ESTIMATES OF OBESITY-ATTRIBUTABLE MORTALITY IN THE
UNITED STATES

Presented in Partial Fulfillment of the Requirement for
The Degree Master of Science in the
Graduate School of The Ohio State University

By
Lingyi Lu, B.Sc.

* * * * *
The Ohio State University
2005

Master’s Examination Committee

Dr. Melvin Moeschberger
Dr. Amy Ferketich

Approved By

Advisor
School of Public Health
ABSTRACT

The increasing prevalence of obesity over the past decades has generated considerable concern about health problems and subsequent economic burden. Approximately 64% of U.S. adults are estimated to be overweight or obese. Over 30% are obese and 5% are extremely obese among them. Obesity has been associated with a variety of chronic diseases and with increased risk of all-cause mortality. The number of annual deaths in the U.S. attributable to obesity estimated from epidemiological cohorts ranges widely from 26,917 to 385,000. The objective of the current study is to estimate a more accurate number of excess deaths attributable to obesity. A sub-sample of 6,913 subjects from the First National Health and Nutrition Examination Survey (NHANES I, 1971-1975) data set and corresponding mortality information from the National Health and Nutrition examination Survey I Epidemiologic Follow-up Study (NHEFS,1992) was employed for building a Cox’s proportional hazards model. The number (2,374,029) of total deaths age from 25 and older in the U.S. came from the National Center for Health Statistics 2003 National Vital Statistics Report. Body weight was categorized based upon body mass index (BMI) which is defined as Weight (kg) / Height$^2$ (m$^2$). “Underweight” was defined as BMI < 18.5; “Normal weight” was defined as 18.5 ≤ BMI < 25; “Overweight” was defined as 25 ≤ BMI < 30 ; “Obesity” was defined as 30 ≤ BMI < 35; and “severe obesity” was defined as BMI ≥ 35. A Cox’s proportional hazard model
adjusted for gender, race, education and smoking status was applied to estimate the relative risk of dying associated with obesity and severe obesity. Both time-on-study and left truncation to adjust delayed entry were used to construct Cox's proportional hazard model. The Population Attributable Risk (PAR) was calculated based on Bruzzi's [1] formula which takes into account multiple exposure levels and adjusted relative risk. The extra deaths attributable to obesity were calculated by multiplying the total number of deaths by the PAR. Our results showed that the estimated extra deaths attributable to obesity and severe obesity were 280,708 per year based on the time-on-study model and 260,041 per year for the left truncation model. We also found that the relative risk of death for underweight was as high as severe obesity, whereas the relative risk of death for overweight was less than but very close to that of normal weight. Even though the same NHANES I data set was employed in the two previous studies, our estimated numbers of extra deaths attributable to obesity were higher than the estimated number 184,670 reported by Allison [2], but lower than the number 298,808 estimated by Flegal [3]. The difference in sample size, reference group, target population, model in control of later entry, PAR formula may be the underlying reasons accounting for the different estimated numbers.
Dedicated to my husband Shihua Wang and my son Leonardo Wang
ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my advisor Dr. Melvin Moeschberger. It is his scientific guidance, intellectual support, great patience and encouragement that made this thesis possible.

I wish to thank my thesis committee member, Dr. Amy Ferketich. Her advice, energy and time enabled me to complete this thesis.

I thank Dr. Debra Burr Doss and Dr. Steven K. Clinton for the financial support during my first year graduate study.

I thank my husband, my parents. It is their love and support which made my success possible.
VITA

Oct 26, 1977.................Born – Zhejiang, China

1999.........................B.S. Computer Science,
Zhejiang Science and Technology University

1999-2001....................Associate Engineer,
Hangzhou Bell Telecommunication Co., Ltd.

2002-2003....................Master Student,
Computer Science, Kent State University

2003-Present...................Master Student,
School of Public Health, The Ohio State University

FIELDS OF STUDY

Major Fields: Public Health
# TABLE OF CONTENTS

PAGE

Abstract ........................................................................................................ ii
Dedication ...................................................................................................... iv
Acknowledgments ......................................................................................... v
Vita ................................................................................................................. vi
List of tables.................................................................................................. vii
List of figures................................................................................................. x

Chapters

1. Introduction ............................................................................................... 1
   1.1 Increased prevalence of obesity and caused problems in U.S. ................. 1
   1.2 Previous estimation of extra death attributable to obesity ...................... 2
   1.3 Limitations in previous studies ............................................................... 3
       1.3.1 Selected study population not representative of target population .... 3
       1.3.2 Incorrect time-scale and later entry controlling ................................ 4
       1.3.3 Inconsistency in the definition of reference group ........................... 4
       1.3.4 Incorrect equation for population attributable risk calculation .......... 5
   1.4 Proposed investigation .......................................................................... 5
   1.5 Definition of terms ............................................................................... 6

2. Literature Review ....................................................................................... 7
   2.1 Introduction of population attributable risk ........................................... 7
   2.2 Commonly used equations for estimating population attributable risk .... 9
   2.3 Crude estimation of population attributable risk ..................................... 10
   2.4 Adjusted point estimation of population attributable risk...................... 14
       2.4.1 Adjusted estimates of PAR based on stratification approaches .......... 15
       2.4.2 Model-based adjusted estimation of PAR ............................... 16

-vii-
2.5 Choice of threshold for risk factor ................................................. 18
2.6 PAR in survival analysis ................................................................. 19
2.7 Conceptual problems in PAR ......................................................... 19

3. Study Design and Methodology ...................................................... 20
   3.1 Study cohorts: NHANES I and NHEFS ........................................ 21
   3.2 Descriptive Analysis ................................................................ 22
   3.3 Cox's proportional hazard model coding and building .................... 22

4. Study Design and Methodology ...................................................... 25
   4.1 Demographic Statistics in Study Cohort ...................................... 25
   4.2 Time-on-study Models ............................................................... 26
   4.3 Left Truncation Models .............................................................. 30

5. Discussion ..................................................................................... 32
   5.1 Comparison with previous studies ............................................. 32
   5.2 Choice of time scale ................................................................. 40
   5.3 Association between overweight and death .................................. 42
   5.4 Limitation and uncertainty of the study ...................................... 43
   5.5 Conclusions ............................................................................ 44
   5.6 Recommendations for Future Study ........................................... 44
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Code table for each covariate</td>
<td>23</td>
</tr>
<tr>
<td>2. Demographic statistics for the study cohort</td>
<td>26</td>
</tr>
<tr>
<td>3. Test for proportional hazard assumption for each covariate in univariate model</td>
<td>27</td>
</tr>
<tr>
<td>4. Inferential statistics for the time-to-event model</td>
<td>29</td>
</tr>
<tr>
<td>5. Coefficients from model with adding time dependent variables</td>
<td>31</td>
</tr>
<tr>
<td>6. Relative risk derived from six large cohorts in Allison et al.' study</td>
<td>36</td>
</tr>
<tr>
<td>7. The comparing of the coefficient estimation between the two models</td>
<td>42</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>The univariate Kaplan-Meier survival curves for BMI</td>
<td>50</td>
</tr>
<tr>
<td>A.2</td>
<td>The univariate Kaplan-Meier survival curves for gender</td>
<td>51</td>
</tr>
<tr>
<td>A.3</td>
<td>The univariate Kaplan-Meier survival curves for smoking</td>
<td>52</td>
</tr>
<tr>
<td>A.4</td>
<td>The univariate Kaplan-Meier survival curves for education</td>
<td>53</td>
</tr>
<tr>
<td>A.5</td>
<td>The univariate Kaplan-Meier survival curves for race</td>
<td>54</td>
</tr>
<tr>
<td>A.6</td>
<td>The cumulative hazard curves for BMI</td>
<td>55</td>
</tr>
<tr>
<td>A.7</td>
<td>The cumulative hazard curves for gender</td>
<td>56</td>
</tr>
<tr>
<td>A.8</td>
<td>The cumulative hazard curves for smoking</td>
<td>57</td>
</tr>
<tr>
<td>A.9</td>
<td>The cumulative hazard curves for education</td>
<td>58</td>
</tr>
<tr>
<td>A.10</td>
<td>The cumulative hazard curves for race</td>
<td>59</td>
</tr>
<tr>
<td>A.11</td>
<td>Cumulative hazard for all-cause mortality as a function of age in NHANES I and 1992 NHEFS</td>
<td>60</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Increased prevalence of obesity and associated problems in U.S.

Obesity or overweight for both adult men and women is classified by the body mass index (BMI). BMI is defined as body weight in kilograms (kg) divided by the square of height in meters (m²) (BMI = weight [kg] / height² [m²]). Underweight is defined as BMI ≤ 18.5 kg/m². Normal weight has a BMI between 18.5 and 24.9 kg/m². A BMI ≥25 kg/m² but less than 30 kg/m² is regarded as overweight. A BMI ≥ 30 kg/m² is defined as obesity [4]. The prevalence of overweight and obese has been increasing in both genders and among all population groups since 1960 [5, 6]. Recent estimates indicate that approximately 64% of U.S. adults are overweight or obese, over 30% are obese and 5% are extremely obese [5]. The prevalence of overweight, obesity and extreme obesity is expected to continue to increase [7].

Overweight and obesity pose a major public health challenge in the United States. Overweight and obesity are associated with a variety of co-morbid conditions including osteoarthritis, heart disease, hypertension, high cholesterol, stroke, diabetes, gallbladder disease, asthma, sleep apnea [8], depression [9], and certain cancers including colorectal,
prostate, endometrial, gallbladder, cervical, ovarian, and postmenopausal breast cancer [10]. Obesity is associated with higher levels of upper- and, especially, lower-body disability [11], and diminished quality of life [12]. Obesity is associated with increased risk in all-cause mortality and is the second only after smoking as the preventable cause of death [13].

Overweight and obesity and their associated health problems have substantial economic consequences for the U.S. health care system. The increasing prevalence of overweight and obesity is associated with both direct and indirect costs. In 1995, the total costs attributable to obesity amounted to an estimated $99 billion [14]. In 2000, the total cost of obesity was estimated to be $117 billion [15].

1.2 Previous estimation of extra death attributable to obesity

The estimated number of extra deaths attributable to obesity can be used to assess the impact of obesity in the U.S. population. Several studies have estimated the number of extra deaths attributable to obesity, but those estimated numbers are dramatically different. McGinnis et al. [16] analyzed data from articles published between 1977 and 1993, government reports, compilation of vital statistics and surveillance data, and estimated that 300,000 deaths a year in the U.S. were due to all aspects of diet and physical activity but not specifically to obesity. Allison et al. [2] estimated the number of obesity attributed extra deaths by employing data from six large prospective epidemiologic cohorts (the Alameda Community Health Study, the Framingham Heart Study, the Tecumseh Community Health Study, the American Cancer Society Cancer
Prevention Study I, and the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study and the Nurses’ Health Study). Their results indicated that a range of 236,111 to 341,153 estimated deaths per year could be attributed to obesity. Mokdad et al. [7] estimated the extra deaths attributable to obesity based upon the prevalence of obesity reported in NHANES 1999-2000 and the relative risks reported from previous studies. Their data revealed an estimated 385,000 extra deaths due to obesity in 2000 [7, 17]. A more recent study used First National Health and Nutrition Examination Survey (NHANES I), NHANES II with follow-up through 1992, and NHANES III with follow-up through 2000 to estimate the excess deaths caused by obesity [18]. Annual deaths attributable to obesity ranged widely from 26,917 to 298,808 when different source of data sets were applied [18].

1.3 Limitations in previous studies

There are several limitations in estimating obesity-attributable deaths in previous studies. These will be summarized next.

1.3.1 Selected study population not representative of target population

If the study population can not represent the target population, the relative risk of death associated with overweight and obesity will not be accurately estimated. For example, among these six large cohort studies (Alameda County Health Study, Tecumseh Community Health Study, Framingham Heart Study, American Cancer Society cancer prevention study, Nurses’ Health Study, and NHANS I Epidemiologic Follow-up Study)
applied in Allison's study [2], only the NHANES I Epidemiologic Follow-up Study represents the U.S. population. It is inappropriate to apply the relative risks derived from other five cohorts to the U.S. population.

1.3.2 Incorrect time-scale and later entry controlling

Cox’s proportional hazards model was initially developed to analyze the survival data in clinical trials. Subjects in clinical trials usually start exposure to a condition (i.e. drug under investigation) at the time of enrollment. These subjects are followed up from the time of diagnosis or receiving the treatment to the occurrence of death or relapse. Under such a condition, Cox’s proportional hazards model with time-to-event as a time scale includes age-at-entry as a covariate to adjust the effect of age. However, subjects in epidemiological cohort studies are usually exposed to the risk factor before they were enrolled into the studies. This late entry bias can not be controlled by using the standard analysis. For this reason, it is inappropriate to use the traditional analysis to construct Cox’s proportional hazards model in epidemiological studies [19].

1.3.3 Inconsistency in the definition of reference group

The risk factor of body weight was categorized based on different values of BMI. However, BMI is a continuous variable, and there is no self-evident threshold to define reference groups for normal weight and risk groups for overweight, obesity and severe obesity. Although many organizations give out the classification of overweight and obesity according to BMI, there is no absolute consistency among them. From the
National Heart, Lung, and Blood Institute classifies body weight status as follow: underweight (<18.5); normal (18.5-24.9); overweight (25.0-29.9); obesity stage I (30.0-34.9); obesity stage II (35.0-39.9); obesity stage III (≥ 40). The World Health Organization (WHO) gives a similar definition of body weight status classification except normal weight is defined as BMI falling in the range from 20 to 25. In reality, investigators may classify body weight status by using their own criteria. PAR estimation can be affected by the choice of the reference group and cut-point, and correspondingly the estimated number of extra deaths which is calculated from PAR will also be influenced. When different reference groups or cut-point of body weight status are employed in different studies, it is difficult to compare the data among these studies.

1.3.4 Incorrect equation for population attributable risk calculation

After Levin first generated the concept of the population attributable risk (PAR), several authors provide formulas to estimate PAR for more complicated applications. We should be cautious when applying these formulas. Some of them are only used for unadjusted relative risks (RR), while others can only be used for adjusted relative risks. The most common mistake is to use the adjusted RR in the formula which is only suitable for an unadjusted formula. As a result, the estimate of the PAR will be an overestimate or an underestimate.

1.4 Proposed investigation
In the present study, the NHANES I and NHEFS cohort study data sets were employed to estimate the number of extra deaths attributable to obesity. Cox’s proportional hazards model was applied to estimate the relative risk of deaths associated with obesity adjusted for some related covariates. Age with left truncation was used as the time scale in Cox’s proportional hazards model. The formula used to compute PAR was based on Bruzzi’s et al. formula [1].

1.5 Definition of terms

1. Population Attributable Risk: The proportion of cases in a population that can be attributed to exposure to the risk factor.

2. Obesity-attributable mortality: The number of lives in the population that could be saved by eliminating obesity.

3. Relative Risk: The probability of an event in the group of interest divided by the probability of the event in the control group.
CHAPTER 2

LITERATURE REVIEW

2.1 Introduction of population attributable risk

A attributable risk is commonly applied to evaluate the impact of a risk factor on exposed individuals. Attributable risk describes the proportion of a population among exposed individuals that would be eliminated if the exposure were eliminated. On other occasions, epidemiologists may focus on the impact of exposure on a total population which consists of both exposed and unexposed subjects. Although relative risks and odds ratios summarize the association between exposure and disease, they provide inadequate information about the impact of exposure on disease at the population level. Risk factors with high relative risks may not be a serious public health problem if a small number of the population is exposed. On the other hand, low risk factors will be a serious public health problem if a large number of the population is exposed. For example, certain industrial workers who are exposed to inhale dangerous carcinogens will have a much higher relative risk, as high as 50 times, to develop lung cancer than those who not. However the total number of lung cancer cases due to this occupational exposure to carcinogens in the population is small because the prevalence of exposure is small at the
population level. Compared to industrial carcinogens exposure, smoking induces much lower relative risk to develop lung cancer but contributes a much larger number of cases of lung cancer because of the higher prevalence of smoking in the population. So it is not sufficient to describe the impact of exposure to a disease at the population level only by using relative risk or odds ratio alone. For this reason, both relative risk and prevalence should be taken into account at the same time to assess the impact of a disease in public health at the population level.

Population attributable risk (PAR), instead of attributable risk, is more appropriate to evaluate the impact of the exposure on a disease in a population rather than only on the exposed subjects. PAR can answer the question of how much of the disease burden in a population could be eliminated if the concerned risk factor were eliminated. PAR, unlike the relative risk and odds ratio, measures the impact of exposure on the disease at the population level. In 1953, Levin first proposed the concept of PAR to assess the impact of smoking on lung cancer occurrence [20]. This concept has been widely employed to evaluate the consequences of an exposure factor the disease at the population level rather than at the individual level. After Levin proposed the concept of PAR, numerous interchangeable phrases have been used to refer to the same definition. These terms are “Population Attributable Risk”, “Population Attributable Fraction”, “Excess Fraction” and “Etiologic Fraction”. Among them the most popular terms used by authors are “Population Attributable Risk” and “Population Attributable Fraction”. The widely accepted definition of the PAR is “the proportional reduction in average disease risk over a specified time interval that would be achieved by eliminating the exposure of interest
from the population while distributions of other risk factors in the population remain unchanged” [21].

2.2 Commonly used equations for estimating population attributable risk

\[
PAR = \frac{(I_t - I_o)}{I_t} \quad (2.2.1)
\]

\( I_t \) = Incidence rate of the disease in a population

\( I_o \) = Incidence rate of the disease in the unexposed people in a population

This definition is valid only when no confounding exists [20,21].

\[
PAR = \frac{P_e \times (RR-1)}{(P_e \times (RR-1) + 1)} \quad (2.2.2)
\]

\( P_e \) = proportion of a population exposed to the risk factor of interest

\( RR \) = relative risk comparing the exposed level to the unexposed level

This definition is valid when there is no an exposure-disease confounding association [20,21].

\[
PAR = \frac{1}{\sum_{i=0}^{k} P_i \times (RR_i)} \quad (2.2.3)
\]

\( P_i \) is the proportion of a population in the \( i^{th} \) exposure level.

\( RR_i \) is the relative risk comparing the \( i^{th} \) exposed level to the unexposed level (\( i = 0 \)). This equation is used for multi-category exposures. This definition is not valid when there are confounders [21,22, 23].

- 9 -
\[ PAR = P \cdot (RR - 1) / RR \]  
\hspace{2cm} (2.2.4)

\( P \) = proportion of cases exposed to risk factor. This equation is valid when confounding exists and when adjusted relative risk must be used [22, 24].

\[ PAR = \sum_{i} P_i \cdot (RR_i - 1) / RR_i \quad (i=0 \text{ to } k) \]  
\hspace{2cm} (2.2.5)

\( P_i \) = proportion of cases falling into \( i \)th exposure levels.

\( RR_i \) is relative risk comparing \( i \)th exposed level with unexposed level (\( i=0 \)).

This is an extension of equation 2.1.4 to use with multi-category exposures. This is valid when confounding exists and when adjusted relative risk must be used [1, 21].

2.3 Crude estimation of population attributable risk

As mentioned in Section 2.1, there are two methods to assess the impact of an exposure on the public's health: attributable risk and PAR. Attributable risk describes the proportion of disease among the exposed that would be eliminated if the exposure were eliminated. It is the most relevant measure when making decisions for individuals. Attributable risk is the incidence rate difference between the exposed persons and unexposed persons divided by the incidence rate in exposed persons. Attributable risk can be expressed as follows:
\[
AR = \frac{(Ir - Io)}{Ir} \quad (2.3.1)
\]

where \(Io\) is the incidence rate in the unexposed group and \(Ir\) is the incidence rate in the exposed group.

Dividing both the numerator and denominator of the right side of the equation 2.3.1 by \(Io\), we can express the equation as

\[
AR = \frac{(RR - 1)}{RR}, \quad (RR = \frac{Ir}{Io}, \text{ given } RR > 1) \quad (2.3.2)
\]

where \(RR\) is the relative risk. Relative risk is the ratio of the incidence rate in exposed persons to that in people who are unexposed.

PAR is applied to measure the impact of an exposure on a disease in the whole population. PAR can be calculated as the number of excess cases in a population due to the association with the exposed risk factor divided by the total number of cases in the population. Let \(N\) be the total population size; \(I_t\) be the incidence rate in total population; \(Io\) be the incidence rate of unexposed population, then PAR can be expressed as below,

\[
PAR = \frac{Extra \ cases \ due \ to \ exposure \ in \ population}{Total \ cases \ in \ population}
\]

\[
PAR = \frac{((N * I_t) - (N * Io))}{(N * I_t)} = \frac{(I_t - Io)}{I_t} \quad (2.3.4)
\]
Let $P$ be the proportion of the population exposed and $Ie$ be the incidence rate in exposed population. The total number of cases occurring among the unexposed population is $N \times (1 - P) \times Io$. The total number of cases occurring among the exposed population is $N \times P \times Ie$. The incidence rate in population $It$ is the number of cases in the population divided by the number in the population. The incidence rate in the population $It$ can be expressed as,

$$It = \frac{\text{Number of cases in exposed} + \text{Number of cases in unexposed}}{\text{Number in population}}$$

$$It = \frac{N \times P \times Ie + N \times (1 - P) \times Io}{N = P \times Ie + (1 - P) \times Io} \quad (2.3.5)$$

The equation of 2.3.4 will be rewritten if $It$ was replaced by the equation from 2.3.5.

After replacement the population attribute risk (PAR) can be expressed as follows:

$$PAR = \frac{(It - Io)}{It} = \frac{(P \times Ie + (1 - P) \times Io - Io)}{(P \times Ie + (1 - P) \times Io)}$$

$$= \frac{(P \times RR + (1 - P) - 1)}{(P \times RR + (1 - P))}$$

$$= \frac{P \times (RR - 1)}{(P \times (RR - 1) + 1)}$$

$$PAR = P \times (RR - 1) / (P \times (RR - 1) + 1) \quad (2.3.6)$$

The equation 2.3.6 is used to estimate PAR under dichotomous exposure and when no confounding exist [20].
In the case of multiple levels of exposure, let $P_i$ be the proportion of $i$th exposure level in population ($i=0$ is the reference group); $I_i$ be the incidence rate in the $i$th exposed level in population. The expression of the PAR will be the same as that in equation 2.3.4. The incidence rate of the population with multiple exposure level can be expressed as below,

$$I_i = \sum_j (N_j * P_i * I_j) / N = \sum_j (P_i * I_j)$$

(2.3.7)

where $i = 0$ to 1, $i = 0$ is the reference group.

Replacing $I_I$ in equation (2.3.4) by the equation (2.3.7) PAR can be rewritten as,

$$\text{PAR} = (\sum_i (P_i^* I_0) - I_0) / \sum_i (P_i^* I_0)$$

(i = 0 to 1)

$$= (\sum_i (P_i^* I_0 + P_0^* I_0 - I_0)) / (\sum_i (P_i^* I_0 + P_0^* I_0))$$

(i = 1 to I)

$$= (\sum_i (P_i^* R_{RI}) - (1-P_0^*) R_{RO}) / (\sum_i (P_i^* R_{RI}) + P_0^* R_{RO})$$

(i = 1 to I)

$$= (\sum_i (P_i^* R_{RI} - \sum_i P_i) / (\sum_i (P_i^* R_{RI}) + (1-\sum_i P_i))$$

(i = 1 to I)

$$= (\sum_i P_i^* (R_{RI}-1)) / (1+(\sum_i P_i^* (R_{RI}-1)))$$

(i = 1 to I)

$$= (\sum_i P_i^* (R_{RI}-1)) / (1+(\sum_i P_i^* (R_{RI}-1)))$$

(i = 0 to I)

So, the final expression of PAR is

$$\text{PAR} = \sum_i [P_i (R_{RI} - 1)] / [1 + P_i (R_{RI} - 1)]$$

(i = 0 to 1)            

(2.3.8) [22, 23]
Equations 2.3.6 and 2.3.8 are only valid to estimate PAR under the assumption that there is no confounding of the exposure-disease association.

2.4 Adjusted point estimation of population attributable risk

The estimation of the PAR as stated above takes into account dichotomous and multiple levels of one exposure under the assumption that there is no confounding between disease and exposure. They are unadjusted estimates since they fail to take into account possible confounding between exposure and disease. Generally, unadjusted PAR estimates will introduce bias [23, 25, 26,29]. Adjusted PAR estimation will be consistent with the unadjusted PAR estimation when some constraints are met. In 1980, Walter gave the precise conditions under which adjusted PAR estimates will be consistent with unadjusted PAR estimates [27, 29]. Let X1 and X2 denote two dichotomous factors and X1 is the interested risk factor. Then the unadjusted and adjusted estimates of PAR will be close if and only if at least one of the two conditions (a) and (b) are true: (a) X1 and X2 are independently distributed in the population (b) Exposure to X2 alone does not increase disease risk [27, 29].

2.4.1 Adjusted estimates of PAR based on stratification approaches

As discussed above the estimates of the PAR depend on the estimates of the relative risk associated with exposure and the prevalence of exposure in a population. The
Mantel–Haenszel (MH) approach and the weighted-sum approach can yield valid estimates of relative risks [28, 29]. These two approaches are stratification based.

If no interaction exists between the exposure and confounding factors, employing the MH approach has favorable properties in that bias is very small. Another stratification based approach is the weighted-sum approach. Under the weighted-sum approach, PAR is written as a weighted-sum over all stratified levels of each level-specific PAR, namely

$$\text{PAR} = \sum w_j \text{PAR}_j,$$

where $w_j$ denotes the weight of specific level $j$ and $\text{PAR}_j$ represents PAR value of corresponding level. The weighted-sum approach allows adjustment for one or more polytomous factors forming $J$ levels or strata as the MH approach. Setting $w_j$ as the proportion of diseased individuals (cases) in level $j$ yields an asymptotically unbiased estimator of PAR. The relative risks are estimated separately for each adjustment level $j$. No restriction is placed on them, which corresponds to a fully saturated model for exposure and adjustment factors. From these separate relative risk estimates, separate PAR estimates are obtained for each level of adjustment. Thus, the weighted-sum approach not only accounts for confounding but also for interaction. In general, the MH approach is based on the assumption of no interaction between exposure and confounding factors. On the other hand, the weighted-sum approach allows for full interaction between exposure and adjustment factors [29].

2.4.2 Model-based adjusted estimation of PAR

The MH approach and the weighted-sum approach are stratification based approaches as discussed above. The MH approach takes into account only main effects.
The weighted-sum approach which is a fully saturated model takes into account not only main effects but all interactions among exposure and confounders. These two approaches are two extreme examples [29].

Walter and Fleiss first suggested using regression models for AR estimation [23]. Compared to stratification based approaches, regression models are more flexible. Regression models provide efficient parameter estimation and hypothesis testing. Regression models allow to taking into account main effects as well as some or all of the interactions. Moreover, regression models yield maximum likelihood estimators [29]. The regression approach was first exploited by Bruzzi et al. [1] who expressed PAR as

\[
\text{PAR} = 1 - \sum \sum \rho_{ij} / \text{RR}_{ij} \quad (i=0 \text{ to } I, \ j=0 \text{ to } J) \quad (2.4.2.1)
\]

The quantity \( \rho_{ij} \) represents the proportion of cases in the \( i \) exposure level and the \( j \) adjusted confounder level, while \( \text{RR}_{ij} \) represents the relative risk for the level \( i \) of exposure given the level \( j \) of adjustment factor [1, 29].

Under certain situations, it is believed that if some strata share the same level of the risk factor named A, but differ only in the levels of confounder named C, then these strata have the same value of the relative risk. The relative risk for any given stratum does not depend on the levels of confounder C. Then the calculation of equation \( \text{PAR} = 1 - \sum \sum \rho_{ij} / \text{RR}_{ij} \) can be simplified by considering a cross-classification only on risk factor A [1]. Letting \( q \) index the strata so formed,
\[ \text{PAR} = 1 - \sum q (\rho_q / \text{RR}_q) \quad (q = 0 \text{ to } Q) \]  

(2.4.2.2)

where RR_q is an adjusted estimate of the relative risk for the levels of a represented in stratum q, adjusted on C; \( \rho_q \) = proportion of cases falling into qth exposure levels [1].

The model-based approach has some favoring aspects. Although at first Bruzzi et al. derived the equation for case-control studies, it can also be applied for cohort and cross-sectional studies [29]. Cox’s regression can provide a maximum-likelihood estimate for RR_q. The estimate for \( \rho_q \) is the observed proportion of cases falling into the q exposure level. In addition, regression models based approaches provide convenience for testing hypotheses and selecting models as well as estimating parameters. Partial interaction terms can be incorporated in the model depending on the result of the test [29]. At least, the MH and the weight sum approaches are only two special cases in the model-based approach. The MH approach corresponds to a model with exposure and adjustment factors without any interactions. The weighted-sum approach corresponds to a fully saturated model with all interactions. While a model-based approach has many advantages to obtain a point estimate for PAR, it is difficult to obtain a variance estimate because it involves covariances between estimates of \( \rho_q \) and RR_q [1, 29].
2.5 Choice of threshold for risk factor

When the exposure of interest is a continuous variable, there is not always a clear threshold for defining of exposure and reference groups [30]. PAR estimation, whether adjusted or not, can be greatly affected by the choice of the reference level. If a more stringent definition is adopted for the reference group, then a larger proportion of subjects will correspondingly be in the exposed group. This will change the relative risk for the exposure and affect the PAR value eventually. Therefore, PAR estimates must be reported with respect to a clearly defined reference level in order to be validly interpreted [29].

2.6 PAR in survival analysis

From Bruzzi's [1] formula 2.4.2.2,

\[ \text{PAR} = 1 - \sum q (p \cdot q / RR \cdot q) \]

The proportion of cases falling into qth exposure level and the relative risk of having disease associated with the risk factor both need to estimate the PAR. Relative risk is commonly estimated by employing survival analysis.

2.7 Conceptual problems in PAR

It is important to interpret and communicate the estimated PAR with the aim of making a public health policy decision. PAR is usually defined as the proportion of
disease cases that will be eliminated if the exposure was eliminated. The most frequent mistake is to interpret the PAR as the proportion of cases exposed.
CHAPTER 3

STUDY DESIGN AND METHODOLOGY

3.1 Study cohorts: NHANES I and NHEFS

Our current study employed the data sets from the First National Health and Nutrition Examination Survey (NHANES I 1971-1975) and the National Health and Nutrition examination Survey I Epidemiologic Follow-up Study (NHEFS 1992). The data of the risk factor body mass index (BMI), confounding covariates (age, gender, race, education, smoking status) and the mortality outcome were extracted from the above study cohorts. Annual death number was from 2003 National Vital Statistics Report.

The National Center for Health Statistics, part of the Centers for Disease Control and Prevention, conducted NHANES I to obtain information about the health and nutritional status of the population in the United States during the years of 1971-1974 [32]. Based on a U.S. national probability sample, approximately 32,000 non-institutionalized persons aged 1 to 74 years old were initially enrolled in the survey. A total of 23,808 subjects were enrolled and finished the medical examination, interview.
questionnaire and laboratory determination. Among those 23,808 enrolled subjects, 18,836 subjects were 25 to 74 years old. As an extension between July 1974 and October 1975, a sub-sample of 6,913 persons from 18,836 adult subjects aged 25 to 74 received a more detailed health examination and additional health histories questionnaires. This sub-sample is representative of the United States population aged 25-74 during the time of NHANES I.

The NHANES I Epidemiologic Follow-up Study (NHEFS) was designed to investigate the relationship between clinical, nutritional, and behavioral factors assessed in NHANES I and subsequent morbidity and mortality [33]. The NHEFS cohort included all persons 25-74 years of age who completed a medical examination at NHANES I in 1971-75 (n= 4,407). Three follow-ups were conducted in the NHEFS during the years 1982–84, 1986–87, and 1992. The first one (NHEFS 1982-84) was conducted for all members of the NHEFS cohort. The NHEFS 1986 was conducted for subjects of 55–74 years age at their baseline examination and not known to be deceased at the 1982–84 NHEFS (n=3,980). The 1987 NHEFS was conducted for the entire non-deceased NHEFS cohort (n=11,750) with the same questionnaire as the 1986 survey. The third one, the 1992 NHEFS, included the entire non-deceased NHEFS cohort (n=11,195). Subjects who had died were followed up by proxy interview. Of the 14,407 participants eligible for follow-up, 96% had been successfully traced at some points through the 1992 follow-up, and 91% to 96% of those traced completed the follow-up interview.

The study cohort used in our current study was the sub-sample (n=6,913) in NHANES I. This cohort was aged 25 to 74 and received a detailed health examination
and additional health histories questionnaires. Smoking status was available for persons in this sub-sample. Subjects' height (in inches) and weight (in pounds) were measured by trained technicians. The risk factor body weight status was defined by body mass index (BMI). The equation employed to calculate BMI was based on the formula: \[ \text{BMI} = \frac{\text{Weight (kg)}}{\text{Height}^2 \ (\text{m}^2)} \]. Potential confounders selected for analysis the risk of death included smoking, age, race, gender and education. Outcome variable, all-cause deaths, came from NHEFS 1992 mortality data tape.

3.2 Descriptive analysis

As described above, a sub-sample of NHANES I data set with 6,913 subjects was applied as our study cohort. Mortality data came from the 1992 NHEFS. Before building a Cox proportional hazards model, subjects with missing data values were deleted. We assumed that the missing data were missing at random. A total of 6,869 subjects were finally included in this study. Descriptive statistics included gender, age, smoking status, BMI, education, race, and the all-cause death outcome among all subjects. SAS version 8.1 (SAS Institute Inc, Cary, NC) was used for data analysis.

3.3 Cox's proportional hazards model coding and building

A Cox proportional hazards model was used to estimate the relative risk of mortality on risk of body weight status (defined by BMI), while adjusting for entry age, gender, race and education. Additional potential confounding effects of smoking status were considered for their possible impact on the relationship between body weight and
mortality. The BMI definition that was used is from the National Institutes of Health and the World Health Organization definition. The groups were formed the following way: “Normal weight” was defined as $18.5 \leq \text{BMI} < 25$; “Underweight” was defined as $\text{BMI} < 18.5$; “Overweight” was defined as $25 \leq \text{BMI} < 30$; “Obesity” was defined as $30 \leq \text{BMI} < 35$; and “Severe obesity” was defined as $\text{BMI} \geq 35$. Smoking status was categorized as former smoker, current smoker, and nonsmoker based on their self-report on the NHANES I questionnaire.

Dummy variables for Gender, BMI, smoking, education and race were created before building the model. They were listed in Table 1.

| Gender:          | 0: Male  
|                 | 1: Female  
| BMI:             | 0: Normal Weight $18.5 \text{ kg/m}^2 < \text{BMI} < 25 \text{ kg/m}^2$  
|                 | 1: Underweight $\text{BMI} < 18.5 \text{ kg/m}^2$  
|                 | 2: Overweight $25 \text{ kg/m}^2 \leq \text{BMI} < 30 \text{ kg/m}^2$  
|                 | 3: Obesity $30 \text{ kg/m}^2 \leq \text{BMI} < 35 \text{ kg/m}^2$  
|                 | 4: Severe Obesity $35 \text{ kg/m}^2 \leq \text{BMI} \text{ kg/m}^2$  
| Education:      | 0: less than high school  
|                 | 1: high school  
|                 | 2: above high school  
| Race:           | 0: White  
|                 | 1: Black and Other  
| Smoking Status: | 0: Never smoker  
|                 | 1: Former smoker  
|                 | 2: Current smoker  

**Table 1:** Code Table for Each Covariate

We applied the left truncation method to adjust the age at entry to construct a Cox’s proportional hazards model. As a control, the time-on-study model was also built to observe the magnitude of the bias. Each covariate was modeled individually to check
proportional hazards assumption. AIC criteria were used to choose covariates into the final model.
CHAPTER 4

RESULTS

4.1 Demographic statistics in study cohort

Our current study employed a sub-sample with a total of 6,869 subjects aged from 25 to 74 extracted from the NHANES I and NHEFS Study Cohorts. In addition to the independent risk factor BMI and outcome of deaths, several possible confounders were also included in this study, such as age as a delayed entry time, gender, race, education and smoking status (Table 2). Among these subjects, 3,142 (45.7%) were male and 3,727 (54.3%) were female. The equation $\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height}^2 (\text{m}^2)}$ was used to define the body weight status in this study. Based on BMI criteria, 3.4% subjects were “underweight” (BMI < 18.5 kg/m$^2$), 46.0% were “normal body weight” (18.5 kg/m$^2$ ≤ BMI < 25 kg/m$^2$), 34.1% were “overweight” (25 kg/m$^2$ ≤ BMI < 30 kg/m$^2$), 11.6% were “obese” (30 kg/m$^2$ ≤ BMI < 35 kg/m$^2$), and 4.8% were “severe obese” (BMI > 35 kg/m$^2$). Smoking status was considered as a potential risk factor for deaths. We categorized smoking status into 3 classes: never smoker, former smoker and current smoker. The proportion of never smoking, former smoking and current smoking in this study cohort
were 40.9%, 21.6% and 37.5%, respectively. A total of 1,930 (28.1%) died during this cohort study.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Frequency (n)</th>
<th>Percent (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3142</td>
<td>45.7</td>
</tr>
<tr>
<td>Female</td>
<td>3727</td>
<td>54.3</td>
</tr>
<tr>
<td>Smoking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current Smoking</td>
<td>2376</td>
<td>37.5</td>
</tr>
<tr>
<td>Former Smoking</td>
<td>1485</td>
<td>21.6</td>
</tr>
<tr>
<td>Never Smoking</td>
<td>2808</td>
<td>40.9</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>5940</td>
<td>86.5</td>
</tr>
<tr>
<td>Other</td>
<td>929</td>
<td>13.5</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>2802</td>
<td>40.7</td>
</tr>
<tr>
<td>High school</td>
<td>2332</td>
<td>34.0</td>
</tr>
<tr>
<td>Above high school</td>
<td>1735</td>
<td>25.3</td>
</tr>
<tr>
<td>BMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under Weight (&lt; 18.5 kg/m²)</td>
<td>238</td>
<td>3.4</td>
</tr>
<tr>
<td>Normal Weight (18.5 to &lt; 25 kg/m²)</td>
<td>3162</td>
<td>46.0</td>
</tr>
<tr>
<td>Overweight (25 to &lt; 30 kg/m²)</td>
<td>2342</td>
<td>34.1</td>
</tr>
<tr>
<td>Obesity (30 to &lt; 35 kg/m²)</td>
<td>797</td>
<td>11.6</td>
</tr>
<tr>
<td>Severe Obesity (≥ 35 kg/m²)</td>
<td>330</td>
<td>4.8</td>
</tr>
<tr>
<td>Death Outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Censored</td>
<td>4939</td>
<td>71.9</td>
</tr>
<tr>
<td>Death</td>
<td>1930</td>
<td>28.1</td>
</tr>
</tbody>
</table>

Table 2: Demographic Statistics for the Study Cohort

4.2 Time-on-study Models

We first built the time-on-study model in the traditional way which controls the effect of age by adding age-at-entry as a covariate.

The univariate Kaplan-Meier survival curves and cumulative hazard curves for smoking, BMI, race, gender and education were plotted and can be found in the Appendix Figures 1-10. The survival curves for each BMI level showed that subjects of severe obesity or underweight had the worst survival rate compared with that of overweight or normal weight groups which had the best and similar survival rate, where
obese demonstrated an intermediate survival rate between the above groups. As expected, the survival curve of the never-smoking group was above the other two smoking groups. The current smoking group had the lowest survival rate. The survival curves for the three education groups were close. The survival curve for female was above the male group. The survival curve for white race group was above the other race groups. Cumulative hazard curves were plotted to give a general view of proportional hazards assumptions for the risk factor BMI and each covariate. These figures are presented in the Appendix Figures 6-10. Based on the cumulative hazard plots, we suspected that race did not satisfy the proportional hazard assumption. The Wald test was then employed to test the proportion hazards assumption. The p-value of Wald test for each covariate is given in Table 3.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Wald test</th>
<th>Degree of freedom</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at entry*</td>
<td>86.6530</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Gender</td>
<td>0.1256</td>
<td>1</td>
<td>0.7231</td>
</tr>
<tr>
<td>Education</td>
<td>1.1436</td>
<td>2</td>
<td>0.5645</td>
</tr>
<tr>
<td>Smoking*</td>
<td>10.0691</td>
<td>2</td>
<td>0.0065</td>
</tr>
<tr>
<td>BMI</td>
<td>4.5218</td>
<td>4</td>
<td>0.34</td>
</tr>
<tr>
<td>Race*</td>
<td>15.5045</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 3: Test for proportional hazards assumption for each covariate in univariate model

Table 3 shows that age-at-entry, smoking status and race violated the proportional hazards assumption. Time dependent variables were incorporated into the model to handle the violation of the proportional hazards assumption. The time dependent variables were created as log (time to event) * covariate.
First we only incorporated age-at-entry and BMI into the model before adjusting for gender, race, education and smoking status. This was done to show how incorporating additional confounders can change the results. We call this model the base time-on-study model. The relative risks of dying associated with underweight, overweight, obesity, and severe obesity were 2.331, 0.973, 1.202 and 1.481, respectively. The number of total deaths in 2003 aged 25 and older is 2,374,029 from National Vital Statistics Report. The relative risk of overweight was less than one, so it would not be included to calculate the PAR. The proportions of death fall into normal weight, obesity, and severe obesity levels estimated from NHANES I were 64.48%, 24.83% and 10.69%, respectively.

PAR was calculated from the equation 2.4.2.2 derived by Bruzzi et al. [1]. The equation is

\[ \text{PAR} = 1 - \sum_j \left( \frac{\rho_j}{RR_j} \right) \quad (RR_j > 1) \tag{2.4.2.2} \]

where \( RR_j \) is an adjusted estimate of the relative risk for the levels of a represented in stratum \( j \), adjusted on other covariates; \( \rho_j \) denotes proportion of cases falling into \( j \)th exposure levels (\( j = 0 \) to \( k \)).

Applying the above numbers to equation (2.4.2.2) we get

\[ \text{PAR} = 1 - (0.6448/1+0.2483/1.202+0.1069/1.481) = 0.07644666 \]

Extra Deaths = The number of total deaths * PAR

\[ = 2,374,029 \times 0.07644666 \approx 181,486 \]
<table>
<thead>
<tr>
<th>Models</th>
<th>Covariates</th>
<th>Coefficient</th>
<th>P-value</th>
<th>Hazard Ratio</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unadjusted for gender, race, smoking status</strong></td>
<td>Age-at-Entry</td>
<td>0.07641</td>
<td>&lt;.0001</td>
<td>6.61</td>
<td>30673.140</td>
</tr>
<tr>
<td></td>
<td>Ln (Age-at-Entry)</td>
<td>0.00637</td>
<td>0.0091</td>
<td>2.331</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Underweight</td>
<td>0.84611</td>
<td>&lt;.0001</td>
<td>2.331</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overweight</td>
<td>-0.02759</td>
<td>0.6026</td>
<td>0.973</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Obesity</td>
<td>0.18411</td>
<td>0.0080</td>
<td>1.202</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Severe obesity</td>
<td>0.39258</td>
<td>&lt;.0001</td>
<td>1.481</td>
<td></td>
</tr>
<tr>
<td><strong>Adjusted for gender, race and smoking status</strong></td>
<td>Age-at-Entry</td>
<td>0.08174</td>
<td>&lt;.0001</td>
<td>6.61</td>
<td>30300.604</td>
</tr>
<tr>
<td></td>
<td>Ln (Age-at-Entry)</td>
<td>0.00816</td>
<td>0.0015</td>
<td>2.125</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Underweight</td>
<td>0.75371</td>
<td>&lt;.0001</td>
<td>2.125</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overweight</td>
<td>-0.01195</td>
<td>0.8227</td>
<td>0.988</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Obesity</td>
<td>0.33385</td>
<td>&lt;.0001</td>
<td>1.396</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Severe obesity</td>
<td>0.59255</td>
<td>&lt;.0001</td>
<td>1.809</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gender(female=1)</td>
<td>-0.56543</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Race(white=1)</td>
<td>0.60000</td>
<td>&lt;.0001</td>
<td>4.321</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ln (race)</td>
<td>-0.13871</td>
<td>0.0301</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Former smoker</td>
<td>0.13157</td>
<td>0.4251</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ln (former smoker)</td>
<td>-0.02660</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Current smoker</td>
<td>0.72882</td>
<td>0.7077</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ln (current smoker)</td>
<td>-0.02511</td>
<td>0.6894</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 4:** Inferential Statistics for the time-on-study model

Second, we included gender, race and smoking status into the base time-to-event model to obtain more accurate estimates of adjusted relative risk. The inferential statistics for this time-on-study models are also shown in Table 4. The relative risk of death for underweight, overweight, obesity and severe obesity were 2.125, 0.988, 1.396 and 1.809, respectively. By using the same equation 2.4.2.2 to calculate PAR, we can get

\[
\text{PAR} = 1 - (0.6448/1 + 0.2483/1.396 + 0.1069/1.809) = 0.1182412
\]

**Extra Deaths = The number of total deaths * PAR**

\[
= 2,374,029 * 0.1182412 = 280,708
\]
4.3 Left Truncation Models

Age with left truncation is a more appropriate way to control age at entry. Before building the left-truncation-model, the proportional hazards assumption for other covariates was checked. According to the Wald test, race (P-value<0.0001, df=1) and education (P-value<0.0001, df=2) violates the assumption. We need to add time dependent variable, log (age-at-event) * covariate, for race and education if they were incorporated into the model.

The AIC criterion was used to select significant covariates into model one by one until the AIC no longer decreased. If the last covariate included in the model was not significant, it was deleted from the model. BMI, smoking, gender, race and education were chosen into the model sequentially. The final model included smoking, gender, BMI, race, education and the relevant time dependent variables. The estimated coefficients are shown in Table 5. The effect of interactions among these selected covariates was also checked. The results of the Wald tests suggest that the interactions effects were not significant. No interaction was included in the model. Our results showed that the relative risk for all BMI levels demonstrated a U shape curve. The relative risk of severe obesity and underweight were much higher than those of other body weight statuses. The relative risk was the lowest for the overweight subjects. The equation 2.4.2.2 was applied to calculate
PAR = 1 - (0.6448 / (1 + 0.2483 / 1.362 + 0.1069 / 1.782)) = 0.1129058

Extra deaths = The number of total deaths * PAR

= 2,374,029 * 0.1129058

= 268,041

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Coefficient</th>
<th>P-Value</th>
<th>Hazard-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking1</td>
<td>1</td>
<td>0.07414</td>
<td>0.2509</td>
<td>1.077</td>
</tr>
<tr>
<td>Smoking 2</td>
<td>1</td>
<td>0.65315</td>
<td>&lt;.0001</td>
<td>1.922</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>-0.57510</td>
<td>&lt;.0001</td>
<td>0.563</td>
</tr>
<tr>
<td>BMI1 (under weight)</td>
<td>1</td>
<td>0.73196</td>
<td>&lt;.0001</td>
<td>2.079</td>
</tr>
<tr>
<td>BMI2 (over weight)</td>
<td>1</td>
<td>-0.01769</td>
<td>0.7412</td>
<td>0.982</td>
</tr>
<tr>
<td>BMI3 (obesity)</td>
<td>1</td>
<td>0.30916</td>
<td>&lt;.0001</td>
<td>1.362</td>
</tr>
<tr>
<td>BMI4 (severe obesity)</td>
<td>1</td>
<td>0.57783</td>
<td>&lt;.0001</td>
<td>1.782</td>
</tr>
<tr>
<td>Race</td>
<td>1</td>
<td>4.32256</td>
<td>0.0014</td>
<td>75.381</td>
</tr>
<tr>
<td>RaceIn</td>
<td>1</td>
<td>-6.96002</td>
<td>0.0026</td>
<td>0.383</td>
</tr>
<tr>
<td>Education 1</td>
<td>1</td>
<td>-5.47600</td>
<td>&lt;.0001</td>
<td>0.004</td>
</tr>
<tr>
<td>Education 2</td>
<td>1</td>
<td>-5.07186</td>
<td>0.0004</td>
<td>0.006</td>
</tr>
<tr>
<td>Education1ln</td>
<td>1</td>
<td>1.22461</td>
<td>0.0001</td>
<td>3.403</td>
</tr>
<tr>
<td>Education2ln</td>
<td>1</td>
<td>1.12858</td>
<td>0.0008</td>
<td>3.091</td>
</tr>
</tbody>
</table>

**Table 5**: Coefficients from left truncation model adding time dependent variables
CHAPTER 5

DISCUSSION

5.1 Comparison with previous studies

The prevalence of obesity has been continuously increasing during the past few decades. Obesity is associated with increased risks of chronic disease incidence and mortality. The deaths attributable to obesity are causing increasing concern in our society. Our current study extracted a representative sub-sample from NHANES I (1971-1975) and the 1992 NHFES cohorts. We applied a Cox’s proportional hazards model to estimate the extra deaths attributable to obesity by employing body weight status (BMI) as the risk factor and time to death as the outcome. Two different approaches, time-on-study and left-truncation, were used to build Cox’s proportional hazards models. Age, smoking status, gender and education level were included as confounding factors. Our results showed that 280,708 estimated extra deaths per year would be attributable to obesity if the time-on-study model was used, and 268,041 estimated extra deaths per year would be attributable to obesity if the left truncation model was used.
Several authors have reported their estimates of the extra deaths attributable to obesity in the United States. Among them, the estimations from Alison et al. (1999) and Flegal et al. (2005) were based on Cox's proportional hazards models. We compare our current results with those from Alison et al. and Flegal et al.

Allison et al. estimated relative risks for each BMI category adjusted for age, sex, and smoking from six large cohort studies [2]. These relative risks, combined with the prevalence of BMI categories and mortality data from the 1991 U.S. population, were used to estimate the number of extra deaths attributable to overweight and obesity. The estimated number of extra deaths attributable to overweight and obesity ranges from 236,111 to 341,153 deaths per year when different cohort data set was applied. Then they combined all the subjects in the six study cohorts into one large data set to estimate the extra death attributable to overweight and obesity as 280,184 deaths per year. One of the critical limitations of Allison et al. (1999) study is that selected populations are not representative to U.S. population. Six study cohorts employed in their study were the Alameda Community Health Study, the Framingham Heart Study, the Tecumseh Community Health Study, the American Cancer Society Cancer Prevention Study I, the NHEFS 1992 and the Nurses Health Study. Among these six study cohorts, only NHEFS 1992 which used NHANES I data set as the baseline can represent the U.S. population.

Based on NHEFS data, 236,111 estimated extra deaths per year attributable to overweight and obesity and 184,670 estimated extra deaths per year associated with obesity were reported in their study. This estimated number may still not be accurate due to the following limitations. First, subjects in sampled study population aged between 25 and 75,
whereas the target population with age of 18 years or older was applied. Second, the reference group did not include all subjects of "normal weight" or of BMI between 18.5 to 25 kg/m². The reference group in their study was the group with a BMI of 23-25 kg/m². Third, Allison et al. used time-on-study as the time scale with age at entry as a covariate. The time-on-study model could introduce bias when estimating relative risks. The left truncation approach is a better way to control the effect of entry age at enrollment (discussed in more detail in the next section). Forth, an inappropriate formula was used to calculate PAR in their study. That particular formula is only applied with unadjusted relative risks, not with adjusted relative risks, which were calculated in their study. This mistake will introduce bias when calculating the final number of extra deaths. With a sub-sample of 6,913 subjects from the same NHANES I baseline data set and NHEFS 1992 mortality data set, our current study improved the limitations that appeared in their study. We used the proper target population with age between 25 and 75 years, and we used the whole group of "normal weight" (BMI between 18.5 and 25 kg/m²) subjects as the reference group, applied the left-truncation approach to build the proportional hazards model and employed an appropriate PAR equation. All these improvements led to the difference in the estimated number of extra deaths in our study from that of Allison's study. We estimated 268,041 extra deaths per year associated with obesity. Even with the same time-on-study model, we estimated 280,708 extra deaths per year associated with obesity in our study. Allison's study used all 14,407 subjects aged 25 to 75 in NHANES I, whereas only 6,913 subjects were included in our current sub-sample. The total number of deaths in the Allison target population was 2,110,687. The total number of deaths in
our target population was 2,374,029. There was a difference of 263,342 in the number of total deaths between the two target populations. The differences in sample size and the target population should not have much influence on the estimated number of extra deaths associated with obesity (see explanations in the following section). We conclude that the estimated number of extra deaths attributable to deaths from NHEFS data set in Allison’s study may be under-estimated.

Flegal et al. also employed Cox’s proportional hazards models to estimate the relative risk associated with overweight and obesity. The adjusted confounders in their model were gender, race, education, smoking status and alcohol consumption. The baseline data in their study came from NHANES I, NHANES II and NHANES III which were conducted during 1971-1975, 1976-1980 and 1988-1994, respectively. The subsequent mortality data was from 1992 for NHANES I and NHANES II, and 2000 for NHANES III. These three NHANES surveys were all based on nationally representative cross-sectional samples of the United State population. They combined the relative risks, total number of deaths in year 2000, and distribution of covariates in NHANES 1999-2000 cross-sectional survey data to estimate the extra death attributable to overweight and obesity. The estimated extra deaths attributable to obesity calculated from data sets NHANES I, NHANES II, NHANES III were 298,808, 26,917 and 43,650 per year, respectively. The estimated number from NHANES I was almost 9 times as high as that from NHANES II, and 7 times as that from NHANES III. When all subjects in the three NHANES were combined into one large data set, the estimated number of extra deaths attributable to obesity was 111,909 which was almost one third of that estimated from
NHANES I dataset. Their data suggest that the relative risks of death associated with obesity decreases though the prevalence of obesity has continuously increased during the past thirty years. Flegal et al. explained the dramatic decrease in the estimated number of extra deaths attributable to obesity from NHANES I data set NHANES II, or III data set. They claimed that the improvement of medical care prevented the deaths, especially cardiovascular disease related deaths, and therefore decreased the number of extra deaths attributable to obesity. However, this explanation was not convincing since NHANES II (1976-1980) was conducted immediately after NHANES I (1971-1975), thus the improvement in medical care should not have had a great impact (90% decrease in extra deaths attributable to obesity). In addition, their data showed that the relative risks of death associated with obesity or severe obesity were less than 1.0 or very close to 1.0 in all age categories based on the NHANES II or III data sets [18]. Although several limitations existed in Allison’s study, all six large cohort studies employed in their study estimated a relative risk that was greater than 1.3 for obesity and further increased in the severe obesity group [2]. Table 6 summarizes the estimated

<table>
<thead>
<tr>
<th>Relative risk</th>
<th>30 ≤ BMI &lt; 35</th>
<th>BMI &gt;35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alameda County Health Study</td>
<td>1.36</td>
<td>2.79</td>
</tr>
<tr>
<td>Framingham Heart Study</td>
<td>1.6</td>
<td>1.94</td>
</tr>
<tr>
<td>Tecumseh Community Health Study</td>
<td>1.45</td>
<td>1.87</td>
</tr>
<tr>
<td>American Cancer Society Cancer Prevention Study I</td>
<td>1.35</td>
<td>1.72</td>
</tr>
<tr>
<td>Nurses Health Study</td>
<td>1.49</td>
<td>1.89</td>
</tr>
<tr>
<td>NHANES I Epidemiologic Follow-up Study</td>
<td>1.33</td>
<td>1.68</td>
</tr>
</tbody>
</table>

Table 6: Relative Risk Derived from Six Large Cohorts in Allison et al.’ Study
relative risk of death associated with obesity and severe obesity in Allison's study. We suspect that the results from NHANES II and III in Flegal's study may not be reliable. We speculate that the shorter follow-up period of NHANES II (16 years) and NHANES III (12 years) cohorts compared with NHANES I (21 years) cohort may attribute their low number of estimated extra deaths attributable to obesity. Since obesity is not related to immediate death, the obese subjects of younger ages may still survive during the short follow-up period, and thus fewer deaths can be observed. Consistent with our speculation, Flegal's data showed a 30.10% mortality rate recorded in the NHANES I cohort, whereas only 23.1% and 18.64% of subjects dead in the NHANES II and III cohorts, respectively.

Based on the discussion above, we only compared our results with those from the NHANES I cohort in Flegal's study. Both our study and their study used the same NHANES I data set. The NHEFS 1992 was also applied to acquire subsequent mortality data. Subjects with BMI of 18.5 to < 25 kg/m² were applied as the reference group. Our current study differs from Flegal's study in that a different study population was used as well as different target population and a different method to calculate the PAR. Their study cohort included all 12,655 subjects aged 25 to 74 in NHANES I. Our study cohort was a sub-sample of 6,894 subjects from NHANES I adult subjects 25 to 74. In our study, the total number of deaths (2,374,029) for aged 25 and older was derived from the target population of year 2003. Flegal's study used the total number of deaths (2,331,261) from target population of year 2000. They used formula \( \frac{(R-R^*)}{R} \) to calculate PAR. Where \( R \) is mortality rate for a given age group and \( R^* \) is mortality rate at the reference-weight.
category for the same age group. Their equation to calculate the mortality rate for a given age group is shown as below

\[ R = I \sum ri \ p_i \]

\( ri \): Relative risk corresponding to each combination of BMI level and the level of other covariates.

\( p_i \): Prevelence of each combination of BMI level and the levels of other covariates

\( I \): Population baseline mortality rate

The formula they use to calculate reference group mortality rate is shown as below

\[ R^* = I \sum ri^* \ p_i \]

\( ri^* \): Relative risk in the i stratification where BMI level is at the reference level but other covariates levels are same.

\( p_i \): Prevalence of the risk factor combination

\( I \): Population baseline mortality rate

So the equation to calculate PAR in Flegai et al. can be expressed as below,

\[ PAR = (R-R^*)/R = 1-R^*/R = 1-(I^* \sum ri^* \ p_i) / (I^* \sum ri^* \ p_i) = 1 - (\sum ri^* \ p_i) / (\sum ri \ p_i) \]
Flegal et al. (2005) claimed that they used age as time scale in their model. However, the left truncation method seemed not to be applied in the control of late entry effects since age-at-entry was still included as a covariate in their model. The authors tried to avoid the violation of proportion hazards assumption across age by stratifying all subjects into three groups: 25 to 59 years, 60 to 69 years, and above 70 years. Our estimated number of extra deaths attributable to obesity was 280,708 per year which is 6.1% smaller than the number estimated from Flegal's study. This result suggests that different sample size, different target population and different method to calculate PAR will not account for a big difference in the number of estimated extra deaths associated with obesity.

When the left truncation method was applied in our study, the estimated number of extra deaths attributable to obesity decreased to 268,041 per year, a 4.5% further decrease compared to the number derived from the adjusted time-on-study model. Our results suggest that, compared to the left truncation model, the adjusted time-on-study model will lead to a slight increase in the estimated number of extra deaths attributable to obesity. Our finding suggests that the number reported by Flegal group may be overestimated since the left truncation approach was not applied in their study. In summary, we conclude that approximately 270,000 extra deaths per year may be associated with obesity.
5.2 Choice of time scale

In 1989, Cnaan and Ryan suggested that traditional survival analysis with time on study as the time scale may under or over-estimate the survival distribution of interest [31]. The direction of the bias will depend on the shape of the true underlying hazard. Similar results hold for the coefficients from a proportional hazards model. If left truncation is ignored, length bias may occur since patients must survive long enough to enter a trial. If the observation started at some time after the onset or diagnosis of disease, the estimated survival curve will be different from the one in which the observation started at same time as the onset or diagnosis of the disease. Therefore, bias will exist for estimation derived from time-on-study model unless the survival function is an exponential distribution.

In 1996, Korn et al. assessed the magnitude of the error that occurs when using time-on-study as the time scale in large scale health surveys. They stated that large scale health surveys offer the ability to minimize selection bias while at the same time length bias may exist. Korn et al. presented two sufficient conditions under which Cox’s proportional hazards model built with the left truncation approach will be the same as Cox’s proportional hazards model built with time-to-event as the time-scale adjusting for age-at-entry. First, if the hazard function belongs to a family of exponential distributions then the two approaches will yield identical estimates. Second, if the risk factor and age-at-entry are independent, then the estimates from the two approaches should be similar. They also stated if the hazard function roughly approximate an exponential distribution, the bias of estimation from the time-on-study model will not be serious. They illustrated
their theory by giving two examples based on the women in the 1987 NHEFS cohort. The two outcomes they had chosen were all cause mortality and removal of the ovaries. They showed that the cumulative hazard for all cause mortality as a function of age was approximately exponential, but removal of the ovaries was not. They evaluated the magnitude of the bias in the estimation of the coefficients in by fitting conducting three proportional hazards models. The first proportional hazards model was fit by using the left truncation method and was stratified by age. The second model was the proportional hazards model with left truncation but without stratification on age group. The third model was the traditional time-on-study proportional hazards model including age at entry as a covariate. For all-cause mortality, the three models yielded almost identical estimates as expected because of the roughly exponential cumulative hazard function. For removal of the ovaries, the first and second models yielded almost identical estimates whereas the estimates from the third model was quite different from the first two models since the cumulative hazard function is not anything like an exponential function.

Cnaan and Korn presented the theory to help explain the impact of the bias if the left truncation is not used. According to sections 4.2 and 4.3, the estimated numbers of the extra deaths attributed to obesity are 280,708 per year and 268,041 per year respectively. The difference between the two methods seems modest. Our study revealed that the distribution of the cumulative hazard for all-cause mortality for both males and females in NHANES I and 1992 NHEFS was roughly an exponential function. Therefore, the time-on-study method will not introduce severe bias compared to the left truncation
method. The cumulative hazard function is showed in Appendix Figure A 11. The estimated coefficients are summarized in Table 7.

<table>
<thead>
<tr>
<th>Risk Factors</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time-on-study with baseline age as a covariate</td>
</tr>
<tr>
<td>Underweight</td>
<td>0.75371</td>
</tr>
<tr>
<td>Overweight</td>
<td>-0.01195</td>
</tr>
<tr>
<td>Obesity</td>
<td>0.33385</td>
</tr>
<tr>
<td>Severe obesity</td>
<td>0.59255</td>
</tr>
</tbody>
</table>

Table 7: The comparing of the coefficient estimation between the two models

5.3 Association between overweight and death

According to our study, the point estimate of the relative risk for overweight was less than 1.0 but very close to 1.0 (P-value=0.74, CI=[0.93, 1.04]). The result suggests there is no evidence to support the association between overweight and mortality. One of the possible explanations is the effect of overweight on death is not immediate. Overweight people who were told they have a chronic disease might try to switch to a healthy diet and lifestyle. Overweight people are more likely to lose or control their weight than obese people during a long period of observation time. The risk of mortality should decrease if overweight persons succeed in returning to normal weight. Unfortunately our study did not account for the change of weight in subjects during the follow-up.
5.4 Limitation and uncertainty of the study

There were several limitations in this study. First, we only used the baseline measure of weight and height to calculate BMI under the assumption that enrolled subjects would not change their weight during the follow up to the year 1992. Second, the important potential confounding factor smoking status was based on self report in NHANES I. The third limitation was the time gap between the study population and the target population. There were two estimates needed to estimate PAR. They were the relative risk of death due to obesity and the proportion of deaths which fell into each BMI category. The relative risk derived from NHANES I and NHEFS represents the relative risk of death associated with obesity for the United States population in the middle point of the follow up time. When we applied this older estimate of the relative risk to the contemporary target population, we assumed the relative risk was the same for the population. Fourth, our study did not take the sample design into account. The series of NHANES are based on a complex, multi-stage probability sample design. Several aspects of the NHANES design should be taken into account in the analysis of the data, including the sample weights and the complex survey design. Because of the limitation in the SAS, survey designs can not be considered at the same time when building Cox’s proportional hazard model. The ignoring of the survey design information may not applicable in statistical analysis.
5.5 Conclusions

The goal of our study was to estimate annual extra deaths attributable to obesity among the U.S. population. The relative risks associated with overweight and obesity was derived from Cox's proportional hazards model. Age was used as time scale with left truncation in the models to control the delayed entry. The covariates included gender, race, education and smoking status. The annual death number was from the 2003 National Vital Statistics Report. Our study indicated that obesity and severe obesity are associated with increased risk of death. We estimated approximately 270,000 extra deaths in the United States attributable to obesity and severe obesity. However, overweight is not associated with increased mortality risk. The relative risk of underweight is as high as severe obesity.

5.6 Recommendations for Future Study

In our study we assumed that the BMI will not change overtime. But it is common that one's weight will change over time. It would be better to consider BMI as a time dependent variable. Our study included demographic confounders (gender, education and race) and one behavior confounder (smoking status) into the model but no clinical confounders. We suggest that future research should examine clinical confounders. We also suggest using an updated data set which can represent the recent U.S. population to estimate the relative risk of death. We also suggest taking into account sample designs in survival analysis.
LIST OF REFERENCES


15. *The Surgeon General’s call to action to prevent and decrease overweight and obesity, 2001*.


Figure A.1: The univariate Kaplan-Meier survival curves for BMI
**Figure A.2:** The univariate Kaplan-Meier survival curves for gender
Figure A.3: The univariate Kaplan-Meier survival curves for smoking
Figure A.4: The univariate Kaplan-Meier survival curves for education
Figure A.5: The univariate Kaplan-Meier survival curves for race
Figure A.6: The cumulative hazard curves for BMI
Figure A.7: The cumulative hazard curves for gender
Figure A.8: The cumulative hazard curves for race
Figure A.9: The cumulative hazard curves for smoking
Figure A.10: The cumulative hazard curves for education
Figure A.11: Cumulative hazard for all-cause mortality as a function of age in NHANES I and 1992 NHEFS