Annals of Epidemiology 25 (2015) 907–911

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Effects of categorization and self-report bias on estimates of the association between obesity and mortality

Samuel H. Preston PhD [a], Ezra Fishman MA [a], Andrew Stokes PhD [b,*]

[a] Population Studies Center, University of Pennsylvania, Philadelphia
[b] Department of Global Health, Boston University, Boston, MA

ARTICLE INFO

Article history:
Received 5 June 2015
Accepted 25 July 2015
Available online 19 August 2015

Keywords:
Body mass index
Obesity
Body weight
Mortality
Epidemiologic methods

ABSTRACT

Purpose: The health consequences of obesity are often assessed using categorical, self-reported data on body mass index (BMI). This article investigates the combined effects of categorization and self-report bias on the estimated association between obesity and mortality.
Methods: We used the National Health and Nutrition Examination Survey (1988–2008) linked to death records through 2011. Cox models and age-standardized death rates were used to evaluate the effects of categorization and self-report bias on the mortality risks and percent of deaths attributable to obesity.
Results: Despite a correlation between measured and self-reported BMI of 0.96, self-reports miscategorized 20% of adults. Hazard ratios using self-reports were overstated for the obese 1 (BMI, 30–35 kg/m^2) and obese 2 (BMI $\geq$ 35 kg/m^2) categories. The bias was much smaller using a continuous measure of BMI. In contrast, the percent of deaths attributable to excess weight was lower using self-reported versus measured data because self-reports led to systematic downward bias in the BMI distribution.
Conclusions: Categorization of BMI and self-report bias combine to produce substantial error in the estimated hazard ratios and percent of deaths attributable to obesity. Future studies should use caution when estimating the association between obesity and mortality using categorical self-reported data.

In 2011–2012, an estimated 34.9% of American adults aged 20 years and more were obese based on measured height and weight collected in the National Health and Nutrition Examination Survey (NHANES) [1]. In contrast, recent estimates from the Behavioral Risk Factor Surveillance System and Gallup Poll, based on self-reported height and weight, place the national obesity rate in adults ages 18 years and more at 29.4% and 27.7%, respectively [2,3].

The lower rates reported in these surveys are in part attributable to systematic misreporting of height and weight, which has been documented in numerous studies [4–7]. A comprehensive review of the accuracy of self-reported weight and height concluded that both sexes have a tendency to overreport their height and underreport their weight, with body mass index (BMI) underestimated as a result [8].

Misreporting may create especially large analytic errors in epidemiologic studies when combined with categorization of BMI. Misreporting that incorrectly transfers someone across one of the conventional five-unit BMI categories connotes large changes in an individual's implied weight. For someone 5′10″, such a transfer would imply a movement of 35 pounds from the midpoint of the actual class to the midpoint of the adjacent reported class, even though weight might be misreported by as little as a pound.

Despite the biases in self-reports, many of the largest and most influential studies of the mortality consequences of obesity are based on self-reported weight and height [9–13]. Keith et al. [14] explored the differences in estimated relations between BMI and mortality when categorical self-reported data are used rather than measured data. The authors described the biases from using self-reports as "complicated and inconsistent." We build on that study by taking advantage of a more recent data set, by relating patterns of misreporting to biases in estimated relative risks, by using age-standardized death rates in addition to hazard ratios, and by introducing a continuous measure of BMI. The continuous measure enables us to demonstrate how misreporting and categorization combine to produce bias. Finally, we also examine the effect of misreporting and categorization on the fraction of deaths attributable to obesity.

Methods

NHANES is a series of nationally representative samples of the noninstitutionalized US population conducted by the National

* Corresponding author. Department of Global Health, Boston University, 801 Massachusetts Ave., CT362, Boston, MA 02118. Tel.: +1-7179652145; fax: +1-6174141261.
E-mail address: acstokes@bu.edu (A. Stokes).

http://dx.doi.org/10.1016/j.annepidem.2015.07.012

Center for Health Statistics. The survey includes an examination component in which extensive medical data, including height and weight, are collected by trained nurses in mobile clinics or in home visits [15]. Survey design and sample characteristics are described in detail elsewhere [16,17]. We used data from NHANES III (conducted 1988 to 1994) and NHANES Continuous, 1999–2008, which are linked to National Death Index death records through 2011 [18]. NHANES data from 2009 to 2010 were not incorporated because we delay analysis of mortality for three years after baseline to reduce the effects of reverse causality [19]. We used an age range of 40 to 75 years and terminated follow-up when individuals reached their 80th birthday because of declines with age in the mortality risks of obesity [20,21]. Younger individuals were excluded because of the small number of deaths in the sample at these ages.

There were 20,580 nonpregnant individuals aged 40 to 74 years at baseline. We excluded subjects who lacked measured height or weight ($n = 323$) or self-reported height or weight ($n = 781$). We further excluded individuals with very low (BMI <15 kg/m^2) or very high (BMI ≥75 kg/m^2) values for measured or self-reported BMI ($n = 24$). Subjects missing data on smoking status ($n = 9$) or education ($n = 53$) were also excluded. To reduce the possibility that illness was driving weight loss (reverse causality), we excluded individuals who reported a diagnosis of the smoking-related health conditions mentioned below: emphysema ($n = 527$) or cancers of the bladder, esophagus, kidney, larynx, lung, mouth/tongue/lip, or pancreas ($n = 117$) [19]. We also excluded persons who had lost significant weight in the 10-year period before baseline based on self-reported weight, using a cut-off of the 75th percentile of weight loss among weight losers, or 4.0 BMI units ($n = 1046$). Finally, after incorporating mortality follow-up data, we excluded 15 subjects who lacked a date of death or censoring information.

Given our focus on the relation between excess weight and mortality, individuals with measured BMI less than 18.5 were excluded from survival models using measured data ($n = 204$) and individuals with self-reported BMI less than 18.5 were excluded from survival models using self-reported data ($n = 154$). For models using self-reports, there were 16,970 subjects with 1979 deaths in 128,616 person-years of observation. For models using measured BMI, there were 16,928 subjects with 1974 deaths in 128,333 person-years of observation. Median follow-up time was 8.4 years.

We used Cox proportional hazards models to estimate the relation between BMI and mortality, using age as exposure time and compared coefficients estimated using self-reported data to those estimated using measured data. The continuous model used a BMI variable of (BMI 25.0); individuals with BMI values less than 25 were excluded because the focus was on the incremental risks associated with overweight and obesity. This approach was also used in the two largest analyses of pooled data [10,22]. The Prospective Studies Collaboration pooled data on approximately 900,000 individuals in 57 studies and demonstrated that death rates rose exponentially (i.e., linearly on a log scale) above a BMI of 25 [22]. Such a pattern lends itself to efficient estimation of the mortality hazards associated with overweight and obesity by using linear BMI in a hazards model.

In models using discrete categories of BMI, we grouped people into the following categories recommended by the World Health Organization: normal weight: 18.5 ≤ BMI < 25, overweight: 25 ≤ BMI < 30; obese 1: 30 ≤ BMI < 35; and obese 2: BMI ≥ 35. We conducted two sensitivity analyses of the model. In one, the sample was restricted to nonsmokers to assess the impact on estimates of an unusual pattern of BMI misreporting among smokers. The second dropped the 3-year exclusion of observations between survey and mortality.

Covariates in the survival models were gender, race and/or ethnicity (non-Hispanic black, Hispanic, white, and other), smoking status (current, former, and never), and educational attainment (less than high school, high school, and more than high school). To supplement the analysis of the relative hazards of death associated with excess BMI, we estimated age-standardized death rates by BMI category, standardizing in 5-year age groups to the age distribution of the 2000 US census population.

To assess the fraction of deaths attributable to overweight and obesity, we used the following formula for the Population Attributable Fraction [23]:

$$PAF = \sum_{i=0}^{k} pd_i \left(\frac{HR_i - 1}{HR_i} \right) \quad (1)$$

where pd_i is the proportion of decedents in BMI category i and HR_i is the hazard ratio with respect to mortality for an individual in category i. Individuals in the normal and underweight categories were assigned a hazard ratio of 1.0. To implement this formula in our continuous models, we used 0.5-unit wide BMI categories and associated each category with the predicted hazard ratio at the midpoint of the category.

All analyses were weighted to the US civilian noninstitutionalized population and adjusted for the complex survey design of NHANES using the *svy* routine in Stata 13 (StataCorp, TX). Variances were estimated using Taylor Series Linearization, as recommended by NHANES [15].

Results

The correlation between measured BMI and BMI based on self-reported weight and height was 0.96 for each sex. Despite this high correlation, 19.5% were reported to be in a different 5-unit BMI category from their correct category, 18.5% of men and 20.3% of women. Table 1 shows a cross-classification of BMI by reported and measured values. Among those who belong in obese 2, 27.0% were reported to be in a lower category. Among those measured to be in obese 1, 32.0% were reported in the wrong category, 28.9% in a lower category and 3.1% in a higher category. Except for the very small underweight category, obese 1 had the lowest percentage of people who belonged in the class who correctly reported themselves to be in the class as well as the lowest percentage of people who reported themselves in the class who belonged in the class.

Table 2 shows the mean of self-reported and measured BMI values within various self-reported BMI categories. In every reporting category for both sexes, the mean measured BMI was higher than the mean self-reported BMI, consistent with a general tendency to underreport BMI. The difference between the two means was greater in BMI classes above normal, which is likely to impart an upward bias into estimates of the relative risk of death in higher self-reported BMI categories.

Table 1
Joint distribution of population by self-reported and measured BMI

Measured class	Self-reported class					
	Under (%)	Normal (%)	Over (%)	Obese 1 (%)	Obese 2 (%)	Total (%)
Under	0.74	0.54	0.00	0.02	0.00	1.29
Normal	0.33	28.08	2.31	0.02	0.00	30.74
Over	0.01	5.24	29.61	1.50	0.03	36.39
Obese 1	0.00	0.09	5.43	13.22	0.61	19.35
Obese 2	0.00	0.01	0.21	3.13	8.87	12.22
Total	1.08	33.94	37.56	17.90	9.52	100.00

Definitions of classes: BMI: under, BMI <18.5 kg/m^2; normal: 18.5 ≤ BMI<25; over: 25 ≤ BMI < 30; obese 1: 30≤ BMI < 35; and obese 2: BMI ≥ 35. "Self-reported BMI" is BMI calculated based on self-reported weight and height.