary bike | 7.5 (01015) |
| Other aerobic exercise | 7.3 (03015) |
| Weight training or lifting. | 3.5 (02054) |
| | |
| **Sedentary Behaviors** | |
| Sitting watching television | 1.3 (07020) |
| Sitting or driving in a car, bus, or train | 1.3 (16015) |
| Other sitting | 1.3 (09030) |
| | |
| **Sleep** | |
| Sleeping at night or napping during the day | 1.0 (07030) |

Table A.1: Physical activity variables are expressed as MET-hrs per week, which is the MET-value for each activity multiplied by the reported hours per week. Molded correlation coefficients indicate statistical significance at P<0.05 Thresholds for Moderate intensity is ≥3 METs, Vigorous intensity is ≥ 6 METs.Reprinted with permission from Wolters Kluwer Health Inc.

# APPENDIX B

# SUPPLEMENTARY MATERIAL FOR CHAPTER 3

## B.1 R Code and Data

The R programs used in the simulations of Section 4.5 and the data analysis of Section 4.6 are available by request or through Github at `https://github.com/kravitzel/Composite_Scores/`

We do not have permission to distribute the actual data involved in Section 4.6: such data can be obtained via a data transfer agreement with the National Cancer Institute, see `https://epi.grants.cancer.gov/Consortia/cohort_projects.html`.

## B.2 Proof of Theorem 1

By Theorem 2 of Wang and Leng,

$$\lim_{n\to\infty} \text{pr}(\widehat{A} \subseteq A_T) = 1, \tag{B.1}$$

and by Theorem 1 of Wang and Leng,

$$\lim_{n\to\infty} \text{pr}(\widehat{A} \subset A_T) = 0. \tag{B.2}$$

Then (B.1) and (B.2) imply that

$$\lim_{n\to\infty} \text{pr}(\widehat{A} = A_T) = 1. \tag{B.3}$$

### B.3 Proof of Theorem 2

Denote $\widehat{\Theta}(A)$ as the estimates from regressing Y on the subset of $\Theta$ specified by $A$ and $\widehat{\Theta}(A_T)$ as the estimates from regressing Y on the true subset of coefficients. Then (B.3) means that

$$\lim_{n\to\infty} \text{pr}\{\widehat{\Theta}(\widehat{A}) = \widehat{\Theta}(A_T)\} = 1. \tag{B.4}$$

We have,

$$n^{1/2}\{\widehat{\Theta}(A_T) - \Theta_T\} \to N(0, \Sigma_A).$$

Thus for any vector a,

$$n^{1/2}[a^{\mathrm{T}}\{\widehat{\Theta}(A_T) - \Theta_T\}] \to N(0, \sigma^2).$$

where $\sigma^2 = a^{\mathrm{T}}\Sigma_A a$. Thus,

$$pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(A_T) - \Theta_T\}/\sigma \leq z] \to \Phi(z).$$

where $\Phi(\cdot)$ is the normal cumulative distribution function. Since (B.4) holds, we have

$$pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(\widehat{A}) - \Theta_T\}/\sigma \leq z] \to \Phi(z).$$

This can be expressed as,

$$\begin{aligned}
& pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(\widehat{A})-\Theta_T\}\,\sigma \leq z, \widehat{A}=A_T]\} + pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(\widehat{A})-\Theta_T\}\,\sigma^2 \leq z, \widehat{A}\neq A_T] \\
= \;& pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(A_T)-\Theta_T\}\,\sigma \leq z, \widehat{A}=A_T]\} + pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(\widehat{A})-\Theta_T\}\,\sigma^2 \leq z, \widehat{A}\neq A_T]\} \\
= \;& pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(A_T)-\Theta_T\}/\sigma \leq z, \widehat{A}=A_T] + o_p(1) \\
= \;& pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(A_T)-\Theta_T\}/\sigma \leq z] - pr[n^{1/2}a^{\mathrm{T}}\{\widehat{\Theta}(A_T)-\Theta_T\}/\sigma \leq z, \widehat{A}\neq A_T] + o_p(1) \\
= \;& \Phi(z)+o_p(1)
\end{aligned}$$

### B.4 Proof of Theorem 3

Theorem 3 is a direct application of Wang and Leng's Theorem 4.

### B.5 Variance Calculation of Rescaled Coefficients

Let $H=(h_1,\ldots,h_J)^{\mathrm{T}}$, where $h_j(\boldsymbol{\alpha})=\alpha_j/(\mathbf{c}_{max}^{\mathrm{T}}\boldsymbol{\alpha})$. Define $D=(d_1,\ldots,d_J,)^{\mathrm{T}}$ where $d_j=(\partial h_j/\partial\alpha_1,\ldots,\partial h_j/\partial\alpha_J)$. The diagonals of matrix D are given by $D_{ii}=(\sum_{k\neq i}c_k\alpha_k)/(c_{max}^{\mathrm{T}}\alpha)^2$. For $i\neq j$, $D_{ij}=(c_j\alpha_i)/(c_{max}^{\mathrm{T}}\alpha)^2$. The delta method states that the covariance of $H$ is given by cov$\{H(\alpha)\}\approx D\Sigma_\alpha D^{\mathrm{T}}$.

Now, we move on to calculating the variance of $\beta_{k\ell}$. For $k=1,...,K$ and $\ell=1,...,L$, the logits are

$$\beta_{k\ell}X^{\mathrm{T}}\alpha+Z^{\mathrm{T}}\theta_{k\ell},$$

with the initial constraint is that $\beta_{11}=-1$ for identifiability.

Denote $c(\boldsymbol{\alpha})=c_{max}^{\mathrm{T}}\boldsymbol{\alpha}$. After model (3.1) converges replace each $\alpha$ by $\boldsymbol{\alpha}^*=\boldsymbol{\alpha}/c(\boldsymbol{\alpha})$, Then the logits become

$$\beta_{k\ell}^*X^{\mathrm{T}}\alpha^*+Z^{\mathrm{T}}\theta_{k\ell}=\beta_{k\ell}d(\alpha)X^{\mathrm{T}}\alpha_*+Z^{\mathrm{T}}\theta_{k\ell}.$$

For $k=\ell=1$, this means that $\beta_{k\ell}^*=-c(\boldsymbol{\alpha})$, By the delta method, var$(\widehat{\beta}_{k\ell}^*)\approx$ cov$\{c(\widehat{\alpha})\}\approx$

$c_\alpha(\widehat{\alpha})^{\mathrm{T}}\mathrm{cov}(\widehat{\alpha})d_\alpha(\widehat{\alpha})$, where the subscript $\alpha$ indicates the $J \times 1$ vector of derivatives of $d(\alpha)$.

If $(k, \ell) \neq (1, 1)$, we have that $\widehat{\beta}^*_{k\ell} = \widehat{\beta}_{k\ell}c(\widehat{\alpha})$. We can express this as $\beta^*_{k\ell} = g(\beta_{k\ell}, \boldsymbol{\alpha})$ and use the delta method to get var$(\widehat{\beta}^*_{k\ell}) \approx \nabla g(\widehat{\beta}_{k\ell}, \widehat{\boldsymbol{\alpha}})^{\mathrm{T}} cov(\widehat{\beta}_{k\ell}, \widehat{\boldsymbol{\alpha}})\nabla g(\widehat{\beta}_{k\ell}, \widehat{\boldsymbol{\alpha}})$. Here $\nabla g(\widehat{\beta}_{k\ell})$ is the gradient of $g(\cdot)$ with respect to $\beta_{k\ell}$ and $\boldsymbol{\alpha}$.

# APPENDIX C

# SUPPLEMENTARY MATERIAL FOR CHAPTER 4

## C.1 R Code and Data

The R programs used in the simulations of Section 4.5 and the data analysis of Section 4.6 are available by request or on Github at `https://github.com/kravitzel/Sample_Splitting`.

We do not have permission to distribute the actual data involved in Section 4.6: such data can be obtained via a data transfer agreement with the National Cancer Institute, see `https://epi.grants.cancer.gov/Consortia/cohort_projects.html`.

### C.1.1 Assumptions

Suppose $\widehat{\theta}$ is the solution to

$$0 = n^{-1}\textstyle\sum_{i=1}^{n}\Psi(Y_i, \theta), \tag{C.1}$$

where $\Psi(\cdot)$ is a known influence function that does not depend on $i$ and is continuous and twice differentiable with respect to $\theta$ and $\beta$. Suppose $\widehat{\beta}$ is the solution to

$$0 = n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{K}\{Y_i, f(\mathbf{X}, \widehat{\theta}), \beta)\},$$

where $\mathcal{K}(\cdot)$, as in (C.1), is a known function which does not depend on $i$, $\mathcal{K}(\cdot)$ twice differentiable with respect to $\beta$, and differentiable with respect to $f(\mathbf{X}, \theta)$. Assume $f(\mathbf{X}, \theta)$ is differentiable with respect to $\theta$.

We the estimators from different splits are exchangeable. We assume that they are also unbiased. Therefore, $E(\theta_k) = E(\theta_\ell) = \theta$ and $E(\beta_k) = E(\beta_\ell) = \beta$ for all $k, \ell$.

### C.1.2 Taylor Expansions from Sections 4.4.1 and 4.4.2

Since $\widehat{\theta}_b$ and $\widehat{\beta}_b$ are asymptotically linear they have expansions

$$\begin{aligned} n^{1/2}(\widehat{\theta}_b - \theta) &= -\Omega_{nb}^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} \delta_{ib} \Psi(W_i, \theta) + o_P(1). & \text{(C.2)} \\ n^{1/2}(\widehat{\beta}_b - \beta) &= -\Lambda_{nb}^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}_{ib}\{Y_i, f(\mathbf{X}_i, \widehat{\theta}_b)\mathbf{Z}_i, \beta\} + o_P(1). & \text{(C.3)} \end{aligned}$$

Taking a Taylor expansion of (C.3) around $\theta$ we have

$$\begin{aligned} & n^{-1/2} \textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}\{Y_i, f(\mathbf{X}_i, \widehat{\theta}_b), \beta\} \\ & \quad = n^{-1/2} \textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}\{Y_i, f(\mathbf{X}_i, \theta), \beta\} \\ & \quad + n^{-1} \textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}_f\{Y_i, f(\mathbf{X}_i, \theta), \beta\} f_\theta^{\mathrm{T}}(\mathbf{X}_i, \theta) n^{1/2}(\widehat{\theta}_b - \theta) + o_P(1). \quad \text{(C.4)} \end{aligned}$$

Define

$$\Delta_{nb} = n^{-1} \textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}_f\{Y_i, f(\mathbf{X}_i, \theta), \beta\} f_\theta^{\mathrm{T}}(\mathbf{X}_i, \theta).$$

Taking a Taylor expansion from (C.4) and plugging (C.2) in place of $n^{1/2}(\widehat{\theta}_b - \theta)$ we get

$$\begin{aligned} n^{1/2}(\widehat{\beta}_b - \beta) &= -\Lambda_{nb}^{-1} n^{-1/2} [\textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}\{Y_i, f(\mathbf{X}_i, \theta), \beta\} \\ & \quad - \Delta_{nb} \Omega_{nb}^{-1} \textstyle\sum_{i=1}^{n} \delta_{ib} \Psi(W_i, \theta)] + o_P(1). \end{aligned}$$

The asymptotic expansion of $\widehat{\theta} = B^{-1} \sum_{b=1}^{B} \theta_b$ and $\widehat{\beta} = B^{-1} \sum_{b=1}^{B} \beta_b$ are the averages over $B$ of the previous expansion,

$$\begin{aligned} n^{1/2}(\widehat{\theta}_b - \theta) &= -B^{-1} \textstyle\sum_{b=1}^{B} \Omega_{nb}^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} \delta_{ib} \Psi(W_i, \theta) + o_P(1), & \text{(C.5)} \\ n^{1/2}(\widehat{\beta}_b - \beta) &= -B^{-1} \textstyle\sum_{b=1}^{B} \Lambda_{nb}^{-1} n^{-1/2} [\textstyle\sum_{i=1}^{n} (1-\delta_{ib}) \mathcal{K}\{Y_i, f(\mathbf{X}, \theta), \beta\} \\ & \quad - \Delta_{nb} \Omega_{nb}^{-1} \textstyle\sum_{i=1}^{n} \delta_{ib} \Psi(W_i, \theta)] + o_P(1). & \text{(C.6)} \end{aligned}$$

### C.1.3 Theory for Infinite Splits from Section 4.4.3

We note of two technical details. First

$$
\begin{aligned}
\widehat{\Omega}_{nb}(\widehat{\theta}_b) &= \pi E\{\partial\Psi(W,\theta)/\partial\theta^{\mathrm{T}}\} + O_P(n^{-1/2}) = \Omega + O_P(n^{-1/2})\\
\widehat{\Lambda}_{nb}(\widehat{\theta}_b,\widehat{\beta}_b) &= (1-\pi)E\{\partial\mathcal{K}(Y,\beta,\theta)/\partial\beta^{\mathrm{T}}\} + O_P(n^{-1/2}) = \Lambda + O_P(n^{-1/2}).\\
\widehat{\Delta}_{nb}(\widehat{\theta}_b,\widehat{\beta}_b) &= (1-\pi)E[\partial\mathcal{K}\{Y,\beta,f(\mathbf{X};\theta)\}/\partial\theta^{\mathrm{T}}] + O_P(n^{-1/2}) = \Delta + O_P(n^{-1/2}).
\end{aligned}
$$

Also, as in Carroll et al. (1996), the $o_P(1)$ terms in the expansion in (C.2) and (C.3) are $O_P(n^{-1/2+a})$ for any $a > 0$. Combining these, we have,

$$
\begin{aligned}
n^{1/2}(\widehat{\theta}_b - \theta) &= \{-\Omega^{-1} + O_P(n^{-1/2})\}n^{-1/2}\textstyle\sum_{i=1}^{n}\delta_{ib}\Psi(W_i,\theta) + O_P(n^{-1/2+a})\\
&= -\Omega^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\delta_{ib}\Psi(W_i,\theta) + O_P(n^{-1/2+a})\\
n^{1/2}(\widehat{\beta}_b - \beta) &= -\Lambda^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}[(1-\delta_{ib})\mathcal{K}\{Y_i, f(\mathbf{X}_i,\theta),\beta\}\\
&\quad +\Lambda^{-1}\Delta\Omega^{-1}\delta\Psi(Y_i,\theta)] + O_P(n^{-1/2+a}).
\end{aligned}
$$

Since the $\delta_{ib}$ are independent of the data, $B^{-1}\sum_{b=1}^{B}\delta_{ib} = \pi + o_P(1)$, and since $B = B_n = o_P(n^{-1/2+a})$ by assumption, the expansion in (C.5) and (C.6) become

$$
\begin{aligned}
n^{1/2}(\widehat{\theta} - \theta) &= -\Omega^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\pi\Psi(W_i,\theta) + o_P(1)\\
n^{1/2}(\widehat{\beta} - \beta) &= -\Lambda^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}(1-\pi)\mathcal{K}(Y_i,\beta,\theta)\\
&\quad +\Lambda^{-1}\Delta\Omega^{-1}n^{-1/2}\textstyle\sum_{i=1}^{n}\pi\Psi(W_i,\theta) + o_P(1).
\end{aligned}
\tag{C.7}
$$

### C.1.4 Asymptotic Equivalence to Stacked Estimating Equations

Both $\theta$ and $\beta$ can be estimated by the stacked estimating equations instead of the sample splitting of the previous sections. The stacked estimating equations are given by,

$$\textstyle\sum_{i=1}^{n} \begin{pmatrix} \Psi(Y_i, \mathbf{X}_i, \theta) \\ \mathcal{K}\{W_i, f(\mathbf{X}_i, \theta), \beta\} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Define

$$\Phi = \begin{pmatrix} \Phi_{11} & 0 \\ \Phi_{21} & \Phi_{22} \end{pmatrix} = E \begin{pmatrix} \partial\Psi(Y_i, \mathbf{X}_i, \theta)/\partial\theta^{\mathrm{T}} & \partial\Psi(Y_i, \mathbf{X}_i, \theta)/\partial\beta^{\mathrm{T}} \\ \partial\mathcal{K}\{W_i, f(\mathbf{X}_i, \theta), \beta\}/\partial\theta^{\mathrm{T}} & \partial\mathcal{K}\{W_i, f(\mathbf{X}_i, \theta), \beta\}/\partial\beta^{\mathrm{T}} \end{pmatrix}$$

and

$$\Phi^{-1} = \begin{pmatrix} \Phi_{11}^{-1} & 0 \\ -\Phi_{22}^{-1}\Phi_{21}\Phi_{11}^{-1} & \Phi_{22}^{-1} \end{pmatrix}.$$

Then it follows that the stacked estimating equations estimator of $\beta$ has influence function

$$\mathcal{G}_i = \left[\Phi_{22}^{-1}\mathcal{K}\{W_i, f(\mathbf{X}_i, \theta), \beta\} - \Phi_{22}^{-1}\Phi_{21}\Phi_{11}^{-1}\Psi(Y_i, \mathbf{X}_i, \theta)\right] + o_P(1) \tag{C.8}$$

Since $\Phi_{22} = (1-\pi)^{-1}\Lambda$, $\Phi_{11} = \pi^{-1}\Omega$ and $\Phi_{21} = (1-\pi)^{-1}\Delta$, then (C.8) simplifies to (C.7). Therefore as $B, n \to \infty$ and when $\pi = 1/2$, the sample splitting estimator becomes equivalent to the estimator from stacked estimating equations.