School of Economic Sciences

Working Paper Series
WP 2013-10

Bias in Measuring Smoking Behavior

By

Vidhura Tennekoon and Robert Rosenman

August 2013

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.

BIAS IN MEASURING SMOKING BEHAVIOR

VIDHURA TENNEKOON[a] AND ROBERT ROSENMAN[b]

August 2013

---

[a] Lecturer, Department of Economics, 308 Cate Center Drive, University of Oklahoma, Norman, OK 73019, USA. vtennekoon@ou.edu. 1-405-325-3614. Fax 1-405-325-5842. Corresponding author.

[b] Professor, School of Economic Sciences, Washington State University, Pullman, WA 99164, USA. yamaka@wsu.edu. 1-509-335-1193. Fax 1-509-335-1173.

BIAS IN MEASURING SMOKING BEHAVIOR

ABSTRACT

Researchers often use the discrepancy between the self-reported and biochemically assessed active smoking status to argue that self-reported smoking status is not reliable, ignoring the limitations of biochemically assessed measures and treating it as the gold standard in their comparisons. Here, we employ recent advances in econometric techniques to compare self-reported and biochemically assessed smoking data taking into account errors with both methods. Our results suggest that biochemical measures may not always be more reliable than self-reported data.

Key words: smoking prevalence, misclassification, social desirability, biochemical assessments.
JEL codes: C13, C18, I10, I13, I18.

BIAS IN MEASURING SMOKING BEHAVIOR

I. INTRODUCTION

The reliability of self-reported smoking behavior reported at surveys is questioned widely. Social desirability and other biases may lead respondents to misrepresent their smoking status. When the reported smoking status is linked to a direct financial incentive as in the case of insurance premiums, a smoker has additional reasons to misreport. As a result, biochemical measurements, considered more objective, are commonly used to predict smoking behavior. However, biochemical measurements are subject to a variety of measurement and interpretation errors and thus do not provide the surety often attributed to them[1]. In this paper we employ recent advances in econometric techniques to estimate the accuracy of each measure in predicting true smoking status. This is the first attempt to estimate the probability that the predicted behavior is in error and use those estimates to evaluate the reliability of self-reported and biochemically measured smoking status.

Tobacco use remains the single largest preventable cause of death in the US as well as globally. Each year 443,000 deaths in the US (USDHHS, 2010) and nearly 6 million deaths globally (WHO, 2011) are attributed to tobacco use. Controlling tobacco use is a policy priority of most governments and having accurate measures of tobacco use is an important prerequisite for measuring the success of such efforts. The patterns of smoking across different sub-populations, for example, can be used to optimally allocate resources between various prevention programs. Accurately predicting smoking status is not only a concern of policymakers. Insurance service providers and many employers too desire accurate reports of individual smoking behavior.

The reliability of self-reported smoking data has widely been questioned. In particular, in the aftermath of increased anti-tobacco legislation and more hostile social norms against smoking, some survey respondents are believed not likely to feel comfortable admitting that they currently smoke. Some groups such as pregnant women and parents of young children are more reluctant than others to admit that they are smoking as their smoking behavior is even more socially undesirable (Florescu et al., 2009). Smokers typically pay a higher insurance premium than nonsmokers and face unfavorable labor market outcomes including higher unemployment and wage penalties (Levine et al., 1997; van Ours, 2004; Auld, 2005; Grafova and Stafford, 2009; Cowan and Schwab, 2011) providing financial motivation for smokers to hide their true status.

[1] We visit some of these limitations below.

Biochemical measures of smoking behavior are often used instead of self-reported data because of this intrinsic bias. Among the biochemical measures used to identify active smokers are the levels of carbon monoxide, NNAL (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol), and cotinine in various body fluids. Cotinine is the most popular biomarker for identifying smokers due to its perceived high accuracy. Nicotine, the main addictive ingredient in tobacco is metabolised into cotinine within the body in addition to being available directly in tobacco. Its longer half-life compared to nicotine makes it a better candidate to detect tobacco use more accurately (Perez-Stable et al., 1995). Cotinine concentration is typically measured in blood, urine or saliva samples and occasionally using breast milk or hair (Florescu et al., 2009). Blood cotinine concentration, in particular, is considered a reliable indicator of exposure to tobacco smoke.

There is a wealth of research assessing the relationship between self-reported smoking data as referenced to biochemical assessments. Any discrepancies between the two measures are almost always attributed to the unreliability of self-reported data (West et al. 2007; Gorber et al., 2009) and only in few instances to the limitations of the biochemical measure (Yeager and Krosnick, 2010, for example). Thus, the observed variations are typically explained using the factors that would affect self-reported data including the sample characteristics (adults, adolescents, pregnant women), survey environment (home, school), survey method (face-to-face, online), and survey design (direct and indirect questions). The majority of research comparing the two measures has found little discrepancy between the two approaches, although biochemical measures generally report more smokers than self-assessments. However, many exceptions show large discrepancies one way or the other, almost always attributing the difference to misreporting by true smokers (Gorber et al., 2009).

The accuracy of biochemically assessed tests is often expressed in terms of sensitivity, specificity and sometimes using the positive predictive value. These statistical measures of biochemically assessed smoking tests depend on the biomarker used as well as the type of body fluid tested. Various measures that use different biomarkers are highly correlated but do not always produce the same result.

Sensitivity and specificity as well as the positive and negative predictive values of a biochemically assessed smoking test depend on the chosen threshold point as this value moderates the trade-off between type I and type II errors. There are no universally agreed upon standards for these values. For example, 20 studies that use serum cotinine concentration to identify active smokers (Gorber et al., 2009) had threshold values that varied from 8 ng/mL to 50 ng/mL. The manual for the laboratory procedures used for the National Health and Nutrition Examination Survey (Gunter et al.,1996) suggests two thresholds, a serum

cotinine concentration less than 5 ng/mL as an indication of a nonsmoker and a concentration of more than 15 ng/mL as an indication of an active smoker. The range in between, according to the manual, may indicate exposure to passive smoking.

In spite of broad differences between various biochemical measures used to identify active smokers some researchers have assumed their chosen measure as a gold standard; deviations from these measures and self-reported smoking status is considered a bias in the self-reported data . For example, West et al. (2007) dispute the usual assumption that the estimates based on self-reported data are 'sufficiently accurate for policy purposes', arguing that 'this assumption has not been adequately tested' but ignores the potential errors in biochemical measures. Comparing data used to compute national prevalence estimates in the US, UK and Poland with Cotinine concentration in serum for US and in saliva for UK and Poland they estimate that the national prevalence rates in the US, UK and Poland are underestimated by 0.6%, 2.8% and 4.4% respectively. Gorber et al. (2009) systematically review 54 previous studies on adult smoking behavior each comprising the self-reported and biochemically assessed smoking prevalence estimates. Assuming that the biochemical measure is correct, the authors find an overall trend of underestimation when self-reported smoking status is used to derive smoking prevalence rates. While 40 of the 54 studies reviewed by Gorber et al. show that self-reported prevalence is lower than the biochemical measure, 11 studies report more smokers with self-reported data. Only 3 studies show no discrepancy between the two measures. Yeager and Krosnick (2010), using the data from National Health and Nutrition Examination Surveys conducted during 2001–2002 to 2007–2008, estimate the discrepancy in prevalence rates based on self-reported smoking status and serum cotinine concentration levels to be slightly less than 1%. They, however, believe that the self-reported data could be more accurate and attribute this discrepancy to errors in the biochemical measurement.

The wide variation in results reported in previous studies that compares self-reported and biochemical measures show the dependence of each measure on underlying attributes. The accuracy of self-reported data depends on the characteristics of the target population, survey method, framing of the questionnaire and the survey environment. The accuracy of the biochemical measure depends on the biomarker used, type of body fluid tested and, perhaps most importantly, the threshold used to separate smokers and nonsmokers. While these factors could explain the observed differences between self-reported and biochemical measures in the rates of smoking, they do not give any indication of the bias that results with either method, and thus we emphasize the fallacy of comparing self-reported and biochemically measured data to investigate the accuracy of one or the other.

The only study we found that evaluates the self-reported and serum-cotinine based measures of smoking status without comparing them to each other is Perez-Stable et al. (1995). Their strategy is to compare each measure to other biochemical measures such as hemoglobin, red and white blood cells, iron, lead, cholesterol, vitamin A and vitamin E, physical examination results including body mass index, pulse rate and blood pressure and depression assessments. Using a sample of 743 Mexican Americans they show that their outside anchors have stronger correlations with serum cotinine level than the self-reported smoking intensity. In addition to being not a population level study, they assume that their outside anchors can be used to identify true smokers based on previous studies that document such association. However, since the association of these measures with smoking intensity could be a result of the elevated cotinine level rather than the smoking behavior itself, it still raises the identification problem, although once removed.

In this study, we use an econometric approach to predict the probabilities of misclassification in self-reported and biochemically assessed data in absolute terms. The paper contributes to the literature in several ways. First, we propose a method to estimate the extent of misreporting in self-reported smoking data without requiring any biochemical assessments for comparison. The method measures two types of misreporting probabilities (smokers reporting as nonsmokers and vice versa) separately, not just the net effect at the population level, by predicting the likelihood that a particular respondent misreports, again without using any 'gold standard'. Second, by using the same method on biochemical measures of smoking behavior, we estimate the proportion of type I and type II errors. Finally, the results provide some insights when one should rely on self-reported data and when such data should be validated using a separate biochemical measure. The results will be useful for insurance service providers and any employers who would like to recruit nonsmokers as our method helps to identify not only the causal factors of being a current smoker but also the causal factors of a smoker being misclassified when self-reported data or a biochemical assessment is used, as well as for policy makers who need accurate estimates of smoking prevalence.

In section II, we explain our study sample, discuss the characteristics of self-reported and biochemically measured data and compare those directly as previous researchers do. In section III, we present the model used to estimate the covariate dependent misclassification of smoking status. In section IV, we present our estimates of two types of error probabilities in self-reported and biochemically assessed smoking status data. The results are compared with the relative errors presented in section II. We discuss the relevant policy issues and conclude in section V.

II. THE STUDY SAMPLE

We use data from National Health and Nutrition Examination Survey (NHANES) for our analysis. The survey is a continuous program of the US National Center for Health Statistics that examines a national sample of about 5000 persons each year and has both a survey component and a laboratory examination component. The availability of self-reported answers to smoking related questions as well as the levels of serum cotinine concentrations which can be used to construct biochemically assessed measure of smoking behavior makes this dataset a perfect choice for our study. The sampling procedure of NHANES is complicated as certain categories (for example Mexican Americans and other Hispanics) are oversampled. Accordingly, we use survey weights in all our estimations and analysis.

We used data from NHANES 2009-10, the most recent, and eliminated the observations without both objective and self-reported measures[2]. The final sample included 5712 observations from adults (aged 20 years or older) of both genders. The survey was administered as a face to face interview for this sample. We define a self-reported smoker based on the answers to two survey questions. First, the respondents are asked 'Did you smoke at least 100 cigarettes in your entire life?'. If they answered 'Yes', they are asked a second question 'Do you now smoke cigarettes?' for which they can choose one of the three answers, 'everyday', 'some days' or 'not at all'. If someone answered 'everyday' or 'some days' to the second question we coded that person's reported smoking status (R_i) as 1. If they answered 'No' to the first question or 'not at all' to the second question the variable was coded 0.

We followed the guidelines of the laboratory manual for NHANES and used the threshold value of 15 ng/mL to define tested smoking status (T_i) based on the measured serum cotinine concentration. The variable was coded as 1 if the cotinine measurement exceeded the threshold and as 0 otherwise. When coded in this manner there were 1,247 (21.8%) reported smokers and 1,391 (24.4%) tested smokers. After applying survey weights, the percentages of reported and tested smokers in the population were estimated as 20.35% and 24.26% respectively.

[2] If the respondents with missing observations are systematically different in a way that it affects their smoking behavior, eliminating of these observations could introduce a selection bias. We tested this possibility using the Heckman 2-step procedure by including the 'Inverse Mills Ratio' from a first stage probit regression as a covariate and found no evidence of selection bias with other estimates being qualitatively similar. The results reported here do not include these selection correction terms.

Table 1: Descriptive statistics

Variable	Mean	Weighted mean
Measures of current smoking status		
Reported smoker	0.218	0.203
Tested smoker (cotinine level > 15 ng/mL)	0.244	0.243
Gender		
Male	0.484	0.483
Race/Ethnicity		
Mexican American	0.183	0.085
Other Hispanic	0.103	0.050
Non Hispanic Black	0.171	0.107
Non Hispanic White	0.487	0.688
Mixed or other race/ethnicity	0.056	0.070
Age		
20-35 years	0.251	0.275
35-50 years	0.267	0.294
50-65 years	0.244	0.257
65-75 years	0.130	0.102
Over 75 years	0.109	0.072
Education		
College graduate or above	0.203	0.278
Some college or AA degree	0.280	0.303
High school graduate/ GED or equivalent	0.228	0.228
9-11th grade	0.161	0.126
Less than 9th grade	0.128	0.065
Marital status		
Married	0.519	0.566
Widowed	0.086	0.059
Divorced	0.109	0.099
Separated	0.033	0.023
Never married	0.172	0.175
Living with a partner	0.081	0.076
Body structure		
Underweight	0.016	0.019
Normal	0.262	0.288
Overweight	0.336	0.330
Obese	0.386	0.363
Other		
Employed	0.538	0.628
Pregnant	0.011	0.011
Failing/weak kidney or a liver condition	0.051	0.040
Exposed to tobacco smoke at work	0.074	0.082
Early initiator (before 14 years)	0.072	0.068
Consume Alcohol	0.657	0.701

In our sample 48.4% were male, 18.3% were Mexican American, 10.3% were other Hispanic, 48.7% were non-Hispanic White and 17.1% were non-Hispanic Black. There were 5.6% reported as belong to other or mixed race/ethnicity. Over one-half (51.9%) were married while 10.9% were divorcees, 8.6% were widows, 3.3% were separated, 8.1% lived with a partner and 17.2% were never married. There were 20.3% with graduate degrees or above, 28.0% were with some college or associate degrees, 22.8% were high school graduates with GED or equivalent and 16.1% were educated up to grade 9-11. Additional summary statistics including the weighted averages of these variables are presented in Table 1.

Using the rules discussed above, 20.35% of the sample are self-reported smokers, while 24.26% are classified as smokers using the biochemical measure. Because neither method is fully objective, this information is not sufficient to infer whether the self-reported data is underreported, the tested data comprises of a large number of false positives, or it is a combination of these two possibilities. Both measures produce identical results for 93.48% of the sample; 74.44% are nonsmokers and 19.04% are smokers by both measures. For the remaining 6.52% the two measures disagree, with 5.22% self-reported nonsmokers identified as smokers by the biochemical measure and 1.31% self-reported smokers being classified as nonsmokers by the biochemical assessment.

Table 2: Comparison of each measure assuming the other measure is correct

	Reported data	Test results
Estimated probabilities		
Being a smoker	0.2426	0.2035
Classified correctly	0.1904	0.1904
Misclassified	0.0522	0.0131
Being a nonsmoker	0.7574	0.7966
Classified correctly	0.7444	0.7444
Misclassified	0.0131	0.0522
Statistical measures		
Sensitivity	0.9356	0.7850
Specificity	0.9345	0.9827
Positive predictive value	0.7850	0.9356
Negative predictive value	0.9827	0.9345

Three approaches have been used to reconcile differences between self-reported and biochemical assessment of "true" smoking behavior (S_i):

- Assume that the biochemical measure is true and attribute the entire discrepancy to underreporting of self-reported smoking so $(S_i = T_i)$. Under this assumption, the self-reported data has a sensitivity and specificity each approximately equal to 0.935 as reported in Table 2.

- Assume that the self-reported data is accurate and attribute the deviations to the limitations of the biochemical test, hence $(S_i = R_i)$. Under this assumption the biochemical measure has a sensitivity of 0.785 and a specificity of 0.983.
- Assume that when either of the measures identifies a smoker it is correct but a person is considered a nonsmoker only if both self-reported and cotinine based measures indicate the person is not a smoker, i.e., $(S_i = (T_i = 1) \cup (R_i = 1))$. This is the usual practice of insurance service providers.

As the researcher never observes true smoking status, verifying the validity of any assumption above is a challenge. In fact, S_i can take any value irrespective of the values of T_i and R_i. Even if T_i and R_i perfectly agree for all observations, it does not prove that S_i is accurately measured since both measures could be wrong. Our approach here is to independently estimate the expected error probabilities of each measure with respect to the true value using an econometric approach.

III. IDENTIFYING COVARIATE DEPENDENT MISCLASSIFICATION IN SMOKING BEHAVIOR

Our approach is based on the MLE for misclassified binary dependent variables presented in Hausman et al. (1998) and generalized to the systematically-misclassified case presented in Tennekoon and Rosenman (2011). Let S_i^0 is the reported smoking behavior (the self-reported or biochemically assessed smoking status) equal to 1 if classified as a smoker and 0 otherwise. In addition, let $\Phi(X_i\beta) = \Pr(S_i = 1)$ be the propensity to be a smoker where S_i is the true (unobserved) smoking status and X_i is a vector of causal factors affecting the smoking status. We have two types of errors; $\Phi(Z_i^0\gamma_0) = \Pr(S_i^0 = 1 \mid S_i = 0)$ is the probability of a true nonsmoker being classified as a smoker (false positive), and $\Phi(Z_i^1\gamma_1) = \Pr(S_i^o = 0 \mid S_i = 1)$ is the probability of a true smoker being classified as a nonsmoker (false negative). The vectors Z_i^0 and Z_i^1 are, respectively, causal factors affecting the probability of a false positive or false negative. In all cases $\Phi(\bullet)$ denotes the standard normal cumulative probability distribution function. Finally $(\beta, \gamma_0, \gamma_1)$ are the vectors of parameters to be estimated[3]. The sensitivity

[3] An interested reader is directed to Tennekoon and Rosenman (2011) for additional technical details of the estimator.

and specificity of the observed measure are given by $1-\Phi\left(Z_i^1\gamma_1\right)$ and $1-\Phi\left(Z_i^0\gamma_0\right)$ respectively. As shown in Tennekoon and Rosenman (2011) we have the likelihood function

(1)

$$\mathcal{L}\left(\beta,\gamma_0,\gamma_1\right)=n^{-1}\sum_{i=1}^{n}\left(\begin{array}{l} S_i^0\ln\left[\Phi\left(Z_i^0\gamma_0\right)+\left(1-\Phi\left(Z_i^0\gamma_0\right)-\Phi\left(Z_i^1\gamma_1\right)\right)\Phi(X_i\beta)\right]+ \\ \left(1-S_i^0\right)\ln\left[1-\Phi\left(Z_i^0\gamma_0\right)-\left(1-\Phi\left(Z_i^0\gamma_0\right)-\Phi\left(Z_i^1\gamma_1\right)\right)\Phi(X_i\beta)\right] \end{array}\right).$$

When $Z_i^0\neq Z_i^1$, the parameters of the model can be identified through non-linearity. The maximum likelihood estimator is,

(2) $$\left[\hat{\beta},\hat{\gamma}_0,\hat{\gamma}_1\right]=\arg\max\,\mathfrak{L}\left(\beta,\gamma_0,\gamma_1;X_i,Z_i^0,Z_i^1\right)$$

Once the parameters of (1) are estimated, the fractions of type I and type II errors can be estimated probabilistically as follows.

(3) $$E\left(\Pr\left[\left(S_i^o=1\right)\cap\left(S_i=0\right)\right]\right)=n^{-1}\sum_{i=1}^{n}\frac{S_i^o\Phi\left(-X_i\hat{\beta}\right)\Phi\left(Z_i^0\hat{\gamma}_0\right)}{\Phi\left(-X_i\hat{\beta}\right)\Phi\left(Z_i^0\hat{\gamma}_0\right)+\Phi\left(X_i\hat{\beta}\right)\Phi\left(Z_i^1\hat{\gamma}_1\right)}$$

(4) $$E\left(\Pr\left[\left(S_i^o=0\right)\cap\left(S_i=1\right)\right]\right)=n^{-1}\sum_{i=1}^{n}\frac{\left(1-S_i^o\right)\Phi\left(X_i\hat{\beta}\right)\Phi\left(Z_i^1\hat{\gamma}_1\right)}{\Phi\left(X_i\hat{\beta}\right)\Phi\left(Z_i^1\hat{\gamma}_1\right)+\Phi\left(-X_i\hat{\beta}\right)\Phi\left(-Z_i^0\hat{\gamma}_0\right)}$$

Finally, the estimated error rates can be used to derive sensitivity, specificity and positive and negative predictive values of each observed indicator variable in comparison to the true (but still unobserved) smoking status. When estimating the error rates of self-reported data we calibrate the model allowing $S_i^o=R_i$. When estimating the error rates of the cotinine based test results we allow $S_i^o=T_i$. Because true smoking behavior does not depend on whether it is measured by self-reported or biochemical measures, we use the same variables in vector X_i in both models; these variables comprise the socio-demographic and behavioral factors that affect the decision to be a current smoker.

However, the factors causing smoking behavior to be misreported do depend on how it is assessed. When data is self-reported, smokers are more likely to underreport if they are not 'expected' to be smokers, for example when pregnant. We do not see any reason for a nonsmoker to report as a smoker[4] except for random reporting errors. When the serum cotinine level is used as an indicator of current smoking behavior, it is likely that some of the nonsmokers with a high level of environmental tobacco exposure to be classified as smokers. Misclassification is also a possibility when a person has a deficiency with her system of metabolism (Hukkanen et al., 2005). Accordingly, the variables included in vectors Z_i^0 and Z_i^1 of each model were chosen considering the causal factors underlying each misclassification process. These factors are discussed in the next section.

IV. EMPIRICAL SPECIFICATION AND ESTIMATION RESULTS

However smoking behavior is measured, there should be no difference in what factors predispose someone to smoke. Hence, whether we use self-reported or biochemically assessed behavior to indicate whether an individual smokes, we use the same set of explanatory variables for the X_i in equation, covering age, race/ethnicity, education level, marital status, body structure, employment, whether or not the respondent consumes alcohol, and whether the respondent was an early smoker. The point is to identify those socioeconomic characteristics which best help predict smoking behavior. All of these variables have been used in earlier studies to identify predisposition to smoking (Oh et al., 2010; Hosseinpoor et al., 2011). Means of these variables are reported in Table 1. As all are measured as categorical data, the variances are not reported.

Although predisposition to smoking is independent of how smoking behavior is measured, misclassification of smokers and nonsmokers is not – it could depend on the measure used. When smoking behavior is self-reported, social desirability bias may make it more likely that a smoker reports as a nonsmoker. We include two variables that may exacerbate this bias – whether the respondent is pregnant, and whether the respondent is over 75 years of age. Pregnant women are cautioned that smoking could harm their fetus, and elderly individuals may receive particular pressure to cease smoking as it compounds many of the infirmities that accompany growing older. We treat any misclassification of a nonsmoker self-reporting as a smoker as a random error, since we see no incentive for nonsmokers to misreport that status.

[4] Some nonsmoking adolescents are likely to think that they will look cool if report as smoking. Since our sample only includes adults aged 20 or older this possibility is unlikely.

Table 3: Estimation results using self-reported data

Variable	Parameter		Std. Dev	P-value
Equation 1 : Being a smoker				
Constant	-0.479	***	0.156	0.002
Gender				
Male	0.157	***	0.056	0.005
Age (Excluded: 35-50 years)				
20-35 years	-0.017		0.071	0.812
50-65 years	-0.320	***	0.078	0.000
Over 65 years	-1.026	***	0.159	0.000
Race/Ethnicity (Excluded: Non Hispanic White)				
Mexican American	-0.496	***	0.113	0.000
Other Hispanic	-0.491	***	0.129	0.002
Non Hispanic Black	0.043		0.082	0.600
Mixed or other race/ethnicity	-0.111		0.124	0.371
Education (Excluded: Less than 9^{th} grade)				
College graduate or above	-1.279	***	0.190	0.000
Some college or AA degree	-0.463	***	0.123	0.000
High school graduate/ GED or equivalent	-0.116		0.119	0.326
$9\text{-}11^{th}$ grade	0.181		0.125	0.147
Marital status (Excluded: Married)				
Widowed	0.533	***	0.162	0.001
Divorced	0.573	***	0.095	0.000
Separated	0.492	***	0.158	0.002
Never married	0.369	***	0.082	0.000
Living with a partner	0.823	***	0.119	0.000
Body structure (Excluded: Normal)				
Underweight	0.392	**	0.196	0.046
Overweight	-0.237	***	0.073	0.001
Obese	-0.429	***	0.077	0.000
Other				
Employed	-0.296	***	0.062	0.000
Consume Alcohol	0.367	***	0.072	0.000
Early initiator	0.872	***	0.125	0.000
Equation 2:				
Being a smoker and reporting as a nonsmoker				
Constant	-1.414	***	0.540	0.009
Pregnant	1.145	*	0.626	0.067
Over 75 years	1.816	***	0.563	0.001
Equation 3:				
Being a nonsmoker and reporting as a smoker				
Constant	-1.814	***	0.133	0.000
Number of observations				5712
Log likelihood				-2402.83
Adjusted R-squared (McFadden)				0.1894

*** p<0.01; ** p<0.05; *p<0.10

When the biochemical measure is used to assess smoking behavior, errors beyond false positives or false negatives, which we consider random, are primarily dependent on environmental and health factors. False positives (being a nonsmoker and being assessed biochemically as a smoker) have been shown to be related to the respondent is exposed to tobacco smoke at work or if the respondent has a kidney or liver condition. These two factors are used to explain false positives. No physical or environmental factors are thought to mask smoking use, so false negatives (smoking but being classified as a nonsmoker when biochemically assessed) are treated as random.

The results of the two estimations we did are reported in Tables 3 and 4. We used self-reported smoking status as the dependent variable for producing the results presented in Table 3 and a dependent variable based on the serum cotinine level to produce the results presented in Table 4. The results of both models are in agreement about the causal factors affecting current smoking behavior. Males are more likely to be smokers than females while Mexican Americans and other Hispanics are less likely than non-Hispanic Whites. Propensity to be a current smoker decreases with age, education and body mass index. Married people are less likely to be smokers than the people with a different marital status. Regular consumers of alcohol and those who initiated smoking before the age of 14 years are more likely to be current smokers and employed respondents are less likely to smoke.

The results in Table 3 also show that both pregnant women and those who are older than 75 years are more likely to underreport if they smoke; corroborating our hypothesis that these two groups face marginally stronger societal disapproval of smoking. With other smokers this probability is random. As noted, we treated the misclassification probability of all nonsmokers as random. On average, these results indicate that 9.18% of smokers and 1.64% of nonsmokers were misrepresented in self-reported data. Despite a smoker being significantly more likely to be misclassified as a nonsmoker than a nonsmoker is to be misclassified as a smoker, the fraction of misclassified smokers in the overall population is only 1.93%, compared to the 1.29% misclassified nonsmokers when reported data is used because there are nearly 4 nonsmokers for each smoker. As a result, the smoking prevalence rate is only underestimated by 0.64% when self-reported data are used.

When a cotinine based indicator is used (Table 4) a smoker is very unlikely to be misclassified (less than one in ten thousand) but on average 4.16% of nonsmokers are likely to be misclassified. Nonsmokers with exposure to environmental tobacco smoking at their workplace and those with a failing kidney or a liver condition that could slow down their cotinine metabolism, in particular, are likely to have a serum cotinine level over 15ng/mL and be classified as smokers when the objective measure is used. This leads to 3.27% false positives

and virtually no false negatives to ultimately overstate smoking prevalence by 3.27% when the objective measure is used.

Table 4: Estimation results using biochemical test results

Variable	Parameter		Std. Dev	P-value
Equation 1 : Being a smoker				
Constant	-0.521	***	0.153	0.001
Gender				
Male	0.357	***	0.057	0.000
Age (Excluded: 35-50 years)				
20-35 years	-0.052		0.075	0.488
50-65 years	-0.294	***	0.075	0.000
Over 65 years	-1.345	***	0.155	0.000
Race/Ethnicity (Excluded: Non Hispanic				
Mexican American	-1.088	***	0.155	0.000
Other Hispanic	-0.722	***	0.144	0.000
Non Hispanic Black	0.012		0.081	0.887
Mixed or other race/ethnicity	-0.248	*	0.133	0.062
Education (Excluded: Less than 9^{th} grade)				
College graduate or above	-1.428	***	0.207	0.000
Some college or AA degree	-0.431	***	0.131	0.001
High school graduate/ GED or equivalent	-0.080		0.129	0.523
$9\text{-}11^{th}$ grade	0.209		0.130	0.108
Marital status (Excluded: Married)				
Widowed	0.511	***	0.153	0.001
Divorced	0.698	***	0.088	0.000
Separated	0.616	***	0.156	0.006
Never married	0.336	***	0.083	0.000
Living with a partner	0.763	***	0.097	0.000
Body structure (Excluded: Normal)				
Underweight	0.392	**	0.164	0.017
Overweight	-0.218	***	0.071	0.002
Obese	-0.441	***	0.073	0.000
Other				
Employed	-0.361	***	0.062	0.000
Consume Alcohol	0.397	***	0.071	0.000
Early initiator	0.676	***	0.089	0.000
Equation 2: Being a smoker and testing				
Constant	-3.748		8.683	0.666
Equation 3: Being a nonsmoker and testing positive				
Constant	-1.551	***	0.082	0.000
Exposed to tobacco smoke at work	0.652	***	0.116	0.000
Failing/weak kidney or a liver condition	0.330	**	0.166	0.046
Number of observations				5712
Log likelihood				-2629.96
Adjusted R-squared (McFadden)				0.1620

*** p<0.01; ** p<0.05; *p<0.10

Both models agree on the fraction of active smokers in the population to be 20.98%. This estimate is higher than and closer to the percentage of self-reported smokers (20.35%) and smaller than the percentage of tested smokers (24.26%). This indicates that the self-reported data seems to underestimate smoking prevalence as agreed by most researchers. It also suggests that the biochemical measure may overestimate smoking prevalence by a wider margin, something not investigated previously. These results are summarized in Table 5.

Table 5: Comparison of each measure relative to the true value

	Reported data	Test results
Estimated Probabilities		
Being a smoker	0.2098	0.2098
Classified correctly	0.1905	0.2098
Misclassified	0.0193	0.0000
Being a nonsmoker	0.7902	0.7902
Classified correctly	0.7773	0.7575
Misclassified	0.0129	0.0327
Statistical measures		
Sensitivity	0.9082	0.9999
Specificity	0.9836	0.9586
Positive predictive value	0.9365	0.8652
Negative predictive value	0.9758	1.0000

In our study sample, 3.22% of observations are misclassified in self-reported data which includes 1.93% type I errors (smokers classified as nonsmokers) and 1.29% type II errors (nonsmokers classified as smokers). The misclassification using the biochemical measure, however, is one-sided and includes 3.27% type II errors and a negligibly small percentage of type I errors. The difference between the self-reported and biochemically assessed data is almost equally explained by the misclassified test results and the misclassified self-reported data.

When self-reported data was evaluated assuming the biochemical measure is correct, as is often done, the sensitivity and specificity are both approximately 0.935. If this is correct, there should be approximately 4 misclassified nonsmokers for each misclassified smoker as nonsmokers outnumber smokers by that ratio. But, when we estimate the error rates reference to true status, the sensitivity and specificity are found to be 0.908 and 0.984 respectively. At these values the two types of errors partially cancel out leading to a small net effect on prevalence estimates. The cotinine based indicator, on the other hand, has a very high sensitivity of 0.9999 and a lower specificity of 0.959. If we use the accepted serum-cotinine threshold to identify current smokers, we end up having a

substantially upward biased smoking prevalence rate[5]. The bias when the biochemical measure is used (3.27%) is significantly larger than the bias when self-reported data is used (-0.64%) to estimate smoking prevalence.

Which measure is more appropriate may depend on why smoking behavior is being measured. For example, the cost of a misclassified smoker is substantially higher than a misclassified nonsmoker for an insurance service provider or a potential recruiter which makes it more important to not miss a smoker when screening, even if the process ultimately labels some nonsmokers as smokers. Of course, the individual cost for nonsmokers being labeled as smokers may be large. Misclassification may result in a welfare transfer as well as efficiency loss. The cotinine based indicator with a high sensitivity helps in this case, despite its low positive predictive value because some real smokers may try to self-identify as nonsmokers, arguing they are part of the 13.5% of nonsmokers who test positive. To offset the high sensitivity of cotinine based test results the specificity too has to be improved further, probably by raising the threshold used. Self-reported data, in comparison, has a better balance between sensitivity, specificity and positive and negative predictive values.

V. DISCUSSION AND CONCLUSIONS

Smoking remains one of the leading preventable causes of premature death and a major burden on healthcare budgets. A sizable amount of money is spent on tobacco prevention programs every year. Reliable indicators of current smoking status are needed if the efficacy of these programs is to be accurately assessed, and self-reported smoking status from national or regional surveys is most commonly used to identify active smokers. However, self-reported smoking behavior is often believed to be underreported. As a result, biochemical assessment, thought to be a more objective measure of smoking status is increasingly used by policy makers and insurance service providers.

Our results show that biochemical assessment may not be superior to self-assessment when trying to measure smoking behavior; it depends primarily on the use of the information and how much it matters that some individuals may be falsely and unfairly treated as smokers. Although our findings confirm that self-reported smoking is underreported we do not find that the biochemically assessed measure we studied is clearly a better indicator. We found the percentages of errors in self-reported data and cotinine based results to be close, 3.22% for the former, and 3.27% for the latter. However, the bias in self-reported data is two-sided and cancels our largely when prevalence rate is estimated. The bias with the

[5] If we use a looser threshold the bias will be even bigger. Hence, we do not consider lower threshold values.

cotinine test results, on the other hand, is one-sided causing prevalence estimates to be overestimated, amplifying (incorrectly) the presumed superiority of the assumed objective test relative to self-reports, and often imposing an unfair burden on those falsely classified as smokers. Differences between the two measures often considered due to misreporting in most research may, in fact, be explained more (and almost equally) by the errors that occur in both measures. It may be reasonable to correct self-reported data statistically in order to eliminate the bias, instead of switching to a potentially more unreliable biochemical assessment.

Our results are particularly useful when the interest is not the prevalence rate but the current smoking status of a given individual. Whether the observed indicator is self-reported or biochemically assessed, the econometric technique that we propose here can be used to identify the true propensity to be a smoker as well as to estimate the probability that the observed data could be misclassified. When determining an appropriate insurance premium, for example, a provider may use the estimated propensities to calculate a customized contract amount for each individual based on the estimated risk rather than proposing one of the two values pre-assigned for smokers and nonsmokers.

The choice of measure is not simply an academic exercise; false positives and false negatives on smoking behavior have real and important economic impacts – both with regards to efficiency and the distribution of welfare. With regard to the distribution of welfare, individuals falsely classified as smokers through biochemical assessment face higher insurance costs and fewer employment opportunities, while those smokers who either lie when self-reporting or show up as false negatives with a biochemical assessment unfairly impose costs on insurance companies and employers. This misallocation of risk carries efficiency impacts, of course, but even at the aggregate level, efficient policies against smoking are best served with an accurate assessment of smoking prevalence. Thus, understanding how much misreporting occurs with both self-reporting and biochemical assessment of smoking behavior will allow a better and more efficient allocation of resources.

REFERENCES

Auld MC. 2005. Smoking, Drinking, and Income. *Journal of Human Resources* **40:** 505-518.

Cowan BW, Schwab B. 2011. The Incidence of the Healthcare Costs of Smoking. *Journal of Health Economics* **30(5):** 1094-1102.

Florescu A, Ferrence R, Einarson T, Selby P, Soldin O, Koren G. 2009. Methods for Quantification of Exposure to Cigarette Smoking and Environmental Tobacco Smoke: Focus on Developmental Toxicology. *Therapeutic Drug Monitoring* **31:**14–30.

Grafova IB, Stafford FP. 2009. The Wage Effects of Personal Smoking History. *Industrial and Labor Relations Review* **62:** 381-393.

Griesler PC, Kandel PB, Schafran C, Hu M, Davies M. 2008. Adolescents' Inconsistency in Self-reported Smoking. *Public Opinion Quarterly*. **72(2):** 260-290.

Gunter EW, Lewis BG, Koncikowski SM. 1996. Laboratory Procedures Used for the Third National Health and Nutrition Examination Survey (NHANES III), 1988-1994. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Environmental Health, Public Health Service and Hyattsville, MD: National Center for Health Statistics.

Gorber SC, Schofield-Hurwitz S, Hardt J, Levasseur G, Tremblay M. 2009. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine &Tobacco Research*. **11(1):**12-24.

Hausman JA, Abrevaya J, Scott-Morton FM. 1998. Misclassification of the Dependent Variable in a Discrete-Response Setting. *Journal of Econometrics* **87:** 239-269.

Heckman J. 1979. Sample selection bias as a specification error. *Econometrica* **47** (1): 153–61.

Hosseinpoor AR, Parker LA, Tursan d'Espaignet E, Chatterji S. 2011. Social Determinants of Smoking in Low- and Middle-Income Countries: Results from the World Health Survey. *PLoS ONE* 6(5): e20331. doi:10.1371/journal.pone.0020331

Hukkanen J, Jacob P III, Benowitz NL. 2005. Metabolism and disposition kinetic of nicotine. *Pharmacological Reviews* **57:** 79-115.

Levine PB, Gustafson TA, Velenchik AD. 1997. More Bad News for Smokers? The Effects of Cigarette Smoking on Wages. *Industrial and Labor Relations Review* **50:** 493-509.

Oh DL, Heck JE, Dresler C, Allwright S, Haglund M, Del Mazo SS, Kralikova E, Stucker I, Tamang E, Gritz ER,Hashibe M. 2010. Determinants of smoking initiation among women in five European countries: a crosssectional survey. *BMC Public Health* **10:**74.

Perez-Stable EJ, Benowitz NL, Marin G. 1995. Is serum cotinine a better measure of cigarette smoking than self-report? *Preventive Medicine* **24:** 171-179.

Tennekoon V, Rosenman R. 2011. Systematically misclassified binary dependent variables. School of Economic Sciences Working Paper 2011-09, Washington State University.

U.S. Department of Health and Human Services. 2010. *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General.* Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2010.

van Ours J C. 2004. A Pint a Day Raises a Man's Pay; but Smoking Blows That Gain Away. *Journal of Health Economics* **23**: 863-886.

West R, Zatonski W, Przewozniak K, Jarvis MJ. 2007. Can we trust national smoking prevalence figures? Discrepancies between biochemically assessed and self-reported smoking rates in three countries. *Cancer Epidemiology, Biomarkers & Prevention.* **16(4):**820-2.

World Health Organization. 2011. *WHO Report on the global tobacco epidemic, 2011: Warning about the dangers of tobacco*. Geneva, Switzerland: World Health Organization.

Yeager DS, Krosnick JA. 2010.The Validity of Self-Reported Nicotine Product Use in the 2001–2008 National Health and Nutrition Examination Survey. *Medical Care* **48(12):**1128-32.