TOPICS IN CONSUMER HEALTH DECISIONS

By

LYUDMYLA KOMPANIYETS

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Economic Sciences

DECEMBER 2017

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of LYUDMYLA KOMPANIYETS find it satisfactory and recommend that it be accepted.

_____

Robert Rosenman, Ph.D., Chair

_____

Ron Mittelhammer, Ph.D.

_____

Mark Gibson, Ph.D.

TOPICS IN CONSUMER HEALTH DECISIONS

Abstract

by Lyudmyla Kompaniyets, Ph.D.
Washington State University
December 2017

Chair: Robert Rosenman

This dissertation discusses three topics in health economics focusing on consumer health decisions, as well as the application of econometric techniques to misclassified self-reported data. The first chapter attempts to estimate the relationship between smoking and morbidity in the United States. The second chapter focuses on identifying the relationship between participation in the Supplemental Nutrition Assistance Program and BMI, when the binary independent variable is endogenously misclassified. The third chapter presents a new look at the problem of asymmetric information in the life insurance market. We analyze the relation between mortality risk and individual term life insurance ownership, in an attempt to understand whether individuals with a higher risk of mortality are also more likely to own individual term insurance.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

Dedication

To my Mom and Dad

CHAPTER ONE: SMOKING AND MORBIDITY

ABSTRACT

Smoking is a direct cause of many diseases. As a result, many public policy initiatives, including cessation programs, excise taxes, and informational campaigns, target smoking. While there is extensive literature relating these initiatives to smoking-related mortality, evidence on whether smoking-related morbidity responds similarly is lacking. Our paper uses data on the number of smokers and nine smoking-related diseases in each of the U.S. states in the years 2005-2010 to measure the marginal effects of these diseases to the number of smokers. In all cases, we find that the marginal effects are positive but less than one. This means that reducing the number of smokers by 1% has a less than 1% effect on the reduction in morbidity. We find that larynx and pharynx cancers are the most sensitive to the reduction in smoking incidence among the cancer diseases. Among the non-cancer diseases, heart attack, coronary disease and stroke proved to have more sensitivity to the reduction in smoking incidence, compared to other diseases. There results have implications for the effectiveness of smoking cessation programs and laws in reducing the incidence of smoking-related diseases and costs associated with them.

INTRODUCTION

Smoking is a primary cause of numerous diseases (bronchitis, cardio-vascular disorders, emphysemas, and malignant tumors in the lung, trachea, larynx, esophagus, and other organs) and has been shown to increase the use of medical and health services. Hence, smoking is the direct cause of significant health care costs. It is estimated that from 2000–2004, cigarette

smoking cost more than $193 billion: $96 billion in direct health care expenditures and $97 billion in lost productivity (CDC, 2009).

Public policies, seeing smoking as a cause of preventable disease, aim to reduce smoking to help control health care costs and to improve health. These policies include limitations on advertisements and on places where smoking is allowed, smoking cessation programs, informational campaigns against smoking, and fiscal measures, such as raising excise taxes.

The question addressed in this paper is how disease incidences respond to reduction in smoking. Many recent studies have looked at how smoking reduction affects mortality. The issue that has not been explored yet is how it affects morbidity. Our study attempts to answer this question by estimating the marginal effect of smoking on morbidity. When coupled with estimated effectiveness of smoking-cessation policies, this research provides some of the information needed for informed policymaking.

There is a vast literature on the association of smoking and health-related issues. Cigarette smoking is shown to cause respiratory disorders, cardiovascular hazards, and cancers (Sherman 1991, Hummer, Nam & Rogers 1998, Bartal 2001, Sasco, Secretan & Straif 2004). In particular, studies find association between smoking and over twelve types of cancer, including cancer of the lung, oral cavity, pharynx, larynx, urinary bladder etc (Carborne 1992, Sasco, Secretan & Straif 2004, Gandini et al. 2008). In addition it is well documented that cigarette consumption contributes to coronary heart disease, emphysema and diabetes (Haire-Joshu, Glasgow, & Tibbs 1999, Pope et al. 2004, Willi et al. 2007, Erhardt 2009). Smoking in women is linked to child mortality and morbidity (Rantakallio 1978).Because smoking is the leading cause of preventable death in the U.S. and abroad (Wald & Hackshaw 1996), most research has studied the impact of smoking on mortality, including estimates of the elasticity of mortality with respect

to smoking (Yuan et al.1996; Hummer, Nam & Rogers 1998; Pokorski 2000; Rogers et al. 2005; Ho & Elo 2013). Only a few studies explore the reduction of mortality and morbidity that results from smoking cessation programs (Elixhauser 1990; Moore 1995; Escario & Molina 2004), and most do not estimate the elasticity of morbidity with respect to smoking. Of those studies that explore the effects of smoking on morbidity, most focus on small clinical trials (see, for example, Yuan, et. al. 1996). We expand on their analysis using population-level data.

If smoking is the primary cause of a disease and the disease incidence randomly strikes smokers, or if the predilection to smoke (and quit) is unrelated to predilection towards smoking-induced diseases, a reduction in smoking should result in proportionate decreases in mortality and morbidity. However, studies have shown that the elasticity of mortality with respect to smoking is less than one. (Evans & Farrelly 1998, Escario & Molina 2004), possibly due to compensating behavior or residual effects on health that persist even after smoking has ceased. Another explanation is that smoking, while contributing to disease incidence, is not the only cause of the disease – so the incidence decrement is not proportional to smoking cessation. What is heretofore unknown is whether morbidity responds similarly. If morbidity, too, is less than responsive to smoking incidence we add to the puzzle, and it informs better that the anti-smoking policies which target existing smokers may have less value than alternative policies designed to prevent smoking from ever starting.

The remainder of the paper is organized as follows: Section 2 discusses the research methodology. Section 3 describes the data, while section 4 presents the empirical results. In Section 5, we present conclusions and policy implications.

METHODOLOGY

This study measures the relationship between the rate of smoking and rates of several smoking-related diseases. We separately estimate how the rate of smoking affects the rate of each disease. We find the morbidity coefficients, which are marginal effects of smoking incidence on of the disease incidence.

The dependent variable in our model (rate of disease) is a proportion. Because of its flexibility and being constrained between 0 and 1, we assume a beta distribution in our empirical analysis, making any linear model unsuitable for estimation (Kieschnick & McCullough 2003, Paolino 2001). Hence we use the General Estimating Equations model, which is a semiparametric model (Papke & Wooldridge 2008).

We estimate separately nine equations for each of the smoking-related diseases with data at the state level. Let $Y = (Y_{it})$ be a 255×1 vector of observations of the disease rates (for state i at year t). Let $X = (X_{it})$ be a set of explanatory variables, and β is a vector of unknown regression coefficients.

To relate the marginal response $\mu_{it} = E(y_{it})$ to a linear combination of the covariates

$$g(\mu_{it}) = x_{it}^T \beta$$

where the response and the explanatory variables are connected through the logit link function $g(\cdot)$.

Let the variance of y_{it} be a function of the mean

$$V(y_{it}) = \upsilon(\mu_{it})\phi$$

where ϕ is a possible unknown scale parameter,

$\upsilon(\cdot)$ is a known variance function.

Since the response and explanatory variables are linked through a logit link function:

$$g(\mu_{it}) = \log[\mu_{it}/(1-\mu_{it})]$$
$$\upsilon(\mu_{it}) = \mu_{it}/(1-\mu_{it})$$
$$\phi = 1$$

The functional form is given by:

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\phi]$$

where $\theta_{it} = h(\eta_{it}),\ \eta_{\text{it}} = x_{it}\beta$. Then the first two moments of y_{it} are given by

$$E(y_{it}) = a'(\theta_{it}),\ \text{var}(\text{y}_{\text{it}}) = a''(\theta_{it})/\phi$$

Assuming an n by n working correlation matrix $R(\alpha)$,

$$\Delta_i = diag(d\theta_{i,j}/d\eta_{i,j}) \quad \text{n} \times \text{n matrix}$$
$$A_i = diag\{a''(\theta_{i,j})\} \quad \text{n} \times \text{n matrix}$$
$$\text{S}_\text{i} = y_i a'(\theta_i) \quad \text{n} \times 1 \text{ matrix}$$
$$D_\text{i} = A_i \Delta_i \text{X}_\text{i} \quad \text{n} \times p \text{ matrix}$$
$$V_\text{i} = A_i^{1/2} R(\alpha) A_i^{1/2} \quad \text{n} \times \text{n matrix}$$

So the "model-based" GEE variance becomes:

$$V(\hat{\beta}) = (\sum_{i=1}^{m} D_i^T \hat{V}_i^{-1} D_i)^{-1}$$

We use the robust or "empirical" variance is:

$$V(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1}$$
$$M_0 = \sum_i^m D_i' \hat{V}_i D_i$$
$$M_1 = \sum_i^m D_i' \hat{V}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{V}_i^{-1} D_i$$

DATA

Our data come from BRFSS - The Behavioral Risk Factor Surveillance System (BRFSS 2014), an ongoing survey of health behaviors coordinated by the US government and supported by different state surveys (CDC BRFSS 2014). The data contain 306 observations collected for each state of the U.S. and the District Columbia in 2005-2010. Cancer data are obtained from United States Cancer Statistics gathered by National Program of Cancer Registries (NPCR 2014). These data are rates are per 100,000 persons and age-adjusted to the 2000 U.S. standard population (19 age groups - Census P25-1130). Data on smoking-related diseases (other than cancer), obesity, alcohol consumption and demographic characteristics are obtained from BRFSS Prevalence and Trends Dataset. These data are percentages of respondents. In both cases sample only contains individuals who are older than 18 years of age.

We are interested in how the rate of smoking affects morbidity (the rate of smoking-related diseases). "Morbidity" takes into account a number of diseases. A morbidity index could be calculated in several ways: using weighted sum, average prevalence etc. In the case of smoking-related morbidity, calculating such index presents many complications. One of the most prevalent is that some people have more than one disease, so a weighted sum would over count the number of ill people. On the other hand, average prevalence may severely under count the number of ill people. We finesse building a morbidity index by studying major disease categories: lung cancer, larynx cancer, oral cavity and pharynx cancer, urinary cancer, asthma, heart attack, coronary disease, stroke and diabetes. Rates of these diseases constitute the dependent variables in our estimation. Descriptions of these variables are presented in Table 1.1.

Independent variables include rates of smokers, drinkers and obesity in each state, as well as standard demographic variables (Table 1.2).

On average, states have a population that is 49 percent male and 33 percent with college education. The average state has 19% current smokers and 25 % former smokers. The data indicate that on average 48 percent of a state's population earns over $50,000 in annual income, about 35 percent of the population is overweight, and 26 percent suffer from obesity. On average, 5 percent of a state's population are heavy drinkers, and 15 percent are binge drinkers.

Additional details of the sample are presented in Table 1.3. Cancer data were available for only 294 out of 306 observations because cancer data in select states and years are absent from the United States Cancer Statistics dataset. The missing observations were replaced by the mean of the rest of the sample. Although the missing observations can be approximated using various statistical procedures, a conservative approach is to replace them with the sample average, especially when the proportion of missing data is relatively small.

Comparisons of sample descriptive statistics to the 2010 American Community Survey (US Census Bureau, 2010), is reported in Table 1.4. Results suggest that our sample is well represented in terms of gender. Our age data are not representative of the US population because the category under 18 is not represented. Although 24% of the US population are under 18, we have 0% in our sample; hence the percentages of other age categories are larger in the sample than in population although they are representative of the over-18 population. Individuals with higher educational attainment are overrepresented in the BRFSS data and states with lowest income seem to be somewhat underrepresented while individuals with medium income (15,000 – 49,999) are slightly overrepresented in the sample. Hence, our state averages reflect these facts.

RESULTS

The results of the estimation are presented in Table 1.5. In the preliminary estimation, we included variable that showed the percentage of former smokers per state in the model. This is because we wanted to compare the possible effects of current smoking and past smoking on health. Separating these two effects would give us an opportunity to compare the potential effectiveness of smoking cessation programs to smoking prevention programs. However, we had to omit the "former smoker" variable from the model due to collinearity issues.

As expected, we find that number of current smokers positively and significantly correlates with larynx cancer, pharynx and oral cavity cancer, heart attack, coronary disease, and stroke.

All morbidity coefficients with respect to smoking are less than 1. It means that states with 1% fewer smokers have less than 1% fewer disease incidence, on average. In the case of cancer, the morbidity coefficients with respect to the number of smokers are very close to zero, even when they are significant at conventional p-values. Although the evidence is that smoking strongly contributes to the likelihood of getting these diseases, it has only a very small impact on the rates of these diseases. This small impact may be explained by the low average cancer rates and by the high death rate associated with cancer, leaving few people sick with the disease.

The coefficient of morbidity for larynx cancer and pharynx cancer is 0.0000979 and 0.0001273, respectively. This means that states with 1% fewer smokers have 0. 0000979% less larynx cancer and 0.0001273% less pharynx cancer incidence. In the case of heart attack, the morbidity coefficient with respect to the number of current smokers is significant and positive. States with 1% fewer smokers have 0.08% smaller asthma rates on average. In the case of

coronary disease and stroke, the coefficients of morbidity are positive and similar in magnitude (0.06 for coronary disease, and 0.03 for stroke).

In order to understand the sensitivity of the disease rates to the number of smokers, we calculate elasticities at the mean. (Table 1.6). We find that the elasticities of larynx and pharynx cancer are higher (0.488 and 0.225, respectively), while the elasticities of lung and urinary cancer are close to zero (0.087 and 0.091, respectively). Among non-cancer diseases, elasticity of heart attack is 0.278, elasticity of angina is 0.25, and elasticity of stroke is 0.25. Elasticities of asthma and diabetes are close to zero (0.008 and 0.045, respectively). This implies that cancer incidence may be more sensitive to the reduction of smoking in the population in the case of pharynx and larynx cancer, as well as heart attack, coronary disease (angina) and stroke.

We find that overall the prevalence of smoking is a small marginal contributor to the incidence of these diseases. Hence smoking cessation programs will reduce the incidence of these diseases, but since the marginal effect is small, the effect on incidence is also small. On the other hand, small changes in incidence may lead to big savings. In 2020 predicted 18.1 million cancer survivors, with cancer care costing $157 billion, or about $8700 per case (Mariotto et al 2011). If by smoking cessation lowers all cancers by 0.5%, that would save $.758 billion. These are only savings from reduction in cancer incidence, while the total health care costs saved by the reduction in smoking incidence will be much larger.

Other lifestyle and behavior variables have impact on morbidity as well. Exercising has a negative correlation with larynx cancer, urinary cancer and diabetes incidence. Heavy drinking has positive and significant correlation with lung cancer incidence (0.00767), coronary disease incidence (0.13), and stroke incidence (0.05), while binge drinking has surprising negative correlation with lung cancer, urinary cancer, heart attack, coronary disease, stroke and diabetes

incidence. We find significant positive correlation between being overweight and heart attack, coronary disease, and stroke incidence. Obesity is positively correlated with larynx, pharynx cancer, heart attack, coronary disease, stroke, and diabetes incidence.

States with higher male population have lower cancer incidence. This result holds for lung cancer, larynx cancer, coronary disease, and diabetes.

Table 1.5 finds elasticities of morbidity to the percentage of smokers, calculated at the mean. We find that elasticities of cancer diseases are closer in magnitude to the elasticities of non-cancer diseases. In fact, larynx cancer has the highest elasticity (0.48), and pharynx cancer is elasticity is lower (0.2). Among non-cancer diseases, heart attack has the elasticity of 0.38, and the elasticities of stroke and coronary disease are both around 0.25. In all cases, we see that the elasticities are not greater than zero. The implication of this finding is that the smoking cessation programs may have less value than the smoking prevention programs in reducing smoking-related morbidity and health care costs.

The obtained coefficients of morbidity can be useful in being able to weigh the cost-effectiveness of smoking cessation programs and smoking prevention programs. Smoking cessation may have positive economic effect on the reduction in larynx cancer, pharynx cancer, heart attack, stroke and coronary disease.

The coefficients of all diseases to smoking are positive, but less than one. This means that reducing the number of smokers by 1% may cause less than 1% reduction in disease incidence. The coefficients of cancer to smoking are positive, but very close to zero. However, to observe the total effect of smoking on cancer rates, the obtained morbidity coefficients should be used in combination with mortality coefficients. This is due to higher mortality associated with cancers compared to other diseases in question. In the case of heart attack, stroke, and coronary disease,

smoking cessation and smoking prevention may be effective. Specifically, smoking cessation may be more effective than smoking prevention in reducing asthma, heart attack, stroke and coronary disease incidence.

CONCLUSIONS

This paper attempts to find coefficients of smoking-related diseases to the number of current and former smokers. Smoking is the leading cause of preventable morbidity and mortality in the United States. In this study, we focus on the relation of smoking and morbidity. One reason for this is that not all smoking-related diseases are fatal. While decreasing mortality, reduction in smoking may increase morbidity and healthcare costs, since people live longer. Another reason is that smokers may exhibit compensating behavior – they may substitute smoking risk with other types of risk. In this case the marginal effects of healthcare costs with respect to smoking is less than one, which may make it not a good place to spend policy money.

In this paper we choose nine diseases, which are known to be caused by smoking. Then we use generalized estimating equations model to estimate the coefficient of each disease with respect to the number of smokers. We find that the magnitude of cigarette consumption effect on morbidity is ambiguous but smaller than one. In most cases smoking rate is positively correlated with the rates of smoking-related diseases.

Magnitude of morbidity coefficients varies by disease type. Coefficients are positive and significant for states with current smokers in the case of larynx and pharynx cancer, heart attack, stroke, and coronary disease.

In all cases, cancer coefficients are very close to zero. However, due to the high mortality risk of cancer diseases, the obtained morbidity coefficients should be used in combination with

mortality elasticities, to see the overall effect of smoking reduction. In the cases of non-cancer diseases, coefficient, while being less than one, is still more economically significant. The policy implication of these results is that policies targeting smoking may be more effective in preventing heart attack, stroke, and coronary disease, rather than cancer.

We find elasticities of disease rates to the number of smokers, and find that larynx and pharynx cancer are, on average, more sensitive to the reduction in the number of smokers than the other types of cancer. We also show that heart attack, stroke and coronary disease are more sensitive than other non-cancer diseases.

The main result of this study is that morbidity coefficient with respect to smoking in all cases is less than one. This result is similar to that of smoking-related mortality, which was found to be less than one. These findings create important implications for policy-makers, as they may be used to demonstrate that anti-smoking policies that target existing smokers may have less value than smoking prevention policies.

In our further research we intend to compare the costs of smoking cessation programs to the cost of curative care of smoking-related diseases. We can use coefficients obtained from this paper to estimate the morbidity elasticities and cost effectiveness of prevention or cessation programs.

REFERENCES

Barendregt, J. J., Bonneux, L., & van der Maas, P. J. (1997). The health care costs of smoking. New England Journal of Medicine, 337(15), 1052-1057.

Bartal M. (2001) Health effects of tobacco use and exposure. Monaldi Arch Chest Dis. 2001 Dec;56(6):545-54.

Carbone D. (1992). Smoking and cancer. The American Journal of Medicine, Volume 93, Issue 1, Supplement 1, 15 July 1992, Pages S13-S17

CDC – Behavioral Risk Surveillance System – Homepage. http://www.cdc.gov/brfss/. [accessed 2014 September 10]

CDC – Smoking-Attributable Mortality, Years of Potential Life Lost, and Productivity Losses—United States, 2000–2004. Morbidity and Mortality Weekly Report 2008;57(45):1226–8 [accessed 2014 September 25].

Elixhauser A. (1990). The Costs of Smoking and the Cost Effectiveness of Smoking-Cessation Programs. Journal of Public Health Policy, Vol. 11, No. 2 (Summer, 1990), pp. 218-237

Erhardt L. (2009). Cigarette smoking: An undertreated risk factor for cardiovascular disease, Atherosclerosis, Volume 205, Issue 1, July 2009, 23-32

Escario, J. J., & Molina*, J. A. (2004). Will a special tax on tobacco reduce lung cancer mortality? Evidence for EU countries. Applied Economics, 36(15), 1717-1722

Evans, W. N., & Farrelly, M. C. (1998). The compensating behavior of smokers: taxes, tar, and nicotine. The Rand Journal of Economics, 578-595

Gandini S, Botteri E, Iodice S, Boniol M, Lowenfels AB, Maisonneuve P, Boyle P. (2008). Tobacco smoking and cancer: a meta-analysis. Int J Cancer. 2008 Jan 1;122(1):155-64

Haire-Joshu, D. E. B. R. A., Glasgow, R. E., & Tibbs, T. L. (1999). Smoking and diabetes. Diabetes care, 22(11), 1887-1898

Ho, J. Y., & Elo, I. T. (2013). The Contribution of Smoking to Black-White Differences in U.S. Mortality. Demography, 50(2), 545-568. doi:http://ntserver1.wsulibs.wsu.edu:2102/10.1007/s13524-012-0159-z

Hummer, R. A., Nam, C. B., & Rogers, R. G. (1998). Adult Mortality Differentials Associated with Cigarette Smoking in the USA. Population Research And Policy Review, 17(3), 285-304

Kieschnick, R. & McCullough, B.D. (2003). Regression analysis of variates observed on (0,1): percentages, proportions and fractions. Statistical Modelling, 3, 193–213

Lahiri K. and Song J.G. (2000) The effect of smoking on health using a sequential self-selection model. Health Econ. 2000 Sep;9 (6), 491-511

Mariotto, A. B., Yabroff, K. R., Shao, Y., Feuer, E. J., & Brown, M. L. (2011). Projections of the cost of cancer care in the United States: 2010–2020. Journal of the National Cancer Institute.

Max W. (2001) The Financial Impact of Smoking on Health-related Costs: A Review of the Literature. American Journal of Health Promotion: May/June 2001, Vol. 15, No. 5, 321-331

Moore, M. J. (1995). Death and tobacco taxes (No. w5153). National Bureau of Economic Research.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. Political Analysis, 9, 325–346

Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. Journal of Econometrics, 145(1), 121-133

Pokorski, R. J. (2000). Excess Mortality in Asia Associated with Cigarette Smoking. North American Actuarial Journal, 4(2), 101-113

Pope CA 3rd, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ. (2004). Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. Circulation. 2004; 109: 71–77

Rantakallio, P. (1978), Relationship of Maternal Smoking to Morvidity and Mortality of the Child up to the Age of Five. Acta Paediatrica, 67: 621–631

Rogers, R. G., Hummer, R. A., Krueger, P. M., & Pampel, F. C. (2005). Mortality Attributable to Cigarette Smoking in the United States. Population And Development Review, 31(2), 259-292

Sasco AJ, Secretan MB, Straif K. (2004) Tobacco smoking and cancer: a brief review of recent epidemiological evidence, Lung Cancer. 2004 Aug;45 Suppl 2:S3-9

SEER Stat Fact Sheets: Larynx Cancer. National Cancer Institute – Homepage: http://seer.cancer.gov/statfacts/html/laryn.html [accessed on October 15, 2014]

Sherman CB. (1991) Health effects of cigarette smoking. Clin Chest Med. 1991 Dec;12(4):643-58

Wald, N. J., & Hackshaw, A. K. (1996). Cigarette smoking: an epidemiological overview. British medical bulletin, 52(1), 3-11

Warner, K. E., Hodgson, T. A., & Carroll, C. E. (1999). Medical costs of smoking in the United States: estimates, their validity, and their implications. Tobacco Control, 8(3), 290-300

Willi, C., Bodenmann, P., Ghali, W. A., Faris, P. D., & Cornuz, J. (2007). Active smoking and the risk of type 2 diabetes. JAMA: the journal of the American Medical Association, 298(22), 2654-2664

Yuan J, Ross RK, Wang X, Gao Y, Henderson BE, Yu MC (1996). Morbidity and Mortality in Relation to Cigarette Smoking in Shanghai, China: A Prospective Male Cohort Study. JAMA. 1996;275(21):1646-1650. doi:10.1001/jama.1996.03530450036029

Table 1.1 Description of Dependent Variables

Dependent Variable (units- percent)	**Description**
Lung Cancer	Rate of lung cancer in each state
Larynx Cancer	Rate of larynx cancer in each state
Pharynx and Oral Cavity Cancer	Rate of pharynx and oral cavity cancer in each state
Urinary Cancer	Rate of urinary cancer in each state
Asthma	Rate of individuals who answered "yes" to the question: "Do you currently have asthma?"
Heart Attack	Rate of individuals who answered "yes" to the question: "Have you ever had a heart attack?"
Angina and Coronary Disease	Rate of individuals who answered "yes" to the question: "Have you ever had angina or coronary disease?"
Stroke	Rate of individuals who answered "yes" to the question: "Have you ever had a stroke?"
Diabetes	Rate of individuals who answered "yes" to the question: "Have you ever been told that you have diabetes?"

Table 1.2 Description of Independent Variables

Independent Variable (units – percent)	**Description**
Smoker	Percentage of respondents who are current smokers in each state
Male	Percentage of male population among responders in each state
Income	
less than 15000	Percentage of individuals with annual income less than $15,000
15000 to 24999	Percentage of individuals with annual income $15,000- $24,999
25000 to 34999	Percentage of individuals with annual income $25,000- $34,999
35000 to 49999	Percentage of individuals with annual income $35,000- $49,999
over 50000	Percentage of individuals with annual income more than $50,000
Age	
18 to 24	Percentage of individuals 18-24 years of age
25 to 34	Percentage of individuals 25-34 years of age
35 to 44	Percentage of individuals 35-44 years of age
45 to 54	Percentage of individuals 45-54 years of age
55 to 64	Percentage of individuals 55-64 years of age
over 65	Percentage of individuals over 65 years of age
College	Percentage of individuals with college education
Exercise	Percentage of individuals who exercised at least once in the past month
Heavy Drinking	Percentage of heavy drinkers (adult men having more than two drinks per day and adult women having more than one drink per day)
Binge Drinking	Percentage of binge drinkers (males having five or more drinks on one occasion, females having four or more drinks on one occasion)
Overweight	Percentage of individuals who are overweight (bmi 25.0 - 29.9)
Obese	Percentage of individuals who are obese (bmi 30.0 - 99.8)

Table 1.3 Descriptive statistics for the variables in the estimation (unit: percent)

	Count	Mean	Standard deviation	Min	Max
Lung cancer	294	0.067	0.01212	0.0255	0.102
Larynx cancer	295	0.0039	0.0097	0.0016	0.0064
Pharynx and oral cavity cancer	294	0.0110	0.00131	0.0072	0.0162
Urinary cancer	294	0.0214	0.00359	0.0123	0.0312
Asthma	306	8.56	1.10	5.9	11.1
Heart attack	306	4.25	0.88	1.9	7.7
Angina and coronary disease	306	4.28	0.92	2	8.3
Stroke	306	2.67	0.60	1.4	4.7
Diabetes	306	8.18	1.55	4.4	13.2
Current Smoker	306	19.45	3.40	9.1	28.7
Former Smoker	306	25.08	2.72	14.3	32.2
Male	306	48.71	0.95	46	52.2
Income					
less than 15000	306	9.00	2.78	4	19.7
15000 to 24999	306	15.49	3.30	7.6	27.4
25000 to 34999	306	11.69	2.14	7	19.4
35000 to 49999	306	15.83	2.14	9.4	21.5
over 50000	306	47.99	7.83	30.2	67.2
Age					
18 to 24	306	12.11	2.26	5	18.5
25 to 34	306	17.80	2.34	10.6	29.2
35 to 44	306	18.96	2.22	14.6	29.6
45 to 54	306	19.36	1.37	16.2	29.4
55 to 64	306	14.67	1.11	10.6	17.9
over 65	306	17.09	2.02	9.4	22.8
College	306	33.41	6.41	19.8	62.2
Exercise	306	76.00	4.16	52.7	85.8
Veggies	153	23.55	3.65	14.6	32.5
Heavy Drinking	306	5.03	1.19	1.9	8.2
Binge Drinking	306	15.10	3.15	6.6	24.3
Overweight	306	34.93	3.92	19	40.7
Obese	306	26.23	3.31	17.8	35.4

Table 1.4 Demographic Characteristics

Variable	U.S.A.	
	Sample	Population
Male (%)	48.75	49.20
Age distribution (%)		
Under 18	0	24.00
18-24	12.11	9.90
25-34	17.80	13.30
35-44	18.96	13.30
45-54	19.36	14.58
55-64	14.67	11.81
Over 64	17.09	13.00
College (%)	33.41	28..2
Income (%)		
Under 15,000	9.00	13.00
15,000-24,999	15.49	11.90
25,000-34,999	11.69	11.10
35,000-49,999	15.83	14.10
50,000 and up	47.99	49.60

Note 1: U.S. population statistics, such as age, education, gender, is based on the 1-year estimates of the 2010 American Community Survey (U.S. Census Bureau). Income statistics is based on 2009 Census data.

Note 2: U.S. statistics on education attainment is based on population over 25 y.o., while the sample contains respondents of 18 y.o. and older.

Table 1.5 Marginal effects (coefficients) of smoking and state characteristics on morbidity using GEE approach (Years 2005-2010)

	(1) Lung cancer	(2) Larynx cancer	(3) Pharynx and oral cavity cancer	(4) Urinary cancer	(5) Asthma	(6) Heart attack	(7) Angina and coronary disease	(8) Stroke	(9) Diabetes
Smoker	.0003318	.0000979***	.0001273**	.0000817	.0033629	.0825487***	.0555439**	.0346185***	.0189245
	(.0001864)	(.0000277)	(.0000399)	(.0000799)	(.0324451)	(.0147063)	(.0187407)	(.0102523)	(.025946)
Male	-.0031731*	-.00054***	-.000087	-.0003255	-.076543	-.0017656	-.1449461*	-.0468467	-.2625906*
	(.0013976)	(.0001013)	(.0001743)	(.0005904)	(.1401801)	(.0675171)	(.0677918)	(.0418563)	(.1064117)
Income									
under 15000	-.0002733	1.03e-06	-2.91e-07	-.0001045	.1346658***	.0230783	-.0257806	.0460858**	.0347279
	(.0003477)	(.0000308)	(.0000706)	(.0001361)	(.0398057)	(.0172151)	(.0309724)	(.017254)	(.0297794)
15000 to 24999	-.0001961	-.0000142	-.0000165	-.0001916	.0582912	.0059064	-.0601127	.0102018	-.0352974
	(.0003611)	(.000029)	(.0000833)	(.0001312)	(.0653596)	(.0265792)	(.0353801)	(.0209274)	(.04026)
35000 to 49999	-.0002557	.0000269	-.0000284	-.0002085	.1142594	-.0198876	-.089066**	.0155261	-.124356**
	(.0002809)	(.0000405)	(.0000828)	(.0001181)	(.0583395)	(.0205631)	(.0337545)	(.0286037)	(.0406072)
over 50000	.0000571	-2.60e-06	-.0000226	-.0001177	.0963103**	.0264664	-.0216736	.0105533	-.0068269
	(.0002404)	(.0000196)	(.0000641)	(.0001087)	(.0330146)	(.0158357)	(.0249686)	(.0161867)	(.0237239)
Age									
18to24	.000098	7.51e-06	.0000815	-.0001656*	.0257059	-.0045552	.0080655	-.0056379	-.07447***
	(.0000941)	(.0000174)	(.0000492)	(.0000741)	(.026821)	(.0151409)	(.014305)	(.0088988)	(.0199633)
25to34	-.0001439	-.0000431*	-7.91e-06	-.0001313	.0146463	-.0431835**	-.06405***	.0189554	.0105488
	(.0001334)	(.0000209)	(.0000275)	(.0000878)	(.0250497)	(.0147444)	(.0155606)	(.0121099)	(.0230132)
45to54	-2.36e-06	8.69e-06	-.000148***	.0000312	.0456837	-.0129785	-.0311893	.0301966	.0077484
	(.0001888)	(.000022)	(.000035)	(.0000907)	(.0266415)	(.0149693)	(.0232051)	(.0187499)	(.0231772)
55to64	-.0028262***	7.09e-06	.0003005*	-.0006566*	.2402291**	-.0250891	-.0769579	-.0329903	.1271714
	(.000738)	(.0000696)	(.0001301)	(.0002684)	(.0891124)	(.0350543)	(.0428585)	(.0298655)	(.0685114)

	(1) Lung cancer	(2) Larynx cancer	(3) Pharynx and oral cavity cancer	(4) Urinary cancer	(5) Asthma	(6) Heart attack	(7) Angina and coronary disease	(8) Stroke	(9) Diabetes
over65	-.0008296	-.0000934	.0000458	.000082	-.0140804	.1645933***	.142903***	.0850384***	.1162589*
	(.0007708)	(.0000589)	(.0001096)	(.0001682)	(.0821587)	(.0299658)	(.0391699)	(.015901)	(.0552094)
College	-.0002551	2.67e-06	.0000761	-.0000687	.0354298	-.048859***	-.04909***	.0025264	-.0305035
	(.0001398)	(.0000213)	(.0000393)	(.0000853)	(.0289083)	(.0107695)	(.0148135)	(.0135315)	(.0180626)
Exercise	-.000012	-.0000325***	-3.46e-06	-.000100**	.013737	-.0035848	-.008219	-.0120893	-.038498**
	(.0000864)	(8.79e-06)	(.0000265)	(.0000343)	(.0185925)	(.0084581)	(.0140475)	(.0064089)	(.0123822)
Heavy Drinkers	.000767*	-.0000558	.0000579	-.0000173	-.0039123	.058905	.132153***	.0533251*	.0644443
	(.0003393)	(.0000628)	(.0001071)	(.0001208)	(.0820297)	(.0343805)	(.0349515)	(.0264659)	(.0540826)
Binge Drinkers	-.0004968*	.0000193	-.0000447	.0001468*	.0311381	-.055819***	-.06235***	-.037653***	-.085181**
	(.0002064)	(.000021)	(.0000526)	(.0000749)	(.0313601)	(.0141595)	(.0153093)	(.0103024)	(.0295508)
Overweight	9.31e-06	-5.91e-06	.0000158	-.0000335	.0134288	.01852***	.030149***	.0107168*	-.0079443
	(.0000343)	(6.81e-06)	(.000018)	(.0000276)	(.0089237)	(.0053634)	(.005279)	(.0046619)	(.0075201)
Obese	.0000156	.0000442*	.0001114**	-.0001389	.0151086	.0412332**	.0468206**	.0609148***	.174837***
	(.0001526)	(.0000214)	(.0000353)	(.0000717)	(.0401697)	(.0133697)	(.0148091)	(.0145844)	(.0213678)
_cons									
N	294	295	294	294	306	306	306	306	306

standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.6 Elasticities of current smoking on morbidity

	(1) Lung cancer	(2) Larynx cancer	(3) Pharynx and oral cavity cancer	(4) Urinary cancer	(5) Asthma	(6) Heart attack	(7) Angina and coronary disease	(8) Stroke	(9) Diabetes
Smoker	0.08709	0.48824***	0.22509**	0.09089	0.00773	0.37778***	0.25241**	0.25218***	0.04494

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

SNAP PARTICIPATION AND OBESITY IN THE PRESENCE OF MISCLASSIFICATION BIAS

ABSTRACT

There is a lack of general consensus in the literature about the relationship between SNAP participation and obesity. While some studies claim that SNAP participants are more likely to be overweight and obese, others find no link between obesity and participation in SNAP. Additionally, the majority of the studies rely on self-reported SNAP participation, which is severely underreported in most national surveys. This paper investigates the link between Body mass index (BMI) and participation in the Supplemental Nutrition Assistance Program (SNAP) in the United States. We address the issue of underreported SNAP participation and its effect on the estimated link between BMI and SNAP. We use the data from the 2011-2012 and 2013-2014 waves of the National Health and Nutrition Survey (NHANES), in order to understand the link between BMI and SNAP and the effect of misclassification bias on this link. Our results show a positive relationship between BMI and SNAP participation in women, but not in men. We also find that a significant degree of misreporting is present in self-reported SNAP, and that correcting for misclassification reduces the magnitude of the marginal effect of SNAP on BMI. These findings have implications for the USDA policy decisions pertaining to limiting SNAP-eligible purchases.

INTRODUCTION

The United States government has introduced welfare programs to help different segments of the population. The Supplemental Nutrition Assistance Program (commonly known

as "food stamps") and other food support programs were introduced with the goals of increasing food security and reducing hunger. However, as SNAP allows participants wide latitude in the food they choose, there is considerable disagreement in the literature about whether SNAP contributes to obesity. Some studies find significant support linking food stamp participation and obesity (Townsend et al., 2001; Gibson, 2003; Chen, Yen & Eastwood, 2005; Wang & Beydoun, 2007; Robinson & Zheng, 2011; Baum, 2011; Condon et al., 2015). Other papers, including Vassilopoulos et al. (2011), Gundersen (2015) & Almada, McCarthy & Tchernis (2015), argue that there is no evidence of a causal link. In contrast, recent studies about WIC (Women, Infant and Children), a program that more closely limits choice to healthy foods, find no causal link between participation and obesity (Ver Ploeg, Mancino & Lin, 2007).

In its report on calorie intake of SNAP participants, the USDA finds that adult SNAP participants of both sexes consume more calories and are more likely to be obese than either their income-eligible or higher-income nonparticipant counterparts (USDA, 2015).

One potential explanation of why the literature linking SNAP participation to obesity is inconclusive is that SNAP participation is often misreported. There is considerable evidence that this is the case, with most studies finding substantial under-reporting of SNAP participation (Bollinger & David, 1997; Bollinger & David, 2001; Bollinger & David, 2005; Kreider, Pepper & Roy, 2012).[1] This misreporting, (usually termed misclassification given its binary measurement) of SNAP participation can bias (downward) and render inconsistent the parameter

[1] Misreporting is thought to be a problem in many major datasets, including the Survey of Income and Program Participation (Bollinger & David, 1997; Marquis & Moore, 2010), Community Population Survey (Cody & Tuttle 2002; Meyer et al., 2009), American Community Survey (Meyer et al., 2009), and National Health Interview Survey (Bitler, 2014). With SNAP in particular, misreporting participation was found to occur more often among men, unmarried individuals, and individuals with higher income, due to a stigma or lack of knowledge (Bollinger & David, 1997). At the same time, several studies relating SNAP to obesity attempt to estimate the impact of misclassification errors on their findings, all with different results (Vassilopoulos et al, 2011; Mittag, 2013; Almada, McCarthy & Tchernis, 2015).

estimate on SNAP participation when it is used as an explanatory variable for obesity (Hausman, 2001).

In this paper, we use the data from the National Health and Nutrition Survey 2011-2014 to find if USDA findings of the positive link between SNAP participation and obesity hold after correcting for misclassification. We employ statistical methods (Hausman, Abrevaya and Scott-Morten, 1998; Tennekoon and Rosenman, 2014) to correct for misclassification in SNAP participation, providing what should be more reliable measurements of the SNAP participation impact on obesity.

We find that, among males, there is no evidence of a relationship between SNAP participation and Body Mass Index. However, among females, we find a significant positive relationship between SNAP participation and BMI. We also find that correcting SNAP for misreporting reduces the marginal effect of SNAP participation on BMI. This finding is important, because without controlling for misspecification, the underreporting of SNAP participation causes a severe bias in the estimates.

Our results have important implications for the USDA policy on limiting SNAP-eligible purchases. The results are consistent with several previous studies in its finding that SNAP participation is positively correlated with BMI in women. This finding needs to be explored further. One possible explanation is that male SNAP recipients consume fewer calories than female SNAP recipients. This is consistent with previous USDA findings from 2007-2010 NHANES. Our future research will focus on the calorie intake and dietary quality in male and female participants in SNAP.

In the next section we show our research design. We then describe the data and variables of interest. Finally, we present our results and discuss their implications and areas of future research.

METHODOLOGY

To estimate the relationship between BMI and SNAP participation, we use a multivariate model

$$BMI_i = \beta_0 + \beta_1 SNAP_i + \beta_2 W_i + \varepsilon_i \quad (1)$$

where BMI is body mass index (continuous or categorical), $SNAP_i$ is a predicted current participation after correcting for misclassification, W is a vector of other control variables and $\beta_i,\ i = (1,2,3)$ are parameters.

We correct for possible misclassification of SNAP using Hausman, et al. (1998) (HAS) and its generalization to systematic misclassification by Tennekoon and Rosenman, 2016 (GHAS), then use the predicted value of SNAP participation when estimating the relationship between BMI and SNAP.

Let $SNAP_i^0$ be an unobserved latent variable for SNAP participation given by

$$SNAP_i^0 = X_i^{'} b + \eta_i \quad (2)$$

where X_i is a vector of variables explaining SNAP participation, b is a vector of parameters, and η_i is an i.i.d. error term with a known common distribution. We should observe a binary variable $SNAP^*_i$:

$$SNAP_i^* = 1\,(SNAP^0{}_i \geq 0)$$

If no misclassification is present, we observe $SNAP^*{}_i$. However, when misclassification is present, we observe variable $SNAP_i$ that includes some true '1's classified as '0's, and some true '0's classified as '1's.

Let F be the cdf of the error term and $SNAP_i^*$ be true participation. HAS assumes the probability of misclassification is random. Specifically, in HAS, the misclassification probabilities are:

$$\alpha_0 = \Pr\left(SNAP_i = 1 | SNAP_i^* = 0\right),$$

$$\alpha_1 = \Pr\left(SNAP_i = 0 | SNAP_i^* = 1\right)$$

where α_0 is the probability that a zero is misclassified as a one (a non-participant claims to participate in SNAP) and α_1 is the probability that a one is misclassified as a zero (a SNAP participant claims to not participate in SNAP). The conditional expected value of (recorded) SNAP participation is

$$E\left(SNAP_i | X_i\right) = \Pr\left(SNAP_i | X_i\right) = \alpha_0 + \left(1 - \alpha_0 - \alpha_1\right) F\left(X_i' b\right) \qquad (3).$$

When there is no misclassification, $\alpha_0 = \alpha_1 = 0$, and the above expression becomes $F\left(X_i' b\right)$

We estimate (α_0, α_0 and b) by maximum likelihood estimation for the log likelihood function

$$\mathcal{L}(\alpha_0, \alpha_1, b) = n^{-1} \sum_{i=1}^{n} \{SNAP_i \ln\big(\alpha_0 + (1 - \alpha_0 - \alpha_1) F(X_i' b)\big) +$$

$$(1 - SNAP_i) \ln\big(\alpha_0 + (1 - \alpha_0 - \alpha_1) F(X_i' b)\big)\}$$

over (α_0, α_0, b).

GHAS generalizes the probabilities α_0 and α_1 by allowing them to depend on covariates:

$$\alpha_0\left(Z_i^0\right) = \Pr\left(SNAP_i = 1 | SNAP_i^* = 0, Z_i^0\right) = F_0\left(Z_i^0\gamma_0\right)$$

$$\alpha_1\left(Z_i^1\right) = \Pr\left(SNAP_i^0 = 1 | SNAP_i^* = 0, Z_i^1\right) = F_1\left(Z_i^1\gamma_1\right)$$

where Z_i^0 and Z_i^1 may be, but are not necessarily subsets of X_i, and F_0 and F_1 are the cumulative distribution functions of stochastic components that determine the underreporting and overreporting of SNAP (Tennekoon & Rosenman, 2016). In this case the conditional expected value of (recorded) SNAP participation is

$$E\left(SNAP_i | X_i\right) = \Pr\left(SNAP_i | X_i\right) = F_0\left(Z_i^0\gamma_0\right) + \left(1 - F_0\left(Z_i^0\gamma_0\right) - F_1(Z_i^1\gamma_1\right)F\left(X_i^{'}b\right) \quad (4)$$

which we also estimate with maximum likelihood methods. Equations (3) and (4) can be used to generate instruments for SNAP participation that correct for potential misclassification. More specifically, we use (3) or (4) to predict $SNAP_i$ for each observation i and use it in the equation (1) to estimate the impact of SNAP on obesity.

One question is whether (3) or (4) is the correct model. Hausman, et al. (1998) shows that even a small amount of misclassification can bias ordinary probit estimates by 15-20%, and the problem grows as the amount of misclassification increases. Tennekoon and Rosenman (2016) go further, demonstrating that HAS may increase the bias and reduce efficiency of the primary parameter estimates if the misclassification is systematic and large. At the same time, Tennekoon and Rosenman show that if the probability of misclassification is not covariate-dependent, there is little cost of using GHAS instead of HAS to fix misclassification; the estimates of the primary equation, the β in equation (1), are only less efficient. Nonetheless, for comparison we use both instruments for SNAP participation.

A common issue when studying the relationship between SNAP and obesity is whether SNAP is endogenous to obesity. It is argued that there may be unobserved characteristics that affect both the decision to participate in SNAP and BMI. Hence we use IV-GMM approach to correct for the endogeneity issues, with Household Size and Poverty Index as excluded instruments for SNAP participation. We find that our selected instruments are strong and valid predictors of SNAP participation.

DATA

The data are from the 2011-2012 and 2013-2014 waves of the National Health and Nutrition Survey (NHANES) a nationally representative sample of about 5,000 persons each year, which assesses the health and nutritional status of persons of all ages in the United States. The survey is unique in that it combines interviews and physical examinations. The variables of interest come from the Questionnaire, Demographics and Examination parts of the NHANES study. Descriptions of all variables used in our analysis are given in Table 2.1.

The dependent variable for equation (1) is Body Mass Index. The NHANES dataset uses measured height and weight, as well as survey-based SNAP participation. By using measured height and weight, we avoid mismeasurement in the reporting of BMI. Independent variables in equation (1) include reported or predicted participation in SNAP anytime in the past 12 months, and demographic and lifestyle variables. SNAP participation in the past 12 months is our independent variable of interest, because it reflects current and recent participation in the SNAP program. Lifestyle variables include variables that may affect BMI, specifically: moderate-intensity work, vigorous-intensity work, vigorous and moderate exercise, exercise duration, as well as biking or walking to work. Demographic variables of interest are age, gender, income,

college education, race, and marital status. All variables are described in Table 2.1. Table 2.2 provides descriptive statistics of all variables used in the analysis. Most participants (86%) are U.S. citizens. The sample includes an almost equal number of males and females. In the sample, the average household size is 3, the average age is 45, and the average household income is $48,000.

According to the USDA (USDA 2014) SNAP participation rates of eligible adults were about 79% in 2011 and 83% in 2012. One criterion for eligibility is income. In the NHANES data, average SNAP participation is about 23% of the whole (eligible and non-eligible) population. SNAP eligibility is not recorded in NHANES data, but we do have income. According to the Department of Health and Human Services guidelines, income-eligible nonparticipants are defined as "individuals from households with monthly income less than or equal to 130 percent of … poverty (level)." (Condon et al, 2015). We find reported participation of 47% among adults with income-to-poverty ratio of less than 1.3, reported participation of 22% among adults with income-to-poverty ratio between 1.3 and 1.85, and reported participation of 6% among adults with income-to-poverty ratio over 1.85. This indicates that SNAP participation may indeed be under reported in the NHANES.

RESULTS

Table 2.3 reports the estimates for equations (3) and (4): HAS estimates are reported in column 2 of the table, and GHAS in column 3. Based on the previous studies that find certain individual characteristics more prevalent in misclassification of SNAP participation (Bollinger, 1997; Bollinger, 2005), we select these individual characteristics to instrument the probability of

over- and under-reporting SNAP using GHAS. Specifically, we use age, household size, poverty index categories, household income, education and gender as the variables explaining the likelihood of a specific observation being misclassified. When GHAS is the correct model, using HAS may bias the estimated coefficients of equation (1), which helps to explain the differences found using the different instruments. More specifically, poverty category 1 (lower-income) has a significantly positive effect on α_1, while household income and education are both statistically significant (at a p-value<0.001) in the equations explaining α_0 with GHAS.[2]

We find that estimated effects of individual characteristics on SNAP participation are consistent between HAS and GHAS. The results imply that the probability of SNAP participation is higher with bigger household size, age and acquisition of citizenship. Probability of SNAP participation is lower in men, and has a significant negative relationship with education, income, married status.

We find that the expected values of α_0 (probability of over-reporting SNAP) and α_1 (probability of under-reporting) differ between HAS and GHAS. Specifically, the expected values translate to $\alpha_1 = .04$ and $\alpha_0 = 0.005$ when using HAS, and much higher values of $\alpha_1 = .26$ and $\alpha_0 = 0.043$ when using GHAS. The difference in the expected probability of underreporting estimated through HAS and GHAS is large, with GHAS-associate probability higher and closer to the proportion estimated in the previous literature (Almada, McCarthy & Tschernis, 2015). The significant effects of individual characteristics (income, household size, education) on the probability of misclassification lead us to believe that misclassification of SNAP is systematic, rather than random. Despite the difference in the expected probabilities of

[2] This significance indicates that GHAS may be more appropriate than HAS to correct for misclassification in SNAP participation.

misreporting via HAS and GHAS, both methods yield similar estimated marginal effects on the SNAP variable.

Using the results from Table 2.3, we create two SNAP participation variables: one corrected using the HAS model, another corrected using the GHAS model. Then, we estimate the effect of the SNAP variable on Body Mass Index. For comparison purposes, we also estimate equation (1) using the observed value for SNAP. These results are presented in Table 2.4.

Table 2.4 reports the estimated effect of SNAP participation on BMI obtained using the IV-GMM approach. The independent variable of interest is SNAP participation in columns (1) – (3), HAS-corrected SNAP participation in columns (4) - (6), and GHAS-corrected SNAP participation in columns (7) - (9).

We find that in a pooled IV-GMM estimation, there is a positive significant effect of participation in SNAP on BMI, among women. The estimated effect ranges from 3.2 to 4.6 kg/m^2, depending on whether observed or corrected SNAP was used. In all cases, we do not find significant effect of SNAP on BMI in men.

In unpooled IV-GMM estimation (separate samples for men and women), we find a positive significant effect of SNAP participation on BMI in women. The estimated effect ranges from 2.1 to 3.4 kg/m^2, depending on whether observed or corrected SNAP was used. We do not see a significant impact of SNAP participation on men.

While the effect of SNAP on BMI in women is consistently positive across all models, we find that correcting for misclassification reduces the magnitude of this effect.

Among other results, we see that employment and age have an upward effect on BMI, particularly in women. Vigorous- and moderate-intensity work sometimes has a positive impact on BMI. We also find that time spent in a sedentary position, watching TV or on a computer has

an upward impact on BMI, while vigorous and moderate recreational activities have a negative effect on BMI. As expected, income has a negative effect on BMI. We also see that married respondents had a somewhat higher BMI than the unmarried ones. Our results show that most variables have a different effect on BMI in men vs. women, which reinforces our decision to separate the sample by gender.

Tests of instruments used for SNAP participation variable are presented in Table 2.5. In the pooled sample, the endogenous variables are SNAP Participation and the interaction between SNAP Participation and Male. Thus we use the following instruments: Poverty Index, Household Size, and the interaction between Male and Poverty Index. In the separate samples, the endogenous variable is SNAP Participation, which we instrument using Poverty Index and Household Size. In all models, the sets of instruments are valid and strong predictors of the endogenous variables.

Our findings are partly consistent with those of Vassilopoulos et al. (2011), but have different conclusions. Using different waves of the same dataset (NHANES), Vassilopoulos et al. (2011) find a positive link between SNAP benefits and obesity, but claim that the consistency of SNAP-obesity relationship holds with SNAP measurement errors less than 10%. As evidence of higher measurement errors, Vassilopoulos et al. (2011) refer to the finding of 10% - 15% under-reporting of SNAP participation in older studies (Marquis & Moore, 1990; Bollinger & David, 1997). It is worth noting that these studies are based on a different, older dataset (1984 Survey of Income and Program Participation), and a much narrower sample (individuals of asset-eligible couples 18 – 24 years of age). Using HAS and GHAS correction algorithms, we find much higher probabilities of under-reporting of SNAP participation in NHANES study.

Our study takes a different approach to SNAP-BMI relationship, compared to Almada et al. (2015) who use a longitudinal survey (National Longitudinal Survey of Youth, 1979) to study the link of SNAP and obesity. We focus on the fact of receiving SNAP benefits, while Almada et al. (2015) use the dollar amount of SNAP benefits as their variable of interest. Secondly, in their study Almada et al. use self-reported weight and height, while we use measured weight and height. There is evidence of survey respondents under-representing their weight and sometimes over-representing their height, which would lead to a lower BMI. Thirdly, we use the full sample of data, while Almada et al. (2015) restrict their sample to only SNAP-eligible respondents, and subsequently – to SNAP-eligible families with at least one child over 5 years of age. Similarly to our finding of positive SNAP-BMI relationship, they find a positive link between the amount of SNAP benefits amount and BMI, in the sample of SNAP-eligible respondents. However, we show that positive SNAP-BMI link only holds in women, but not in men. Almada et al. focus on families with at least one child over 5 years of age, and their main finding of negative relationship between SNAP and BMI only holds in that restricted sample. This approach is problematic, as it disregards a large part of the sample, and cannot be extrapolated on the whole population of SNAP recipients.

Our findings are consistent with other earlier studies that find a positive relationship between obesity and SNAP participation among women, but not among men (Townsend et al., 2001; Gibson, 2003; Chen et al., 2005; Baum, 2011). However, we find that this estimated effect weakens after correcting misclassification in self-reported SNAP.

CONCLUSIONS

In this paper, we answer the research question: what is the relation between food assistance program participation and obesity in the United States, in the presence of misreporting? The goal of food assistance programs is to provide the necessary nutrition to the population in need. If participation in such programs contributes to obesity, it means that these programs do not provide the necessary nutrition for their recipients. However, if the analysis does not account for misreporting errors in self-reported data, these findings may be wrong. In this paper, we correct for this bias and then estimate the relation between program participation and obesity.

We use two waves (2011-2012 and 2013-2014) of National Health and Nutrition Survey. This dataset uses measured height and weight, as well as survey-based SNAP participation. By using measured height and weight, we avoid endogeneity issues in BMI reporting. We use Hausman-Abrevaya-Scott and Generalizes Hausman-Abrevaya-Scott techniques to correct for misreporting errors in SNAP participation (Hausman, Abrevaya & Scott-Morton, 1998; Tennekoon & Rosenman, 2016). After applying corrections, we estimate the effect of SNAP participation on obesity in the United States.

We find that, among males, there is no evidence of a relationship between SNAP participation and Body Mass Index. However, among females, we find a significant positive relationship between SNAP participation and BMI. We also find that correcting SNAP for misreporting reduces the marginal effect of SNAP participation on BMI. This finding is important, because without controlling for misspecification, the underreporting of SNAP participation causes severe bias in the estimates.

Earlier papers discover positive relationship between SNAP participation and BMI in women. In our study, we find that this relationship holds even after correcting for misclassification of SNAP. While several recent papers argue that the positive relationship between SNAP and BMI should disappear after correcting for SNAP misclassification, we only find a smaller positive effect of SNAP participation on BMI in women.

Our paper has several limitations. First, we find rather low goodness-of-fit of the OLS models (R-squared of 0.1), even in the presence of significant marginal effects. This means that while there is a significant trend in the model, there is also high variability and noise in the data. Second, positive relationship between SNAP and BMI in women does not prove causality. One possibility is that female SNAP recipients are more overweight due to SNAP participation. Another possibility is that higher consumption of food in certain low income women simultaneously leads to higher BMI and to higher food expenditures and, in turn, SNAP participation.

Our results have important implications for the USDA policy on limiting SNAP-eligible purchases. The results are consistent with several previous studies in its finding that SNAP participation is positively correlated with BMI in women. This finding needs to be explored further. One possible explanation is that male SNAP recipients consume fewer calories than female SNAP recipients. This is consistent with previous USDA findings from 2007-2010 NHANES. Our future research will focus on the calorie intake and dietary quality in male and female participants in SNAP. Other future areas of research include investigating the effects of SNAP participation on adult food insecurity, poor physical and mental health outcomes, in the presence of misclassification bias.

REFERENCES

Almada, L., McCarthy, I. M., & Tchernis, R. (2015). What can we learn about the effects of food stamps on obesity in the presence of misreporting?. Available at SSRN 2563822.

Baum, C. L. (2011). The effects of food stamps on obesity. Southern Economic Journal, 77(3), 623-651.

Bitler, M. (2014). The Health and Nutrition Effects of SNAP: Selection into the Program and a Review of the Literature on its Effects.

Bollinger, C. R., & David, M. H. (1997). Modeling discrete choice with response error: Food Stamp participation. Journal of the American Statistical Association, 92(439), 827-835.

Bollinger, C. R., & David, M. H. (2001). Estimation with response error and nonresponse: food-stamp participation in the SIPP. Journal of Business & Economic Statistics, 19(2), 129-141.

Bollinger, C. R., & David, M. H. (2005). I didn't tell, and I won't tell: dynamic response error in the SIPP. Journal of Applied Econometrics, 20(4), 563-569.

Bolton-Smith, C., Woodward, M., Tunstall-Pedoe, H., & Morrison, C. (2000). Accuracy of the estimated prevalence of obesity from self reported height and weight in an adult Scottish population. Journal of Epidemiology and Community Health, 54(2), 143-148.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, http://wwwn.cdc.gov/nchs/nhanes/search/nhanes_continuous.aspx

Chen, Z., Yen, S. T., & Eastwood, D. B. (2005). Effects of food stamp participation on body weight and obesity. American Journal of Agricultural Economics, 87(5), 1167-1173.

Cody, S., & Tuttle, C. (2002). The Impact of Income Underreporting in CPS and SIPP on Microsimulation Models and Participating Rates. Washington, DC: Mathematica Policy Research, Inc, July, 24.

Condon, Elizabeth, Susan Drilea, Keri Jowers, Carolyn Lichtenstein, James Mabli, Emily Madden, and Katherine Niland. (2015). Diet Quality of Americans by SNAP Participation Status: Data from the National Health and Nutrition Examination Survey, 2007–2010. Prepared by Walter R. McDonald & Associates, Inc. and Mathematica Policy Research for the Food and Nutrition Service.

Dauphinot, V., Wolff, H., Naudin, F., Gueguen, R., Sermet, C., Gaspoz, J. M., & Kossovsky, M. P. (2009). New obesity body mass index threshold for self-reported data. Journal of Epidemiology and Community Health, 63(2), 128-132.

Economic Research Service (ERS), U.S. Department of Agriculture (USDA). National Household Food Acquisition and Purchase Survey (FoodAPS). http://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey.aspx

Eslami, E. (2014). Trends in Supplemental Nutrition Assistance Program Participation Rates: Fiscal Year 2010 to Fiscal Year 2012 (No. 35927b5532964e72b1190b443c39caf0). Mathematica Policy Research.

Gibson, D. (2003). Food stamp program participation is positively related to obesity in low income women. The Journal of nutrition, 133(7), 2225-2231.

Gundersen, C. (2015). SNAP and Obesity. SNAP Matters: How Food Stamps Affect Health and Well Being.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. Journal of Economic Perspectives, 15(4) 57-67.

Hausman, J.A., Abrevaya, J., Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. Journal of Econometrics, 87:239–269.

Heckman, J. J. (1977). Dummy endogenous variables in a simultaneous equation system.

Kreider, B., Pepper, J. V., & Roy, M. (2012). Identifying the effect of WIC on very low food security among infants and children.

Marquis, K. H., & Moore, J. C. (2010). Measurement errors in SIPP program reports. Survey Methodology, 1.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2009). The under-reporting of transfers in household surveys: its nature and consequences (No. w15181). National Bureau of Economic Research.

Millar, W. J. (1986). Distribution of body weight and height: comparison of estimates based on self-reported and observed measures. Journal of Epidemiology and Community Health, 40(4), 319-323.

Mittag, N. (2013). A Method of Correcting for Misreporting Applied to the Food Stamp Program. US Census Bureau Center for Economic Studies Paper No. CES-WP-13-28.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., & Marks, J. S. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. Jama, 289(1), 76-79.

Nieto-Garcia, F. J., Bush, T. L., & Keyl, P. M. (1990). Body mass definitions of obesity: sensitivity and specificity using self-reported weight and height. Epidemiology, 146-152.

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2014). Prevalence of childhood and adult obesity in the United States, 2011-2012. Jama, 311(8), 806-814.

Perry, G. S., Byers, T. E., Mokdad, A. H., Serdula, M. K., & Williamson, D. F. (1995). The validity of self-reports of past body weights by US adults. Epidemiology, 6(1), 61-66.

Robinson, C. A., & Zheng, X. (2011). Household Food Stamp Program Participation and Childhood Obesity. Journal of Agricultural and Resource Economics, 1-13.

Tennekoon, V., & Rosenman, R. (2016). Systematically misclassified binary dependent variables. Communications in Statistics-Theory and Methods, 45(9), 2538-2555.

Townsend, M. S., Peerson, J., Love, B., Achterberg, C., & Murphy, S. P. (2001). Food insecurity is positively related to overweight in women. The Journal of nutrition, 131(6), 1738-1745.

Vassilopoulos, A., Drichoutis, A., Nayga, R., & Lazaridis, P. (2011). Does the Food Stamp Program Really Increase Obesity? The Importance of Accounting for Misclassification Errors.

Ver Ploeg, M., Mancino, L., & Lin, B. H. (2007). Food and nutrition assistance programs and obesity: 1976-2002 (No. 55965). United States Department of Agriculture, Economic Research Service.

Wang, Y., Beydoun, M. A., Liang, L., Caballero, B., & Kumanyika, S. K. (2008). Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. Obesity, 16(10), 2323-2330.

Table 2.1 Variable description

Variable name	Description	Source
Body Mass Index	Body Mass Index, measured at time of screening, calculated as kg/m^2 (from 12.4 to 82.1)	Examination
SNAP	Have you participated in SNAP in the last 12 months? (1=Yes, 0=No)	Questionnaire
Household Income	Annual household income (thousand dollars). Recoded from categorical variable to continuous (2,500 to 100,000)	Demographics
Education	Education level (1= less than 9^{th} grade, 2 = 9^{th} -11^{th} grade, 3 = High school, GED or equivalent, 4 = Some college or AA degree, 5 = College graduate or above)	Demographics
College	College education	
Household Size	Total number of people in the household (1-7)	Demographics
Married	Marital status = married (1=Yes, 0=No)	Demographics
Male	Male gender (1=Yes, 0=No)	Demographics
Age	Age in years at screening (16 – 80)	Demographics
White	Race = white (1=Yes, 0=No)	Demographics
Black	Race=black (1=Yes, 0=No)	Demographics
Citizen	U.S. citizen (1=Yes, 0=No)	Demographics
Poverty index	Family monthly poverty index, a ratio of monthly family income to the HHS poverty guidelines specific to family size (0 – 4.99)	Questionnaire
Poverty category	Family poverty level index category (1= index under 1.30; 2 = index between 1.30 and 1.85; 3 = index over 1.85)	Questionnaire
Walk-bike	Do you walk or use a bicycle for at least 10 minutes continuously to get to and from places? 1=Yes, 0=No	Questionnaire
Moderate-intensity work	Does your work involve moderate-intensity activity for at least 10 minutes continuously? 1=Yes, 0=No (moderate-intensity activities are activities that require moderate physical effort and cause small increases in breathing or heart rate)	Questionnaire
Vigorous-intensity work	Does your work involve vigorous-intensity activity for at least 10 minutes continuously? (vigorous-intensity activities are activities that require hard physical effort and cause large increases in breathing or heart rate)	Questionnaire
Sedentary	Minutes sedentary activity (0 – 1380) How much time do you usually spend sitting on a typical day (excluding sleeping)?	Questionnaire
Moderate recreational activity	Do you do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming or golf for at least 10 minutes continuously? (1=Yes, 0=No)	Questionnaire

Variable name	Description	Source
Days moderate recreational activities	In a typical week, on how many days do you do moderate-intensity sports, fitness or recreational activities? (1 to 7)	Questionnaire
Vigorous recreational activity	Do you do any vigorous-intensity sports, fitness, or recreational activities that cause a large increase in breathing or heart rate like running or basketball for at least 10 minutes continuously? (1=Yes, 0=No)	Questionnaire
Days vigorous recreational activities	In a typical week, on how many days do you do vigorous-intensity sports, fitness or recreational activities? (1 to 7)	Questionnaire
Hours watch TV/videos past 30 days	Over the past 30 days, on average, how many hours per day did you sit and watch TV or videos? (0 to 5)	Questionnaire
Hours use computer past 30 days	Over the past 30 days, on average, how many hours per day did you use a computer or play computer games outside of work or school? Include Playstation, Nintendo DS or other portable video games. (0 to 5)	Questionnaire

Table 2.2 Descriptive statistics for the variables used in the estimation

	Count	Mean	Standard deviation	Min	Max
Wave	12146	1.51	0.50	1	2
Body Mass Index	11527	28.68	7.05	13.4	82.9
SNAP	12005	0.23	0.42	0	1
Household Income (thousands)	12146	48.04	31.54	2.4	100
Education	12134	3.42	1.24	1	5
Household Size	12146	3.28	1.69	1	7
Married	12136	0.46	0.50	0	1
Female	12146	0.51	0.50	0	1
Age	12146	46.84	18.91	17	80
Employed	12139	0.53	0.50	0	1
White	12146	0.39	0.49	0	1
Black	12146	0.24	0.42	0	1
Citizen	12108	0.86	0.34	0	1
Poverty index	11103	2.41	1.66	0	5
Poverty category	11390	2.09	0.92	1	3
Walk-bike	12144	0.30	0.46	0	1
Moderate-intensity work	12141	0.32	0.47	0	1
Vigorous-intensity work	12143	0.17	0.37	0	1
Sedentary	12086	394.38	201.02	0	1380
Moderate recreational activity	12145	0.41	0.49	0	1
Days moderate recreational activities	12143	1.47	2.13	0	7
Vigorous recreational activity	12145	0.24	0.43	0	1
Days vigorous recreational activities	12144	0.80	1.64	0	7
Hours watch TV/videos daily	12135	2.44	1.64	0	5
Hours use computer daily	12142	1.09	1.51	0	5

Table 2.3 Hausman-Abrevaya-Scott-Morton (HAS) and Generalized HAS (GHAS) correction for misspecification

	(1) HAS (dependent variable: SNAP)	(2) GHAS (dependent variable: SNAP)
Household size	0.299***	0.332***
	(0.0167)	(0.018)
Household Income	-0.0272***	-0.021***
	(0.00134)	(0.001)
Education	-0.174***	-0.163***
	(0.0162)	(0.021)
Married	-0.519***	-0.528***
	(0.0415)	(0.071)
Male	-0.0592	-0.006
	(0.0325)	(0.071)
Black	0.506***	0.578***
	(0.0477)	(0.058)
White	0.223***	0.270***
	(0.0441)	(0.053)
Citizen	0.433***	0.515***
	(0.0557)	(0.066)
Age	0.131***	0.154***
	(0.0204)	(0.025)
Age^2	-0.00226***	-0.003***
	(0.000450)	(0.001)
Age^3	0.0000109***	0.000**
	(0.00000305)	(0.000)
_cons	-2.546***	-3.214***
	(0.279)	(0.352)
a1		
Poverty Category 1		-1.290***
		(0.159)
Male		0.183
		(0.160)
_cons	-1.758***	-0.259*
	(0.349)	(0.118)
$E(\alpha_1)$	.040	.267
a0		
Poverty Category 3		-5.012
		(370.03)
Household Income		-0.076***
		(0.013)

Education		-0.182***
		(0.013)
_cons	-2.568***	0.262**
	(0.190)	(0.264)
$E(\alpha_0)$	.005	.043
N	11,954	11,338

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.4 Results of IV-GMM estimation (dependent variable: Body Mass Index)

Model:	(1) Pooled IV	(2) Male IV	(3) Female IV	(4) Pooled IV	(5) Male IV	(6) Female IV	(7) Pooled IV	(8) Male IV	(9) Female IV
Variable of Interest:	observed SNAP	observed SNAP	observed SNAP	corrected SNAP using HAS	corrected SNAP using HAS	corrected SNAP using HAS	corrected SNAP using GHAS	corrected SNAP using GHAS	corrected SNAP using GHAS
SNAP	4.613***	0.326	3.412***	3.256***	0.305	2.308***	3.193***	0.278	2.071**
	(0.747)	(0.809)	(1.208)	(0.569)	(0.648)	(0.871)	(0.527)	(0.584)	(0.776)
Male * SNAP	-5.498***			-3.951***			-4.072***		
	(0.679)			(0.476)			(0.485)		
Male	0.212			0.340			0.474*		
	(0.204)			(0.213)			(0.226)		
Employed	0.807***	0.250	1.174***	0.693***	0.239	1.003***	0.682***	0.237	0.976***
	(0.154)	(0.196)	(0.244)	(0.146)	(0.192)	(0.221)	(0.146)	(0.192)	(0.219)
Age	0.0278***	0.00800	0.0424***	0.0261***	0.00879	0.0391***	0.0261***	0.00882	0.0388***
	(0.00480)	(0.00600)	(0.00787)	(0.00462)	(0.00596)	(0.00721)	(0.00460)	(0.00597)	(0.00713)
Vigorous-intensity work	0.514***	0.350	0.709*	0.520***	0.332	0.737**	0.530**	0.332	0.745*
	(0.195)	(0.215)	(0.362)	(0.193)	(0.215)	(0.358)	(0.193)	(0.215)	(0.358)
Moderate-intensity work	0.664***	0.906***	0.410*	0.655***	0.918***	0.390*	0.652***	0.916***	0.384
	(0.151)	(0.188)	(0.235)	(0.150)	(0.188)	(0.234)	(0.150)	(0.187)	(0.234)
Days vigorous recreational activities	-0.284***	-0.222***	-0.394***	-0.291***	-0.220***	-0.414***	-0.290***	-0.220***	-0.414***
	(0.0434)	(0.0508)	(0.0725)	(0.0424)	(0.0500)	(0.0710)	(0.0423)	(0.0499)	(0.0710)
Days moderate recreational activities	-0.0140	-0.0445	0.0116	-0.0299	-0.0464	-0.0123	-0.0306	-0.0463	-0.0138
	(0.0312)	(0.0385)	(0.0492)	(0.0307)	(0.0385)	(0.0479)	(0.0307)	(0.0385)	(0.0478)
Walk-bike	-1.015***	-1.100***	-1.005***	-1.062***	-1.085***	-1.091***	-1.054***	-1.083***	-1.085***
	(0.145)	(0.181)	(0.229)	(0.144)	(0.179)	(0.226)	(0.144)	(0.178)	(0.226)
Sedentary	0.00240***	0.00227***	0.00257***	0.00237***	0.00225***	0.00250***	0.00236***	0.00225***	0.00250***
	(0.00035)	(0.00045)	(0.00053)	(0.00035)	(0.0004)	(0.00053)	(0.000351)	(0.000452)	(0.000533)
TV hours	0.468***	0.423***	0.506***	0.469***	0.416***	0.514***	0.470***	0.416***	0.515***
	(0.0433)	(0.0542)	(0.0666)	(0.0426)	(0.0538)	(0.0657)	(0.0426)	(0.0538)	(0.0657)
Computer hours	0.327***	0.264***	0.368***	0.323***	0.253***	0.372***	0.323***	0.252***	0.370***
	(0.0456)	(0.0572)	(0.0701)	(0.0450)	(0.0569)	(0.0695)	(0.0449)	(0.0568)	(0.0694)
Income	-0.0094***	-0.00085	-0.0184***	-0.00973**	-0.00064	-0.0188***	-0.0111**	-0.000901	-0.0213***
	(0.00366)	(0.00424)	(0.00628)	(0.00400)	(0.00481)	(0.00656)	(0.00358)	(0.00436)	(0.00577)
Married	0.656***	0.913***	0.678***	0.672***	0.928***	0.676***	0.663***	0.926***	0.662**

Model:	(1) Pooled IV	(2) Male IV	(3) Female IV	(4) Pooled IV	(5) Male IV	(6) Female IV	(7) Pooled IV	(8) Male IV	(9) Female IV
Variable of Interest:	observed SNAP	observed SNAP	observed SNAP	corrected SNAP using HAS	corrected SNAP using HAS	corrected SNAP using HAS	corrected SNAP using GHAS	corrected SNAP using GHAS	corrected SNAP using GHAS
	(0.147)	(0.189)	(0.228)	(0.145)	(0.189)	(0.225)	(0.144)	(0.188)	(0.223)
Education	-0.0827	-0.0265	-0.121	-0.0764	-0.00154	-0.125	-0.0820	-0.00165	-0.133
	(0.0631)	(0.0773)	(0.101)	(0.0642)	(0.0799)	(0.101)	(0.0637)	(0.0797)	(0.0996)
White	-0.163	0.0616	-0.469**	-0.169	0.0464	-0.448**	-0.168	0.0467	-0.454*
	(0.136)	(0.171)	(0.212)	(0.135)	(0.170)	(0.210)	(0.135)	(0.170)	(0.210)
Wave	0.179	0.00849	0.330*	0.146	-0.0134	0.300	0.143	-0.0134	0.293
	(0.130)	(0.165)	(0.200)	(0.129)	(0.164)	(0.198)	(0.129)	(0.164)	(0.198)
_cons	24.39***	25.20***	24.13***	24.54***	25.10***	24.50***	24.59***	25.11***	24.74***
	(0.670)	(0.723)	(1.226)	(0.698)	(0.805)	(1.178)	(0.661)	(0.772)	(1.095)
N	11337	5577	5760	11424	5620	5804	11424	5620	5804

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
SNAP in columns 4 – 9 is expected probability of SNAP participation (ranging from 0 to 1)

Table 2.5 Tests of Instruments for SNAP participation

		Test Statistic (P-Value)								
Diagnostic Test	Null Hypothesis	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Valid/strong instruments test										
Under identification (Kleibergen-Paap)	Instruments are under-identified	657 (0.00)	391 (0.00)	259 (0.00)	1414 (0.00)	1003 (0.00)	664 (0.00)	2415 (0.00)	1294 (0.00)	1108 (0.00)
Weak identification (Cragg-Donald)	Instruments are weak	232.08 (0.00)	209.7 (0.00)	135 (0.00)	536 (0.00)	608 (0.00)	374 (0.00)	1019 (0.00)	838 (0.00)	682 (0.00)
Over identifying restrictions (Hansen J)	Instruments are valid, and excluded instruments are correctly excluded	2.138 (0.14)	1.845 (0.17)	0.715 (0.39)	0.998 (0.32)	2.055 (0.15)	0.127 (0.72)	2.156 (0.14)	2.05 (0.15)	0.750 (0.38)

Notes: The table presents test statistics (Chi-squared statistics for the tests of under- and over-identification, F statistic for the test of weak identification). The corresponding P-values are presented in parentheses.

CHAPTER THREE

ADVERSE SELECTION IN TERM LIFE INSURANCE MARKETS: A NEW LOOK

ABSTRACT

Adverse selection hypothesis argues that individuals with a higher mortality risk are more likely to purchase life insurance. Theory of advantageous selection argues the opposite. Recent empirical studies have not reached consensus on whether adverse selection or advantageous selection is more prevalent in term life insurance markets. Specifically, two studies use one dataset and come to opposing conclusions about the sign of the relationship between life insurance ownership and mortality risk. Our study offers a new look at the same problem, by using a comprehensive set of regressors and by addressing misclassification in self-reported insurance ownership. In contrast to the previous studies, we find a positive link between mortality and individual term life insurance ownership, which leads us to believe that adverse selection prevails in the individual life insurance market. Additionally, we find evidence of advantageous selection in the group term life insurance market. This implies that under-insurance from adverse selection in the individual term life insurance market must be considered as well as the over-insurance from advantageous selection in the group term life insurance market.

INTRODUCTION

Economic theory discusses health and life insurance markets as markets characterized by information asymmetry (Rothschild & Stiglitz, 1976; Chiappori, 2000). Studies find mixed evidence on the nature of the relationship between the insurance coverage and risk outcome. There are two conflicting theories on life insurance purchase decisions: adverse selection

hypothesis and advantageous selection hypothesis. Adverse selection theory predicts people with a high risk of death are more likely to own life insurance (Akerlof, 1970; Rothschild & Stiglitz, 1976). The adverse selection theory maintains that consumers of life insurance may possess more knowledge of their risk of death than insurance companies or agents. Advantageous selection theory (Hemenway, 1990) argues that risk-adverse behavior is the cause of both a lower mortality risk and a higher demand for life insurance purchase.

Previous empirical studies have found mixed evidence on the correlation between demand for life insurance and mortality, and therefore about the existence of asymmetric information in the life insurance markets. Some studies find evidence of higher mortality rates in life insurance holders (He, 2009). Other studies find the same or lower mortality rates in life insurance holders, maintaining that insurance companies can assess mortality risks better than individuals themselves (Cawley & Philipson, 1996; McCarthy & Mitchell, 2003; Hedengren and Stratmann, 2016).

The difference might be due to different perspectives on life insurance purchases. Both Cawley and Philipson (1996) and He (2009) use the Health and Retirement Study (HRS). But Cawley and Philipson focus on demand-side regressors like income, marital status, and number of children, to find a negative or neutral relationship between mortality risk and life insurance purchases. They argue their results signify no evidence of asymmetric information exists in the life insurance market. He (2009) uses the same dataset, but excludes all demand-side variables. She argues that demand-side variables do not affect the pricing of life insurance by the insurance providers. Instead, she controls for variables that influence the price of health insurance – factors like age, gender, smoking status and measures of health, which she argues determine the supply side of the market. After limiting her sample to potential new buyers (those who reported having

no life insurance in the first wave) she finds a positive relation between life insurance and mortality– and interprets it as evidence of asymmetric information in the life insurance markets.

Hedengren and Stratmann (2016) use a unique dataset merging administrative and survey records, and find that people with high death risk are less likely to own life insurance. They further study the market for employer-sponsored (group) life insurance and find no significant relationship between ownership and mortality risk, which they claim to be the evidence of price discrimination overcoming adverse selection in this market. Relevant to our study, Hedegren and Stratmann (2016) criticize He's findings, asserting that systematic misreporting of responses to life insurance questions in surveys may cause the wrong interpretation of the relationship between life insurance coverage and subsequent death.

In this paper, we revisit the relationship between mortality and individual term life insurance ownership[3], melding the models of Cawley and Philipson (1996), and He (2009). Insurance purchase is an equilibrium decision (Beck & Webb, 2003), so both supply and demand side variables may influence the outcome. Hence, our models includes both supply-side and demand-side determinants of life insurance ownership, and we look at insurance ownership in 1992 as in Cawley and Philipson, and new buyers in 1992, as in He. As an extension, we address the concern of Hendengren and Stratmann (2016), statistically correcting for the potential of misreported responses to life insurance questions. Doing so allows us to see if misreporting compounds the shortcomings we find with the specifications used in the earlier

[3] Our focus is mainly on individual term life insurance and, as an extension, on group term life insurance. Term life insurance carries no cash value and pays death benefit only if the insured dies within a specified period (California Department of Insurance, 2011). Individual term life insurance is a term life insurance policy where a single contract covers a single insured, while group term life insurance is offered to a group (often through an employer).

studies. Our approach also allows us to identify if the potential misreporting is random or dependent on specific individual characteristics.

For the most part, our results contradict both earlier studies. Contrary to Cawley and Philpson, we find that individuals who die by 2004 are, on average, more likely to own individual term insurance in 1992. This positive effect of mortality on ownership increases in magnitude and significance when we correct for misreported owning. However, when looking only at potential new buyers, we find little support for He's findings. Using the raw data there is no evidence of a positive effect of dying by 2004 on buying term life insurance; one form correcting for misreporting yields a positive relationship between mortality by 2004 and buying a new term life insurance policy at conventional levels of statistical significant, but the generalization that allows systematic misclassification does not. Furthermore, we find a negative effect of mortality on group term life insurance ownership, which may be evidence of advantageous selection in the group term life insurance market.

The next section discusses the data used in our analysis. We then give a brief review of our research design. Following that, we present and discuss our results, and offer conclusions and limitations in a final section.

METHODOLOGY

We proceed in two steps. First we estimate a model that incorporates demand-side regressors (from Cawley & Philipson) and supply-side regressors (from He). To further facilitate comparison to the earlier studies, we mimic Cawley and Philipson with the full sample, and then He with the sample restricted to potential buyers. We follow He (2009) in selecting actual

observed mortality by 2004 as our independent variable of interest[4]. Following up on He's argument that both her results and those of Cawley and Philipson may be inconsistent if self-reported life insurance purchase is misreported (He, 2008; Hedegren & Stratmann, 2016), we then correct for possible misclassification in the reported life insurance ownership using Hausman, et al., 1998 (HAS), and its generalization to systematic misclassification by Tennekoon and Rosenman, 2016 (GHAS) and repeat our analysis.

We estimate the relationship between mortality risk and individual term life insurance, using the following logit model:

$$\log\left(\frac{IndInsurance_i}{1-IndInsurance_i}\right) = \beta_0 + \beta_1 Mortality_i + \beta_2 W_i + \varepsilon_i \quad (1)$$

where a binary variable IndInsurance indicates whether individual i reported owning individual term life insurance, Mortality is a binary variable indicating whether the individual had died by 2004, W is a vector of control variables, and β_j, $j = (0,1,2)$ are parameters.

We estimate this model twice. First, following Cawley and Philipson, we use the entire sample from 1992. Then, following He, we restrict the sample to potential buyers by using only individuals who reported not owning individual term life insurance in 1992 and looking whether or not term life insurance was owned in 1994. In the second estimating, then, the dependent variable is characterized using new buyers and the model becomes:

$$\log\left(\frac{NewBuyer_i}{1-NewBuyer_i}\right) = \beta_0 + \beta_1 Mortality_i + \beta_2 W_i + \varepsilon_i \quad (2)$$

[4] Cawley and Philipson (1996) use mortality data for the period 1992 – 1994, and compute mortality risk as the fitted value of a logit regression of experienced mortality between 1992 and 1994 on demographic and health characteristics Meanwhile, He (2009) uses observed actual mortality by 2004.

where binary variable NewBuyer is 1 if individual i reported owning individual term life insurance in 1994 but not in 1992, and zero if the individual reported having individual term life insurance in neither 1992, nor 1994. All else stays the same as in (1).

The vector of control variables W includes demand-side variables (income, wealth, marital status, whether has children), pricing controls, including basic demographics (age and gender), health behaviors (smoking status -- including whether the individual smokes now or ever and current drinking status -- whether drinks now), and health indicators (whether the individual has been diagnosed with diabetes, high blood pressure, cancer, heart disease, arthritis, lung disease, stroke, asthma, kidney disease, ulcer, high cholesterol, or back pain; whether she had a hospital stay in the previous 12 months; whether BMI indicates underweight, overweight, or obese; and whether father and mother had died before age 60). To address possible misclassification, we use HAS and GHAS to correct the reported variables IndInsurance and NewBuyer for possible misreported responses.[5]

Because access to group term life insurance can be an important factor in an individual's decision to obtain or forego individual term life insurance, we perform one additional check of our model by estimating a reduced form bivariate probit model, where the dependent variables are individual term life insurance ownership and group term life insurance ownership. This model accounts for the fact that the decisions to own these two types of insurance may be jointly dependent. For this model, we assume there are underlying latent variables measuring predisposition to purchasing each type of insurance given by

[5] The details of how HAS and GHAS, and the effect on measured life insurance ownership, are provided in Appendix A.

$$
\begin{aligned}
IndInsurance_i^* &= \gamma_{10} + \gamma_{11} Mortality + \gamma_{12} W + \varepsilon_1 \\
GroupInsurance_i^* &= \gamma_{20} + \gamma_{21} Mortality + \gamma_{22} W + \varepsilon_2 \\
\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right]
\end{aligned} \tag{3}
$$

where the * indicates the variable is latent (unobserved). Mortality again indicates whether the individual had died by 2004, W is a matrix of control variables, ρ denotes the tetrachoric correlation between the dependent variables and γ_{jk}, $j = (1,2), k = (0,1,2)$ are parameters. Since we observe only if the individual has purchased each type of insurance, that is,

$$
\begin{aligned}
IndInsurance_i &= 1(IndInsurance_i^* > 0) \\
GroupInsurance_i &= 1(GroupInsurance_i^* > 0).
\end{aligned}
$$

Here $IndInsurance_i$ is whether individual i reported owning individual term life insurance in 1992, and $GroupInsurance_i$ is whether the individual reported owning group term life insurance in 1992 we estimate (3) using a bivariate probit model via maximum likelihood estimation.

DATA

We use the Health and Retirement Study (HRS) dataset. The HRS is a nationally representative longitudinal survey whose objective is to provide information about the U.S. population over age 50 through biennial surveys with samples of that population. It contains data on health status, insurance coverage, financial status, behavioral characteristics, demographics, and family structure. Following He (2009), our analysis uses the HRS cohort, which consists of

individuals born between 1931 and 1941. This cohort has been interviewed biennially since 1992.

To facilitate comparison with the earlier studies, we use data from two early waves, 1992 and 1994, the same waves are used in Cawley and Philipson (1996) and He (2006). HRS divides a respondent's vital status in each wave into one of five categories: alive in current wave, presumed alive in current wave, death reported in the current wave, death reported in a prior wave, and vital status unknown. We code a respondent as alive in 2004 if they fall into category 1 or 2 – and dead in 2004 if they fall into category 3 or 4 – in wave 2004.

Table 3.1 provides summary statistics of the relevant variables in our sample, based on the information from the first (1992) wave. Twenty-six percent of the HRS cohort owned individual term life insurance in 1992 and twenty seven percent owned it in 1994. Nineteen percent of potential new buyers obtained individual term life insurance between 1992 and 1994. Thirty five percent reported owning group term life insurance in 1992, and twenty five percent owned it in 1994. By 2004, about 15% of the cohort had died. The sample was rather balanced in gender (48% are male), and 74% of the respondents were married. The most commonly reported medical conditions were arthritis, high blood pressure, and back pain. About 10% sample had a hospital stay in the past year. Less than 10% of the sample was diagnosed with diabetes, cancer, lung disease, stroke, or asthma. Thirty two percent of the sample had healthy weight, 44% were overweight, and 22% were obese.

RESULTS

Results from estimating the relationship between mortality risk (dying by year 2004) and individual term life insurance ownership are presented in Table 3.2, columns 1 - 3. We find that higher risk individuals are, on average, more likely to purchase individual term insurance. This positive effect increases in magnitude and significance when we correct for misclassification using HAS or GHAS correction techniques.

Columns 4-6 of Table 3.2 show the estimates when the sample is restricted to potential new buyers, as in He (2009). Although we observe the same (positive) sign of the marginal effect, only with the HAS correction for misclassification do we find a significant positive correlation between mortality by 2004 and new purchase of life insurance.

Table 3.3 presents a complete set of marginal effects from the estimation above. Of most interest in these estimates is that that marital status and income are significant predictors of insurance ownership. Specifically, married individuals and those with a higher income are more likely to purchase individual life insurance. This reinforces our use of the demand-side variables in the model, as omission of these relevant variables may have caused omitted variable bias in the previous papers.

Overall, our results are different from those in Cawley and Philipson (1996) and He (2009) who find no significant relation between life insurance purchase and mortality in the full sample. In fact, He (2009) claims that restricting the sample to potential buyers would address the sample selection problem induced by potential mortality differences between those with and those without coverage. We show that the positive sign is consistent across both full sample and

the sample restricted to potential buyers, but with a loss of significance after imposing the restriction.

Measurement of mortality risk is one of the main explanations for the difference in findings between our study and Cawley & Philipson (1996). Cawley & Philipson (1996) use estimated mortality between 1992 and 1994, which constitutes approximately 1.5% of the sample. This lack of observed mortality data over the years likely explains their failure to find any significant relationship between mortality and individual life insurance ownership. We follow He (2009) by using data on actual mortality by 2004, which includes about 15% of the sample.

Our use of a more comprehensive set of life insurance factors sets our study apart from Cawley & Philipson (1996) and He (2009). Cawley & Philipson (1996) omit important supply-side variables, while He (2009) omits most demand-side variables. Resulting omitted variable bias likely causes the authors to find zero or negative marginal effect of mortality on life insurance ownership. Specifically, variables income and wealth (omitted by He, 2009) have a positive correlation with the dependent variable (life insurance ownership) and a negative correlation with the independent variable (mortality), which causes a downward bias on the estimated coefficient of mortality on life insurance ownership. Furthermore, variables omitted by Cawley & Philipson (1996), like hospital stay, cancer, parent dying before the age of 60, have a negative correlation with the dependent variable (life insurance ownership) and a positive correlation with the independent variable (mortality), which also causes a downward bias on the estimated coefficient of mortality on life insurance ownership.

After comparing our findings in the full sample to the previous studies, we also see a difference between our results and those in He (2009) after imposing a sample restriction to

potential buyers. We find that this approach presents several problems beyond just reducing the significance of the estimated marginal effect of mortality on insurance ownership. He's sample restriction only applies to respondents with no individual term life insurance coverage in 1992 and disregards other types of life insurance coverage in that year. In fact, 66.5% of the restricted sample actually owned one or more life insurance of a different kind in 1992 (37.8% owned group term life insurance, and 38.4% owned whole life insurance). This means that He treats respondents with no life insurance coverage in 1992 the same as those who have life insurance (just not individual term life insurance) in 1992. It is unclear whether this approach is useful in eliminating the sample selection bias.

Unlike Cawley & Philipson (1996) and He (2009), we address misclassification in self-reported life insurance ownership. We indeed find evidence of under- and over-reporting of life insurance, without an apparent systematic pattern. Our findings show that correcting for misclassification presents only a slight increase in the magnitude of marginal effects. Thus, in contrast to Hedegren & Stratmann's claim (2016), misclassification of reported life insurance in HRS survey does not seem to bias the magnitude of the results. However, HAS correction does have an effect on the significance of the results: the marginal effects of mortality after HAS correction are significant on a 5% level. This matters especially in the restricted sample (potential buyers), where the marginal effect of mortality is not significant before correction. Determining whether the marginal effect is significantly different from zero is crucial to our conclusions about the prevalence of one of the two opposing hypotheses (adverse selection vs advantageous selection theory).

A potential weakness of this study (and the two previous studies) is that our data only reflect if an individual owns a type of insurance or not. But the choice to purchase individual

term insurance may depend on an individual's access to other types of life insurance, such as group life insurance. The full range of insurance coverage is decided jointly. Hence, as an extension, we estimate a reduced-form bivariate probit model, where the dependent variables are individual term insurance ownership (IndInsurance) and group term insurance ownership (GroupInsurance).

The results from the bivariate probit model are presented in Table 3.4. We find that the results are consistent with the results from our main model. Specifically, we observe a marginal effect of 0.028 (compared to 0.029 in the main model), which is significant on a 10% level. The positive effect means that individuals with a higher mortality risk are more likely to have individual term life insurance. The results also show a highly significant and negative marginal effect of mortality on group term life insurance (-0.038). This means that individuals with higher mortality risk are less likely to own group term life insurance coverage. The likelihood ratio test of the correlation between the two residuals of the bivariate model (H0: rho=0) is significant on a 5% level, and the estimated correlation between the two equations (rho) is -0.22. This means that the log likelihood for the bivariate probit models is not identical to the sum of the log likelihoods of the two univariate probit models, and our use of bivariate probit model is justified.

CONCLUSIONS

Our study presents a new look at the relationship between mortality and individual term life insurance. We use the same data as two conflicting previous studies, but change the specification by simultaneously incorporating both supply- and demand-side determinants of life insurance. Moreover, we correct for potential misclassification of self-reported life insurance ownership. We find that correcting for misclassification yields only slightly bigger marginal

effects of mortality on life insurance ownership. However, appropriately correcting misclassification has a big effect on the significance of the results, which is key in determining the presence of adverse selection in the life insurance market. Finally, as an extension and a robustness check to the main model, we estimate and compare the relationship between mortality and two types of term life insurance (group and individual).

Our finding of a significant positive relation between mortality and individual term life insurance coverage means that individuals with higher mortality risk are more likely to purchase individual term life insurance. When we follow He (2009) and restrict our dataset to potential buyers of individual life insurance, this effect is only significant when corrected with HAS, but not significant otherwise. We identify several issues with He's sample restriction, including the fact that many "potential buyers" in her sample actually have another type of life insurance, which does not necessarily solve the sample selection problem.

Our results differ from those by Cawley and Philipson (1996) who do not find any evidence of adverse selection in the life insurance market. They also differ from those by He (2008) who only finds evidence of adverse selection by restricting the sample to potential buyers of life insurance. We argue that the difference in our findings and those in the two previous studies can be explained by our use of actual mortality data (as opposed to Cawley and Philipson, 1996), addressing misclassification of self-reported life insurance ownership, and the use of a model that includes both supply-side and demand-side determinants of life insurance.

Our results signify the presence of possible adverse selection in the individual life insurance market. We see that individuals with a shorter life span and a higher mortality risk are more likely to own individual term life insurance. This would be a contrast to many recent theoretical and empirical studies that reject adverse selection in the life insurance market and

find insurance sellers possessing more information on consumers' mortality risk or imposing higher premiums to high-risk consumers. Our results indicate that this may not be the case, and the providers of individual term life insurance may not possess enough information to price the high-mortality candidates out of the market. Thus if an individual possesses asymmetric information about their own health and mortality risk, they may be more motivated to purchase individual term life insurance, even at a higher price. At the same time, we find no evidence of adverse selection in the group term insurance market. In fact, individuals with lower mortality risk are more likely to have group term insurance coverage, possibly because they tend to be more risk-averse. This negative relationship may signify advantageous selection and/or price discrimination in the group term life insurance market.

The implications of our results are that, contrary to recent studies, there is asymmetric information in the term life insurance market. Specifically, the presence of adverse selection in the individual term life insurance market means that buyers of individual term life insurance have incentive and ability to misuse their knowledge of their health status and mortality risk, which is not fully alleviated by price discrimination on the part of the insurance providers. Furthermore, the information asymmetry in the life insurance market may be different depending on the type of life insurance. Thus under-insurance from adverse selection in the individual term life insurance market must be considered as well as the over-insurance from advantageous selection in the group term life insurance market.

REFERENCES

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. The quarterly journal of economics, 488-500.

Beck, T., & Webb, I. (2003). Economic, demographic, and institutional determinants of life insurance consumption across countries. The World Bank Economic Review, 17(1), 51-88.

California Department of Insurance (2011) Life insurance guide. Retrieved from http://www.insurance.ca.gov/01-consumers/105-type/95-guides/07-life/life-ins-guide.cfm

Call, K. T. (2016, November). Health Insurance Coverage Reporting Accuracy in the American Community Survey. In 2016 Fall Conference: The Role of Research in Making Government More Effective. Appam.

Call, K. T., Davidson, G., Davern, M., & Nyman, R. (2008). Medicaid undercount and bias to estimates of uninsurance: new estimates and existing evidence. Health Services Research, 43(3), 901-914.

Cawley, J., & Philipson, T. (1996). An empirical examination of information barriers to trade in insurance (No. w5669). National bureau of economic research.

Chiappori, P. A. (2000). Econometric models of insurance under asymmetric information. In Handbook of insurance (pp. 365-393). Springer Netherlands.

He, D. (2009). The life insurance market: Asymmetric information revisited. Journal of Public Economics, 93(9), 1090-1097.

Hedengren, D., & Stratmann, T. (2016). Is There Adverse Selection in Life Insurance Markets?. Economic Inquiry, 54(1), 450-463.

Hemenway, D. (1990). Propitious selection. The Quarterly Journal of Economics, 105(4), 1063-1069.

Hill, S. C. (2007). The accuracy of reported insurance status in the MEPS. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 44(4), 443-468.

Klerman J. A., Ringel J. S., Roth B. (2005). Under-Reporting of Medicaid and Welfare in the Current Population Survey. RAND Working Paper WR-169-3. Santa Monica: RAND

McCarthy, D., & Mitchell, O. S. (2003). International adverse selection in life insurance and annuities (No. w9975). National Bureau of Economic Research.

Pascale, J. (2016, November). Health Insurance Coverage Reporting Accuracy in the Current Population Survey Annual Social and Economic Supplement. In 2016 Fall Conference: The Role of Research in Making Government More Effective. Appam.

Rothschild, M., & Stiglitz, J. (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. The quarterly journal of economics, 629-649.

Table 3.1 Descriptive statistics for the variables in the estimation

Variable	Definition	Number of observations	Mean	Standard deviation	Min	Max
Individual Term (in 1992)	whether reported owning individual term life insurance in 1992	8368	0.26	0.44	0	1
Individual Term (in year 1994)	whether reported owning individual term life insurance in 1994	8626	0.27	0.44	0	1
New Buyer	=1 if reported owning individual term life insurance in wave1994, but not in 1992; =0 if reported owning in neither waves.	8551	0.14	0.34	0	1
Group Term (1992)	whether reported owning group term life insurance in 1992	8368	0.35	0.48	0	1
Group Term (1994)	whether reported owning group term life insurance in 1994	8368	0.28	0.45	0	1
Mortality (entire 1992 sample, 2004)	whether dead by wave 2004	8326	0.15	0.36	0	1
Mortality (potential new buyers, 2004)	whether dead by wave 2004	6003	0.15	0.36	0	1
Age	Age	8642	55.73	3.11	51	61
Gender	=1 if male, =0 if female	8642	0.46	0.50	0	1
Smoke_ever	whether smoke now	8642	0.64	0.48	0	1
Smoke_now	whether smoke ever	8642	0.27	0.44	0	1
Drink	whether drink now	8642	0.60	0.49	0	1
Diabetes	whether diagnosed with diabetes	8642	0.11	0.31	0	1
HBP	whether diagnosed with HBP	8642	0.40	0.49	0	1
Cancer	whether diagnosed with cancer	8642	0.05	0.23	0	1
Heart	whether diagnosed with heart disease	8642	0.13	0.34	0	1

Arthritis	whether diagnosed with arthritis	8642	0.39	0.49	0	1
Lung	whether diagnosed with lung disease	8642	0.08	0.27	0	1
Stroke	whether diagnosed with stroke	8642	0.029623	0.169554	0	1
Asthma	whether diagnosed with asthma	8642	0.06075	0.238885	0	1
Kidney	whether diagnosed with kidney disease	8642	0.106341	0.308292	0	1
Ulcer	whether diagnosed with ulcer	8642	0.10	0.29	0	1
Cholesterol	whether diagnosed with high cholesterol	8642	0.25	0.43	0	1
Back_pain	whether suffering from back pain	8642	0.15	0.36	0	1
Hospital_stay	whether had a hospital stay in the previous 12 months	8642	0.11	0.31	0	1
BMI	Body mass index	8642	27.19	5.16	12.76	102.71
Underweight	whether BMI<=18.5	8642	0.01	0.12	0	1
Healthyweight	whether BMI<=24.5 and BMI>18.5	8642	0.31	0.46	0	1
Overweight	whether BMI<=30 and BMI>24.5	8642	0.44	0.50	0	1
Obese	whether BMI>30	8642	0.24	0.42	0	1
History_father	whether father died before 60	8642	0.20	0.40	0	1
History_mother	whether mother died before 60	8642	0.13	0.33	0	1
Income	household income	8604	49723.02	48609.34	-1078	1309000
Wealth	household wealth	8604	235133.70	518951.30	-745000	8734700
Married	whether married	8642	0.74	0.44	0	1
Children	children (1= yes, 0 = no)	8642	0.95	0.22	0	1
Spouse Age	spouse’s age	8529	40.75	24.95	0	85

Table 3.2 Marginal effect of mortality on life insurance ownership, estimated by logit model

	Full Sample			Potential Buyers Only		
	(1) Dep var: Insurance	(2) Dep var: HAS Insurance	(3) Dep var: GHAS Insurance	(4) Dep var: NewBuyer	(5) Dep var: HAS NewBuyer	(6) Dep var: GHAS NewBuyer
	Full Sample	Full Sample	Full Sample	Potential Buyers	Potential Buyers	Potential Buyers
Mortality by 2004	0.0281*	0.0338**	0.0332**	0.00410	0.0285**	0.0128
	(0.0154)	(0.0154)	(0.0161)	(0.0153)	(0.0139)	(0.0145)
Observations	8065	8293	8293	5994	6049	5694

Marginal effects; Standard errors in parentheses
(d) for discrete change of dummy variable from 0 to 1
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3 Full Marginal Effects

	Full Sample			Potential Buyers Only		
	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Insurance	Dep var: HAS Insurance	Dep var: GHAS Insurance	Dep var: NewBuyer	Dep var: HAS NewBuyer	Dep var: GHAS NewBuyer
	Full Sample	Full Sample	Full Sample	Potential Buyers	Potential Buyers	Potential Buyers
Mortality	0.02*	0.03**	0.03**	0.00	0.02**	0.01
by 2004	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01
Age 51	0.01	0.03	-0.05**	-0.06***	-0.03	-0.03
	-0.02	-0.03	-0.02	-0.02	-0.02	-0.02
Age 52	0.02	0.04	-0.05	-0.04	-0.03	-0.02
	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
Age 53	0.00	0.02	-0.06***	-0.04**	-0.02	0.00
	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
Age 54	0.01	0.00	-0.06***	-0.04**	0.01	0.02
	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03
Age 55	-0.04	-0.01	-0.07***	-0.06***	-0.01	0.00
	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Age 56	0.01	0.00	-0.05**	-0.03	0.01	0.01
	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03
Age 57	0.01	0.01	-0.05**	-0.01	-0.01	0.01
	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03
Age 58	0.02	0.02	-0.04	-0.03	0.02	0.05
	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03
Age 59	0.02	0.02	-0.01	-0.02	0.02	0.03
	-0.03	-0.03	-0.03	-0.02	-0.02	-0.03
Age 60	-0.01	0.00	-0.04	-0.02	-0.01	0.01
	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03
Male	-0.01	-0.01	0.00	0.01	-0.02***	-0.02
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Smoke now	-0.01	0.00	0.02	0.03	0.00	0.00
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Smoke ever	-0.01	0.00	-0.01	-0.01	0.00	-0.01
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Mother History	0.00	-0.01	-0.01	-0.01	0.00	0.00

	-0.01	-0.01	-0.02	-0.01	(.)	-0.01
FatherHistory	0.00	0.02	0.01	0.00	0.00	0.00
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Drink	-0.01	-0.04***	-0.04***	-0.02	0.00	0.00
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Diabetes	-0.02	-0.01	-0.01	0.02	0.01	0.01
	-0.02	-0.02	-0.02	-0.02	-0.01	-0.02
HBP	0.01	-0.02	-0.01	0.01	0.00	0.00
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Cancer	-0.01	-0.03	-0.03	-0.04	-0.02	0.00
	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Heart	0.00	0.00	-0.01	0.01	-0.01	0.00
	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01
Arthritis	-0.01	-0.01	0.00	0.00	0.00	-0.01
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Lung	0.00	0.00	0.00	0.00	0.02	0.04
	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Ulcer	0.01	0.02	0.00	-0.01	0.01	0.01
	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Cholesterol	0.00	0.00	0.00	0.02	0.00	0.00
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Stroke	0.01	0.04	0.03	-0.03	0.01	0.03
	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
Hospital stay	0.00	0.00	0.01	-0.02	-0.01	-0.01
	-0.02	-0.02	-0.02	-0.02	-0.01	-0.02
Overweight	0.05	0.08	0.03	0.05	0.04	0.07
	-0.05	-0.05	-0.05	-0.05	-0.04	-0.05
Obese	0.06	0.117**	0.05	0.07	0.03	0.08
	-0.05	-0.05	-0.05	-0.06	-0.04	-0.06
Healthy weight	0.05	0.08	0.03	0.05	0.05	0.09
	-0.05	-0.05	-0.05	-0.05	-0.04	-0.06
Income	0.00***	0.00***	0.00***	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00
Wealth	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00
Married	0.03***	0.06***	0.16***	0.02**	0.00	-0.01
	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Children	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02
	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02
N	8065.00	8293.00	8293.00	5994.00	6049.00	5694.00

Table 3.4. Extension: bivariate probit model estimation (dependent variables: individual term life insurance, group term life insurance; independent variable of interest: mortality by 2004)

	Full Sample	
	(1)	(2)
	Dep var: Individual Term Insurance	Dep var: Individual Term Life Insurance = 1 Group Term Life Insurance = 0
	Coefficient	Marginal Effect
Mortality by 2004	0.088*	0.029*
	(0.046)	(0.015)
	Dep var: Group Term Insurance	Dep var: Individual Term Life Insurance = 0 Group Term Life Insurance = 1
	Coefficient	Marginal Effect
Mortality by 2004	-0.118**	-0.038***
	(0.047)	(0.015)
Rho	-0.224	
	(0.019)	
LR test of H0: Rho=0	Chi2=126.301	Prob > Chi2 = 0.000

Marginal effects; Standard errors in parentheses
(d) for discrete change of dummy variable from 0 to 1
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

APPENDIX

Appendix A

Correction of misreported Life Insurance Ownership

Let $IndInsurance_{it}^{*}$ be the reported ownership of individual term life insurance by individual i in wave t given by

$$IndInsurance_{it}^{*} = X_{it}^{'} b_t + \eta_{it} \qquad (3)$$

Where i indicates a respondent, t indicates a wave (1992, 1994), X_{it} is a vector of variables affecting the decision to purchase life insurance, b is a vector of parameters, and η_i is an i.i.d. error term. Let F be the cdf of the error term and be true life insurance coverage of individual i in wave t.

HAS assumes the probability of misclassification is random. Specifically, the misclassification probabilities are:

$$\alpha_{0,t} = \Pr\left(IndInsurance_{it} = 1 \middle| \widetilde{IndInsurance}_{it} = 0\right),$$

$$\alpha_{1,t} = \Pr\left(IndInsurance_{it} = 0 \middle| \widetilde{IndInsurance}_{it} = 1\right),$$

where $\alpha_{0,t}$ is the probability that a zero is misclassified as a one in wave t (a respondent claims having life insurance when they do not) and $\alpha_{1,t}$ is the probability that a one is misclassified as a zero in wave t (an insurance holder claims not to have insurance). The conditional expected value of (recorded) insurance coverage indicator variable is

$$E\left(IndInsurance_{it} | X_{it}\right) = \Pr\left(IndInsurance_{it} | X_{it}\right) = \alpha_{0,t} + \left(1 - \alpha_{0,t} - \alpha_{1,t}\right) F\left(X_{it}^{'} b_t\right) \qquad (4)$$

When there is no misclassification, $\alpha_{0t} = \alpha_{1t} = 0$, and the above expression becomes $F\left(X_{it}^{'}b\right)$

We estimate (α_{0t}, α_{0t} and b) by maximum likelihood estimation for the log likelihood function

$$L(\alpha_{0,t}, \alpha_{1,t}, b_t) = n^{-1}\sum_{i=1}^{n} IndInsurance_{it} \ln\left(\alpha_{0,t} + (1-\alpha_{0,t} - \alpha_{1,t})\right)F\left(X_{it}^{'}b_t\right) + \\ + \left(1 - IndInsurance_{it}\right)\ln\left(\alpha_{0,t} + (1-\alpha_{0,t} - \alpha_{1,t})\right)F\left(X_{it}^{'}b_t\right) \quad (5)$$

over $(\alpha_{0,t}, \alpha_{1,t}, b_t)$.

GHAS generalizes the probabilities $\alpha_{0,t}$ and $\alpha_{1,t}$ by allowing them to depend on covariates:

$$\alpha_{0,t}(Z_{it}^0) = \Pr(IndInsurance_{it}^0 = 1 \big| IndInsurance_{it} = 0, Z_{it}^0) = F_{0,t}(Z_{it}^0\gamma_{0,t})$$

$$\alpha_{1,t}(Z_{it}^1) = \Pr(IndInsurance_{it}^0 = 0 \big| IndInsurance_{it} = 1, Z_{it}^1) = F_{1,t}(Z_{it}^1\gamma_{1,t})$$

where Z_{it}^0 and Z_{it}^1 might be, but are not necessarily subsets of X_{it}, and $F_{0,t}$ and $F_{1,t}$ are the cumulative distribution functions of stochastic components that determine the underreporting and overreporting of IndInsurance (Tennekoon & Rosenman, 2016). In this case the conditional expected value of reported life insurance indicator variable is

$$E\left(IndInsurance_{it}|X_{it}\right) = \Pr\left(IndInsurance_{it}|X_{it}\right) = F_{0,t}\left(Z_{it}^0\gamma_{0,t}\right) + \\ + \left(1 - F_{0,t}\left(Z_{it}^0\gamma_{0,t}\right) - F_{1,t}(Z_{it}^1\gamma_{1,t}\right)F\left(X_{it}^{'}b_t\right) \quad (6)$$

The conditional expected value of true life insurance variable is:

$$E\left(IndInsurance^{*}_{it}|X_{it}\right) = \Pr\left(IndInsurance^{*}_{it}|X_{it}\right) = F\left(X_{it}^{'}b_t\right) \quad (7)$$

A standard HAS or GHAS correction procedure applies to the main model and is performed in one step. However, HAS/GHAS estimator may fail to converge if the model contains too many regressors. In our case, the estimator did not reach convergence.

We therefore perform the HAS/GHAS correction and estimation in two steps. First, we conduct HAS and GHAS correction techniques on the IndInsurance dependent variable, using a limited set of explanatory variables: age, age of spouse, income, wealth, marital status, whether has children, whether has given some financial aid to children in the past 12 months, whether has whole life insurance, whether spouse has life insurance, whether reported having individual term life insurance in another wave. Using the HAS/GHAS correction, we obtain the expected life insurance ownership variable in each wave (1992, 1994). Expected life insurance ownership is equal to 0 if its predicted value is less than 0.5, and to 1 if its predicted value is greater than 0.5. We also obtain the expected NewBuyer variable (equals 1 if expected true life insurance is 1 in 1994 and 0 in 1992, and equals 0 if expected true life insurance is 0 in both 1992 and 1994). The first step (HAS/GHAS correction) includes only variables relevant to individual's life insurance ownership, but no variables relevant to the individual's mortality (mortality and health variables are only included in the second step). Therefore, there is no concern that the newly created life insurance variable is inherently correlated with mortality.

Table A.1 gives the results of the step 1 analysis. We find significant misreporting of life insurance ownership among the respondents in 1992 and 1994, in both directions. The estimated percentage of misreported observations is consistent across HAS and GHAS techniques, indicating 25 – 27% of expected underreporting of life insurance in 1992, 10% expected overreporting of life insurance in 1992, 22 – 24% expected underreporting of life insurance in

1994 and 10 – 12% expected overreporting of life insurance in 1994. Expected probabilities of over- and underreporting are very similar in magnitude between GHAS and HAS techniques. The misreporting may be random or partly affected by education, objective memory tests and age.

Table A.1 Correction of a misreported variable "life insurance ownership" in two waves using HAS and GHAS

	(1) HAS y = Life Insurance in 1992	(2) HAS y = Life Insurance in 1994	(3) GHAS y = Life Insurance in 1992	(4) GHAS y = Life Insurance in 1994
Age			0.031*	0.003
			(0.018)	(0.00948)
Married			0.6930*	3.897
			(0.366)	(91.655)
Memory Objective			-0.009**	0.037**
			(0.019)	(0.015)
_cons	-0.593***	-0.667***	-3.034***	-4.357
	(0.000)	(0.24)	(1.089)	(91.648)
$E(\alpha_1)$	.277	.252	.23	.242
Male			-0.134**	0.107*
			(0.064)	(0.06)
Education			0.046***	0.006
			(0.012)	(0.009)
Think Quick			-0.011	0.008
			(0.011)	(0.025)
_cons	-1.309***	-1.15	-1.739***	-1.306***
	(0.000)	(0.058)	(0.177)	(0.171)
$E(\alpha_0)$	.097	.125	.099	.121

In the second step, we use a logit model as in equation (1), where the dependent variable is the expected IndInsurance variable from step 1, and a logit model as in equation (2), where the dependent variable is the expected NewBuyer variable from step 1. The set of explanatory variables includes all the same regressors as in equations (1) and (2). The results from this estimation are presented in Table 3.2 and Table 3.3, and discussed in the Results section.